10-21-2023

Ensemble Methods for Consumer Price Inflation Forecasting

Jorge M. Bravo
NOVA IMS, Universidade Nova de Lisboa & Université Paris-Dauphine PSL & MagIC & BRU-ISCTE-IUL & CEFAGE-UE, jbravo@novaims.unl.pt

Afshin Ashofteh
NOVA IMS, Universidade Nova de Lisboa & MagIC, aashofteh@novaims.unl.pt

Ensemble Methods for Consumer Price Inflation Forecasting

Jorge Miguel Bravo, NOVA IMS, Universidade Nova de Lisboa & Université Paris-Dauphine PSL & MagIC & BRU-ISCTE-IUL & CEFAGE-UE, Portugal, jbravo@novaims.unl.pt

Afshin Ashofteh, NOVA IMS, Universidade Nova de Lisboa & MagIC, Portugal. aashofteh@novaims.unl.pt

Abstract

Inflation forecasting is one of the central issues in micro and macroeconomics. Standard forecasting methods tend to follow a "winner-take-all" approach by which, for each time series, a single believed to be the best method is chosen from a pool of competing models. This paper investigates the predictive accuracy of a metalearning strategy called Arbitrated Dynamic Ensemble (ADE) in inflation forecasting using United States data. The findings show that: i) the SARIMA model exhibits the best average rank relative to ADE and competing state-of-the-art model combination and metalearning methods; ii) the ADE methodology presents a better average rank compared to widely used model combination approaches, including the original Arbitrating approach, Stacking, Simple averaging, Fixed Share, or weighted adaptive combination of experts; iii) the ADE approach benefits from combining the base-learners as opposed to selecting the best forecasting model or using all experts; iv) the method is sensitive to the aggregation (weighting) mechanism.

Keywords: Inflation; Time Series Forecasting; Model Combinations; Arbitrating; Stacking.

1. Introduction and Previous research

Inflation forecasting is one of the central issues in micro and macroeconomics. Economic agents base their consumption, expenditure, wage bargaining, price setting, asset allocation, financing, and savings rational decisions on a nominal anchor used to tie down the price level (El Mekkaoui et al. 2021). Regular wage bargaining resulting in nominal wage increases relies on backward- or forward-looking inflation forecasts. The indexation of social security benefits, retirement income benefits, inflation-linked debt instruments, longevity-linked insurance-based or capital market-based securities are tied to national inflation (deflation) measures such as the Consumer Price Index (CPI) or Retail Price Index (RPI) (Bianchi et al. 2021; Simões et al. 2021; Bravo & Herce, 2022; Ayuso et al., 2021a,b; Bravo et al., 2021, 2023). Prices in regulated markets (e.g., energy, water, residential housing markets) tend to follow the dynamics of prices in the economy.

Central banks depend on inflation forecasts to assess, inform, and regulate standard and non-standard monetary policy instruments (e.g., interest rate policy, standing facilities, bank reserve requirement, long-term refinancing operations, asset purchase programmes) targeting price stability, preserving the purchasing power of the currency, supporting general economic policy, and setting inflation

ISSN 2183-489X
http://dx.doi.org/10.18803/capsi.v23.317-336

expectations which enhance policy credibility and efficacy. Moreover, in recent decades an increasing number of central banks (e.g., the Reserve Bank of New Zealand, the European Central Bank, the Bank of England, The U.S. Federal Reserve, Brazilian Central Bank) believe monetary policy should be conducted according to predictable and intertemporally consistent rules and have adopted an inflation targeting approach which consists of adjusting monetary policy to achieve an explicit target for the annual inflation rate (Bernanke & Mishkin, 1997).

Inflation is challenging to forecast, partially because the time series properties of inflation measures often comprise non-stationarities and time-evolving complex structures and have changed substantially over time (Cogley & Sbordone, 2008). Over the last decades, a persistent effort has been made by economists to produce more accurate inflation forecasts, reducing the sizeable welfare costs associated with significant forecast errors. Previous research on inflation forecasting identifies five main types of traditional time series methods. The first group encompasses classical univariate time series methods such as autoregressive integrated moving average (ARIMA), random walk (RW), unobserved components stochastic volatility model (UCSV) or exponential smoothing (ETS) (Stock & Watson, 2007), and multivariate time series methods such as the structural vector autoregressive (SVAR) model and the Bayesian vector autoregressive (BVAR) model (see, e.g., Carriero et al. 2015). The second group refers to Philips-curve type of inflation models fitting data to a pre-specified relationship between input (e.g., past values of the unemployment gap, the NAIRU – the non-accelerating inflation rate of unemployment –, housing prices, cash and credit availability, exchange rates, interest rates, past values of inflation or core inflation, inflation expectations) and output variables, thereby assuming a specific functional form for the stochastic process underlying that variable (Faust & Wright, 2013; Bravo & El Mekkaoui, 2022). The third group produces inflation forecasts from anchored and unanchored inflationary expectations (Gobbi et al., 2019) considering the theoretical channels through which the inflation expectations of households and firms are supposed to impact the actual inflation levels. The fourth group encompasses dynamic stochastic general equilibrium (DSGE) models based on using modern macroeconomic (e.g., the real business cycle model, the New Keynesian monetary models) theory to explain and predict co-movements of aggregate time series over the business cycle (Christiano et al., 2018). The fifth group refers to survey-based methods (Berge, 2018).

What these forecasting methods have in common is that they tend to follow a "winner-take-all" approach by which, for each inflation (or other macroeconomic variables) dataset, a single believed to be the best model is chosen from a pool of competing approaches using some in-sample fitting or out-of-sample metric or criteria (e.g., model confidence set, information criteria, cross-validation, stepwise regression, Bayesian variable selection methods based on decision-theory, shrinkage methods, Extreme Bounds Analysis, s-values, best subset regression) (Steel, 2020). Statistical inference continues conditionally upon the conjecture that the selected forecasting model happens

to be a good estimate of the actual data-generating process, deprecating conceptual uncertainty for statistical inference purposes. Previous research on inflation forecasting suggests that univariate time series models often exhibit superior forecasting accuracy in most economic scenarios and that computing simple averages of past inflation values tend to produce more accurate forecasts than the canonical Phillips curve or other structural models (Atkeson & Ohanian, 2001; Stock & Watson, 2007). Recently, research has investigated the suitability of machine learning models in inflation forecasting (Medeiros et al., 2021).

To cope with conceptual uncertainty and improve the accuracy of prediction, recent research investigated the use of dynamic model combinations in economics and finance time series forecasting (Bravo et al., 2021; Ashofteh et al., 2021, 2022). Ensemble methods integrating multiple heterogeneous learning algorithms can capture more information on the underlying structure of the data and have proved to provide a superior predictive performance relative to single experts (Brown et al. 2005; Clemente et al., 2023).

Notwithstanding the attractive features of model combinations, selecting the model set and the model weights of each learning algorithm in the combination rule and ensuring diversity among the experts are important and non-trivial tasks. First, several windowing strategies for expert combination have been developed, including model selection before aggregation using alternative trimming methods, e.g., discarding a percentage of the worst forecasters in the training set (Jose and Winkler, 2008) and averaging the output of the remaining experts, building a «team of champions» using the model confidence set approach (Hansen et al., 2011; Samuels & Sekkel, 2017). A popular approach is to determine the model weights based on the expert's out-of-sample forecasting accuracy using, e.g., a Bayesian model ensemble approach (Hernández et al. 2018; Bravo & Ayuso, 2021a; Bravo, 2022), determining performance on a window of recent data, or using some forgetting mechanism that values more recent forecasting performance (Timmermann, 2008; Sánchez, 2008).

Second, several methods use metalearning strategies for modelling the learning process of individual forecasting algorithms to improve que predictive accuracy in model combinations. Popular approaches include stacking (Wolpert, 1992), applying multiple regression on the output of the experts (Gaillard & Goude, 2015), or arbitrating, a strategy that combines learners that are selected according to their expertise pertaining to the input data (Ortega et al., 2001).

Against this background, this paper investigates the predictive accuracy of a metalearning strategy denoted as Arbitrated Dynamic Ensemble (ADE) in inflation forecasting. Proposed by Cerqueira et al. (2019), the strategy is based on arbitrating and adaptively combines heterogeneous forecasters by creating an embedded meta-learner for each base algorithm that specializes them across the time series. The underlying assumption of this metalearning strategy is that forecasting models tend to have different areas of expertise capturing, with varying relative performance, the time series

recurring structures and changes in the data distribution. The findings from empirical experiments using time series from several real-world domains provided evidence of the method's competitiveness relative to state-of-the-art approaches. However, the approach was not tested using inflation time series data.

In this preliminary experiment, we consider a single type of measurement for inflation - the Consumer Price Index (CPI), and a single forecasting horizon of 36 months. The Consumer Price Index (CPI) consists of a family of indexes that measure the average change in the price paid by a representative private household for a basket of consumer goods and services (e.g., housing, communication, food, education, transportation, medical care, leisure) between two periods of time. The CPI proxies the average cost-of-living and socioeconomic development in a country or region by estimating the purchasing power of money. The CPI is the key macroeconomic indicator for assessing inflation (or deflation) and is measured in terms of the annual growth rate or an index.

The model space of base learning algorithms considered in the ensemble includes both statistical learning and machine learning methods such as Gaussian processes (Karatzoglou et al., 2004), Support vector regression (Karatzoglou et al., 2004), Projection Pursuit Regression (R Core Team, 2022), Multi-layer Perceptron (Venables & Ripley, 2002), Multivariate adaptive regression spline models (Milborrow, 2012), Generalised linear regression (Friedman et al., 2010), Generalized boosted regression (Ridgeway, 2015), Random Forest (Wright, 2015), Principal components regression (Mevik et al., 2016), Partial least regression (Mevik et al., 2016) and Cubist Rule-based regression (Kuhn et al., 2014). Different parameter specifications are considered for each of the individual forecasters, adding up to 52 different models. The forecasting performance is compared with benchmark univariate time series models (ARIMA, the Exponential Smoothing State Space Model (ETS), Seasonal Naïve (SNAÏVE)) and benchmark state-of-the-art ensemble and metalearning strategies (e.g., Stacking, Arbitrating, weighted adaptive combinations of experts, a forecast combination approach based on an exponentially weighted average of experts, Simple average of base forecasters with model trimming).

The findings show that: i) the SARIMA model exhibits the best average rank relative to ADE and competing state-of-the-art model combination and metalearning methods; ii) the ADE methodology presents a better average rank compared to widely used model combination approaches, including the original Arbitrating approach, Stacking, Simple averaging, Fixed Share, or weighted adaptive combination of experts; iii) the ADE approach benefits from combining the base-learners as opposed to selecting the best forecasting model or using all experts; iv) the method is sensitive to the weighting mechanism. The remainder of this article is structured as follows. Section 2 outlines the key concepts and research methods used in the paper. Section 3 reports and critically discusses the empirical results of the paper. Section 4 concludes.

2. ARBITRATED DYNAMIC ENSEMBLE

This section briefly describes the Arbitrated Dynamic Ensemble (ADE) forecasting methodology proposed by Cerqueira et al. (2019) and used in this paper. Let $Y = \{y_1, \dots, y_t\}$ denote a numerical time series (e.g., CPI values) with components $y_t \in \mathbb{R}$ observed at times $t = 1, \dots, t$. The time series forecasting problem is framed as a regression task, with the past K observations of the time series (embedding vector) and summary statistics on the embedding vector as attributes in the learning of the experts. Each observation comprises a feature vector $x_i \in \mathrm{X} \in \mathbb{R}^{K+N}$ including the past K values and N summary statistics on the embedding vector, and a target vector $y_i \in \mathrm{Y} \in \mathbb{R}$ representing the variable we want to predict. The goal is to obtain an estimate of the approximation $\hat{F}(\boldsymbol{x})$ of the function $F(\boldsymbol{x})$ mapping the unknown functional dependence $x \xrightarrow{F} y$, that minimizes the expected value of some specified loss function $\mathcal{L}\big(y, F(\boldsymbol{x})\big)$ over the distribution of all Y-values, where F denotes the regression function.

The ADE methodology for time series forecasting is implemented in three main steps (Figure 1). The first step consists of training the $m \in \mathcal{M}_0$ heterogeneous base learners included in the initial model space $\mathcal{M}_0$ to forecast future values of $\hat{Y} = \{\hat{y}^1, \dots, \hat{y}^m\}$. The second step involves training the associated meta-learners $\mathcal{B} = \{B^1, \dots, B^m\}$ to model the error of $\mathcal{M}_0$, $\hat{E} = \{\hat{e}^1, \dots, \hat{e}^m\}$, which is then used to dynamically weigh the base learners and select/update the set of superior models $\mathcal{M} \in \mathcal{M}_0$ according to their relative forecasting accuracy.

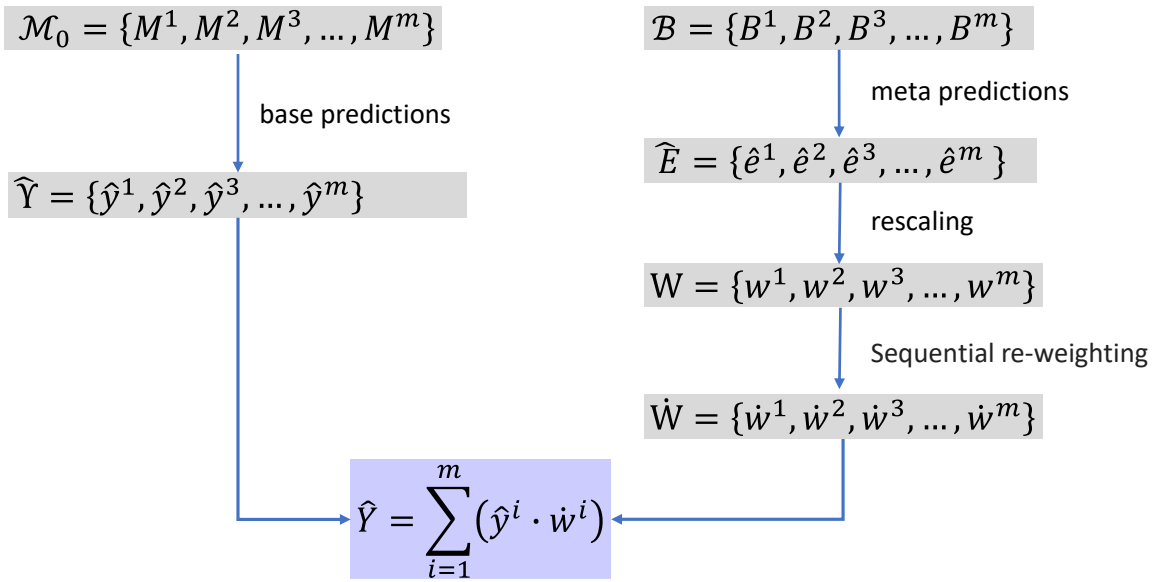

Figure 1 – ADE prediction workflow. The base-learners $\mathcal{M}_0$ are trained to model and predict the next value of the time series $\hat{Y} = \{\hat{y}^1, \dots, \hat{y}^m\}$. In parallel, the associated meta-learners $\mathcal{B} = \{B^1, \dots, B^m\}$ are trained to model the error of $\mathcal{M}_0$, $\hat{E} = \{\hat{e}^1, \dots, \hat{e}^m\}$, which is then used to dynamically weigh the base learners, select and/or update the set of superior models $\mathcal{M} \in \mathcal{M}_0$ according to their relative forecasting accuracy. The final prediction $\hat{Y}$ is computed using a weighted average of the predictions relative to the weights.

The metalearning strategy adopted by ADE is inspired by arbitrating (Ortega et al., 2001), an approach that combines the output of experts according to predictions of the loss that they will produce, and the mixture of experts architecture based on the Divide-and-Conquer principle of Jacobs et al. (1991) by which the problem space is partitioned stochastically into subspaces through a specific error function, with experts becoming specialized on each subspace. Specifically, a meta-learner $B^j, j \in \{1,2,\dots,m\}$ is trained – using the same feature set used by the base learners to predict y_{t+1} – to build the following model:

$$\hat{e}_i^j = f(x_i) \tag{1}$$

where $\hat{e}_i^j$ is the absolute error incurred by M^j in an observation (x_i, y_i), i.e., $\hat{e}_i^j = \left|y_i - \hat{y}_i^j\right|$. The regression analysis on a meta-level is used to apprehend the relationship between each model error and the structure and dynamics of the time series, building on this knowledge to dynamically combine the base learners according to their expected forecasting proficiency.

Contrary to most metalearning approaches for dynamic model selection or combination (including the original arbitrating formulation) that only start the metalearning layer at run-time, using only information from the test set, resulting in few observations to train the meta-learners at the beginning of the time series, the ADE methodology uses the training set to generate out-of-bag predictions which are subsequently considered to compute an unbiased estimate of the loss of each base-learner (Cerqueira et al. 2019). This approach considerably expands the data available for training the meta-learners, ultimately contributing to improving the accuracy of each meta-learner and the ensemble.

The ADE methodology uses a blocked prequential procedure (Figure 2) with a growing lookback window approach to producing out-of-bag samples (Dawid, 1984). Specifically, the embedded time series used for training is divided into β equally sized and time-sequential blocks of contiguous observations. Then, in the first iteration, the first block is used to train the base-forecasters $\mathcal{M}_0$ and the second is used to test them. Then, the second block is merged with the first one for training $\mathcal{M}_0$ and the third block is used for testing.

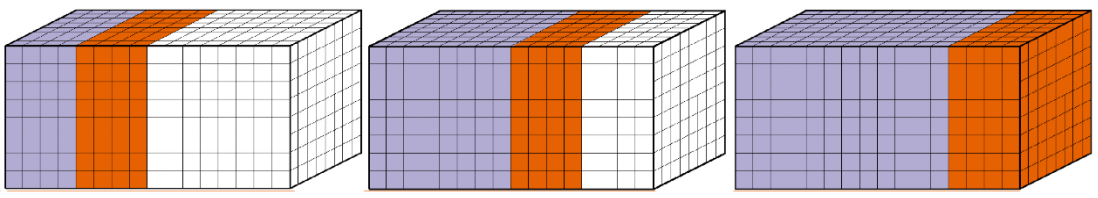

Figure 2 – Time block prequential procedure method with growing window. Blocks of data used for training in lilac; blocks of data used for testing in orange.

We note that in the blocked prequential evaluation procedure, the temporal order is always preserved, i.e., if a block in time is used for testing, then only previous blocks are used for training. The procedure continues until all blocks are verified (except the first one). The motivation for the growing window approach is to profit from all past data for training both the experts (as described above) and the arbiters. This differs from sliding window approaches which try to preserve some consistency in the training size of the different repetitions.

The final step of the ADE methodology consists in predicting the next value of the time series y_{t+1} combining the output of the experts $\hat{Y}$ according to the output of the arbiters $\mathcal{B}$ and the recent correlation among the experts. Contrary to the original arbitrating architecture, which selects the forecaster with the highest confidence as predicted by the arbiters, the ADE approach produces a model combination. However, a windowing strategy is applied before aggregation by forming a committee of best forecasters trimming and suspending a percentage of recently poor forecasters from the combination rule for the upcoming prediction (Jose & Winkler, 2008). If a forecaster is predicted by the arbiters to perform poorly in a given observation relative to the other experts, the methodology assigns a small weight (a zero weight means the learner is not included) in the final ensemble prediction. Formally, ADE selects the Ω% base forecasters (${\mathcal{M}_0}^{\Omega}$) with the lowest mean absolute error (MAE) in the last λ observations, holding (until the next iteration) the remaining ones. The predictions of the meta-level models ($\mathcal{B}^{\Omega}$) are used to weigh the selected forecasters. The weight of the forecaster $M^j \in \mathcal{B}^{\Omega}$ for observation y_{t+1}, w_{t+1}^{j}, is determined by using the following aggregation function on the normalised prediction error produced by the arbiters, i.e.,

$$w_{t+1}^{j} = \frac{min\text{-}max\left(-\hat{e}_{t+1}^{j}\right)}{\sum_{j \in \mathcal{B}^{\Omega}} min\text{-}max\left(-\hat{e}_{t+1}^{j}\right)} \tag{2}$$

where $\hat{e}_{t+1}^{j}$ is the prediction made by $\mathcal{B}_j^{\Omega}$ for the absolute loss that ${\mathcal{M}_0}_j^{\Omega}$ is expected to produce in y_{t+1}, with the min-max scaling function used to normalise the vector of predicted loss into a [0,1] scale. The softmax function assigns larger weights to models with smaller forecasting errors, with the weights decaying exponentially the larger the error. It is commonly used in classification and forecasting exercises (Bravo & Ayuso, 2020, 2021b).

To model the inter-dependence among forecasters, account for model redundancy, and increase model diversity, a sequential re-weighting procedure is implemented by which, first, models are ranked sequentially by their decreasing weight and, second, the correlation among the output of the forecasters in a window of recent observations is used to quantify their redundancy and reweight experts. Intuitively, the higher the correlation between an expert and a higher-ranked forecaster, the lower the model weight. In the limiting case of perfect correlation between forecasters, the weight becomes zero, and the model is effectively withdrawn from the model combination set.

The final ensemble prediction is the weighted average of the predictions made by the individual forecasters $\hat{y}^j$ is computed using a weighted average of the predictions relative to their re-weighted importance $\dot{w}_{t+1}^j$

$$\hat{y}_{t+1} = \sum_{j \in \mathcal{B}^\Omega} \hat{y}_{t+1}^j \cdot \dot{w}_{t+1}^j, \tag{3}$$

with $\sum_{j \in \mathcal{B}^\Omega} \dot{w}_{t+1}^j = 1$.

3. EMPIRICAL STRATEGY

We compare the performance of the ADE methodology in forecasting United States CPI inflation with that of competing state-of-the-art time series model combination approaches. The dataset is provided by the US Federal Reserve and comprises 1315 monthly observations from January 1913 to July 2022 encompassing regimes of low, moderate, and high inflation (Figure 3). The USA inflation exhibits both deterministic and stochastic patterns, with the latter being a dominant factor in the dynamics of the inflation process.

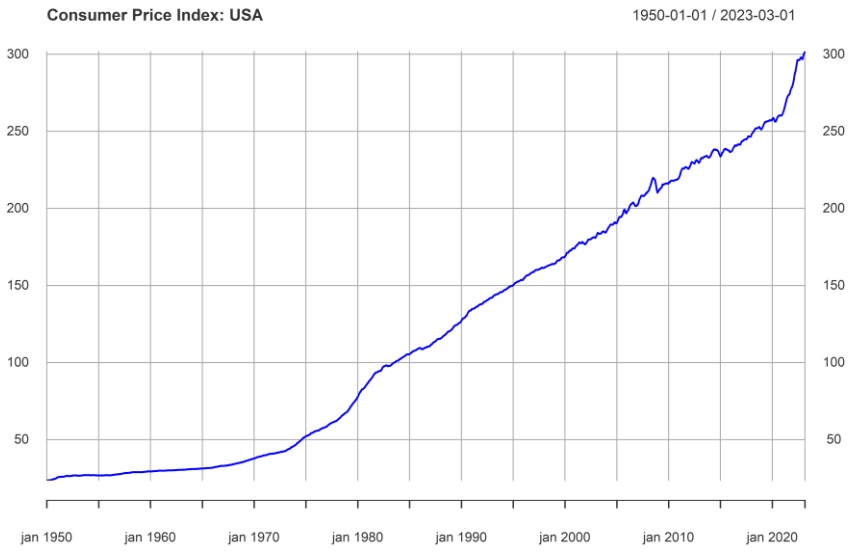

Figure 3 – Consumer Price Index, USA 1950-2022

To account for trend and check for stationarity, we use the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) unit root test (Kwiatkowski et al. 1992) and apply the required difference operators to ensure trend-stationarity. The optimal embedding dimension (K) is estimated using the method of False Nearest Neighbours (Kennel et al., 1992) setting the tolerance of false nearest neighbours to 1%.

Following Cerqueira et al. (2019), the base forecasters and the metalearning models use a feature set that includes the embedding vector (past K values) and several characteristics summarising the overall structure of time series: (i) Local trend, (ii) Skewness, (iii) Mean, (iv) Standard deviation,

(v) Serial correlation, (vi) Long-range dependence using a Hurst exponent estimation with wavelet transform, (vii) Chaos, using the maximum Lyapunov exponent. The final representation of the feature set and target value is as follows:

$$Y_{[n,K]} = \begin{pmatrix} y_1 & y_2 & \cdots & y_{K-1} & y_K & S_{trend_1} & \cdots & S_{chaos_1} & y_{K+1} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ y_{i-K+1} & y_{i-K+2} & \cdots & y_{i-1} & y_i & S_{trend_i} & \cdots & S_{chaos_i} & y_{i+1} \\ \vdots & \vdots & \vdots & \vdots & y_i & \vdots & \vdots & \vdots & \vdots \\ y_{n-K+1} & y_{n-K+2} & \cdots & y_{n-1} & y_n & S_{trend_n} & \cdots & S_{chaos_n} & y_{n+1} \end{pmatrix}$$

Considering the first row as an example, the purpose is to forecast y_{K+1} using as features the previous K values of the time series $\{y_1, y_2, \dots, y_K\}$ along with the corresponding embedded vector summary statistics $\{S_{trend_1}, \dots, S_{chaos_1}\}$. The approach can easily be extended to include other (external) attributes. At the meta-level, the target value is replaced by the absolute loss of a predictive model in that observation.

The forecasting accuracy of base-learners and model combination approaches is evaluated using the root mean squared error (RMSE). In addition, to draw inferences about the differences in model forecasting performance, we implemented a two-step procedure using two post-hoc non-parametric tests. First, the Friedman test (1940) is used to test the null hypothesis that all the forecasters have equivalent performance and, therefore, equal expected average rankings. The test compares the average rankings of the m models across each of the N samples or datasets and is based on the following test statistic measuring the probability of the observed rankings under the null hypothesis.

$$\chi^2_{F,m-1} = \frac{12N}{m(m+1)}\left[\sum_j R_j^2 - \frac{m+1}{2}\right], \quad (4)$$

where R_j^2 is the average rank assigned to method j after applying the methods to N different times series and the statistic is distributed according to a chi-square distribution with $m-1$ degrees of freedom. In the second step, if the null hypothesis is rejected at the selected significance level ($\alpha = 5\%$ in this study), the post-hoc Nemenyi test is used to compare all forecasters to each other. The Nemenyi test is based on a critical difference between the mean rankings among all the models (comparing pairs of models). This measure is computed as

$$CD_{Nemenyi} = q_{\alpha,m} \cdot \sqrt{\frac{m(m+1)}{12N}}, \quad (5)$$

where $q_{\alpha,m}$ is the critical value based on the Studentized range statistic divided by 2.

The estimation method considers a repeated holdout procedure (learning plus testing cycle) in 25 randomly selected testing periods using different but overlapping observations. In our experiment,

each repetition uses 50% of the time series size for training and a lookforward window of 36 months for testing.

The initial model space $\mathcal{M}_0$ considered in this study is summarised in Table 1. The set of base forecasters includes Gaussian processes (Karatzoglou et al., 2004), Support Vector Regression (Karatzoglou et al., 2004), Projection pursuit regression (R Core Team, 2022), Multi-layer perceptron (Venables and Ripley, 2002), Multivariate adaptive regression splines (Milborrow, 2012), Generalized boosted regression (Ridgeway, 2022), Random Forest (Wright, 2023), Principal components regression (Mevik et al., 2023), Partial least squares regression (Mevik et al., 2023) and Cubist Rule-based regression (Kuhn et al., 2023). We consider alternative parameter settings for each of the individual forecasters, totalling 52 different base learners.

ID	Algorithm	Parameter	Value
GP	Gaussian processes (Karatzoglouet al., 2004)	Kernel	{Linear, RBF, Polynomial, Laplace}
SVR	Support Vector Regression (Karatzoglouet al., 2004)	Kernel	{Linear, RBF Polynomial, Laplace}
			Cost ∈ {1, 5, 10}
			ε ∈ {0.001, 0.01}
PPR	Projection pursuit regression (R Core Team, 2022)	No. terms	{2, 5, 10}
		Method	{Super smoother, spline}
MLP	Multi-layer perceptron (Venables & Ripley, 2002)	Hidden units	{5, 10, 15, 30}
		Decay	{0.01, 0.05}
MARS	Multivariate adaptive regression splines (Milborrow, 2012)	Degree	{1, 3, 5}
		No. Terms	{7, 15, 30}
		Forward thresh.	{0.001}
GLM	Generalised linear regression models (Friedman et al., 2010)	Penalty mixing	{0, 0.25, 0.5, 0.75, 1}
		Distribution	{Gaussian}
GBR	Generalized boosted regression (Ridgeway, 2022)	Depth	{5, 10, 15}
		Distribution	{Gaussian, Laplace}
		Shrinkage	{0.1, 0.01}
		No. Trees	{500, 1000}
		Learning rate	{0.1}
RF	Random Forest (Wright 2023)	No. trees	{500, 1000}
PCR	Principal components regression (Mevik et al., 2023)	Default	
PLS	Partial least squares regression (Mevik et al., 2023)	Method	{Kernel PLS, Sijmen de Jong's SIMPLS, Principal Component Regression}
RBR	Cubist Rule-based regression (Kuhn et al., 2023)	No. iterations	{1, 5, 15}

Table 1 – Summary of the base forecasters

The Random Forest model was used as a meta-learner. The time-blocked prequential procedure was run with 10 folds ($\beta = 10$) using Pearson's correlation function for the sequential re-weighting of experts. For each prediction, the selected model space (committee) considers 50% of the experts with the best performance in the last 50 observations, i.e., $\Omega = 0.5$; $\lambda = 50$. These values were selected following the analysis of Cerqueira et al. (2019) on the sensitivity of ADE to different combinations between Ω and λ.

The performance of the ADE methodology in inflation forecasting is compared with that of the following state-of-the-art approaches: (i) Stacking for times series (Wolpert, 1992); (ii) Arbitrating (Ortega et al., 2001); (iii) Simple: model ensemble with individual forecasters averaged using an arithmetic mean (Timmermann, 2008); (iv) SimpleTrim: Simple average of base forecasters with model trimming, with Ω% of the best past performing models are selected to take part in the ensemble committee; (v) LossTrain: static weighted average of forecasters, with weights defined according to the performance of experts in the training set; BestTR: forecasting approach that selects the best performing model in the training data; (vi) EWA: forecast combination method based on an exponentially weighted average of experts; (vii) FixedShare: the fixed share approach adapted for identifying the best forecaster across a time series (Cesa-Bianchi and Lugosi, 2006); (viii) MLpol: the polinomially weighted average forecast combination (Cesa-Bianchi and Lugosi, 2006); (ix) OGD: An approach based on online gradient descent (Zinkevich, 2003); (x) ARIMA: Seasonal Autoregressive Integrated Moving Average (SARIMA) model. The ARIMA model is one of the most commonly used time series forecasting approaches, with remarkable forecasting accuracy and efficiency in multiple domains, including inflation forecasting (Stock and Watson, 2007); (xi) SNAÏVE: approach in which each forecast y_{t+1} is set to be equal to the last observed value from the same season y_{t+1-m}, where m is the seasonal period. Particularly, for monthly time series, we use the value from the previous year; (xii) ExpSmoothing: The exponential smoothing state space model (Hyndman and Athanasopoulos, 2021).

In addition, we investigated the performance of four variants of ADE: (i) ADE-SB: a variant of ADE in which at each time point the best performing model (the one with the lowest predicted loss) is selected to make a prediction; (ii) ADE-ALLM: A variant of ADE without the formation of a committee, i.e., with $\Omega = 100\%$; (iii) ADE-v0: variant of ADE with linear re-weighting of the output of the arbiters instead of the softmax-type function and no sequential re-weighting; (iv) ADE-noSR: A variant of ADE in which there is no sequential reweight of the experts.

4. RESULTS

As an exploratory analysis, Figure 4 shows the distribution of the RMSE of each base learner across the 25 randomly selected testing periods. Figure 5 shows the distribution of the corresponding rank,

together with the Friedman test p-value and the Nemenyi test critical difference. A rank of 1 for a given learner means the model was the best-performing one in a specific testing period.

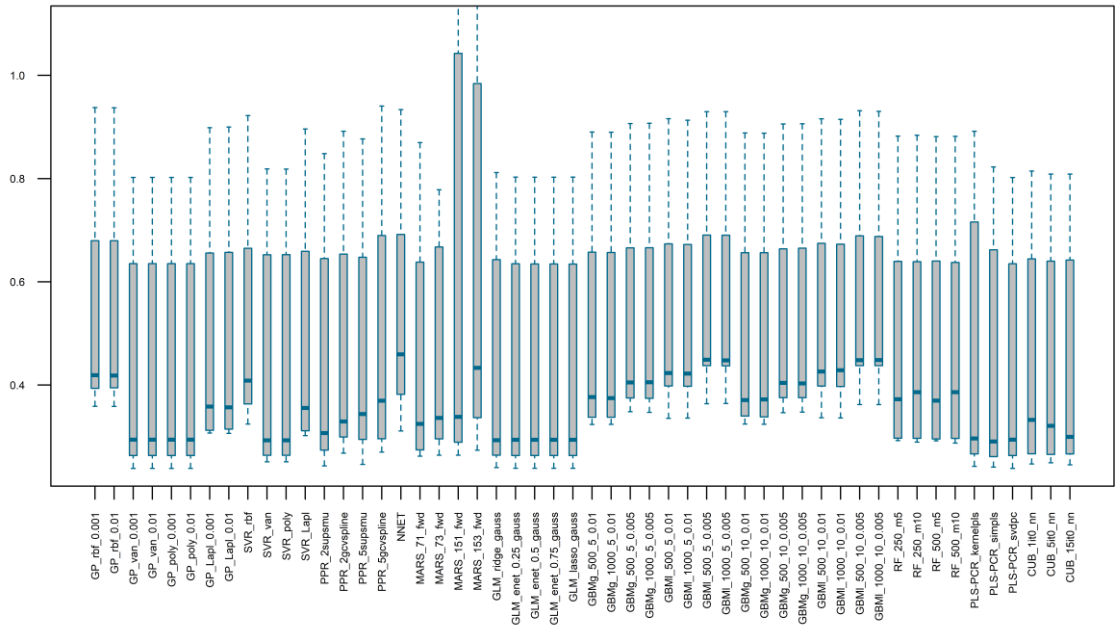

Figure 4 – Distribution of RMSE of the base learners across the testing periods

The findings show that the distribution of the base learners is wide, with generalised linear regression models, gaussian processes, partial least squares regression, and principal components regression among the experts with low mean rank. However, as expected, not all models perform the same, and the critical difference between the mean rankings among all the models is high. This is not a major concern for approaches such as ADE since the trimming procedure used to form the committee before aggregation excludes the worst performers in the model combination used for forecasting.

In Figure 6, we show the boxplot of the distribution of the RMSE error of ADE and its variants and competing state-of-the-art approaches for forecast combinations across the testing periods. Figure 7 compares the average rank of ADE and its variants, together with the corresponding Friedman test p-value and the Nemenyi test critical difference.

The experiment findings show, first, that the standard seasonal autoregressive integrated moving average (SARIMA) model exhibits the best average rank relative to ADE and competing state-of-the-art model combination methods. The SARIMA is considerably better compared to all other approaches. Second, the results show that the ADE methodology presents a better average rank compared to competing widely used model combination approaches, including the original Arbitrating approach, Stacking, Simple, Fixed Share, or OGD. The results of this experiment in Figure 7 show that, except for the SARIMA model, ADE outperforms all other approaches, most in a statistically significant way.

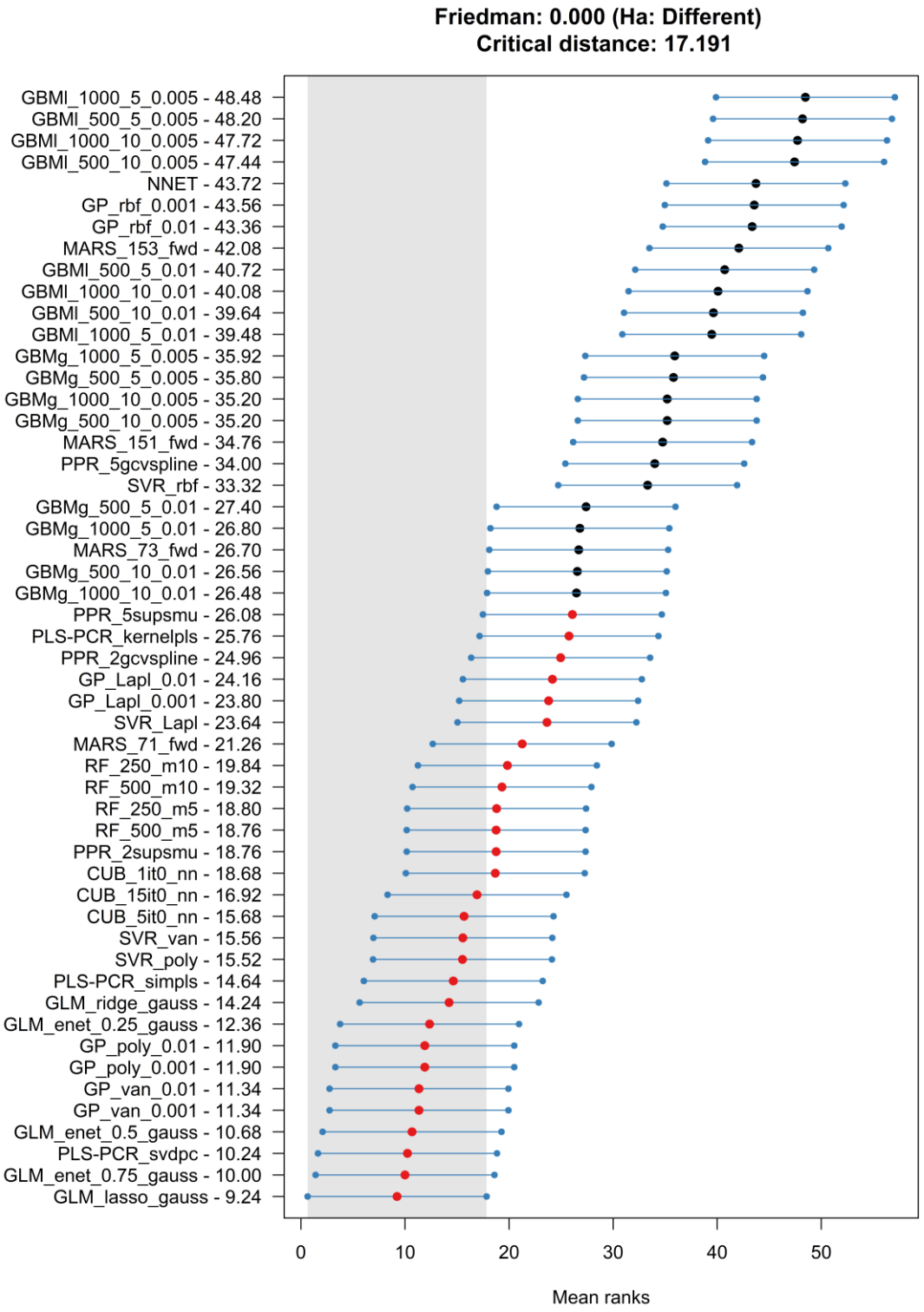

Figure 5 – Distribution of rank of the base learners across the testing periods.

The results suggest that relative to the original Arbitrating architecture, the ADE represents a considerable improvement in CPI inflation forecasting (the average rank of ADE is 5.26 against 9.76 of Arbitrating), suggesting that using the training set to generate out-of-bag predictions, using them

to compute an unbiased estimate of the loss of each base-learner and adopting a blocked prequential procedure contributes to improving the forecasting performance.

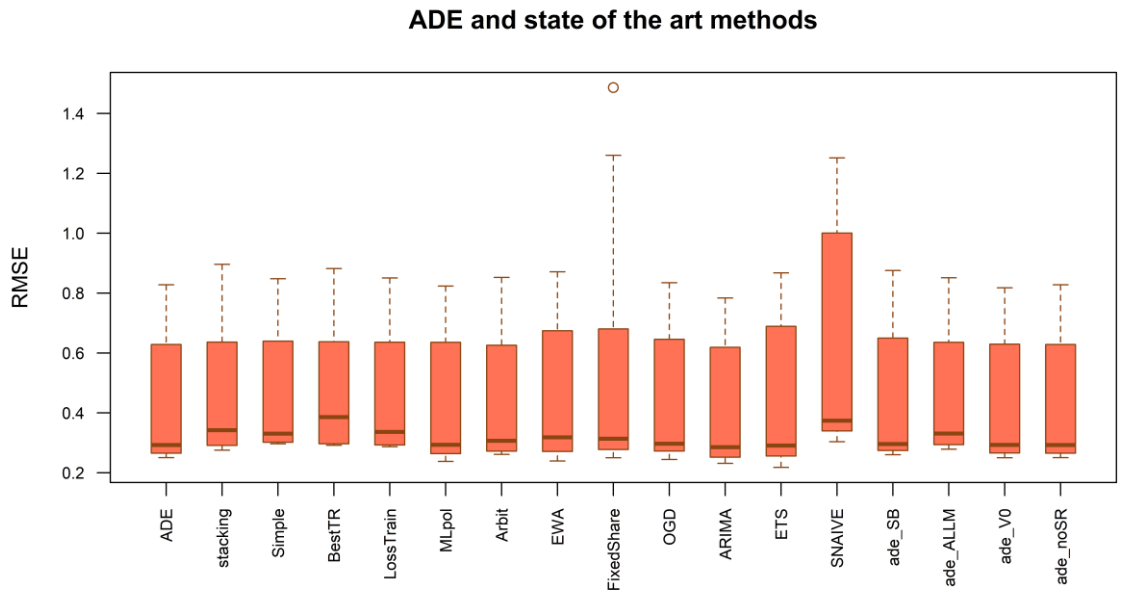

Figure 6 – Distribution of RMSE of ADE and state-of-the-art approaches for forecast combination.

Among the model combination approaches, the Simple averaging approach preceded by a selection of a percentage of recent best past performing models presents the lowest average ranks, which suggests that simple average aggregation coupled with model selection performs poorly in CPI inflation forecasting. In this case, the use of a windowing approach by which weights are computed based on model performance in a window of past recent data does not pay off.

The alternative dynamic ensemble method MLpol, a polynomially weighted average forecaster based on regret minimization exhibits remarkable performance, with an average rank only topped by ADE and some of its variants. The online gradient descent regret minimization approach OGD evidences an interesting performance outperforming consolidated ensemble methods such as Arbitrating or Stacking.

Figure 8 displays the summary results of testing the statistical significance of the forecasting accuracy differences between alternative forecasters using the Friedman and Nemenyi non-parametric tests. All tests were performed for a 95% confidence level.

The Friedman test null hypothesis that the forecasters have similar performance is rejected. The ADE approach and its variant ADE_noSR are the only two forecasting approaches for which we cannot reject the null hypothesis that they exhibit similar performance when compared to the top forecaster seasonal ARIMA method at a 95% confidence level.

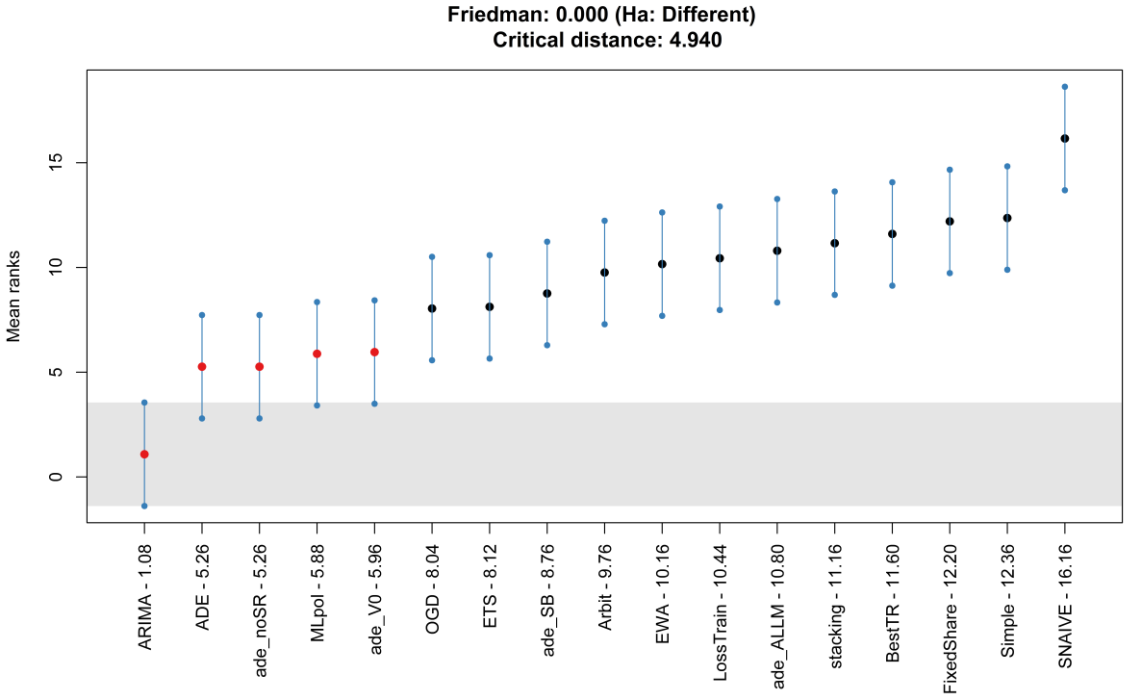

Figure 7 – Distribution of rank of ADE and state-of-the-art approaches across the testing periods.

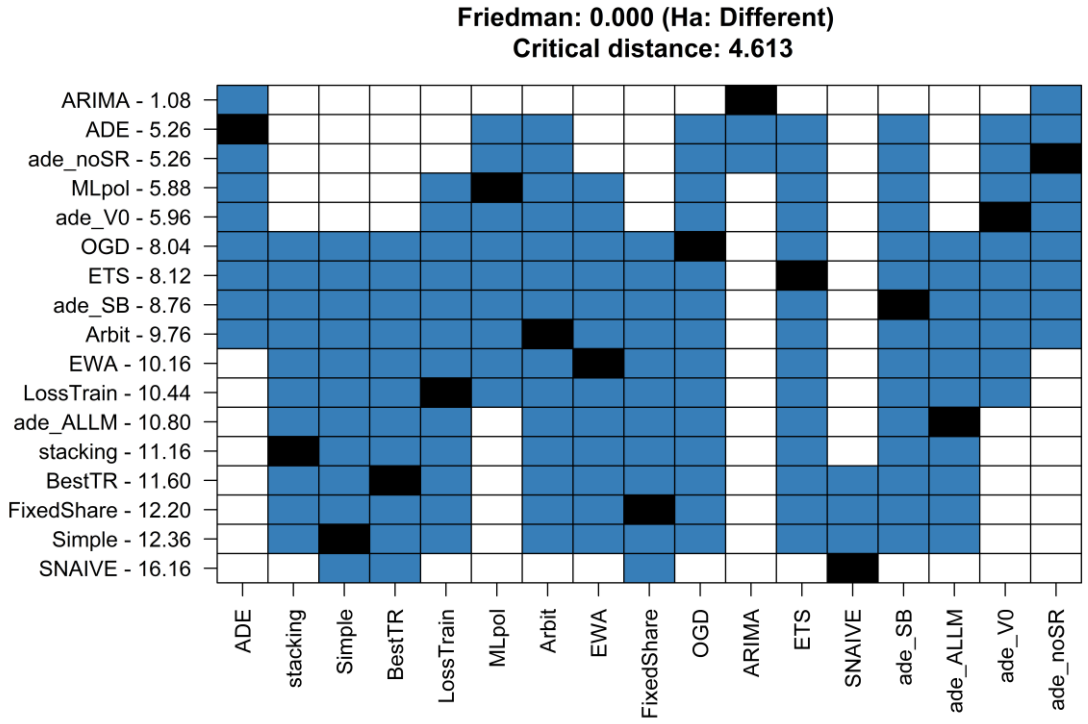

Figure 8 – Statistical significance of accuracy differences using Friedman and Nemenyi non-parametric tests

Regarding the performance of ADE against its variants, Figure 9 displays the average rank of ADE and its variants. The results of this experimental study in inflation forecasting show that ADE exhibits a solid advantage over the performance of ADE methodology excluding model selection,

i.e., including all models in the ensemble committee (ADE-ALLM). The findings also show that the ADE methodology benefits from combining the base learners as opposed to selecting the best forecasting model at each iteration (ADE_SB).

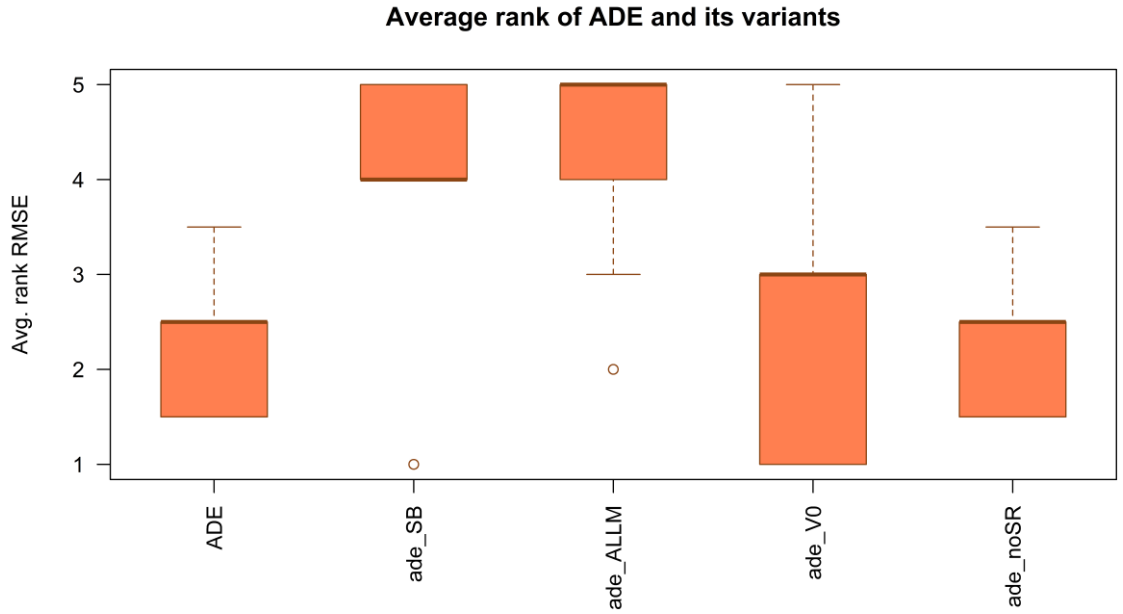

Figure 9 – Boxplot of the average rank of ADE and its variants.

Moreover, the forecasting accuracy metrics of ADE are also superior to ADE-v0, which circumvents the weighting step of the procedure by considering a simple linear re-weighting of the output of the arbiters and bypasses the sequential re-weighting step instead of considering the softmax aggregation function. This conclusion differs from Cerqueira et al. (2019), which concluded that the use of the softmax function does not improve the results over a linear transformation.

The findings also suggest that bypassing the sequential re-weighting of the experts according to the recent correlation between their performance does not contribute to improving the methodology's accuracy in CPI inflation forecasting. However, like Cerqueira et al. (2019), we note that the sequential re-weighting procedure does not jeopardise the overall performance of ADE performance.

Finally, the results of the Friedman and Nemenyi non-parametric tests reported in Figure 8 suggest that the ADE methodology excluding model selection (ADE_AALM) is the only ADE variant that exhibits statistically significant different performance when compared to ADE at a 95% confidence level.

5. CONCLUSIONS

Forecasts of key macroeconomic variables such as inflation, GDP growth, unemployment, or money supply, by national and supranational institutions are critical inputs for short- and central

bank monetary policy assessment, long-term government fiscal planning, and business decision-making. The forecasting exercise in economics is a difficult challenge, partially because of conceptual (model) uncertainty. Model combination and metalearning strategies can be an alternative to classical univariate and multivariate time series methods, Philips-curve type of inflation models, anchored and unanchored inflationary expectations models, dynamic stochastic general equilibrium (DSGE) models, or survey-based methods.

This paper investigates the predictive accuracy of a metalearning strategy called Arbitrated Dynamic Ensemble in inflation forecasting using United States data. The forecasting performance of ADE is compared with benchmark univariate time series models and benchmark state-of-the-art ensemble and metalearning strategies including Stacking, Arbitrating, weighted adaptive combinations of experts, or computing a simple average of base forecasters with model trimming. The model space of base learning algorithms considered in the ensemble includes both statistical learning and machine learning methods. Different parameter specifications are considered for each of the individual forecasters, adding up to 52 different base learners.

The findings show the SARIMA model exhibits the best average rank relative to the second ADE and competing state-of-the-art model combination and metalearning methods, confirming previous research validating the use of univariate time series models in inflation forecasting. The ADE methodology presents a better average rank compared to widely used model combination approaches, including the original Arbitrating approach, Stacking, Simple averaging, Fixed Share, or weighted adaptive combination of experts. The results confirm that the use of the ADE metalearning approach in inflation forecasting benefits from combining the base experts as opposed to selecting the best forecasting model at each iteration or using all experts. The findings also suggest that the method is sensitive to the aggregation (weighting) mechanism.

Several research avenues can be triggered from this first experiment using model combinations and metalearning strategies in inflation forecasting. First, to validate these preliminary results, we plan to extend the analysis to other measurements of inflation and datasets. Second, the inclusion of additional covariates in the feature set used to calibrate the experts and the metalearning strategies accounting for the macroeconomic determinants of inflation (cost push, demand push, monetary policy, fiscal policy, expectations) will be explored. Third, further research should be conducted to investigate the sensitivity of the ADE results to the aggregation method.

FUNDING AND ACKNOWLEDGMENTS

This research was funded by national funds through the FCT—Fundação para a Ciência e a Tecnologia, I.P., grants UIDB/04152/2020—Centro de Investigação em Gestão de Informação (MagIC) and UIDB/00315/2020—BRU-ISCTE-IUL. The authors express their gratitude to two anonymous referees for their careful review and insightful comments that helped to strengthen the quality of the paper.

REFERENCES

Ashofteh, A., Bravo, J. M., & Ayuso, M. (2022). A New Ensemble Learning Strategy for Panel Time-Series Forecasting with Applications to Tracking Respiratory Disease Excess Mortality during the COVID-19 pandemic. *Applied Soft Computing*, 128, 109422.

Ashofteh, A., Bravo, J. M. & Ayuso, M. (2021). A Novel Layered Learning Approach for Forecasting Respiratory Disease Excess Mortality during the COVID-19 pandemic. *CAPSI 2021 Proceedings*. (Atas da 21ª Conferência da Associação Portuguesa de Sistemas de Informação 2021), Volume 2021 – October 2021, Code 183080.

Atkeson, A., Ohanian, E. L. (2001). Are Phillips curves useful for forecasting inflation? *Federal Reserve Bank of Minneapolis Quarterly Review,* 25, 2–11.

Ayuso, M., Bravo, J. M. & Holzmann, R. (2021). Getting Life Expectancy Estimates Right for Pension Policy: Period versus Cohort Approach. *Journal of Pension Economics and Finance*, 20(2), 212–231.

Ayuso, M., Bravo, J. M., Holzmann, R. & Palmer, E. (2021). Automatic indexation of pension age to life expectancy: When policy design matters. *Risks*, 9(5), 96.

Berge, T. J. (2018). Understanding survey-based inflation expectations. *International Journal of Forecasting*, 34 (4), 788-801.

Bernanke, B. S. & Mishkin, F. S. (1997). Inflation Targeting: A New Framework for Monetary Policy? *Journal of Economic Perspectives*, 11 (2), 97-116.

Bianchi, D. Büchner, M. & Tamoni, A. (2021). Bond risk premiums with machine learning. *The Review of Financial Studies*, 34(2), 1046–1089.

Bravo, J. M. (2022). Pricing Participating Longevity-Linked Life Annuities: A Bayesian Model Ensemble approach. *European Actuarial Journal*, 12, 125–159.

Bravo, J. M., & Ayuso, M. (2021a). Linking Pensions to Life Expectancy: Tackling Conceptual Uncertainty through Bayesian Model Averaging. *Mathematics*, 9(24), 1-27.

Bravo, J. M., & Ayuso, M. (2021b). Forecasting the Retirement Age: A Bayesian Model Ensemble Approach. In: Rocha Á., Adeli H., Dzemyda G., Moreira F., Ramalho Correia A.M. (eds) Trends and Applications in Information Systems and Technologies. WorldCIST 2021. Advances in Intelligent Systems and Computing, Volume 1365 AIST, 123 – 135. Springer, Cham. Doi: 10.1007/978-3-030-72657-7_12.

Bravo, J. M., & Herce, J. A. (2022). Career Breaks, Broken Pensions? Long-run Effects of Early and Late-career Unemployment Spells on Pension Entitlements. *Journal of Pension Economics and Finance,* 21(2): 191–217. Doi: 10.1017/S1474747220000189

Bravo, J. M., Ayuso, M. (2020). Mortality and life expectancy forecasts using bayesian model combinations: An application to the portuguese population. *RISTI - Revista Iberica de Sistemas e Tecnologias de Informação*, E40, 128–144. Doi: 10.17013/risti.40.128–145.

Bravo, J. M., Ayuso, M., Holzmann, R. & Palmer, E. (2021). Addressing the Life Expectancy Gap in Pension Policy. *Insurance: Mathematics and Economics,* 99, 200-221.

Bravo, J. M., Ayuso, M., Holzmann, R., & Palmer, E. (2023). Intergenerational Actuarial Fairness when Longevity Increases: Amending the Retirement Age. *Insurance: Mathematics and Economics*, 113, 161-184. Doi: 10.1016/j.insmatheco.2023.08.007.

Bravo, J. M., El Mekkaoui, N. (2022). Short-Term CPI Inflation Forecasting: Probing with Model Combinations. In: Rocha, A., Adeli, H., Dzemyda, G., Moreira, F. (eds). *Information Systems and Technologies. WorldCIST 2022. Lecture Notes in Networks and Systems*, Volume 468. Springer, Cham, 564–578. Doi: 10.1007/978-3-031-04826-5_56

Brown, G., Wyatt, J., Harris, R., & Yao, X. (2005). Diversity creation methods: a survey and categorisation. *Information Fusion,* 6(1), 5-20.

Carriero, A. Clark, T. E. & Marcellino, M. (2015). Bayesian vars: Specification choices and forecast accuracy. *Journal of Applied Econometrics*, 30(1), 46–73.
Cerqueira, V., Torgo, L., Pinto, F., Soares, C. (2019). Arbitrage of forecasting experts. *Machine Learning*, 108, 913–944.
Cesa-Bianchi, N., & Lugosi, G. (2006). *Prediction, learning, and games*. New York: Cambridge University Press.
Christiano, L., Eichenbaum, M. S. & Trabandt, M. (2018). On DSGE Models. *Journal of Economic Perspectives,* 32 (3), 113-40.
Clemente, C., Guerreiro, G. R., and Bravo, J. M. (2023). Modelling Motor Insurance Claim Frequency and Severity using Gradient Boosting. *Risks*, 11(9): 163. Doi: 10.3390/risks11090163
Cogley, T. Sbordone, A. M. (2008). Trend Inflation, Indexation, and Inflation Persistence in the New Keynesian Phillips Curve. *The American Economic Review*, 98 (5), 2101-2126.
Dawid, A. P. (1984). Present position and potential developments: Some personal views: Statistical theory: The prequential approach. *Journal of the Royal Statistical Society. Series A (General),* 147(2), 278–292.
El Mekkaoui de Freitas, N., & Bravo, J. M. (2021). Drawing Down Retirement Financial Savings: A Welfare Analysis using French data. *ICEEG 2021: The 5th International Conference on E-Commerce, E-Business and E-Government. Association for Computing Machinery (ACM),* New York, NY, USA, 152–158. Doi: 10.1145/3466029.3466041
Faust, J., Wright, J. H. (2013). Forecasting Inflation, in: Elliott, G., Granger, C., Timmermann, A. (Eds.*), Handbook of Economic Forecasting, Volume 2,* Elsevier, 2–56.
Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1), 1–22.
Friedman, M. (1940). A comparison of alternative tests of significance for the problem of rankings. *Annals of Mathematical Statistics*, 11(1), 86-92.
Gaillard, P., Goude, Y. (2015). Forecasting Electricity Consumption by Aggregating Experts; How to Design a Good Set of Experts. In: Antoniadis, A., Poggi, JM., Brossat, X. (eds) *Modeling and Stochastic Learning for Forecasting in High Dimensions. Lecture Notes in Statistics,* Volume 217, Springer, Cham. Doi: 10.1007/978-3-319-18732-7_6
Gobbi, L., Mazzocchi, R., & Tamborini, R. (2019). Monetary policy, de-anchoring of inflation expectations, and the new normal. *Journal of Macroeconomics*, 61, 103070, 1-15.
Hansen, P.R., Lunde, A., & Nason, J.M. (2011). The model confidence set. *Econometrica*, 79(2), 453–497.
Hernández, B., Raftery, A., Pennington, S. & Parnell, A. (2018). Bayesian additive regression trees using Bayesian model averaging. *Statistics and Computing,* 28, 869–90.
Hyndman, R.J. & Athanasopoulos, G. (2021). *Forecasting: Principles and Practice*. 3rd Edition, Otexts Publishing.
Jacobs, R. A., Jordan, M. I., Nowlan, S. J., & Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural Computation*, 3(1), 79–87.
Jose, V. R. R., & Winkler, R. L. (2008). Simple robust averages of forecasts: Some empirical results. *International Journal of Forecasting*, 24(1), 163–169.
Karatzoglou, A., Smola, A., Hornik, K., & Zeileis, A. (2004). kernlab—An S4 package for kernel methods in R. *Journal of Statistical Software,* 11(9), 1–20.
Kennel, M.B., Brown R., and Abarbanel H.D.I. (1992). Determining embedding dimension for phase-space reconstruction using a geometrical construction. *Physical Review A*, 45, 3403-3411.
Kuhn, M., Weston, S., & Keefer, C. (2023). *Code for Cubist by Ross Quinlan,* N.C.C.: Cubist: Rule- and Instance-Based Regression Modeling. R package Version 0.4.2.1
Kwiatkowski, D., Phillips, P. C., Schmidt, P., & Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *Journal of Econometrics*, 54(1–3), 159–178.
Medeiros, M., Vasconcelos, G., Veiga, A., & Zilberman, E. (2021). Forecasting inflation in a data-rich environment: the benefits of machine learning methods. *Journal of Business & Economic Statistics*, 39(1): 98-119.
Mevik, B. H., Wehrens, R., & Liland, K. H. (2023). pls: Partial least squares and principal component regression. R package version 2.8-2.
Milborrow, S. (2012). *Earth: Multivariate adaptive regression splinemodels.* Derived from mda:mars by Trevor Hastie and Rob Tibshirani.
Ortega, J., Koppel, M., & Argamon, S. (2001). Arbitrating among competing classifiers using learned referees. *Knowledge and Information Systems*, 3(4), 470–490.
R Core Team. (2022). R*: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Ridgeway, G. (2022) *gbm: Generalized Boosted Regression Models*. R package version 2.1.8.1.
Samuels, J.D., & Sekkel, R.M. (2017). Model confidence sets and forecast combination. International *Journal of Forecasting*, 33(1): 48-60.
Sánchez, I. (2008). Adaptive combination of forecasts with application to wind energy. International *Journal of Forecasting*, 24(4), 679–693.
Simões, C., Oliveira, L. & Bravo, J. M. (2021). Immunization Strategies for Funding Multiple Inflation-Linked Retirement Income Benefits. *Risks*, 9(4), 60.
Steel, M. F. (2020). Model Averaging and Its Use in Economics. *Journal of Economic Literature*, 58 (3): 644-719.
Stock, J. H. & Watson, M. W. (2007). Why has us inflation become harder to forecast? *Journal of Money, Credit and Banking,* 39, 3–33.
Timmermann, A. (2008). Elusive return predictability. *International Journal of Forecasting,* 24(1), 1–18.
Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (4th ed.). New York: Springer.
Wolpert, D. H. (1992). Stacked generalization. *Neural Networks,* 5(2), 241–259.
Wright, M. N. (2023). *Ranger: A fast implementation of random forests*. R package Version 0.15.1
Zinkevich, M. (2003). Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th international conference on machine learning (ICML-03),* 928–936.