Inflation nowcasting in persistently high inflation environments

Richard Schnorrenberger*
Kiel University

Aishameriane Schmidt†
Erasmus Universiteit Rotterdam, Tinbergen Institute and De Nederlandsche Bank

Guilherme Valle Moura‡
Federal University of Santa Catarina

This version: June, 2023.
Preliminary Draft

Abstract

We study the predictive ability of high-frequency macro-financial indicators to nowcast monthly inflation in an environment characterized by persistently high inflation rates, namely the Brazilian economy of the past decades. Using novel machine learning methods within a mixed-frequency framework, we identify two key elements that improve inflation nowcasts upon the survey of professional forecasters, particularly during periods of rising inflation. First, we show that shrinkage-based models combined with timely releases of non-official consumer price indices and market expectations better capture the inflation surge following the Covid-19 pandemic. Second, using the real-time flow of data releases to guide model specifications leads to higher-quality nowcasts.

Key words: inflation nowcasting, machine learning, mixed-frequency modeling.
JEL classification: E31; E37; C53; C55.

*Corresponding author. Institute for Statistics and Econometrics - Kiel University, Germany. **Email**: richard.schn@stat-econ.uni-kiel.de

†Econometrics Institute - Erasmus Universiteit Rotterdam; Tinbergen Institute and De Nederlandsche Bank. **Email**: a.venes.schmidt@dnb.nl

‡Department of Economics - Federal University of Santa Catarina, Brazil. **Email**: guilherme.moura@ufsc.br

We are grateful to Wellington Maiberg who provided excellent research assistance for this project.

Preliminary Version: please, do not cite nor quote.

The opinions expressed in this paper are the authors' personal views and do not necessarily reflect the views of De Nederlandsche Bank or the European System of Central Banks.

1 Introduction

Recent episodes of stress on global supply chains following the Covid-19 pandemic and the war in Ukraine have shown that inflationary waves can unfold extremely fast around the globe while leading to considerable macroeconomic uncertainty as spiraling inflation expectations become a real threat. In this scenario, real-time forecasts of inflation are of utmost importance for central banking, macroeconomic policy and investment decisions.

Knowledge of current inflation is essential for the conduct of monetary policy, since it allows a timely detection of inflationary shocks, and pins down the starting point of long term inflation forecasts. However, as it is often the case with macroeconomic aggregates, current inflation measures become available only with some delay. Therefore, policy makers and economic agents have to rely on nowcasts of inflation. Moreover, Faust and Wright (2013) argue that good forecasts begin with high-quality nowcasts, and the availability of higher frequency data that correlates with current inflation is the ideal ingredient to improve nowcasting models. At the same time, there is a rapidly increasing availability of high-frequency data that comes in handy to monitor the state of the economy in real-time (Evans, 2005; Giannone et al., 2008; Bańbura et al., 2012).

Building on these trends, we investigate the contribution of a large set of high-frequency economic and financial variables to nowcast the headline inflation rate in an environment characterized by persistently high price developments, namely the Brazilian economy of the past decades. We draw on the inflation forecasting literature, which found that machine learning methods are able to capture well the underlying joint dynamics of inflation, and economic and survey indicators (Medeiros et al., 2021; Babii et al., 2021; Garcia et al., 2017a), and compare the prediction accuracy of tree-based algorithms against penalized regression methods. Specifically, for tree-based methods, we compare the random forest from Breiman (2001); the bayesian additive regression trees from Chipman et al. (2012); the generalized random forest from Athey et al. (2019) and the local linear forest from Friedberg et al. (2020). As penalized regressions we use the Elastic Net, the LASSO and Ridge, as well as the sparse-group LASSO (Simon et al., 2013; Babii et al., 2021).

The Brazilian hyperinflation history, together with Brazil's recent decades of persistent and relatively high inflation rates, generated an environment in which different consumer price indexes are released by different agencies during different periods of the month. The official inflation measure targeted by the Brazilian Central Bank, IPCA, is released by the Brazilian statistical bureau (IBGE) usually ten days after the end of a given month (see Figure 1). Additionally, the official Brazilian inflation measure has a mid-month version, IPCA-15, which is computed based on prices observed from the 16th day of the previous month until the 15th day of the current month, and is obviously a leading indicator of end-of-the-month inflation. On a weekly basis, the *Fundação Getúlio Vargas* (FGV) computes

the IPC-S, which is also a national consumer price index. IPC-FIPE is yet another consumer price index released on a weekly basis by *Fundação Instituto de Pesquisa Econômica* (FIPE), although it only covers the São Paulo municipality. Figure 2 shows a timeline of the arrival of new inflation information within a given month. In addition to these intra-month price measures, the Brazilian Central Bank also surveys market participants on a weekly basis about their expectations regarding several economic variables, including their expectations about IPCA for several horizons, and publishes them in the so-called FOCUS survey of professional forecasters. The availability of such data makes Brazil an interesting case to study the potential of nowcasting inflation.

Our target variable, IPCA, is sampled on a monthly frequency, whereas data on a large set of predictors is either available at a weekly basis or released before the target variable, generating a mixed frequency dataset. Often, the issue of mixed frequency data has been addressed by converting the higher-frequency data to the sampling rate of the lower frequency data, for example by temporally aggregating monthly indicators to quarterly, and adding them as regressors in a usual setup. However, this approach does not make use of the high frequency information available during the period of interest, and a common finding is that exploiting intra-period information can reduce forecasting errors. Clark et al. (2022) deal with the mixed frequency characteristics of the dataset using simple averages of the high frequency observations over the low frequency period. We, on the other hand, use the mixed-data sampling (MIDAS) approach put forward by Ghysels et al. (2004), which allows for different weighting schemes of the high frequency data. [1]

The broader literature on nowcasting usually focuses on factor models following the success of Giannone et al. (2008) in real time analysis of GDP. An early use of this methodology for inflation nowcasting is Modugno (2013). However, Knotek and Zaman (2017) show that carefully selecting variables is fundamental to the development of an effective nowcasting model for US consumer prices. Thus, variable selection algorithms like LASSO Tibshirani (1996) might be more promising than factor models when the interest lies on inflation nowcasting. Support for this conjecture comes from the literature on inflation forecasting. For example, Medeiros et al. (2021) support show that variable selection methods outperform factor models in their inflation forecasting application. More specifically, these authors present compelling inflation forecasting results in favor of random forests. On the other hand, Joseph et al. (2021) and Garcia et al. (2017b) show that penalized regression models outperform nonlinear machine learning methods for short horizon forecasts, which are more closely related to nowcasts. Such behavior might indicate that nonlinearities are more prominent in longer horizon forecasts. Building on these evidences, we compare the performance of linear variable selection methods such as penalized regression approaches with that of nonlinear tree-based models for nowcasting Brazilian inflation.

[1] Also formally conceptualized in Ghysels et al. (2007) Andreou et al. (2010).

Our findings indicate that penalized regression methods outperform tree algorithms in predicting monthly headline inflation. In comparison to market expectations, penalized regression models have smaller forecast errors across all out-of-sample forecast periods. In particular, the sparse group LASSO, on average, offers smaller RMSE during the first two weeks of the month while the LASSO is better during the last two weeks of the month. These two models are particularly good during the COVID-19 inflation surge, where professional forecasters had a tendency to underestimate what would be the headline inflation for the current month. We also document the usefulness of having timely releases of non-official consumer price indices, since their inclusion in the models improves the nowcasts. Finally, we show that adjusting model specifications based on the real-time flow of data releases yields better inflation nowcasts compared to a single model based on the entire set of predictors, which potentially assigns high coefficients to variables with no contemporaneous data released by the time of the nowcast.

Our nowcast results are in line with inflation forecasting results of Garcia et al. (2017b) and indicate that linear models with shrinkage and variable selection done via the LASSO outperform tree-based methods. More specifically, both the LASSO and sparse-group LASSO models deliver smaller root mean squared errors than the FOCUS survey of professional forecasters. Cumulative sum of loss differentials indicate that the advantage of LASSO-type models over the FOCUS comes from periods of rising inflation. Consistent with the literature on the use of survey data in forecasting models (Bańbura et al., 2012), we find that the inclusion of the professional survey forecasts improves the nowcasts of all models. We also find that the higher frequency indicators are more relevant in the first half of the month, in comparison to low-frequency variables. This implies that, for our inflation nowcasting exercise, it is important to adjust model specification in real-time depending on the forecast horizon of interest, and the arrival of intra-month data.

The paper proceeds as follows. Section 2 outlines the real-time dataset of the Brazilian macroeconomy and how these variables relate to the CPI inflation series. In Section 3, we illustrate the nowcasting setup and provide an overview of various mixed-frequency modeling strategies to nowcast inflation dynamics. Next, we present our empirical results in Section 4 and provide a guideline for updating CPI figures using the real-time data flow. Finally, Section 5 concludes.

2 Data

To compute weekly nowcasts of inflation figures we need to select predictors that have two features: relatively high correlation with price developments and earlier availability in comparison to official inflation releases. In this sense, our dataset mainly consists of timely

price indicators, financial variables and experts' forecasts that carry predictive content about the current month's inflation rate of the Brazilian economy.[2] More precisely, we construct a real-time dataset of key macro-financial indicators that covers the period from January 2003 up to December 2022 ($T = 240$ monthly observations), whereas information on release dates is only available as of January 2013.[3]

The official inflation measure corresponds to the Brazilian consumer price index (IPCA), which is the reference for the inflation-targeting system in Brazil.[4] The IPCA reflects consumption patterns of urban households in major Brazilian cities that earn from 1 to 40 minimum wages (90% of urban population). The Brazilian statistical office publishes IPCA figures with an average lag of seven workdays after the end of the reporting month. Figure 1 shows the IPCA evolution since mid 2000's - a year after the Brazilian Central Bank adopted the inflation targetting regime. In 2003, with the increase in political risk due to the election of the worker's party representative, there was an outflow of foreign capital, increasing the exchange rate and pressuring domestic prices. This was followed by a relatively calm period, in which the yearly IPCA fluctuated around 5%. It rose again to the double digit figures with the political turmoil that initiated in 2013 and led to the impeachment of president Rouseff in the beginning of 2015. And similar to what was observed in other economies, inflation rose in 2019-2021 as response to pandemic shock.

We organized a dataset containing 20 predictors for IPCA, besides its own lags[5]. These predictors can be divided into four categories: monthly price indicators, weekly price indicators, daily financial variables, and daily expectations of professional forecasters. The data and publication dates are obtained from many sources, including the Brazilian Institute of Geography and Statistics (IBGE), Central Bank of Brazil (BCB), Brazil Stock Exchange (B3), Getulio Vargas Foundation (FGV), Institute of Economic Research Foundation (Fipe), Brazilian National Agency of Petroleum, Natural Gas and Biofuels (ANP) and Bloomberg. Table 1 presents a summary of the selected predictors for IPCA dynamics.

The first group of predictors consists of five indicators of prices primarily collected in

[2] We disregard monthly indicators of real economic activity for two reasons: (i) short or none availability before official releases of the target inflation and (ii) non-significant cross-correlations up to six lags with month-on-month inflation rates. Hence, economic activity variables do not fit our nowcasting purpose.

[3] Although our analysis focuses on price indicators and financial variables as the potential predictors for inflation, the dataset also comprises vintages of hard and survey data for economic activity (e.g., industrial production, unemployment rate, net payroll jobs, PMI manufacturing, retail and services indices, consumer and business confidence, among others) along with publication dates.

[4] Besides, a sizeable number of inflation-linked government bonds use the IPCA as their reference.

[5] Although there are more indicators that potentially correlate with IPCA, we opted for a medium sized dataset following the findings from Carriero et al. (2020), namely small and medium datasets are better for nowcasting inflation than large datasets. Nonetheless, our dataset becomes large in the number of covariates due to the high-frequency data, since each weekly variable needs four coefficients when using the MIDAS structure.

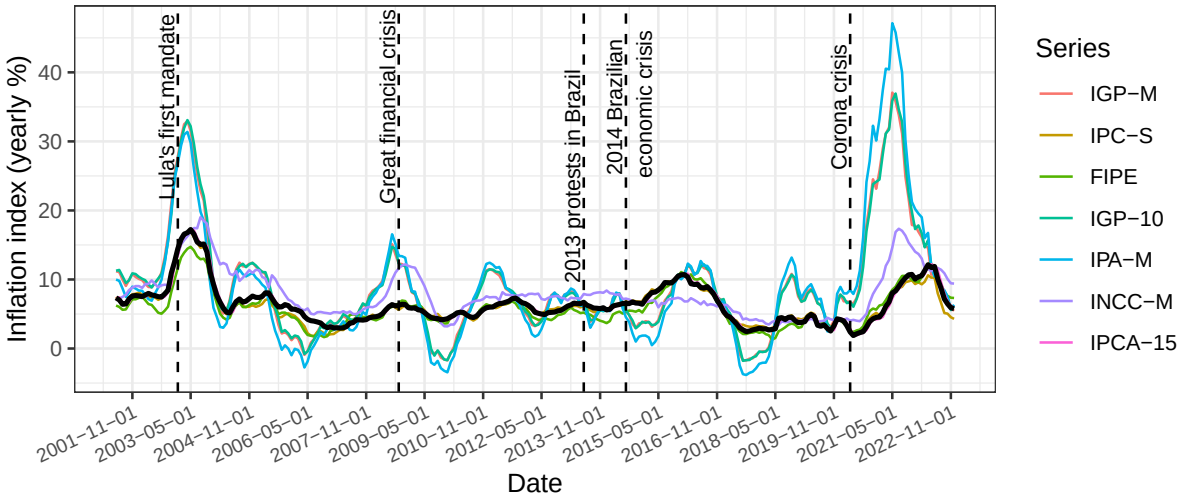


Figure 1: Monthly Brazilian consumer price index (IPCA) from Jun 2001 to Dec 2022 in black. All lines correspond to different inflation indexes.

urban areas of major Brazilian cities. These indices are sampled at the monthly frequency but released before the end of the reporting month, and essentially differ in terms of the reference period and targeted prices. For instance, IPCA-15 mimics IPCA itself, but it reflects prices collected from the 16^{th} of the preceding month to the 15^{th} of the reporting month, allowing for early releases (usually at the beginning of the 4^{th} week). Moreover, IPA-M is a producer price index for agricultural and industrial products, INCC-M reflects house construction costs and both IGP-M and IGP-10 comprise a weighted average of the IPA, INCC and IPC-S indexes, being distinguished only by their reference periods. These indexes are also displayed in Figure 1, and they are largely correlated with IPCA, although some display a larger volatility. For example, IPA-M is largely affected by the exchange rate, and in periods of political turmoil and during the Great Financial Crisis, IPA and the indexes that include it in its calculation (IGP-M and IGP-10) increase more than the other indexes.

Next, we have six timely indicators of prices sampled at the weekly frequency and published with a lag of one or two days after the closing of a given week. As for general consumption baskets, IPC-S accounts for earnings in the range of 1-33 minimum wages while IPC-FIPE accounts for households in São Paulo city that earn from 1 to 10 minimum wages. Moreover, we include prices of major energy components such as diesel, gasoline, ethanol fuel and liquefied natural gas. These data are surveys of the wholesale fuel price practiced by retailers of around 500 cities nationwide.[6]

[6]Compared to information on raw oil prices available in financial markets, these surveys have the ad-

Figure 2 provides a timeline of the real-time data flow related to the above price indicators in December 2022. As shown, IPCA figures have been released on the 10th of January, 2023, but the timing of releases for the selected predictors mostly occurs throughout the reporting month. For example, the first release is available for energy prices on 5 December while the next releases of these prices are provided in the following Mondays. The timing calendar of the high-frequency indices IPC-S and IPC-FIPE is a bit slower compared to energy prices, though very quick for international standards.[7] Turning to the low-frequency indicators, data on IGP-10 and IPCA-15 arrive relatively early in the month, whereas INCC-M, IGP-M and IPA-M follow next before the end-of-month.

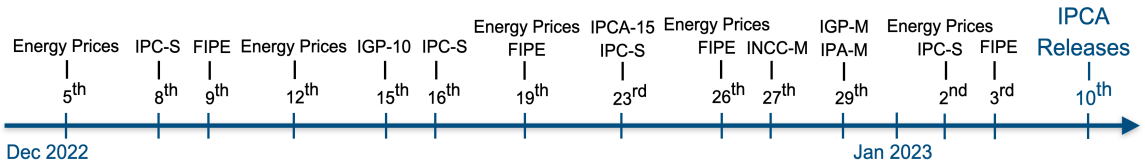


Figure 2: Timeline of data releases for price indicators in the reference period of December 2022.

The third group of predictors contains daily information from financial markets, including interest rates, commodity and stock price indices, exchange rates and credit default swaps. The choice of these variables is motivated by their relation to inflation expectations and findings in the literature on inflation forecasting. For example, Modugno (2013) and Breitung and Roling (2015) show that commodity and crude oil prices are among the most reliable indicators of inflation changes. Furthermore, central banks and practitioners have been monitoring financial variables on a daily basis to forecast the state of the macroeconomy (Andreou et al., 2013).

Finally, we include one daily series of experts' forecasts from the survey produced by the Central Bank of Brazil and published in the subsequent business day. To be more precise, this variable denotes the median of daily nowcasts for IPCA provided by the FOCUS survey of professional forecasters (SPF). There are over 100 active participants in the survey and they can answer, on a daily basis, what is their expectations regarding several price indexes and macroeconomic indicators. The median forecast of the IPCA is closely monitored for the market handout repport, and is released every Monday morning with data up until the previous Friday (Marques, 2012).

vantage that distribution and retail margins are fully accounted for.

[7] It happens because both the IPC-S and IPC-FIPE are based on a four-week collection system, ending on four set dates (07, 15, 22 and end-of-month). Thereby, computation of these indices considers the average of prices collected during the four weeks preceding the closing date.

Table 1: Database

Series	Mnemonic	Reference time span	Publication timing	Avg. delay	Starting date	Source
Target inflation variable						
Broad national CPI	IPCA	full month t	2nd week, following month	7	2003M1	IBGE
Monthly price indicators						
IPCA - extended	IPCA-15	16^{th}_{t-1} to 15^{th}_{t}	3rd/4th week, reporting month	8	2003M1	IBGE
General market CPI	IGP-M	21^{st}_{t-1} to 20^{th}_{t}	last week, reporting month	7	2003M1	FGV
General CPI - 10	IGP-10	11^{th}_{t-1} to 10^{th}_{t}	2nd/3rd week, reporting month	4	2003M1	FGV
Wholesale market PPI	IPA-M	21^{st}_{t-1} to 20^{th}_{t}	last week, reporting month	7	2003M1	FGV
National construction cost	INCC-M	21^{st}_{t-1} to 20^{th}_{t}	last week, reporting month	5	2003M1	FGV
Weekly price indicators						
FGV's CPI	IPC-S	four-week	1st day, following week	1	2003M2	FGV
Fipe's CPI	FIPE	four-week	2nd day, following week	2	2003M1	Fipe
Diesel prices	DIESEL	full week	1st day, following week	1	2004M5W2	ANP
Gasoline prices	GAS	full week	1st day, following week	1	2004M5W2	ANP
Ethanol fuel prices	ETOH	full week	1st day, following week	1	2004M5W2	ANP
Liquefied natural gas prices	LNG	full week	1st day, following week	1	2004M5W2	ANP
Daily financial variables						
Short-term interest rates	SELIC	end of day	real-time	0	2003M1	BCB
Brazilian Real/U$$ forex	FOREX	end of day	real-time	0	2003M1	BCB
Bovespa stock price index	IBOV	end of day	real-time	0	2003M1	B3
Electric utilities index	IEE	end of day	real-time	0	2003M1	B3
DI-rates (10Y maturity)*	DI10	end of day	real-time	0	2004M1	B3
DI-spread (10Y minus 3M)*	SPREAD	end of day	real-time	0	2004M1	B3
Brazil credit default swaps	CDS	end of day	real-time	0	2007M12D19	B3
Bloomberg commodity index	BCOM	end of day	real-time	0	2003M1	Bloomberg
Daily expectations from the FOCUS survey of professional forecasters						
IPCA nowcasts (median)	SPF	full day	subsequent day	1	2003M1	BCB

Note: This table reports the full list of time series selected for the nowcasting exercise. The reference time span relates to the data collection period. The publication timing provides the regular release calendar with respect to the reference period while the average delay stands for the publishing lags (in business days). The variables are not seasonally adjusted and transformed into month-on-month (MoM) % change in order to guarantee stationarity of the time series; the only exceptions are the interest rates series (SELIC, DI10 and SPREAD) which are transformed into monthly changes. MoM transformations for high-frequency variables consider the same reference week or day from the preceding month. *DI-rates are yields of Brazilian interbank deposit future contracts negotiated at B3.

3 Methodology

3.1 Nowcasting setup

Let the month-on-month inflation rate be denoted by y_t while $x_t^{(m)}$ represents a high-frequency macro-financial predictor that can be sampled m times more frequently than y_t. Moreover, denote a monthly price indicator by x_t, which is only available at the low-frequency and released before y_t. In this sense, time indices $t = 1, \dots, T$ act as the common monthly frequency between y_t and covariates $x_t^{(m)}$ and x_t.

Next, we convert daily financial predictors to the sampling rate of weekly predictors andwe assume a fixed month/week frequency ratio of $m = 4$, to focus on a single monthly/weekly mixture in order to avoid a proliferation of parameters due to a larger monthly/-daily frequency mismatch. Therefore, high-frequency predictors in the dataset convey the latest information available at days 8, 15, 22, and the end of each given month, such that frequency alignment with the target variable y_t is achieved. This particular choice of days allows us to control for the issues of overlapping weeks across months and an irregular number of days across different months. Hence, for each time period t, the information set also includes a set of m high-frequency observations $X_t^{(m)} = \left(x_t^{(m)}, x_{t-\frac{1}{m}}^{(m)}, x_{t-\frac{2}{m}}^{(m)}, \dots, x_{t-\frac{m-1}{m}}^{(m)} \right)'$, where $t - \frac{i}{m}$ denotes the i^{th} past high-frequency period with $i = 0, \dots, m-1$. In particular, end-of-month observations correspond to $x_t^{(m)}$, observations from day 22 refer to $x_{t-\frac{1}{m}}^{(m)}$, and so on up to day 8.

Building on these assumptions, we consider the baseline high-dimensional unrestricted MIDAS regression to nowcast the inflation rate on a weekly basis:

$$\phi(L)\, y_{t+h} = c + \sum_{k=1}^{K} B(L^{1/m})\, x_{k,t}^{(m)} + \sum_{j=1}^{J} \alpha_j\, x_{j,t} + \sum_{l=1}^{11} \gamma_l\, D_{l,t+h} + \varepsilon_{t+h}, \tag{1}$$

where $B(L^{1/m}) = \sum_{i=0}^{m-1} \beta_{k,i}\, L^{i/m}$ are unrestricted distributed lag coefficients that aggregate over the m contemporaneous high-frequency lags of predictor $x_{k,t}^{(m)}$, with $L^{i/m} x_t^{(m)} = x_{t-i/m}^{(m)}$.[8] The term $\phi(L) = (I - \rho_1 L)$ denotes the chosen autoregressive polynomial while dummies $D_{l,t+h}$ capture deterministic seasonal patterns. All in all, assuming a forecast horizon of $h = 0$, the baseline model (1) is a direct nowcasting tool for updating predictions

[8] Although we restrict both the low- and high-frequency components of (1) to account only for contemporaneous information of the predictor space, one might include lags of $\boldsymbol{x}_t$ and high-frequency distributed lags in $B(L^{1/m})$ that span over past low-frequency periods.

of the low-frequency target y_{t+h} as new observations of the high-frequency predictors come in.

The fairly small frequency mismatch with $m = 4$ suggests the adoption of the autoregressive distributed lag (ADL) MIDAS structure, but with unrestricted parameters (hereafter U-MIDAS, see Foroni et al., 2015; Ghysels and Marcellino, 2018).[9] This avoids the nonlinear temporal aggregation of high-frequency lags that characterizes a standard MIDAS regression and subsequently complicates estimation and prediction in a relatively high dimension (though penalized versions of MIDAS regressions have been put forward with empirical applications to GDP nowcasting, see Marsilli, 2014; Siliverstovs, 2017; Uematsu and Tanaka, 2019; Mogliani and Simoni, 2021; Babii et al., 2021).[10]

It is worth emphasizing that a particular low-frequency predictor $x_{j,t}$ only enters the model specification (1) when the corresponding data for month t has already been released by the time the nowcast is made . For instance, given that data on IPCA-15 is usually published between the 19$^{\text{th}}$ and 23$^{\text{rd}}$ of the reporting month, such predictor does not enter the baseline specification when nowcasts are made on days 8 and 15. In this sense, we constantly check for real-time data availability of low-frequency predictors by the time of the nowcast and adjust the model specification accordingly. Not adjusting the set of low-frequency predictors based on data availability by the time of the nowcast can lead to very imprecise nowcasts due to non-zero coefficients α_j being combined with non-contemporaneous data releases of $x_{j,t}$.

The baseline model (1) features $mK + J + 13$ parameters and thereby can easily lead to parameter proliferation as the number of high-frequency predictors increases. In the big data setting with many covariates, the effective sample size might be relatively short compared to the number of parameters, leading to high estimation uncertainty. To accommodate potentially large sets of covariates, we implement a broad range of machine learning techniques that apply some form of dimensionality reduction. We divide them into shrinkage methods - penalized regression schemes that shrink coefficient estimates towards zero - and tree-based methods that split the predictor space into a number of simple regions. The key ideas of these methods are outlined next.

[9] See Appendix A.1 for an explicit representation of the high-frequency component of (1) in matrix form.

[10] The MIDAS approach can efficiently address the dimensionality issue arising from the number of high-frequency lags in the model via tightly specified lag polynomials, but it is not suitable when the number of predictors is very large.

3.2 Overview of nowcasting models based on machine learning methods

The methods used to nowcast inflation can be categorized into two groups: shrinkage techniques and tree-based methods. In the first group we have Elastic Net and its two special cases: LASSO and Ridge.Additionally, we use the recently proposed sparse-group LASSO with MIDAS structure. As for the tree-based methods, we use the random forest, the generalized random forest, the local linear forest and the bayesian additive regression trees. In this section, we describe the models by means of their major characteristics. Table 2 provides an overview of the machine learning methods we implement in the nowcasting exercise.

Table 2: Summary of the models used in the paper

Model	Short name	Reference	R function (package)	Tuning parameters/Cross validation/Other criteria
Autoregressive model	AR		ar (stats)	lag order p chosen using AIC
Random Forest	RF	Breiman (2001)	randomForest (randomForest)	number of skip-sampled predictors to split the tree (mtry) using timeslice cross validation
Generalized Random Forest	GRF	Athey et al. (2019)	regression_forest (grf)	sample fraction, mtry, minimum node size, honesty fraction, honest leaves, alpha, imbalance using regular cross validation
Local Linear Forest	LLF	Friedberg et al. (2020)	ll_regression_forest (grf)	sample fraction, mtry, minimum node size, honesty fraction, honest leaves, alpha, imbalance using regular cross validation
Least absolute shrinkage and selection operator	LASSO	Tibshirani (1996)	glmnet (glmnet)	λ using timeslice cross validation
Elastic Net	EN	Zou and Hastie (2005)	glmnet (glmnet)	α, λ using timeslice cross validation
Ridge	Ridge	Hoerl and Kennard (1970)	glmnet (glmnet)	λ using timeslice cross validation
Bayesian Additive Regression Trees	BART	Chipman et al. (2012)	rbart (rbart)	200 trees, 1000 posterior simulations after burn in (100), d=0.95, probability of death = 0.7
Sparse Group LASSO	sg-LASSO	Babii et al. (2021)	cv.sgl.fit (midasml)	α, λ using timeslice cross validation

Note: Time slice cross validation (when used) starts with a 36 month window and subsequent 12 month fold slides.

Shrinkage methods

The main idea of shrinkage methods is to select the relevant predictors from a large matrix of covariates using different penalization schemes such that a higher forecasting precision is achieved (for empirical applications to inflation forecasting, see Garcia et al., 2017b; Joseph et al., 2021; Medeiros et al., 2021; Aliaj et al., 2023, among others). These approaches are particularly helpful in data-rich environments where the dimension of the predictor space is large compared to the sample size. In this regard, shrinkage ensures that forecasting under a high-dimensional mixed-frequency setting such as (1) becomes feasible. Among shrinkage strategies for linear models, the least absolute shrinkage and selection operator (LASSO), Ridge and Elastic Net regression enjoy a growing popularity within economics.

To set the scene for penalized regressions, denote by $\mathbf{y} = (y_1, \ldots, y_t)'$ the target inflation

sequence while the high-frequency predictor space is given by $\boldsymbol{x^{(m)}} = (\boldsymbol{x_1^{(m)}}, \dots, \boldsymbol{x_t^{(m)}})'$ with $\boldsymbol{x_t^{(m)}} = (X_{1,t}^{(m)'}, \dots, X_{K,t}^{(m)'})'$ being the $mK \times 1$ vector of high-frequency data sampled across the m increments of time t. Moreover, let $\boldsymbol{x} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_t)'$ denote the low-frequency predictor space with $\boldsymbol{x}_t$ being the J-dimensional vector of low-frequency data. The baseline matrix of covariates is then given by $\mathbf{X} = (\iota, \boldsymbol{y}_{-1}, \boldsymbol{x}^{(m)}, \boldsymbol{x}, \boldsymbol{d})$, where ι is a t-dimensional vector of ones, $\boldsymbol{y}_{-1}$ is the first lag of $\mathbf{y}$ and $\mathbf{d}$ is a $t \times 11$ matrix of seasonal deterministic dummies.

Given the notation above, the Elastic Net estimator solves the penalized least-squares problem:

$$\hat{\beta} = \min_{\hat{\beta}} ||\mathbf{y} - \mathbf{X}\beta||^2 + \lambda \left(\alpha \left|\beta\right|_1 + \frac{(1-\alpha)}{2} ||\beta||^2 \right), \tag{2}$$

where $\alpha \in (0, 1]$ is a weight parameter that interpolates between LASSO ($\alpha = 1$) and Ridge regression (as $\alpha \to 0$). Hence, LASSO penalizes the sum of absolute coefficients via the shrinking penalty using the ℓ_1 norm while Ridge penalizes the sum of squared coefficients via the ℓ_2-norm. The regularization parameter λ controls the amount of shrinkage in the parameter space β.[11] Hence, estimator (2) shrinks coefficients of irrelevant predictors towards zero or are set exactly to zero. The latter constitutes a LASSO-type shrinkage that performs variable selection and results in a sparse and parsimonious model. On the other hand, coefficients estimated via Ridge regression never become exactly zero, which yields a dense model. All in all, the key benefit comes from achieving a substantial reduction in forecast variances at the cost of a slight increase in bias.

Babii et al. (2021) argue that high-dimensional mixed-frequency representations, such as (1), involve certain data structures that once taken into account should lead to increased performance out-of-sample. These structures relate to groups covering the m relevant lags of a single high-frequency covariate. To that end, the authors leverage on the sparse-group LASSO (sg-LASSO) estimator and show that it selects not only the relevant groups of predictors for the low-frequency target but also the appropriate lag structure within each group. This structured sparsity is the attractive feature of sg-LASSO that aims to improve upon the unstructured LASSO, which does not recognize serial dependence across different high-frequency lags and thereby may be subject to random selection. Zhao and Yu (2006) prove that LASSO selects the true model consistently if and (almost) only if the irrelevant covariates are not highly correlated with the predictors in the true model; a rather strong condition denominated "irrepresentable condition". In penalized U-MIDAS regressions, multicollinearity emerges from the highly correlated unrestricted lags of a given high-frequency covariate $x_{k,t}^{(m)}$. This implies that the unstructured LASSO randomly picks among those strongly correlated lags of $x_{k,t}^{(m)}$, leaving most of the remaining lag coefficients

[11] Note that as $\lambda \to \infty$, the impact of the shrinkage penalty grows. Tuning of parameter λ is critical, and hereby performed in a data-driven way using cross-validation (see Section 3.3).

shrunk to zero.

To describe the estimation procedure of sg-LASSO, let the matrix of covariates now be defined as $\mathbf{X}^L = (\iota, \boldsymbol{y}_{-1}, \boldsymbol{X}^{(m)}, \boldsymbol{x}, \boldsymbol{d})$, where $\boldsymbol{X}^{(m)} = (\boldsymbol{X}_1^{(m)}W, \ldots, \boldsymbol{X}_K^{(m)}W)$ and $\boldsymbol{X}_k^{(m)} = (X_{k,1}^{(m)}, \ldots, X_{k,t}^{(m)})'$ is the $t \times m$ matrix of high-frequency series of the k^{th} predictor for $k = 1, \ldots, K$. The predetermined $m \times L$ matrix of weights W is based on orthogonal Legendre polynomials of degree L that aggregate over the high-frequency lags. For example, the Legendre polynomial of order $L = 0$ attributes same weights to all lags, the $L = 1$ polynomial is an increasing linear function and thereby favours more distant lags, the $L = 2$ polynomial features higher weights to very recent and more distant lags, and so on. These predetermined weights essentially prevent us from paying the price of overparameterization when m is relatively large.

The sg-LASSO estimator then solves the penalized least-squares problem:

$$\hat{\beta} = \min_{\hat{\beta}} ||\mathbf{y} - \mathbf{X}^L\beta||^2 + 2\lambda \left(\alpha |\beta|_1 + (1-\alpha) ||\beta||_{2,1}\right), \tag{3}$$

where $||\beta||_{2,1} = \sum_{G\in\mathcal{G}} |\beta_G|_2$ is the group LASSO norm for a group structure $\mathcal{G}$ that hereby constitutes all lags of a single high-frequency covariate.[12] This implies that sg-LASSO promotes sparsity between and within groups.

Tree-based methods

Decision trees are nonparametric models that work by recursively dividing the covariate space $\mathbb{X}$ into $Q \in \mathbb{N}$ separable regions according to a (pre-determined) splitting rule. Since a decision tree is a collection of dichotomous splittings, the interpretation of predictions from trees is straightforward. However, due to its simplistic structure, single trees are subject to overfitting.

First proposed by Breiman (2001), random forests (RF) are an extension of decision trees in which the results from several non-correlated (or with very small correlation) trees randomly chosen are gathered to form a prediction. The predictions of the trees in a forest are averaged, in such a way that decreases the variance of the final predictions while maintaining the flexibility of the trees. Specifically, for a random forest with B trees, the prediction is given by

$$\hat{y}(\tilde{\boldsymbol{x}}_m) = \frac{1}{B}\sum_{b=1}^{B} \hat{y}_b(\tilde{\boldsymbol{x}}_m), \tag{4}$$

[12] Note that $\alpha \in [0, 1]$ determines the relative importance of LASSO-sparsity and the group structure, whereas $\alpha = 0$ leads to the group LASSO estimator.

where $\hat{y}_b(\tilde{\boldsymbol{x}}_m)$ is the prediction of the b-th tree. RF can deal with high dimensional data without suffering from the curse of dimensionality, but in comparison to a single tree, the forests lack interpretability (James et al., 2013). Nonetheless, random forests have shown superior forecast ability in comparison to other machine learning and traditional econometric models when used to forecast inflation (see Medeiros et al., 2021; Araujo and Gaglianone, 2022, among others).

The Generalized Random Forest (GRF), proposed by Athey et al. (2019), is an enhanced version of the RF. It is a two-step procedure, in which first a random forest is used to generate weights that are later used in a GMM step - weights for points in the same leaf will be higher and thus these points carry more weight in the estimation step Athey and Imbens (2019). To find the weights start from (4):

$$\begin{aligned}\hat{y}\left(\tilde{\boldsymbol{x}}_m\right) &= \frac{1}{B}\sum_{b=1}^{B}\left[\sum_{k=1}^{K_b}\beta_{k,b}\mathbb{1}_{\tilde{\boldsymbol{x}}_m\in\mathcal{J}_{k,b}}\right] = \frac{1}{B}\sum_{b=1}^{B}\sum_{\boldsymbol{x}_i\in\mathcal{J}_b(\tilde{\boldsymbol{x}}_m)}\frac{y_i}{|\mathcal{J}_b(\tilde{\boldsymbol{x}}_m)|} \\ &= \frac{1}{B}\sum_{b=1}^{B}\sum_{i=1}^{n}\frac{y_i\mathbb{1}_{\boldsymbol{x}_i\in\mathcal{J}_b(\tilde{\boldsymbol{x}}_m)}}{|\mathcal{J}_b(\tilde{\boldsymbol{x}}_m)|} = \sum_{i=1}^{n}\alpha_i(\tilde{\boldsymbol{x}}_m)y_i, \end{aligned} \tag{5}$$

where $\mathbb{1}_{\tilde{\boldsymbol{x}}_m\in\mathcal{J}_{k,b}}$ is a indicator function that denotes that $\tilde{\boldsymbol{x}}_m$ belongs to the region $\mathcal{J}_k$ in tree b, and $|\cdot|$ denotes the cardinality of a set. The term $\alpha_i(\tilde{\boldsymbol{x}}_m)$ in (5) is called *forest weight* and denotes the fraction of trees that allocates $\tilde{\boldsymbol{x}}_m$ in the same leaf as the covariate vector $\boldsymbol{x}_i$, and is given by

$$\alpha_i\left(\tilde{\boldsymbol{x}}_m\right) = \frac{1}{B}\sum_{b=1}^{B}\frac{\mathbb{1}_{\boldsymbol{x}_i\in\mathcal{J}_b(\tilde{\boldsymbol{x}}_m)}}{|\mathcal{J}_b(\tilde{\boldsymbol{x}}_m)|}. \tag{6}$$

In Equation (5), the regression forest will assign higher weights to sample points closer to $\tilde{\boldsymbol{x}}_m$ since the prediction is an average over a set of trees. The forests can adapt the weights, such that a covariate that has little relation with y_i will appear less frequently when making splits (Athey et al., 2019). Athey and Imbens (2019) argue that trees can be seen as a way to find weights for the new observation $\tilde{\boldsymbol{x}}_m$, based on the neighbor estimation sample points that fall in the same leaf as $\tilde{\boldsymbol{x}}_m$.

Random forests have characteristics that are a potential issue for macroeconomic forecasting. First, forests inherit the piecewise output from the trees, producing a discontinuous response, so they cannot explore well smoothness in the DGP. Forests' predictions also have a bias for points that are too close to the rectangle's boundaries (Athey and Imbens, 2019). To tackle these limitations, Friedberg et al. (2020) proposed the Local Linear Forest (LLF) model. Similar to the GRF, the LLF also uses RF as a weight generator, but in the second step instead of a GMM regression, one estimates a local linear regression. Specifically, $y(\tilde{\boldsymbol{x}}_m)$ will be the local average, which can be estimated together with a $\theta(\tilde{\boldsymbol{x}}_m)$ through the

following optimization problem:

$$\begin{pmatrix} \hat{y}\left(\tilde{\boldsymbol{x}}_m\right) \\ \hat{\theta}\left(\tilde{\boldsymbol{x}}_m\right) \end{pmatrix} = \underset{y,\theta}{\arg\min} \left\{ \sum_{i=1}^{n} \alpha_i\left(\tilde{\boldsymbol{x}}_m\right) \left(y_i - y\left(\tilde{\boldsymbol{x}}_m\right) - \left(x_i - \tilde{\boldsymbol{x}}_m\right) \theta\left(\tilde{\boldsymbol{x}}_m\right)\right)^2 + \lambda \left\|\theta\left(\tilde{\boldsymbol{x}}_m\right)\right\|_2^2 \right\}. \tag{7}$$

In (7), the term $\hat{y}\left(\tilde{\boldsymbol{x}}_m\right)$ still a prediction for a new point, but with the slope of the local linear regression $\theta\left(\tilde{\boldsymbol{x}}_m\right)$, which corrects for the local trend in $x_i - \tilde{\boldsymbol{x}}_m$. For the local linear forest, the parameter θ is not of interest and predictions are based on the intercept $\hat{y}\left(\tilde{\boldsymbol{x}}_m\right)$. The penalization term $\lambda \left\|\theta\left(\tilde{\boldsymbol{x}}_m\right)\right\|_2^2$ has the role of avoiding overfitting to the local trend and λ is typically chosen via cross-validation. As result, the LLF can approximate well smooth functions as a local regression without becoming infeasible when the number of covariates grows.

The last tree-based model that we use in this work is the bayesian additive regression trees (BART), from Chipman et al. (2012). Like in the RF, BART predictions are also based on the results from several trees, but unlike RF, the trees in BART will be sequentially estimated using as dependent variable the residuals from the previous tree. In general terms, each Bayesian (regression) tree is defined by $\boldsymbol{\mathcal{T}}$, a collection of interior nodes; and $\boldsymbol{\mathcal{M}}$ a set of parameter values that are associated with the terminal nodes. The set $\boldsymbol{\mathcal{T}}$ is also called *tree* structure and contains the information on the topology of the trees: whether a node is terminal or not and how to make splits in non-terminal nodes.

A BART defines a function $g(\boldsymbol{x}, \boldsymbol{\mathcal{T}}, \boldsymbol{\mathcal{M}})$ which maps a row $\boldsymbol{x}$ (from the covariate matrix $\boldsymbol{X}$) to a particular $\theta_j \in \mathcal{M},\ j \in 1, \ldots, |\boldsymbol{\mathcal{M}}|$. Predictions in BART are obtained by sampling from the posterior distribution.In this paper we follow closely the prior specification from Chipman et al. (2012), by choosing the variable for a split using a uniform prior, as well as the cutpoint for the split. We use a conjugate normal prior for the predictions on the terminal nodes and a conjugate inverse χ^2-square for the (constant) error term of the model. Finally, the probability of growing another layer in a tree is given by $\alpha(1+d)^{-\beta}$, where d is the current depth of the tree, and $\alpha \in (0,1)$ and $\beta \in \mathbb{R}^+$ are hyperparameters.

3.3 Tuning of hyperparameters and time series cross-validation

Selection of the regularized regression model parameters λ and α is done via rolling cross-validation. Different from the standard cross-validation procedure in which folds are randomly selected assuming that observations are iid, the rolling cross-validation will take into consideration the time series structure of the data and avoid using future observations to forecast values in the past. In practice, this means that the cross-validation procedure takes place sequentially (time-slice cross-validation). In our empirical exercise, we start with a three-year initial fixed window with folds of one year (Arlot and Celisse, 2010;

Goulet Coulombe et al., 2022; Bergmeir et al., 2018).

For the RF, we tuned via rolling cross-validation the number of covariates to be used at each node in the trees ($mtry$). For the GRF and LLF we used (the standard) cross-validation procedure to find the optimal values of the minimum number of observations in a terminal node, the sample fraction that should be used for estimating a single tree, $mtry$. Furthermore, we also used CV to determine the honesty fraction (what fraction of the sample should be used in the estimation and prediction steps) and whether empty terminal nodes should be eliminated. Finally, α and imbalance penalty are hyperparameters that control the relative size of child and parent nodes with respect to the number of observations that they have assigned.

For BART, we follow closely the recommendations from Chipman et al. (2012). We estimate 200 trees and 1000 posterior draws, with 100 draws as burn-in. For the tree structure, we use $\alpha = 0.95$ and $\beta = 2$, which penalizes bigger trees. For the variance term we assume a unitary prior and for the normally-distributed predictions, we center the prior at zero.

4 Results & Discussion

4.1 Real-time nowcasting exercise

To compute weekly nowcasts for month-on-month inflation rates, we use a sample from January 2003 to December 2022 with $T = 240$ monthly observations based on data availability of high-frequency macro-financial indicators. We run a recursive out-of-sample exercise with rolling windows that span over 120 months such that January 2013 marks the start of the evaluation period. This is based on the fact that publishing dates of price indicators are mostly available from 2013. Moreover, predictors with missing data at the beginning of the sample only enter the set of covariates by the time a full series of 120 observations gets available.

We update our nowcasts on a set of fixed days. Namely, days 8, 15, 22 and end-of-month using the most recent increments of low- and high-frequency data. In this sense, the continuous process of updating a mixed-frequency dataset leads to missing weekly observations of $\boldsymbol{x}_t^{(m)}$ at the end of the sample. This missing data that emerge when nowcasting before the end-of-month generates a time-evolving sample's "ragged edges".[13] To complete

[13] For example, when nowcasting the target inflation on the 15$^{\text{th}}$ of January, we already have up-to-date information from financial variables and inflation expectations while the corresponding data on high-frequency price indicators might only be released in the following week. In the former case, a ragged edge

the weekly dataset, we impute random walk forecasts based on vintages of high-frequency data. Finally, we transform the implied month-on-month predicted rates into forecasts of year-on-year inflation rates to assess Root Mean Squared Errors (RMSE) and Mean Absolute Errors (MAE). However, we also investigate how relative forecast performance changes over time by means of cumulative loss differentials and the fluctuation test of Giacomini and Rossi (2010).

To benchmark our nowcasts based on machine learning techniques, we employ two alternative settings. First, we benchmark our results against SPF's inflation expectations (median across experts' nowcasts). Second, we apply the AR(p) forecast with seasonal deterministic dummies:

$$y_t = c + \sum_{j=1}^{p} \rho_j \, y_{t-j} + \sum_{l=1}^{11} \gamma_l \, D_{l,t+h} + \varepsilon_{t+h}, \tag{8}$$

which captures the autoregressive dynamics of the target variable and often provides competitive forecasts when compared to approaches purely based on the low-frequency flow of information. The number of autoregressive lags p is here set to be a maximum of four and selected using the Akaike Information Criterion (AIC).

4.2 Out-of-sample results

Table 3 reports the forecasting results of competing models in terms of RMSE and MAE relative to the SPF benchmark. The results clearly show that shrinkage methods generally perform better than SPF nowcasts, especially at early month horizons, meaning that penalized regressions better translate the predictive content of the selected predictors into precise nowcasts for the IPCA target. To be precise, the sg-LASSO outperforms in terms of RMSE when nowcasts are made on days 8 and 15 while LASSO beats the benchmark on days 22 and end-of-month.[14] Moreover, nowcasting gains become smaller for horizons approaching the end-of-month, and thus relative performance increases with the horizon. Overall, the average predictive gains of shrinkage-based methods, excluding Ridge, amount to 5% compared to experts' forecasts, which is indeed a tough competitor to beat. On the other hand, tree-based methods are not able to outperform SPF nowcasts. These methods are generally beaten by a considerable amount, even though slightly worse nowcasts might be produced at early month horizons.

emerges for days 22 and end-of-month while the latter gives rise to a ragged edge for day 15 as well.

[14] These overall results also hold in terms of MAE. We hereby do not document the robustness analysis, but nowcasting results have shown to be robust to alternative model specifications and shorter time series windows for both estimation and rolling cross-validation.

Table 3: RMSE and MAE for all out-of-sample periods

Metric	Horizon	SPF	AR	RW	RF	GRF	LLF	LASSO	EN	Ridge	BART	sgLASSO
RMSE	day 8	1	1,33	1,81	1,05	1,09	1,07	0,95	0,95	0,95	1,02	0,92
RMSE	day 15	1	1,72	1,90	1,21	1,26	1,25	0,98	0,98	1,03	1,17	0,96
RMSE	day 22	1	2,35	2,59	1,39	1,53	1,50	0,94	0,94	0,99	1,34	0,99
RMSE	end-of-month	1	2,84	3,13	1,55	1,77	1,79	0,95	0,95	1,03	1,48	0,99
MAE	day 8	1	1,38	1,91	1,05	1,09	1,09	0,96	0,96	0,98	1,04	0,94
MAE	day 15	1	1,74	1,93	1,21	1,24	1,26	0,99	0,99	1,06	1,16	0,97
MAE	day 22	1	2,38	2,63	1,36	1,47	1,45	0,98	0,98	1,02	1,32	0,98
MAE	end-of-month	1	2,87	3,17	1,53	1,71	1,73	1,01	1,02	1,11	1,46	1,06

Note: The table reports the RMSE and MAE for each competing model relative to SPF nowcasts. Smaller values for each metric and nowcasting horizon (days 8, 15, 22 and end-of-month) are indicated by gray-shaded cells.

The findings in Table 3 prompt the question whether relative forecasting errors are constant throughout the evaluation period or largely affected by unusual events. To that end, we report the cumulative sum of squared forecast error (CUMSFE) of model-based nowcasts versus the SPF benchmark for nowcasts made on days 8, 15, 22 and end-of-month. The CUMSFE is given by:

$$\text{CUMSFE}_{t_0,t_1} = \sum_{t=t_0}^{t_1} e^2_{t,M_1} - e^2_{t,M_2} \tag{9}$$

for a benchmark model M_1 (hereby, SPF) versus M_2. A positive value of CUMSFE_{t_0,t_1} indicates an outperformance of M_2 from t_0 up to t_1 in comparison with SPF nowcasts while negative values imply the opposite.

Figure 3 clearly shows that the inflationary period following the Covid-19 pandemic is a game changer in terms of loss differentials. In general, differences in predictive accuracy between shrinkage methods and SPF nowcasts are modest throughout the years preceding the pandemic while large forecasting gains build up from September, 2020. During this period of persistent high inflation, we observe the largest jumps in CUMSFE for nowcasts made on days 8 and 15 using sg-LASSO, but moderate gains are also achieved on days 22 and end-of-month using LASSO or Elastic Net, which mainly drive the results in Table 3. Moreover, early month nowcasts lead to higher CUMSFE by the end of 2022 and these values decrease as we move towards end-of-month horizons. On the other hand, experts' nowcasts are able to consistently outperform Ridge during calm times, meaning that variable selection is key to obtain good quality nowcasts across weekly horizons.

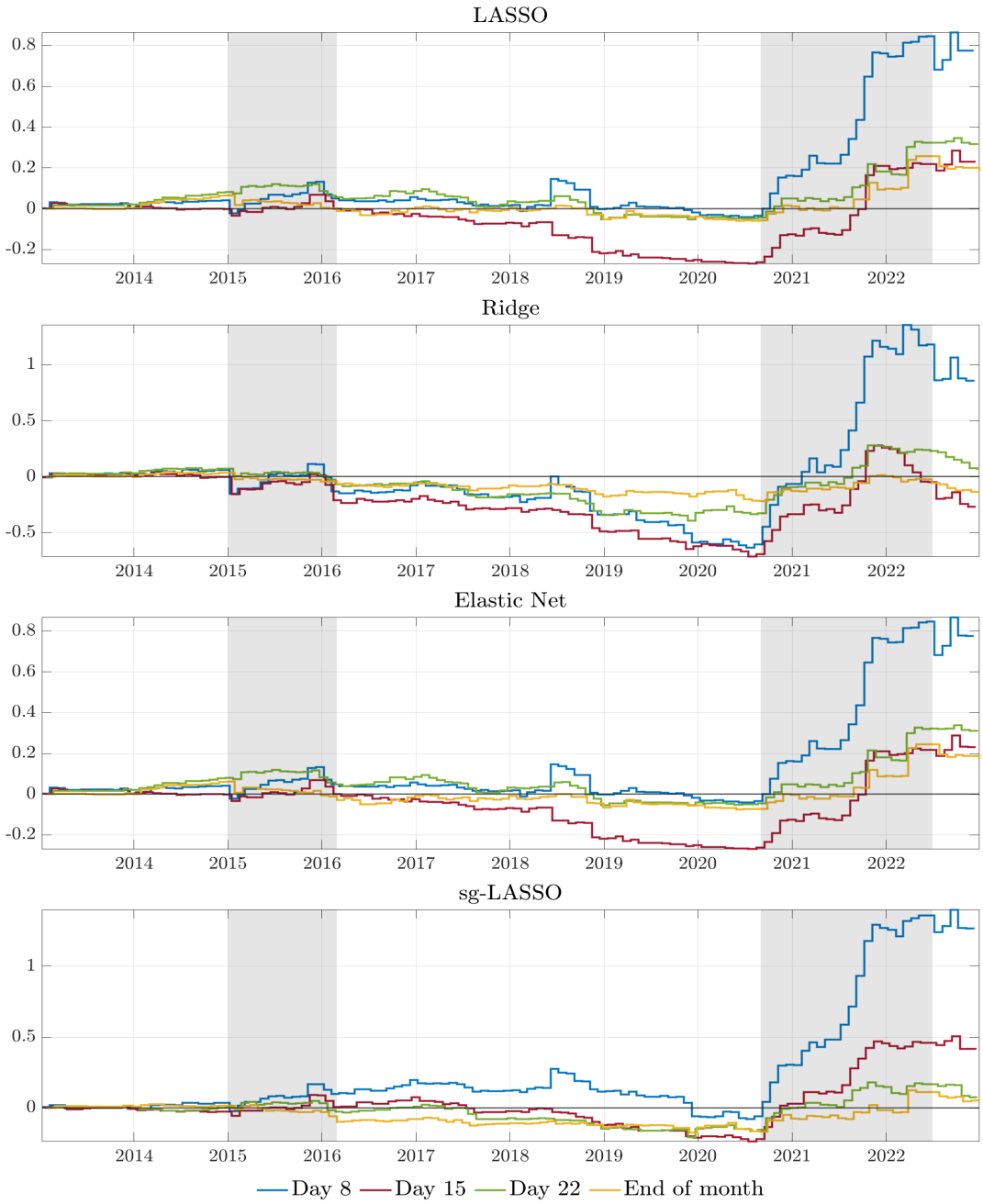


Figure 3: Cumulative sum of loss differentials (CUMSFE) of regularization MIDAS nowcasts (LASSO, Ridge, Elastic Net and sg-LASSO) versus the survey of professional forecasters (SPF, median) on days 8, 15, 22 and end-of-month. The gray shaded areas correspond to rising inflation periods.

These findings on relative performance over time are consistent to results of the fluctuation test in Figure 4. Forecasting gains in comparison to SPF inflation expectations change substantially over time, depending on the model and horizon, nevertheless, improvements in nowcasting accuracy become crystal-clear in the aftermath of the pandemic. The sg-LASSO gains on days 8 and 15 are occasionally significant at the 10% level throughout 2021. Similarly, LASSO forecasts produce a few significant smaller losses in 2021 at horizons closer to the end-of-month. In fact, the picture reveals a clear discrepancy between shrinkage- and tree-based models, as expected from previous results. On top of that, a higher dispersion of prediction accuracy across models can be observed during turbulent times such as the 2014-15 Brazilian economic crisis and the pandemic. At the same time, model-based and SPF nowcasts are statistically equivalent during normal times and typically deliver loss differentials close to zero.

Finally, we investigate the relative importance of the selected predictorsby means of coefficient estimates for each month throughout the evaluation period. In Appendix A.3, we present a heatmap of these period-wise coefficient estimates for sg-LASSO at each horizon. That is, for each one of the four nowcasting moments there is a different heatmap, in which dates are displayed in the x axis and covariates are displayed in the y axis. The numerical value of a coefficient associated with a given variable for a specific date in the sample period will determine the intensity of color in the graph. This gives a view of the evolution of coefficients across all the sample periods and between different nowcasting moments.

Comparing all panels in Figure A.3, we observe that sg-LASSO prompts a fairly sparse structure at early month horizons while a more dense structure prevailing at late month horizons stems from a higher data availability of low-frequency price indicators. In the former case (Figures A2a and A2b), SPF inflation expectations, high-frequency price indicators, and the lagged IPCA(second variable from top to bottom, with a negative sign) are the most relevant variables. At the same time, energy prices and financial variables regularly enter the forecasting model, though with modest coefficients. When it comes to horizons approaching end-of-month, the low-frequency but timely indicator IPCA-15 reveals an enormous importance via coefficient estimates that amount to 0.6 in many cases. On the other hand, SPF inflation expectations now lose a big portion of their relevance. One hypothesis is that professional forecasters adapt their survey responses to the release of this indicator. As for the fuel prices, they do not have an associated coefficient in all nowcasting moments at the beginning of the out-of-sample period, until the first months of 2014. The reason for that is the availability of the data (see Table 1): in order to have a balanced panel, variables only enter in the out-of-sample exercise when they have at least 120 past observations for estimation of the models. The same happens with the credit default swap (CDS), DI-rate and DI-spread.

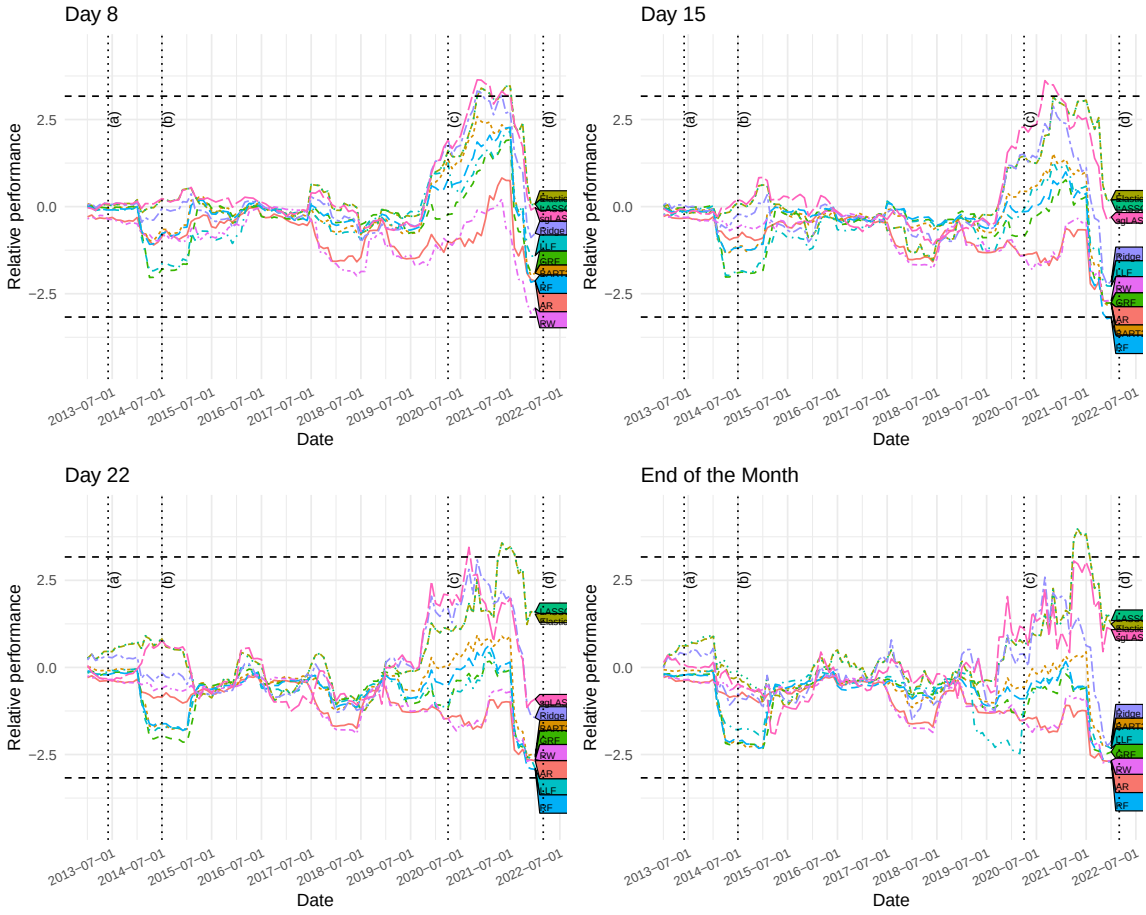


Figure 4: Fluctuation test from Giacomini and Rossi (2010), comparing the squared difference between forecasts and observed values. Each line corresponds to the test statistic comparing a specific machine learning method and SPF nowcasts, and graphs are separated by the corresponding day of the nowcast. Areas between the horizontal dashed lines correspond to the 90% confidence interval of the two-sided test. We used as window parameters of the test $\mu = 0.1$ and five for the number of lags in the variance of the DM test. The dashed vertical lines indicate relevant periods in the Brazilian economy: (a) refers to the civil protests in 2013, (b) is the Brazilian economic crisis that led to predident Rouseff impeachment, (c) is the corona crisis, and (d) is the invasion of Ukraine.

5 Concluding remarks

We have tested many machine learning methods for nowcasting inflation based on high-frequency macro and financial indicators within a Mixed-Data Sampling (MIDAS) structure. Specifically, we compared Shrinkage regression methods and tree-based algorithms forecasting performance in an out-of-sample exercise comprising 120 months using data from Brazil.

Results indicate that linear models with shrinkage and variable selection done via the LASSO outperform tree-based methods in terms of RMSE and MAE. This result is in line with the findings from Joseph et al. (2021). It seems that for shorter data such as macroeconomic time series together with the short forecasting span do not favor the non-linear characteristics of the trees. Fluctuation test shows that the LASSO performs significantly better than the FOCUS survey of professional forecasts in the later period of rising inflation following the Covid-19 pandemic.

Variable selection via the analysis of the coefficients from the sparse group LASSO confirms the relevance of higher frequency indicators. At the beginning of the month, when SPF predictions tend to be worse, the machine learning methods produce better forecasts because they can explore the early information available in real time coming from the weekly data. In the last two weeks of the month, the model and survey forecasts tend to become closer, and low-frequency variables tend to gain higher coefficients. The differences in the releases of the low-frequency indicators, namely the fact that monthly indicators are released at different weeks within the reference month, led us to adapt the estimation strategy. At each week within a month, we reestimated the model's coefficients, in a way that only variables that have been already released in time t would be used for estimating the model at that point in time.

Taking together, the above results suggest that having a rich set of different price indexes combined with survey data within a MIDAS structure seems to allow a rapid detection of inflationary shocks. Furthermore, for monthly nowcasts, it is important to take into account during the estimation procedure the data release differences in the low-frequency variables.

References

Aliaj, T., Ciganovic, M. and Tancioni, M. (2023), 'Nowcasting inflation with lasso-regularized vector autoregressions and mixed frequency data', Journal of Forecasting .
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1002/for.2944*

Andreou, E., Ghysels, E. and Kourtellos, A. (2010), 'Regression models with mixed sampling frequencies', Journal of Econometrics **158**(2), 246–261.

Andreou, E., Ghysels, E. and Kourtellos, A. (2013), 'Should macroeconomic forecasters use daily financial data and how?', Journal of Business & Economic Statistics **31**(2), 240–251.

Araujo, G. S. and Gaglianone, W. P. (2022), 'Machine learning methods for inflation forecasting in Brazil: new contenders versus classical models'.

Arlot, S. and Celisse, A. (2010), 'A survey of cross-validation procedures for model selection'.

Athey, S. and Imbens, G. W. (2019), 'Machine learning methods that economists should know about', Annual Review of Economics **11**, 685–725.

Athey, S., Tibshirani, J. and Wager, S. (2019), 'Generalized random forests', The Annals of Statistics **47**(2), 1148–1178.

Babii, A., Ghysels, E. and Striaukas, J. (2021), 'Machine learning time series regressions with an application to nowcasting', Journal of Business & Economic Statistics **40**(3), 1094–1106.

Bańbura, M., Giannone, D. and Reichlin, L. (2012), 'Nowcasting, in Michael P. Clements, and David F. Hendry, ed.: The Oxford handbook of economic forecasting', pp. 193–224.

Bergmeir, C., Hyndman, R. J. and Koo, B. (2018), 'A note on the validity of cross-validation for evaluating autoregressive time series prediction', Computational Statistics & Data Analysis **120**, 70–83.

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

Breitung, J. and Roling, C. (2015), 'Forecasting inflation rates using daily data: A nonparametric MIDAS approach', Journal of Forecasting **34**(7), 588–603.

Carriero, A., Clark, T. E. and Marcellino, M. (2020), 'Nowcasting tail risks to economic activity with many indicators', Federal Reserve Bank of Cleveland Working Paper (No.20-13).

Chipman, H. A., George, E. I. and McCulloch, R. E. (2012), 'Bart: Bayesian additive regression trees', Annals of Applied Statistics **6**(1), 266–298.

Clark, T. E., Leonard, S., Marcellino, M. and Wegmüller, P. (2022), 'Weekly nowcasting us inflation with enhanced random forests'.

Evans, M. D. (2005), 'Where are we now? Real-time estimates of the macroeconomy', International Journal of Central Banking .

Faust, J. and Wright, J. H. (2013), Forecasting inflation, in 'Handbook of economic forecasting', Vol. 2, Elsevier, pp. 2–56.

Foroni, C., Marcellino, M. and Schumacher, C. (2015), 'Unrestricted mixed data sampling (MIDAS): MIDAS regressions with unrestricted lag polynomials', Journal of the Royal Statistical Society: Series A **178**(1), 57–82.

Friedberg, R., Tibshirani, J., Athey, S. and Wager, S. (2020), 'Local linear forests', Journal of Computational and Graphical Statistics **30**(2), 503–517.

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

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

Ghysels, E. and Marcellino, M. (2018), Applied economic forecasting using time series methods, Oxford University Press.

Ghysels, E., Santa-Clara, P. and Valkanov, R. (2004), 'The MIDAS touch: Mixed data sampling regression models', Discussion paper UNC and UCLA .

Ghysels, E., Sinko, A. and Valkanov, R. (2007), 'Midas regressions: Further results and new directions', Econometric reviews **26**(1), 53–90.

Giacomini, R. and Rossi, B. (2010), 'Forecast comparisons in unstable environments', Journal of Applied Econometrics **25**(4), 595–620.

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

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

Hoerl, A. E. and Kennard, R. W. (1970), 'Ridge regression: applications to nonorthogonal problems', Technometrics **12**(1), 69–82.

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013), An introduction to statistical learning, Springer.

Joseph, A., Kalamara, E., Kapetanios, G., Potjagailo, G. and Chakraborty, C. (2021), 'Forecasting UK inflation bottom up'.

Knotek, E. S. and Zaman, S. (2017), 'Nowcasting us headline and core inflation', Journal of Money, Credit and Banking **49**(5), 931–968.

Marques, A. B. C. (2012), 'Central Bank of Brazil's market expectations system: a tool for monetary policy', IFC Bulletin **36**, 304–324.

Marsilli, C. (2014), Variable selection in predictive midas models, Technical report, Banque de France.

Medeiros, M. C., Vasconcelos, G. F., Veiga, Á. and 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.

Modugno, M. (2013), 'Now-casting inflation using high frequency data', International Journal of Forecasting **29**(4), 664–675.

Mogliani, M. and Simoni, A. (2021), 'Bayesian midas penalized regressions: estimation, selection, and prediction', Journal of Econometrics **222**(1), 833–860.

Siliverstovs, B. (2017), 'Short-term forecasting with mixed-frequency data: a midasso approach', Applied Economics **49**(13), 1326–1343.

Simon, N., Friedman, J., Hastie, T. and Tibshirani, R. (2013), 'A sparse-group lasso', Journal of computational and graphical statistics **22**(2), 231–245.

Tibshirani, R. (1996), 'Regression shrinkage and selection via the lasso', Journal of the Royal Statistical Society: Series B (Methodological) **58**(1), 267–288.

Uematsu, Y. and Tanaka, S. (2019), 'High-dimensional macroeconomic forecasting and variable selection via penalized regression', The Econometrics Journal **22**(1), 34–56.

Zhao, P. and Yu, B. (2006), 'On model selection consistency of lasso', The Journal of Machine Learning Research **7**, 2541–2563.

Zou, H. and Hastie, T. (2005), 'Regularization and variable selection via the elastic net', Journal of the royal statistical society: series B (statistical methodology) **67**(2), 301–320.

Appendix

A.1 Mixed-frequency framework in matrix form

For expositional simplicity, let us reduce the general multiple-predictors case of the baseline model (1) to the one-predictor case for both the low- and high-frequency components and neglect seasonal dummies. From there, assume the latest data release for the target inflation has been released for a given month t. Based on the information set available up to t and a pre-sample observation y_0, our baseline U-MIDAS structure, with a single high-frequency predictor $x_{k,t}^{(m)}$ and low-frequency predictor $x_{j,t}$, can be estimated following the matrix representation

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_t \end{bmatrix} = \begin{bmatrix} 1 & y_0 & x_{k,1}^{(m)} & x_{k,1-\frac{1}{m}}^{(m)} & x_{k,1-\frac{2}{m}}^{(m)} & x_{k,1-\frac{3}{m}}^{(m)} & x_{j,1} \\ 1 & y_1 & x_{k,2}^{(m)} & x_{k,2-\frac{1}{m}}^{(m)} & x_{k,2-\frac{2}{m}}^{(m)} & x_{k,2-\frac{3}{m}}^{(m)} & x_{j,2} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & y_{t-1} & \underbrace{x_{k,t}^{(m)}}_{\text{end-of-month}} & \underbrace{x_{k,t-\frac{1}{m}}^{(m)}}_{\text{day 22}} & \underbrace{x_{k,t-\frac{2}{m}}^{(m)}}_{\text{day 15}} & \underbrace{x_{k,t-\frac{3}{m}}^{(m)}}_{\text{day 8}} & x_{j,t-1} \end{bmatrix} \begin{bmatrix} c \\ \rho_1 \\ \beta_{k,1} \\ \beta_{k,2} \\ \beta_{k,3} \\ \beta_{k,4} \\ \alpha_j \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_t \end{bmatrix} \quad \text{(A1)}$$

Note that (A1) makes explicit that the high-frequency predictor $x_{k,t}^{(m)}$ is sampled m times more frequently than y_t while keeping the low-frequency structure of model (1). This way, frequency alignment between monthly and weekly variables requires exactly $m \times t$ observations for the high-frequency predictor $x_{k,t}^{(m)}$ such that it can be decomposed into m increments within each period t.

Nowcasts for the inflation rate at periods $t+1, \ldots, T$ can then be updated on a regular basis as high-frequency increments become available after t or by the time information on the low-frequency predictor gets available before official releases of the target inflation. Moreover, the process of updating the mixed-frequency dataset leads to missing data of $x_{k,t}^{(m)}$ at the end of the sample when nowcasting before the end-of-month. This sample's ragged edge is supplemented with random walk forecasts based on vintages of high-frequency data by the time of the nowcast.

A.2 Period-wise performance of tree-based models

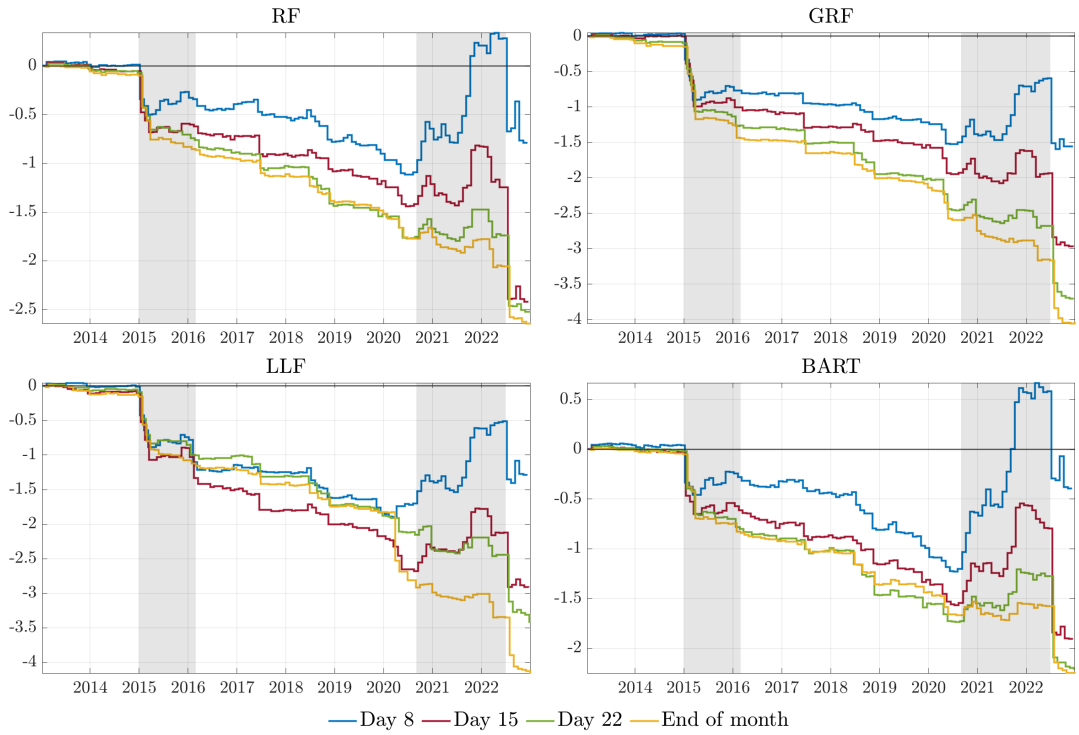
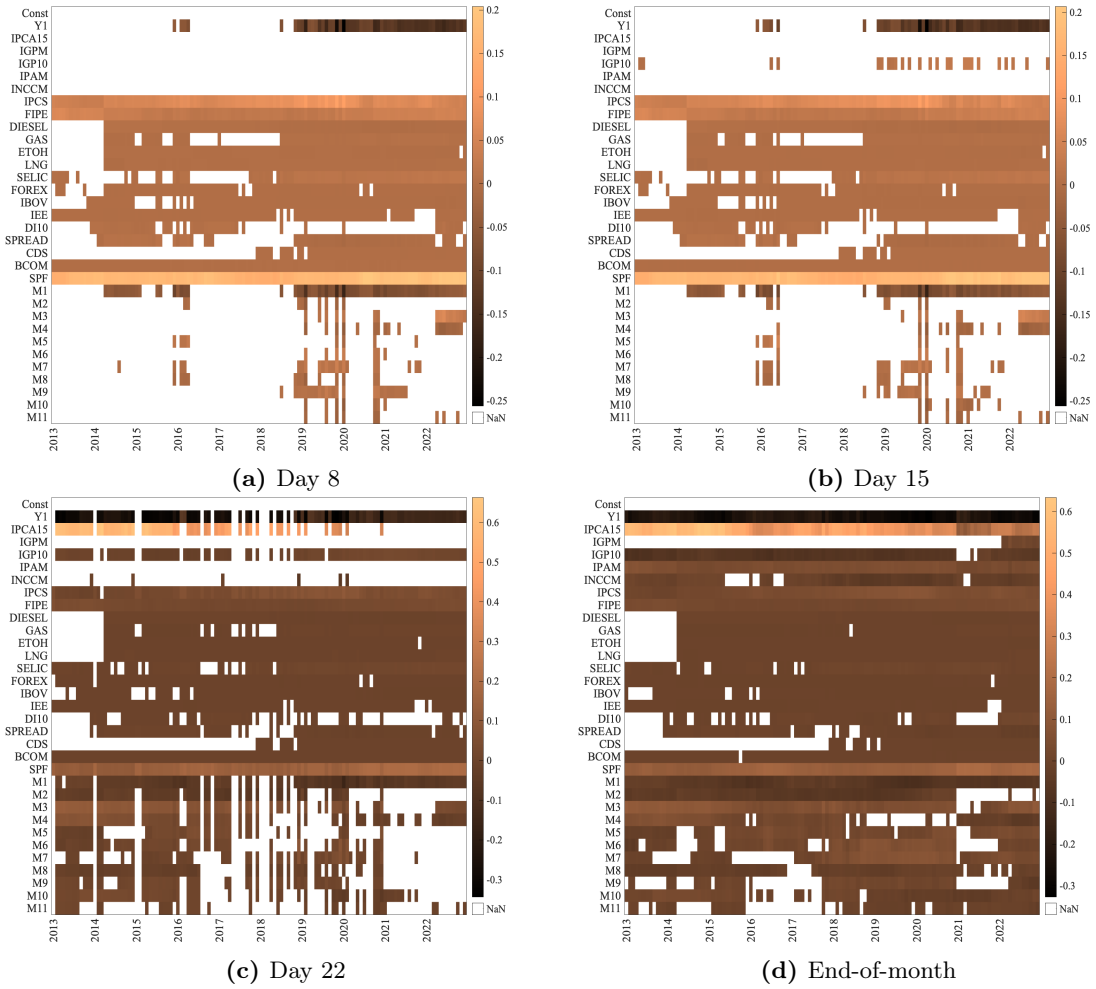


Figure A1: Cumulative sum of loss differentials (CUMSFE) of tree-based MIDAS nowcasts (Random Forest, Generalized Random Forest, Local Linear Forest, Bayesian Additive Regression Trees) versus the survey of professional forecasters (SPF, median) on days 8, 15, 22 and end-of-month. The gray shaded areas correspond to rising inflation periods.

A.3 Variable selection in sg-LASSO



(a) Day 8 **(b)** Day 15

(c) Day 22 **(d)** End-of-month

Figure A2: Heatmap of coefficient estimates using one of the best performing methods: sg-LASSO ($L = 0$). Empty cells represent a coefficient estimate equal to zero, and thereby a predictor that has not been selected for a given period t in the evaluation period.