DNB Working Paper

No 806 / March 2024

Harnessing Machine Learning for Real-Time Inflation Nowcasting

Richard Schnorrenberger, Aishameriane Schmidt, Guilherme Valle Moura

DeNederlandscheBank

EUROSYSTEEM

Harnessing Machine Learning for Real-Time Inflation Nowcasting

Richard Schnorrenberger, Aishameriane Schmidt and Guilherme Valle Moura*

* Views expressed are those of the authors and do not necessarily reflect official positions of De Nederlandsche Bank.

Working Paper No. 806

De Nederlandsche Bank NV
P.O. Box 98
1000 AB AMSTERDAM
The Netherlands

March 2024

Harnessing Machine Learning for Real-Time Inflation Nowcasting

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: February 2024.

Abstract

We investigate the predictive ability of machine learning methods to produce weekly inflation nowcasts using high-frequency macro-financial indicators and a survey of professional forecasters. Within an unrestricted mixed-frequency ML framework, we provide clear guidelines to improve inflation nowcasts upon forecasts made by specialists. First, we find that variable selection performed via the LASSO is fundamental for crafting an effective ML model for inflation nowcasting. Second, we underscore the relevance of timely data on price indicators and SPF expectations to better discipline our model-based nowcasts, especially during the inflationary surge following the COVID-19 crisis. Third, we show that predictive accuracy substantially increases when the model specification is free of ragged edges and guided by the real-time data release of price indicators. Finally, incorporating the most recent high-frequency signal is already sufficient for real-time updates of the nowcast, eliminating the need to account for lagged high-frequency information.

Key words: inflation nowcasting, machine learning, mixed-frequency data, survey of professional forecasters.

JEL classification: E31; E37; C53; C55.

[*] Corresponding author: Richard Schnorrenberger, Kiel University, Institute for Statistics and Econometrics, WSP7 R308, 24118 Kiel, Germany. Email: richard.schn@stat-econ.uni-kiel.de.

We are very grateful for the valuable comments and insightful suggestions provided by the participants at the 43rd International Symposium on Forecasting (Charlottesville, USA), the 38th meeting of the European Economic Association (Barcelona, Spain), the 29th Conference on Computing in Economics and Finance (Nice, France), the BSE Summer Forum on Macroeconometrics and Policy Evaluation (Barcelona, Spain), the 45th Meeting of the Brazilian Econometric Society, the Erasmus School of Economics PhD seminar (Rotterdam, The Netherlands), and the ISÖ Seminar on Statistics and Econometrics (Kiel, Germany). We would also like to thank Wellington Maiberg for his excellent data support. This work previously circulated as "Inflation nowcasting in persistently high inflation environments" and is a substantially revised version.
The opinions expressed in this paper are those of the authors and do not necessarily reflect the views of De Nederlandsche Bank or the European System of Central Banks.

1 Introduction

The inflationary shock that reverberated through global markets following the COVID-19 pandemic highlighted the importance of accurate and timely inflation nowcasts for better-informed monetary policy, business pricing strategies, and portfolio allocation decisions. While official price statistics are only measured at the monthly frequency and released with a significant delay, high-frequency (e.g., weekly or daily) and quickly released data have become particularly useful for anticipating the current state of the inflation process.[1] The relevance of updating inflation nowcasts in a timely fashion extends beyond disruptive environments, such as those witnessed in the aftermath of the pandemic, offering a means to anticipate swift inflationary shocks, as well as inflationary trends that may escalate or dwindle. Moreover, short- and medium-term inflation forecasts highly benefit from taking high-quality nowcasts as a jumping-off point (see, e.g., Faust and Wright, 2013; Krüger et al., 2017).

Machine learning (ML) methods have recently enjoyed great popularity in inflation forecasting under a data-rich environment (Garcia et al., 2017; Medeiros et al., 2021; Joseph et al., 2021; Hauzenberger et al., 2023; Araujo and Gaglianone, 2023; Barkan et al., 2023), exhibiting substantial improvements upon well-established benchmarks (e.g., Atkeson and Ohanian, 2001; Stock and Watson, 2007). However, there remains insufficient guidance on key modeling choices when using ML methods to construct inflation nowcasts in a real-time setup, especially during high inflation periods. The pandemic, in particular, posed challenges to nowcasting frameworks that struggle to anticipate rapidly evolving inflation dynamics not often seen in past data. Furthermore, in a nowcasting setting, the dimensionality challenge is amplified by the presence of high-frequency lags from numerous predictors, which may easily lead to overfitting.

This paper provides clear guidance for inflation nowcasting by evaluating a battery of easy-to-implement ML methods within a mixed-data sampling (MIDAS) approach. We contribute to the nowcasting literature by thoroughly investigating key modeling practices in an environment characterized by persistently high inflation, namely the Brazilian economy of the past decades. Moreover, we assess the predictive value of selected macro-financial predictors for inflation, including informed judgment entailed in a timely survey of professional forecasters (SPF). Specifically, we show that a well-designed unrestricted MIDAS (U-MIDAS) approach (Foroni et al., 2015) combined with linear shrinkage methods, especially the LASSO, produce inflation nowcasts that significantly improve upon SPF expectations. These predictive gains are particularly large at the onset of the COVID-19 crisis, whereas

[1] Giannone et al. (2008) and Bańbura et al. (2013), e.g., provide a comprehensive review of how the rapidly increasing availability of high-frequency data proves invaluable in obtaining early estimates of the current economic landscape while official statistics on key macroeconomic variables are yet to be released.

meaningful off-model information from SPF helps to discipline our model-based nowcasts. The unrestricted mixed-frequency ML structure also facilitates model interpretation and allows us to exploit potential nonlinear dynamics in the data, hereby assessed via tree-based methods.

In the broader field of macroeconomic nowcasting, research has typically focused on the holy grail of constructing high-frequency estimates of GDP, influenced by the success of Giannone et al. (2008). Dynamic factor models and mixed-frequency Bayesian VARs have emerged as popular tools amongst practitioners and policymakers (see, among others, Schorfheide and Song, 2015; McCracken et al., 2015; Carriero et al., 2015; Hindrayanto et al., 2016; Dahlhaus et al., 2017; Cimadomo et al., 2022; Cascaldi-Garcia et al., 2023; Huber et al., 2023). An early use of these econometric frameworks to exploit high-frequency data for inflation nowcasting is presented in Modugno (2013) and Knotek and Zaman (2017). Large-scale factor models, however, are not designed to capture fast-moving inflation dynamics at very short horizons and suffer from the ragged-edge problem that considerably worsens the forecasting properties of the model (see Marcellino and Schumacher, 2010). In addition, Knotek and Zaman (2017) show that inflation nowcasting may benefit from choosing a small number of highly informative predictors in contrast to extracting common factors from a large dataset.

Andreou et al. (2013), Monteforte and Moretti (2013), Breitung and Roling (2015) and Knotek II and Zaman (2023) consider MIDAS regressions with leads to eliminate ragged edges and effectively exploit more daily information of financial markets that are highly correlated with short-term inflation expectations. Although MIDAS regressions gained popularity for their parsimonious treatment of high-frequency lags and successful out-of-sample performance, they struggle with the dimensionality issue posed by numerous high-frequency predictors, which may easily lead to overparameterization.

Boosted by the COVID-19 crisis and the big data boom in economics, this line of research has taken up but with an increased focus on ML methods to guard against overfitting in high-dimensional settings and improve nowcasting accuracy over traditional econometric frameworks.[2] Penalized MIDAS regressions evolved as a suitable strategy for performing variable selection in macroeconomic nowcasting (Marsilli, 2014; Siliverstovs, 2017; Uematsu and Tanaka, 2019; Mogliani and Simoni, 2021; Babii et al., 2021; Kohns and Potjagailo, 2023; Beck et al., 2023; Aliaj et al., 2023). Specifically, Borup et al. (2023) demonstrate that exploiting more recent daily Google Trends data via their proposed combination of the U-MIDAS approach with ML methods can substantially improve predictions of weekly

[2] Non-traditional high-frequency data such as web scraping, Google Search, and scanner data have also become viable sources to nowcast both headline inflation and disaggregated components, such as food prices (see Harchaoui and Janssen, 2018; Powell et al., 2018; Macias et al., 2023; Beck et al., 2023, to name only a few).

initial claims while securing model interpretation in the era of big data. Besides, this mixed-frequency ML structure accommodates nonlinear ML-based predictive relationships, such as those analyzed in Richardson et al. (2021), Clark et al. (2022) and Barbaglia et al. (2023).

Building on these trends, we develop guidelines for key modeling choices for producing accurate weekly nowcasts of inflation using a large set of macro-financial data within the mixed-frequency ML structure. Therefore, from an practical standpoint, providing recommendations for practitioners and policymakers in this domain is the key objective of this paper. In addition, building upon the success of random forest models in forecasting U.S. inflation (Medeiros et al., 2021) and their capability to address temporal nonlinearities when forecasting UK inflation (Joseph et al., 2021), we complement the existing ML applications in macroeconomic nowcasting. Specifically, we evaluate the effectiveness of tree-based methods for inflation nowcasts.

From a practical standpoint, we conduct a real-time empirical exercise based on Brazilian data, which encompasses recent decades marked by persistently high inflation rates. This sets us apart from the predominant focus on U.S. or euro area inflation by existing literature. Notably, the presence of multiple episodes of rising inflation in recent Brazilian history allows us to gain insights that may be extrapolated to advanced economies undergoing unprecedented inflationary shocks not present in past data. To this end, we construct a novel real-time database from the Brazilian macroeconomy, which also features a variety of alternative high-frequency price indicators that are timely released by private agencies and closely monitored by professional forecasters. Furthermore, survey-based expectations have proven valuable to improve model-based nowcasts, particularly during periods of rising inflation (see, e.g., Banbura et al., 2021a,b; Bobeica and Hartwig, 2023). In this sense, we integrate the daily SPF conducted by the Brazilian Central Bank (BCB).

Our empirical exercise produces weekly nowcasts for the monthly developments of the official headline CPI targeted by BCB's monetary policy decisions, which is released with an average delay of seven business days after the reporting month. We select 20 predictors to compose our real-time dataset. The predictors are either available at a higher frequency – and transformed into weekly time series containing the latest month-on-month signal – or sampled monthly but released throughout the reporting month. For model interpretation, we divide them into four categories: monthly price indicators, weekly price indicators, daily financial variables, and daily SPF expectations. To exploit the information in our real-time set of predictors while guarding against overfitting, we compare linear prediction models via shrinkage (the LASSO, Ridge, Elastic Net, and sparse-group LASSO) against tree-based methods (Random Forest, Local Linear Forest, and Bayesian Additive Regression Trees).

Our findings underscore the effectiveness of shrinkage models to nowcasting inflation dynamics, with LASSO consistently surpassing tree-based methods in terms of RMSE. This is consistent with previous results for forecasting Brazilian inflation (see, e.g., Medeiros et al., 2016; Garcia et al., 2017), which indicate that variable selection done via the LASSO outperforms at the very short horizon. Notably, LASSO predictions exhibit exceptional accuracy at longer nowcast horizons compared to SPF expectations. This reflects the tendency of professional forecasters to adjust their expectations more frequently as the information set expands within the reporting month. Additionally, we observe large nowcasting gains building up during the COVID-19 inflation surge, where professional forecasters underestimated the rapidly evolving inflationary environment.

Moreover, our analysis reveals a notable difference in the variables selected by the LASSO depending on the nowcast horizon. Specifically, at longer nowcast horizons, the selection tends to produce a relatively sparse structure with SPF expectations and weekly price indicators as the primary predictors. Conversely, as we approach shorter horizons, a denser model structure emerges, driven by the pronounced relevance of monthly price indicators. This shift reflects the increased availability of accurate contemporaneous inflation signals as the reporting month unfolds. Consequently, data releases on monthly price indicators diminish the relative importance of SPF expectations, although informed judgment remains highly influential, particularly in navigating the challenges posed by the COVID-19 crisis. Overall, while financial variables play a minor role, the combination of timely price indicators with SPF judgments proves critical in producing weekly inflation nowcasts.

Finally, a deeper investigation of key modeling choices within our mixed-frequency ML framework reveals the considerable impact of (i) accounting for SPF data in the predictor set, (ii) eliminating ragged edges, (iii) guiding model specifications by real-time data releases, and to a lesser extent, (iv) focusing solely on the most recent high-frequency signal. A baseline prediction model featuring these key elements yields predictive gains up to 60%. Notably, shrinkage-based predictions can highly benefit from using meaningful judgment in survey data and addressing the ragged-edge problem.

The paper proceeds as follows. Section 2 describes the real-time dataset of the Brazilian macroeconomy and how these macro-financial variables relate to the target variable. Section 3 outlines the nowcasting setup and provides an overview of the mixed-frequency ML strategies. Next, we present our empirical results in Section 4. This section also provides an interpretation of the best-performing fitted model and offers guidance on key modeling choices for constructing accurate weekly nowcasts using the real-time data flow. Finally, Section 5 concludes.

2 Data

To compute weekly nowcasts of inflation figures we select predictors that have two features: significant correlation with price developments and earlier availability in comparison to official inflation releases. We put together a novel real-time database of macro-financial series from the Brazilian economy tailored for inflation nowcasting. In this context, our dataset mainly consists of timely price indicators, financial variables, and experts' forecasts that carry predictive content about the current month's inflation rate.[3] Besides data on the target CPI variable, we organized publicly available information on price indicators released both by public and private institutions, financial indicators, and daily SPF with aggregate predictions for the target variable.[4] Our real-time dataset covers the period from June 2004 up to December 2022 ($T = 222$ monthly observations), whereas information on release dates is available from January 2013 onwards.

The official inflation measure in Brazil is known as the Broad National Consumer Price Index (IPCA), and concurrently, it serves as the reference for the inflation-targeting system in Brazil.[5] The IPCA is designed to reflect 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.

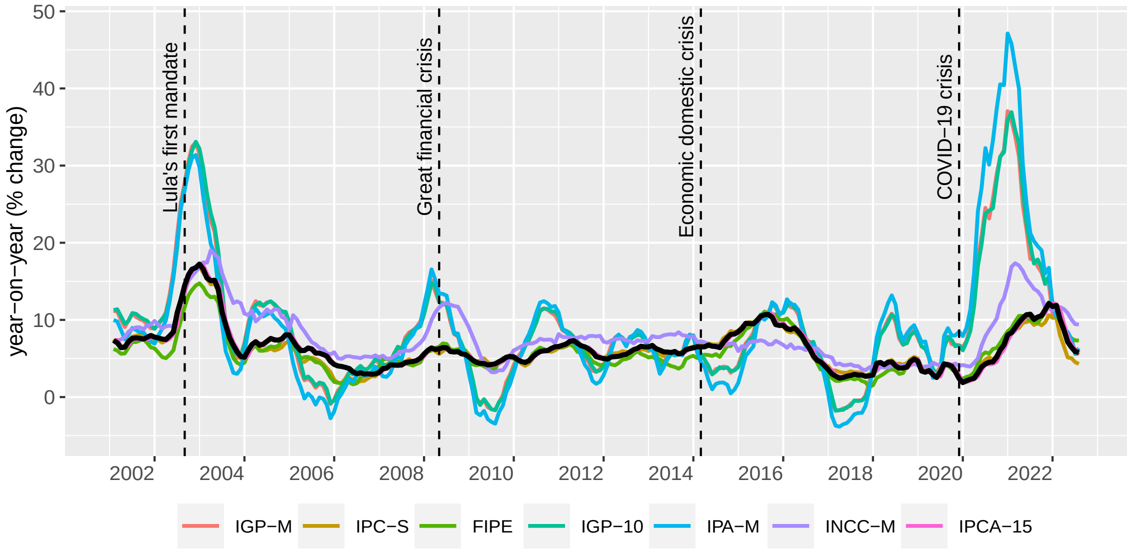
Figure 1 shows the IPCA evolution since mid-2001, shortly after the BCB adopted the inflation targeting regime. The year 2003 witnessed an escalation in political and economic risks following the election of the Workers' Party representative, triggering a foreign capital outflow that led to a strong exchange rate depreciation and domestic inflationary pressure. This was followed by a relatively calm period, marked by annual IPCA fluctuations around 5%. However, a return to double-digit inflation figures occurred during the political turmoil that started in 2013, leading to the impeachment of President Rousseff in early 2015. Following years of price stability with IPCA oscillating close to BCB's target, inflation surged again in the aftermath of the pandemic shock, similar to trends observed worldwide.

[3] We disregard monthly indicators of real economic activity for two reasons: (i) short or no availability before official releases of the target inflation and (ii) non-significant cross-correlations up to six lags with the target month-on-month inflation rate. Hence, economic activity variables do not fit our nowcasting purpose.

[4] Although our analysis focuses on price indicators and financial variables as the potential predictors for inflation, due to their high-frequency and timely attributes, the real-time database also comprises vintages and revisions of hard and survey-based data for economic activity (e.g., industrial production, unemployment rate, net payroll jobs, PMI manufacturing, retail and services indices, consumer and business confidence indicators, among others).

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

Figure 1: Time series of Brazilian price indicators, 2001 – 2022



Notes: The official Brazilian CPI (IPCA) is depicted in black while alternative price indicators are illustrated in colored lines. The full description of each indicator is available in Table 1.

We use a total of 20 predictors in our empirical application, excluding the lags of IPCA[6]. 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), 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, including the sampling period and publication lags.

The first group of predictors consists of five monthly price indicators primarily collected in 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 sampling period and targeted prices. For instance, IPCA-15 mimics IPCA itself in terms

[6] While more indicators could potentially correlate with IPCA, we have chosen a medium-sized dataset. This decision aligns with previous findings in the literature (see, e.g., Carriero et al., 2019, 2020), who show that, for point and density forecasting/nowcasting of GDP growth and inflation, a wider array of predictors do not outperform models with only a few hand-picked predictors. Nonetheless, the potential high-dimensionality issue arising in our application is also connected to the choice of high-frequency lags included in the nowcasting model (see Section 3.1)

of methodology, but it reflects prices collected from the 16th of the preceding month to the 15th of the reporting month. Releases for this mid-month version of the IPCA become available with an average delay of 8 days (usually at the beginning of the 4th week) and thus allow for early signals of IPCA dynamics. Additionally, we include a producer price index (PPI) termed IPA-M, which monitors inter-business transaction prices of agricultural and industrial products, and a construction cost index named INCC-M. The remaining two indices, IGP-M and IGP-10, are the weighted average of the IPA-M (60%), IPC-S (30%, FGV's weekly CPI measure presented below), and the INCC-M (10%), diverging only by their sampling periods.

The above five monthly indices are also displayed in Figure 1. While closely correlated with IPCA, some exhibit greater volatility, especially in turbulent times. For example, IPA-M – and consequently, IGP-M and IGP-10 – are markedly affected by the large volatility in exchange rates observed during the initial year of Lula's administration, the Great Financial Crisis, and the pandemic. The INCC-M, related to construction costs, is in general higher than the IPCA, but presents a lower amplitude than the majority of the other indexes.

The second group of predictors contains six timely indicators of consumer and energy prices sampled at the weekly frequency and published with a lag of one or two days after the closing of a given week. The IPC-S and FIPE intend to closely mirror the IPCA at a higher frequency – as shown in Figure 1 – but respectively accounting for consumption baskets of earnings in the range of 1-33 and 1-10 minimum wages.[7] Moreover, we include prices of major energy components: diesel, gasoline, ethanol fuel, and liquefied natural gas. These prices are collected by surveys of the wholesale fuel price practiced by retailers of around 500 cities nationwide.[8]

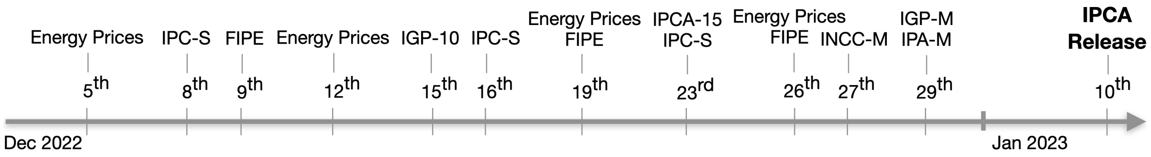
Figure 2 illustrates the timeline of real-time data releases of the above price indicators in December 2022. As shown, IPCA figures came out on the 10th of January 2023, but data releases of the selected predictors mostly occur throughout the reporting month. For example, given that energy prices become available after the closing of a calendar week, the first release is on 5 December while the subsequent numbers are provided on the following Mondays. IPC-S and FIPE become available shortly after the closing of a four-week collection system ending on four set dates (07, 15, 22 and end-of-month).[9] Hereby these numbers are first released on the 8th and 9th, followed by releases on the 16th and 19th, and so on, which is extremely quick for international standards. Turning to monthly indicators, data on IGP-10 and IPCA-15 come out relatively early in the month – around the third week – whereas INCC-M, IGP-M and IPA-M follow next before the month ends.

[7] The sampling procedure of FIPE only accounts for households living in São Paulo city.

[8] Compared to information on raw oil prices available in financial markets, these surveys have the advantage that distribution and retail margins are fully accounted for.

[9] This means that the computation of these indices considers the average of prices collected during the four weeks preceding the closing date.

Figure 2: Release calendar of Brazilian price indicators in December 2022



The third group of predictors contains daily information from financial markets, including movements in the yield curve or interest rate spreads, commodity and stock price indices, and exchange rates.[10] The choice of these financial market variables is motivated by their timely information about short-term inflation expectations and findings in the literature on inflation forecasting. For example, Modugno (2013), Monteforte and Moretti (2013), and Breitung and Roling (2015) show that relevant commodities (e.g., crude oil prices) and financial assets are among the most reliable indicators of inflation changes. Furthermore, central banks and practitioners monitor daily financial variables to forecast the state of the macroeconomy (Andreou et al., 2013).

Finally, we use expert information from the SPF conducted by the BCB, also known as the FOCUS survey. It started in the late 90s, together with the implementation of the inflation-targeting regime in Brazil. Participation in the survey is limited to banks, asset managers, companies linked to real economic sectors, brokers, and consultancies, who have to be pre-screened by the BCB. These institutions can continuously provide their short and long-run expectations regarding key macroeconomic indicators such as GDP, inflation, and exchange rate, among others. The BCB releases daily aggregate statistics of the SPF, with a delay of one business day, as well as a Top 5 ranking with the best-performing forecasting institutions divided across indicators and forecast horizons.

Historically, there are over 100 active participants in the SPF survey. The median of these experts' forecasts for IPCA dynamics is closely monitored by market participants, especially via the weekly handout report released by the BCB every Monday morning with data up to the previous Friday (Marques, 2012). We use the median of SPF expectations as both a predictor in our models as well as a benchmark to compare our nowcasts. As an additional benchmark, we compare our predictions against the median forecast produced by the Top 5 forecasters. The BCB ranks the Top 5 participant institutions based on previous months' performance. Hence, after obtaining the best five institutions, for each indicator and horizon, the BCB averages their forecasts for the current month[11].

[10] The stock index (IBOV) corresponds to the B3 Index, while interest rates are derived from Brazilian interbank deposit future contracts negotiated at B3, ultimately linked to treasury bills issued by the BCB.

[11] Note that there is no "data leakage" given that the ranking is computed based on the past. For instance, it might be that the current Top 5 institutions are not the ones that will produce the best forecasts at the current period.

Table 1: Database

Series	Mnemonic	Reference period	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 price index	IGP-M	21^{st}_{t-1} to 20^{th}_{t}	last week, reporting month	7	2003M1	FGV
General price index - 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
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 period relates to the data collection period. The publication timing provides the regular release calendar for 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 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

Our nowcasting model follows an unrestricted mixed-frequency structure combined with ML methods that guard against overfitting in a high-dimensional setting. Our methodology is essentially divided into two components: (i) the general nowcasting setup, describing the functional form of how the mixed-frequency dataset will be organized and specifying the information used at each nowcast date, and (ii) the classes of ML methods employed to produce the nowcasts. This mixed-frequency ML structure enables us to treat separately the real-time flow of information from each predictor, thereby facilitating model interpretation, while improving nowcasting accuracy by harnessing the power of ML methods.

3.1 Nowcasting setup

To fix ideas, we aim to nowcast monthly inflation rates with predictors sampled at the daily, weekly, and monthly frequencies. Let $\pi_t = 100(P_t/P_{t-1} - 1)$ denote the month-on-month inflation rate, where P_t is the price level in month t. A generic high-frequency (daily or weekly) macro-financial variable is given by $x_t^{(w)}$ and can be sampled w times more frequently than the target π_t. Moreover, x_t represents a generic monthly price indicator, with the sampling process extending over t but disclosed before π_t. In this sense, time indices $t = 1, \dots, T$ act as the common frequency between π_t and predictors $x_t^{(w)}$ and x_t.

Suppose we would like to update our nowcasts at the weekly frequency. Specifically, at four different points within the month: days 8, 15, 22, and end-of-month.[12] Given the mixed-frequency environment, we take a stance on how to incorporate high-frequency information on these four nowcast days. We start by assuming a fixed monthly-to-weekly combination, with a frequency ratio of $w = 4$, to accommodate weekly updates of the nowcast.[13] Hence, at the end of month t, the information set also includes the following K-dimensional vectors of high-frequency predictors: $\boldsymbol{x}_t^{(w)}, \boldsymbol{x}_{t-\frac{1}{w}}^{(w)}, \dots, \boldsymbol{x}_{t-\frac{w-1}{w}}^{(w)}$, where $t-j/w$ denotes the j^{th} past high-frequency period for $j = 0, \dots, w-1$. More precisely, t corresponds to end-of-month observations; $t-1/4$ is the next to end-of-month, and thus day 22; $t-2/4$, day 15; and $t-3/4$, day 8. As a result, the forecast horizon h respectively becomes j/w.

Next, we must address the frequency mismatch between daily and weekly data, along with the non-synchronous nature of macroeconomic data releases. We transform the daily

[12] This particular choice of days allows us to control for the problem of overlapping calendar weeks across consecutive months and the heterogeneous number of days in different months.

[13] The choice for weekly updates of the nowcast with a fixed monthly/weekly mixture also avoids a higher proliferation of parameters arising from a higher frequency mismatch in a model that would combine monthly and daily variables.

information from financial predictors and the SPF data into weekly time series containing the latest month-on-month rates available on the nowcast day.[14] Data on our weekly predictors follow different sampling strategies but become available with a minimal lag of one or two days. Hence, by day 8 of the reporting month t, we assume immediate access to the first week's contemporaneous data. For example, IPC-S data covering the first week of t is reliably published on the first day following the closure of that week – typically on the 8th or the 9th/10th if the closing date is a Friday/Saturday (see Figure 2). Consequently, this data regularly integrates our information set in $t - 3/4$. For the remaining nowcast days within t, we shift forward the latest released contemporaneous month-on-month signal if the corresponding weekly data is not yet published.

The general prediction model for the nowcast horizon $h = j/w$ is given by

$$\pi_{t|t-h} = f^h\left(\pi_{t-1}, \boldsymbol{x}_t^h, \boldsymbol{d}_t, \boldsymbol{x}_{t-h}^{(w)}, \ldots, \boldsymbol{x}_{t-h-p/w}^{(w)}; \boldsymbol{\theta}^h\right) + \varepsilon_t^h, \tag{1}$$

whereas the autoregressive term accounts for temporal dependence in π_t[15]; $\boldsymbol{x}_t^h$ is a horizon-specific J^h-dimensional vector of monthly predictors, thereby sampled at the same frequency as π_t; the set of 11 monthly dummy variables $\boldsymbol{d}_t$ capture potential seasonal patterns in price dynamics; and high-frequency predictors with data up to the nowcast date $t - h$ and corresponding lags are respectively denoted by $\boldsymbol{x}_{t-h}^{(w)}, \ldots, \boldsymbol{x}_{t-h-p/w}^{(w)}$. In addition, $\boldsymbol{\theta}^h$ is a vector of model parameters specific to the prediction function f^h at horizon h; and ε_t^h is a zero-mean disturbance term.

In the general form, model (1) includes $p \geq 0$ relevant high-frequency lags to construct the nowcast at any horizon h. For instance, assuming that we stand at day 8 of month t, the nowcast horizon is $h = 3/4$ and we might use the high-frequency lags $\boldsymbol{x}_{t-3/4}^{(w)}, \boldsymbol{x}_{t-1}^{(w)}, \boldsymbol{x}_{t-5/4}^{(w)}, \ldots, \boldsymbol{x}_{t-p/4}^{(w)}$. Likewise, if end-of-month observations are available, the nowcast horizon is $h = 0$ and the predictors $\boldsymbol{x}_t^{(w)}, \boldsymbol{x}_{t-1/4}^{(w)}, \ldots, \boldsymbol{x}_{t-p/4}^{(w)}$ might be included.[16] Hereby the baseline specification only incorporates the most recent month-on-month high-frequency signal by setting $p = 0$, although one might choose $p = w - 1$ to account for all contemporaneous high-frequency signals when nowcasting at the end-of-month (see Section 4.3), or even $p > w - 1$ to include lags that span over past and distant months.

Since our prediction model assigns individual coefficients to each of the high-frequency predictors in $\boldsymbol{x}_t^{(w)}$ and its associated lags, a linear specification of (1) can be seen as a U-MIDAS model. Foroni et al. (2015) argue that a fairly small frequency mismatch, such as our monthly-to-weekly mixture, favors the adoption of the U-MIDAS over restricted

[14] The month-on-month transformations of daily and weekly data are taken by referencing the same day in the previous month. This also ensures the stationarity of the variables.

[15] If needed, additional lags of π_t can be included.

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

MIDAS regressions with tightly specified lag polynomials that perform nonlinear temporal aggregation of high-frequency lags (see also Ghysels and Marcellino, 2018).[17] Consequently, the U-MIDAS approach allows for flexible estimation of the individual effects of high-frequency lags on the target while facilitating the interpretability of the model.

It is worth emphasizing that $\boldsymbol{x}_t^h$ only incorporates monthly predictors with available contemporaneous information at the time of the nowcast, resulting in a horizon-specific dimension J^h. For example, given that data on IPCA-15 is usually published between the 19th and 23rd of the reporting month t, IPCA-15 will not be included in $\boldsymbol{x}_t^h$ when nowcasts are made on days 8 and 15. This implies that the predictor space will exhibit reduced dimensionality at longer nowcast horizons. Hence, we constantly assess the real-time data availability of monthly predictors by the time of the nowcast and adjust the general specification (1) accordingly. This approach mitigates the risk of generating imprecise nowcasts from assigning non-zero and relevant coefficients to monthly predictors lacking new information to construct the nowcast for π_t.[18] Furthermore, it avoids ragged edges in the modeling.

The flexibility of model (1), however, comes at the cost of overparameterization as the count K of high-frequency predictors and their lags p rise. Specifically, the model features $K(p+1)+J^h+13$ parameters for a specific horizon h, including the intercept. In macroeconomic contexts, the effective sample size might be relatively short compared to the number of parameters, posing challenges for conventional estimation methods and leading to high estimation uncertainty. To address this high-dimensional prediction problem, we implement the mixed-frequency ML strategy (see Borup et al., 2023) by incorporating a wide range of ML methods that allow for flexible estimation of the coefficients while still guarding against overfitting.

3.2 Machine learning methods

The mixed-frequency ML strategy can be applied to both linear and nonlinear prediction models. The ML methods we implement have been enjoying growing popularity within economics and are distinguished between two classes: linear shrinkage and nonlinear tree-based methods. In the first group, we have the Elastic Net (ENet) regression and its two special cases, LASSO and Ridge. As an alternative to these standard methods, we apply

[17] The core idea of MIDAS regressions is to efficiently address the dimensionality issue arising from the numerous high-frequency lags in the model. This is efficiently achieved via tightly specified lag polynomials to ensure parsimonious modeling. However, the adoption of a constrained MIDAS approach in this context would yield only a slight reduction in the number of parameters to be estimated in Eq. (1).

[18] One might also include lags of $\boldsymbol{x}_t^h$ in Eq. (1), but we use the autoregressive term to fully capture the potential serial correlation in π_t.

the sparse-group LASSO estimator with MIDAS structure, a novel approach introduced by (Babii et al., 2021). This method has the advantage of acknowledging the serial dependence across different high-frequency lags. Turning to tree-based methods, we implement the Random Forest (RF), Local Linear Forest (LLF) – both in its solo form and the ensemble prediction with a LASSO pre-selection of predictors – and the Bayesian Additive Regression Trees (BART). Table 2 provides an overview of these methods and the corresponding tuning parameters.

Table 2: Summary of the ML models applied to the general specification (1)

Model	Short name	Reference	R function (package)	Tuning parameters and cross-validation
Least absolute shrinkage and selection operator	LASSO	Tibshirani (1996)	glmnet (glmnet) and trainControl, train (caret)	λ using time series cross-validation
Ridge	Ridge	Hoerl and Kennard (1970)	glmnet (glmnet) and trainControl, train (caret)	λ using time series cross-validation
Elastic Net	ENet	Zou and Hastie (2005)	glmnet (glmnet) and trainControl, train (caret)	α, λ using time series cross-validation
Sparse-Group LASSO	sg-LASSO	Babii et al. (2021)	cv.sgl.fit (midasml)	α, λ using time series cross-validation
Random Forest	RF	Breiman (2001)	randomForest (randomForest)	number of skip-sampled predictors to split the tree (mtry) equal to the maximum between number of predictors divided by three and one
Local Linear Forest	LLF	Friedberg et al. (2020)	ll_regression_forest (grf)	We used default values for sample fraction (0.5), number of trees (2000), mtry (min{number of predictors$^{1/2}$ + $20,$ number of predictors}), minimum node size (5), honesty fraction (0.5), honest prune leaves (1), α (0.05), imbalance penalty (0)
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

To set the stage to formally outline the ML methods, we denote by $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_t)'$ the target inflation series up to t. The low-frequency predictor set specific to horizon h is denoted by $\boldsymbol{x} = (\boldsymbol{x}_1^{h'}, \ldots, \boldsymbol{x}_t^{h'})$. The $t \times K(p+1)$ predictor set of high-frequency data is given by $\boldsymbol{x^{(w)}} = (\boldsymbol{x^{(w)}_{-h}}, \boldsymbol{x^{(w)}_{-h-1/4}}, \ldots, \boldsymbol{x^{(w)}_{-h-p/w}})$, whereas $\boldsymbol{x^{(w)}_{-h}} = (\boldsymbol{x}^{(w)}_{1-h}, \boldsymbol{x}^{(w)}_{2-h}, \ldots, \boldsymbol{x}^{(w)}_{t-h})'$ denotes the $t \times K$ high-frequency set associated with lag $t-h$. The general predictors matrix is then given by $\boldsymbol{X} = (\iota, \boldsymbol{\pi}_{-1}, \boldsymbol{x}, \boldsymbol{d}, \boldsymbol{x^{(w)}})$, where ι accounts for the intercept, $\boldsymbol{\pi}_{-1}$ is the first lag of $\boldsymbol{\pi}$ and $\boldsymbol{d}$ comprises the seasonal deterministic dummies. For convenience, we drop the superscript h from the vector of model parameters $\boldsymbol{\theta}$.

Shrinkage methods

Shrinkage methods are penalized regression schemes that identify the relevant predictors from a large dataset. This targeted selection aims to improve forecasting precision at the cost of a slight increase in bias. The ENet estimator, proposed by (Zou and Hastie, 2005), solves the penalized least-squares problem:

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

where $\alpha \in (0, 1]$ is a weight hyperparameter 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 hyperparameter λ controls the amount of shrinkage in the parameter space $\boldsymbol{\theta}$. Hence, estimator (2) shrinks coefficients of irrelevant predictors toward zero. Because the penalty term of ENet and LASSO include the ℓ_1-norm, they can perform variable selection and thus yielding a sparse and parsimonious model that facilitates interpretation. In contrast, coefficients estimated via Ridge regression never equal zero, yielding a dense model.

Babii et al. (2021) argue that high-dimensional mixed-frequency representations with multiple high-frequency lags ($p > 0$ using our notation) involve certain data structures that once taken into account should lead to increased performance out-of-sample. These structures relate to groups covering the relevant lags of a single high-frequency predictor. In this sense, the sg-LASSO with MIDAS structure selects not only the relevant predictors for the target but also the appropriate lag structure of each high-frequency predictor. This structured sparsity constitutes the key feature of sg-LASSO and a refinement of the unstructured LASSO, which fails to acknowledge serial dependence across high-frequency lags and tends to arbitrarily select one lag from the group (see "irrepresentable condition" in Zhao and Yu, 2006).

The sg-LASSO solves the penalized regression problem:

$$\hat{\boldsymbol{\theta}} = \min_{\hat{\boldsymbol{\theta}}} ||\boldsymbol{\pi} - \boldsymbol{X}\boldsymbol{\theta}||^2 + 2\lambda \left(\alpha \, |\boldsymbol{\theta}|_1 + (1-\alpha) \, ||\boldsymbol{\theta}||_{2,1} \right), \tag{3}$$

where $||\boldsymbol{\theta}||_{2,1} = \sum_{G \in \mathcal{G}} ||\boldsymbol{\theta}_G||$ is the group LASSO norm for a group structure $\mathcal{G}$ that comprises the $p + 1$ lags of each high-frequency predictor.[19] This implies that sg-LASSO promotes sparsity between and within groups.[20]

Moreover, the high-frequency predictor set $\boldsymbol{x}^{(\boldsymbol{w})}$ in (3) is based on orthogonal Legendre polynomials of degree L that aggregate over the high-frequency lags of each predictor. They can be viewed as predetermined weights that alleviate overfitting by reducing the predictor-dimension in $\boldsymbol{x}^{(\boldsymbol{w})}$ from a factor of $(p + 1)$ to L. In our empirical exercise, the sg-LASSO is implemented with $p = 3$ and $L = 1$. This means that four high-frequency lags

[19] $\alpha \in [0, 1]$ determines the relative importance of LASSO-sparsity and the group structure.

[20] Note that our application requires us to assume that each monthly predictor in $\boldsymbol{X}$ represents a whole group in (3).

are considered, which will be aggregated with equal weights for the Legendre polynomial of order $L = 0$ while $L = 1$ features an increasing linear function and thereby favors more distant lags.[21] In addition, we link these high-frequency lags to the contemporaneous information set only, giving rise to missing observations at the end of the sample (ragged-edge problem) when nowcasting on days 8, 15 and 22. Hereby we replace the ragged edges with random-walk updates of the latest month-on-month information available at the time of the nowcast. Therefore, in terms of model specification, sg-LASSO departs from standard shrinkage based on Eq. (2) in two aspects: the number of high-frequency lags ($p = 3$ rather than $p = 0$) and the presence of ragged edges. Finally, having estimated the parameters using either Equation (2) or (3), we can form a nowcast $\hat{\pi}$ by taking a new set of observations $\tilde{\boldsymbol{X}}$ and multiplying by $\hat{\boldsymbol{\theta}}$.

The shrinkage hyperparameters λ and α are tuned in a data-driven manner using time series cross-validation, whereas we set the grid values $(0, 0.25, 0.5, 0.75, 1)$ for α. Differently from the standard cross-validation procedure, in which folds are randomly selected assuming that observations are independently and identically distributed, time series cross-validation splits the training dataset into time slices that retain the chronological order. Therefore, time series cross-validation takes place sequentially and avoids using future observations to fit the model (for a review, see Arlot and Celisse, 2010; Goulet Coulombe et al., 2022; Bergmeir et al., 2018). In our empirical exercise, we start with a 36-month initial fixed window with sequential folds of 12 months.

Tree-based methods

Tree models are based on decision trees, which are nonparametric methods that recursively divide the predictor space according to a pre-determined splitting rule. In our nowcasting exercise, we use the following models: random forest; the local linear forest; the bayesian additive regression tree; and a combination of LASSO with the local linear forest.

First proposed by Breiman (2001), random forests are an extension of decision trees in which the results from several non-correlated (or with very small correlation) trees randomly chosen are aggregated 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, an univariate prediction is given by

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

[21] The choice of $L = 1$ delivers similar results compared to $L = 2$ but at a lower computational cost.

where $\hat{\pi}_b(\tilde{\boldsymbol{X}}_m)$ is the prediction of the b-th tree using new data $\tilde{\boldsymbol{X}}_m$, and m here denotes a subset of all available predictors. 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 to be highly competitive against other ML methods and traditional econometric frameworks when used to forecast inflation (see Medeiros et al., 2021; Araujo and Gaglianone, 2023, among others).

The LLF method proposed by Friedberg et al. (2020) is the combination of a random forest with a local linear regression. In general terms, it combines the RF ability to deal with high-dimensional and nonlinearities with the smoothness of a local linear regression. It is a two-step approach in which the random forest is used to obtain weights for observations that will be later used in the local linear regression with a ridge-type penalty.

To find the weights using a random forest, we start from (4):

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

where $\mathbb{1}_{\tilde{\boldsymbol{X}}_m\in\mathcal{J}_{k,b}}$ is an indicator function denoting that $\tilde{\boldsymbol{X}}_m$ belongs to the region $\mathcal{J}_k$ in tree b, $\theta_{k,b}$ is a parameter, and $|\cdot|$ denotes the cardinality of a set. The quantity π_i denotes a response paired with X_i (from the in-sample information), from which n points are available. The term $\alpha_i(\tilde{\boldsymbol{X}}_m)$ is called forest weight and denotes the fraction of trees that allocates $\tilde{\boldsymbol{X}}_m$ in the same leaf as the predictor vector $\boldsymbol{X}_i$. In Eq. (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 predictor that has little relation with π_i will appear less frequently when making splits (Athey et al., 2019).

The second step is a local linear regression. Specifically, $\pi(\tilde{\boldsymbol{X}}_m)$ will be the local average, which can be estimated together with a $\boldsymbol{\theta}(\tilde{\boldsymbol{X}}_m)$ through the following optimization problem:

$$\begin{aligned}\begin{pmatrix}\hat{\pi}\left(\tilde{\boldsymbol{X}}_m\right)\\ \hat{\boldsymbol{\theta}}\left(\tilde{\boldsymbol{X}}_m\right)\end{pmatrix} = \underset{\pi,\boldsymbol{\theta}}{\arg\min}&\left\{\sum_{i=1}^{n}\alpha_i\left(\tilde{\boldsymbol{X}}_m\right)\left(\pi_i-\pi\left(\tilde{\boldsymbol{X}}_m\right)-\left(\boldsymbol{X}_i-\tilde{\boldsymbol{X}}_m\right)\boldsymbol{\theta}\left(\tilde{\boldsymbol{X}}_m\right)\right)^2\right.\\ &\left.+\lambda\left\|\boldsymbol{\theta}\left(\tilde{\boldsymbol{X}}_m\right)\right\|_2^2\right\},\end{aligned} \tag{6}$$

where $\hat{\pi}\left(\tilde{\boldsymbol{X}}_m\right)$ is still a prediction for a new point but with the slope of the local linear regression $\boldsymbol{\theta}\left(\tilde{\boldsymbol{X}}_m\right)$, which corrects for the local trend in $\boldsymbol{X}_i-\tilde{\boldsymbol{X}}_m$. The LLF prediction is then based on the intercept $\hat{\pi}\left(\tilde{\boldsymbol{X}}_m\right)$ while the parameter vector $\boldsymbol{\theta}$ is neglected at this stage. Note that the penalization term $\lambda\left\|\theta\left(\tilde{\boldsymbol{X}}_m\right)\right\|_2^2$ plays a role in avoiding overfitting to the local trend and λ is typically chosen via cross-validation. As a result, the LLF can effectively approximate smooth functions through local regression without becoming infeasible with a growing number of predictors.

Additionally, we implement a combination of LASSO with LLF, as suggested by Friedberg et al. (2020). The pre-selection step via the LASSO might lead to improved predictive performance given that it helps mitigate the curse of dimensionality and better handles multicollinearity by selecting the most informative predictor among the group. Chinn et al. (2023) offer a broader discussion on multi-step nowcasting approaches composed of pre-selection and factor extraction before the estimation of tree-based models.

Chipman et al. (2012) introduce the BART method, which can be viewed as the Bayesian counterpart to random forests. BART predictions are derived from several trees, but opposed to RF, are here sequentially estimated using the residuals from the preceding tree as the dependent variable. Hence, each subsequent tree attempts to capture the remaining variability not explained by the previous trees. 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 $g(\boldsymbol{X}_i, \boldsymbol{\mathcal{T}}, \boldsymbol{\mathcal{M}})$ which maps a row $\boldsymbol{X}_i$ (from the predictor matrix $\boldsymbol{X}$) to a particular $\theta_j \in \mathcal{M},\ j \in 1, \ldots, |\boldsymbol{\mathcal{M}}|$. Predictions from individual trees form the final BART prediction and are obtained by sampling from the posterior distribution. We closely mirror the prior specification used in Chipman et al. (2012). This implies a uniform prior to determine both the variable for a split and the corresponding cutpoint. A conjugate normal prior is used 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, while $\alpha \in (0,1)$ and $\beta \in \mathbb{R}^+$ are hyperparameters.

In our empirical exercise, we set the hyperparameters of the above tree-based methods to default values respectively recommended by Breiman (2001), Friedberg et al. (2020) and Chipman et al. (2012).[22] To construct BART predictions we estimate 200 trees using 1000 posterior draws, with 100 draws as burn-in. For the tree structure, we use $\alpha = 0.95$ and $\boldsymbol{\theta} = 2$, which penalizes bigger trees. For the conjugate normal prior of the predictions, we

[22] Tuning hyperparameters via time series cross-validation results in a lower nowcasting performance and substantially increased computational burden. These results are available upon request.

centered the prior at 0 and the variance is equal to $0.5/k(\sqrt{m})$ where $k = 2$ and $m = 200$ denote the number of trees. For the variance prior, the hyperparameters ν and λ of the $(\nu\lambda)/\chi_{\nu}$ distribution are obtained from the standard deviation of the response variable in the estimation sample and a factor of 10, respectively.

4 Empirical Results and Discussion

In this section, we investigate the performance of our mixed-frequency ML models for nowcasting Brazilian inflation using a real-time dataset with macro-financial predictors that span from June 2004 to December 2022. For the out-of-sample evaluation, we focus on the interval from January 2013 to the end of our sample using an expanding window scheme. This evaluation sample is constrained by data availability, as the release calendar for the entire dataset is only available from January 2013 onward. Nonetheless, it includes two of the most inflationary periods in Brazil's recent history: the economic domestic crisis of 2014 and the COVID-19 pandemic.

To assess the accuracy of our IPCA predictions, we compare them against SPF expectations – both the median and the Top 5 – published by the BCB. We update our nowcasts on a set of fixed days (8, 15, 22 and end-of-month) using the most recent increments of monthly and high-frequency (weekly and daily) data respecting the release calendar. While model estimation is based on month-on-month transformations of variables, we use year-on-year IPCA rates as our ultimate metric for performance evaluation. Consequently, we adjust our model-derived nowcasts for month-on-month IPCA rates before comparison with actual realizations of the target. Our findings highlight the superior performance of shrinkage methods over tree-based methods. Moreover, a deeper analysis of key modeling choices in Eq. (1) reinforces the importance of eliminating ragged edges in a real-time setup and to account for some degree of informed judgment in SPF data.

4.1 Out-of-sample results

To compare the nowcasting performance across ML models, we use the root mean squared error (RMSE) of a competing model M_i relative to the benchmark SPF's median expectations at the nowcast horizon h. The RMSE is defined as follows

$$\mathrm{RMSE}_{M_i,h} = \sqrt{\frac{1}{t_1 - t_0 + 1} \sum_{t=t_0}^{t_1} e^2_{t,M_i,h}}\,, \tag{7}$$

where $e_{t,M_i,h} = \pi_t - \hat{\pi}_{t|t-h;M_i}$ is the corresponding nowcasting error with information up to $t-h$.[23] To test whether nowcasts generated by the ML model M_i are statistically different from the benchmark, we conduct the Diebold-Mariano (DM) test (Diebold and Mariano, 1995).

Table 3 reports the nowcasting performance of competing models, evaluated in terms of RMSE relative to the benchmark, whereas the rows refer to the nowcast horizon and, consequently, the within-month information set. The results underscore the superior performance of shrinkage methods across all nowcast horizons, highlighting the efficacy of employing penalized regressions alongside a comprehensive real-time dataset. Specifically, standard shrinkage via the LASSO, Ridge and ENet consistently yields lower RMSE values, resulting in statistically significant gains of 8.5% up to 17% compared to the median SPF expectations. While the LASSO generally outperforms, Ridge shows slightly better performance when nowcasts are made on day 22. Relative to the Top 5 SPF benchmark, these predictive gains range from 4% to 15%, indicating a substantial difference despite their status as the best-performing institutions before each nowcasting round. Therefore, we match those results of Medeiros et al. (2016) and Garcia et al. (2017) for forecasting Brazilian inflation, which found that techniques based on LASSO outperform at the very short horizon.

Table 3: RMSE: ML methods relative to the SPF benchmark

	SPF		**Shrinkage-based models**				**Tree-based models**			
Horizon	**Median**	**Top 5**	**LASSO**	**Ridge**	**ENet**	**sg-LASSO**	**RF**	**LLF**	**BART**	**LASSO-LLF**
Day 8	1	0.932	0.830^{**}	0.856^{*}	0.842^{**}	0.930^{*}	1.027	0.965	0.963	0.910
Day 15	1	1.014	0.865^{**}	0.879	0.870^{*}	0.955	1.089	1.035	1.033	1.011
Day 22	1	0.942^{*}	0.833^{*}	0.830^{*}	0.833^{*}	0.920	1.247	1.046	1.134	0.983
End-of-month	1	0.951	0.915	0.974	0.915	0.936	1.399	1.371	1.310	1.042

Notes: The table reports the RMSE for each competing model relative to the survey of professional forecasters (SPF, median). Results for the Diebold and Mariano (1995) test in the event of outperformance relative to the benchmark are indicated by the symbols * (5% level) and ** (1% level).

Across information sets, the performance of our ML models relative to the benchmark generally increases with the nowcast horizon. On days 8 and 15, for example, standard shrinkage methods can respectively cut the nowcast errors by 16% and 13.5% on average. Based on the absolute RMSE, this translates into 4.5 and 2.8 basis points of higher accuracy for tracking the year-on-year inflation target. It is worth noting the significant decline in nowcasting gains for the end-of-month horizon. Exclusively at this horizon, our model-based nowcasts have not exhibited statistically significant differences from the benchmark. This trend aligns with the timing of price indicator releases, which predominantly occur towards

[23] The nowcasting evaluation using the mean absolute error (MAE) slightly changes compared to the RMSE metric. This implies that our results are not affected by a few large errors, making them robust to outliers and asymmetries.

the end of the month, particularly the IPCA-15. Such a dynamic likely prompts professional forecasters to increase the frequency of their updates, narrowing the information advantage exploited by our timely nowcasts at longer horizons.

The dominance of LASSO among shrinkage methods indicates the need for a more aggressive variable selection to nowcasting inflation dynamics. This inherent feature of LASSO is particularly advantageous in dealing with the high degree of collinearity among the price indicators within our dataset. Despite cross-validation tuning of the hyperparameter α in ENet, both Ridge and ENet produce more evenly distributed estimates across those highly correlated predictors, resulting in slightly increased overfitting in this context. Furthermore, the sg-LASSO yields a considerably lower precision relative to standard shrinkage for longer horizons (day 8 up to day 22), possibly affected by the presence of ragged edges (see Section 4.3). For end-of-month horizons, where ragged edges are eliminated, sg-LASSO outperforms Ridge and narrows the gap against LASSO and ENet. This highlights the importance of an assertive selection of predictors and suggests that our baseline setting is not high-dimensional enough for sg-LASSO to thrive.

What drives the poor nowcasting results of tree-based methods? Although the flexibility of these methods allows for potentially detecting turning points and complex nonlinear dynamics in the data, their 'need' for large quantities of data characteristic leads to poor performance in our setup. Notable exceptions are observed for the longest horizon of day 8, where LLF, BART and LASSO-LLF yield RMSE reductions between 3.5% and 9% relative to the benchmark, tough statistical significance is not achieved. The slightly improved performance of LASSO-LLF corroborates the hypothesis that tree-based methods might be ill-equipped to handle the limited samples of macroeconomic time series; pre-selecting strong predictors from a large dataset works best and goes in line with the recommendation from Friedberg et al. (2020) when dealing with large datasets for the LLF. Moreover, these results suggest the absence of relevant temporal nonlinearities in Brazilian data.

The findings in Table 3 prompt the question of whether relative performance is constant throughout the evaluation period or largely affected by inflationary shocks. To gain further insights into the evolution of loss differentials, we report the cumulative sum of squared forecast error:

$$\text{CUMSFE}_{M_i,h} = -\sum_{t=t_0}^{t_1} \left(e^2_{t,M_i,h} - e^2_{t,M_{\text{SPF}},h} \right) . \tag{8}$$

A positive value of CUMSFE indicates an outperformance of the ML model M_i relative to the benchmark median SPF expectations for horizon h and from period t_0 up to t_1. Negative values imply the opposite.

Figure 3: CUMSFE: shrinkage methods versus the SPF benchmark

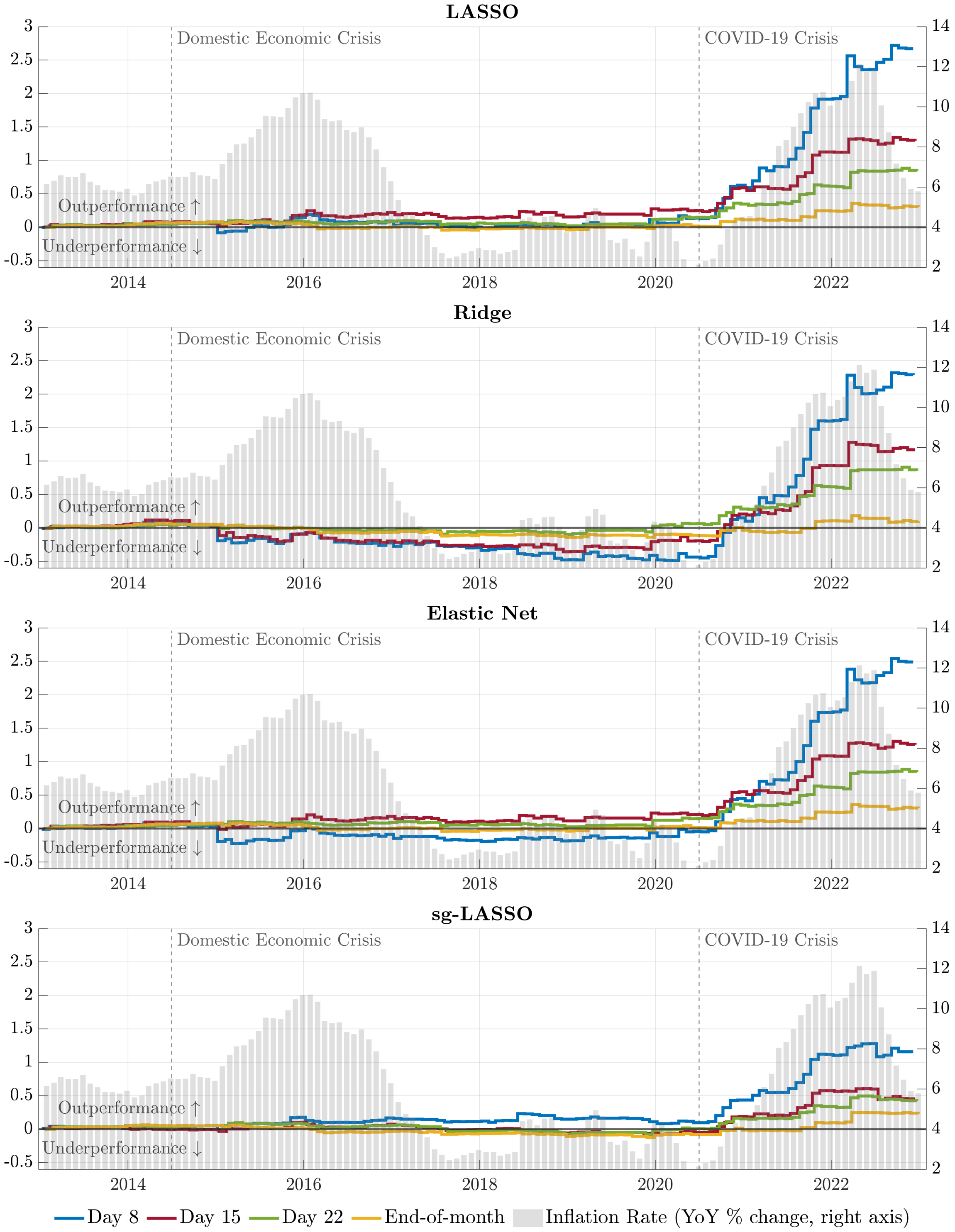


Figure 3 exhibits CUMSFE developments for shrinkage methods across different nowcast horizons. It becomes crystal clear that the inflationary period following the COVID-19 crisis is a game changer in terms of loss differentials. Particularly, large nowcasting gains build up from September 2020. During this period of persistent high inflation, we observe the largest jumps in CUMSFE for nowcasts made on days 8 and 15 using the LASSO. Moderate gains are also achieved when nowcasting at shorter horizons. It is worth emphasizing that these findings during the pandemic mostly drive performance evaluation based on full-sample metrics, as in Table 3. For instance, our previous hypothesis based on the performance across information sets – professional forecasters tend to update their expectations more frequently as the information set increases within the reporting month – is a trend predominantly observed within the context of the COVID-19 crisis.

Turning to the years preceding the pandemic, differences in predictive accuracy between shrinkage methods and SPF expectations are modest across all shrinkage methods. LASSO and sg-LASSO prove the most reliable models by consistently keeping an edge relative to the benchmark for most horizons. In contrast, nowcasts purely based on tree-based models can be highly detrimental during calm times, as shown in Figure B1 of Appendix B. The LASSO-LLF performs roughly on par with SPF expectations, reinforcing the idea that a pre-selection step proves beneficial for tree-based methods. Nonetheless, amidst the COVID-19 crisis, most tree-based models exhibit a rising competitive advantage over the benchmark on days 8 and 15.

Figure B2 in Appendix B reports the fluctuation test, introduced by Giacomini and Rossi (2010), and reaffirms the previous analysis. Predictive gains relative to SPF expectations change substantially over time, depending on the model and horizon, and are prominent in the aftermath of the pandemic. Standard shrinkage via the LASSO, Ridge and ENet deliver occasional significant gains at the 10% level throughout 2021. Other models also produce statistically significant gains during this period: sg-LASSO on day 8, and LASSO-LLF on both day 8 and end-of-month. Besides, the picture reveals a clear discrepancy between shrinkage- and tree-based models, as expected from previous results. Finally, a higher dispersion of prediction accuracy across models can be observed during turbulent times such as the domestic economic crisis of 2014 and the pandemic.

4.2 Interpreting the best-performing model

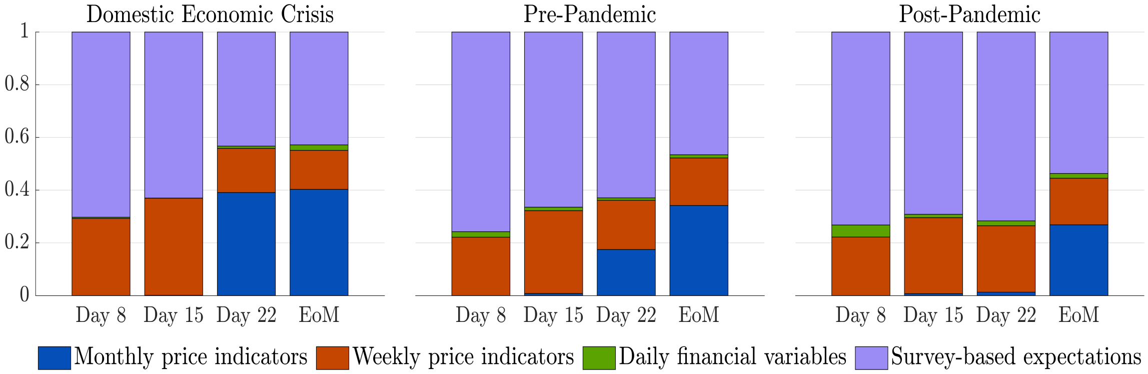
Based on the variable selection performed by our most effective strategy, we investigate the relative importance of the selected predictors fitted via the LASSO. Heatmaps illustrating the evolution of coefficient estimates at each nowcast horizon are presented in Figure B3 of Appendix B. The x-axis denotes a nowcasting round in the evaluation sample, while predictors are displayed in the y-axis. Consequently, the coefficient value associated with a given predictor on a specific date within the evaluation sample determines the color intensity reflected in the graph.

Comparing all panels in Figure B3, we observe that LASSO prompts a fairly sparse structure at higher nowcast horizons while a more dense structure prevails at horizons approaching the end of the month. Two factors contribute to this pattern: (i) increased availability of data on monthly price indicators as the month unfolds, and (ii) signals associated with price developments in the reporting month t become more accurate as month-on-month rates rely less on the information set from $t-1$. Not surprisingly, on days 8 and 15 (panels B3a and B3b), SPF expectations stand out as the primary predictor, with average coefficient estimates near 0.7 but exceeding 0.8 towards the end of the sample. Alongside SPF, the high-frequency price indicators IPC-S and FIPE are consistently selected across the entire evaluation sample, albeit with comparatively smaller coefficients – e.g., on average, 0.16 and 0.1 respectively for day 15. At the same time, energy prices, interest rate variables and commodity prices regularly enter the forecasting model, though with modest coefficient values.

As the horizons approach the end of the month, the low-frequency but timely indicator IPCA-15 takes on enormous importance, with coefficient estimates reaching 0.5 in many cases. Conversely, SPF expectations lose a substantial portion of their relevance. One plausible hypothesis is that professional forecasters adjust their survey responses in response to the release of this indicator.

Using a more aggregate approach, we assess the joint relevance of each class of predictors across different nowcast horizons and sub-periods. As a variable-importance measure, Figure 4 depicts the weighted sum of absolute LASSO estimates grouped into four categories of predictors as described in Table 1: monthly price indicators, weekly price indicators, daily financial variables, and daily SPF expectations. As shown previously, SPF expectations, closely trailed by weekly price indicators (particularly IPC-S and FIPE), exert the most substantial impact on shaping our model-based nowcasts. This suggests that SPF expectations not only incorporate up-to-date information from our dataset but also integrate informed judgment that extends beyond relying solely on hard predictors for inflation. However, as recent data on monthly and weekly price indicators becomes available throughout the reporting month, their relevance in model estimation rises, subsequently diminishing SPF's weights as we approach the end-of-month horizon. Particularly, the availability of contemporaneous data on monthly price indicators after the third week typically elevates their relative importance when nowcasting on day 22 and end-of-month. On the other hand, financial variables play a minor role in shaping our model-based nowcasts due to their limited informativeness regarding current inflation dynamics, especially when compared to the signals already present in the dataset.

Figure 4: Variable relevance via coefficient estimates using LASSO



Notes: This Figure reports the weighted sum of absolute coefficient estimates fitted via the LASSO and grouped into different categories of predictors (see Table 1) on days 8, 15, 22 and end-of-month (EoM). The "Domestic Economic Crisis" covers the period from 2013 to 2016 while March 2020 divides the "Pre-Pandemic" period and the start of the COVID-19 crisis ("Post-Pandemic").

Turning to different sub-periods, the informed judgment in SPF data appears to weigh more heavily on nowcasting the inflation surge following the pandemic. This tendency toward increasing SPF weights is already evident in the calm period preceding the pandemic, particularly when nowcasts are made on day 22. In this case, monthly price indicators lose significant ground relative to SPF and weekly price indicators. This suggests that SPF encompasses more robust signals about the target dynamics as we advance in the sample, possibly stemming from a more timely update of forecasts made by specialists as new information on other predictors is released. Additionally, off-model information proves more valuable during turbulent times, especially if the nature of post-pandemic price spikes differs from the nature of past inflationary shocks in the sample. For example, the inflationary wave induced by 2014's domestic economic crisis was more accurately anticipated by relying solely on hard signals from price indicators.

The natural question that follows is what part of the information set mostly contributes to the outperformance relative to the tough SPF benchmark. To address this question, we replicate our recursive exercise using the SPF nowcasting errors as the dependent variable in LASSO regressions and plot the period-wise estimates in Figure B4 of Appendix B. Surprisingly, professional forecasters mainly overlook recent data increments of FIPE when nowcasting at longer horizons. Other weekly price indicators are also partially disregarded across all horizons but to a smaller extent. As for financial variables, they hardly contribute to explaining expert's errors; except for occasional minor effects of Bloomberg's commodity index and interest rate movements (SELIC, DI10 and SPREAD) for shorter horizons. Although a slightly negative intercept estimate shows up across the board, we observe a consistent downward bias of SPF expectations. Notably during the pandemic period after March 2020, where the average nowcast error based on year-on-year percentage points yields 0.14 for end-of-month nowcasts compared to only 0.003 in the pre-pandemic sample. Moreover, it appears that experts do not fully adjust for their past errors given the significant effect of the lagged dependent variable on day 8 and end-of-month.

Therefore, SPF nowcast errors can be partially predicted with the relevant information set available at the nowcast date, explaining the additional improvements reported in Section 4.1. But what if we add these LASSO-based forecasts for SPF nowcast errors back to the SPF expectations to obtain implied nowcasts for the IPCA target? In terms of relative RMSE as in Table 3, this modeling strategy fares better than our previous nowcasts on days 8 and 15. More precisely, respective predictive gains of 18.5% and 16.5% are now obtained compared to the SPF benchmark. For the remaining horizons, the nowcasting precision significantly drops when competing against the best-performing models in Table 3.

4.3 A deeper assessment of key modeling features: guiding accurate inflation nowcasts

What drives the accuracy of our inflation nowcasts? To explain why the baseline mixed-frequency ML structure introduced in Section 3.1 coupled with a highly informative dataset successfully outperforms tough benchmarks, we assess the value added of key modeling choices in Eq. (1). More specifically, we investigate three aspects: the impact of the high-frequency lag choice; the impact of eliminating ragged edges; and the impact of using SPF expectations in the predictor set. It is noteworthy to recall that our baseline specification of (1), to which we compare the alternative strategies, features the following choices: only the most recent high-frequency data enters the model by setting $p = 0$, there is no ragged-edge problem, and we include the SPF as a predictor.

First, we investigate whether past month-on-month high-frequency regressors carry predictive value beyond the most recent signal available at the nowcast date $t-h$. This is done by extending our baseline choice of $p = 0$ to account for $p = 3$ high-frequency lags. Hence, our alternative specification here includes the four most recent high-frequency information $\boldsymbol{x}_{t-h}^{(w)}, \ldots, \boldsymbol{x}_{t-h-3/4}^{(w)}$. For example, if we stand at 31 December, we include the high-frequency signals stemming from $\boldsymbol{x}_{31\text{ Dec}}^{(w)}, \boldsymbol{x}_{22\text{ Dec}}^{(w)}, \boldsymbol{x}_{15\text{ Dec}}^{(w)}, \boldsymbol{x}_{8\text{ Dec}}^{(w)}$ rather than just $\boldsymbol{x}_{31\text{ Dec}}^{(w)}$. In general terms, we increase the high-frequency predictor set by a factor of 4.

The impact in terms of RMSE from incorporating these additional high-frequency lags can be seen in Figure 5a. The plot indicates that RMSE values generally deteriorate with the inclusion of additional high-frequency lags. This pattern is more pronounced across longer nowcast horizons (days 8, 15 and 22) and tree-based models such as LLF. However, at the end-of-month (yellow dots) a marginal increased gain of 2-5% in RMSE is observed across shrinkage methods, most notably for Ridge regression. Therefore, the last available high-frequency predictor on its own already carries the relevant signal for updating the nowcasts throughout the reporting month, except at end-of-month, whereas including lagged high-frequency information is slightly favorable.

Secondly, we keep the alternative assumption of $p = 3$, however, we redesign the model to incorporate exclusively contemporaneous high-frequency lags $\boldsymbol{x}_t^{(w)}, \ldots, \boldsymbol{x}_{t-3/4}^{(w)}$ at any nowcast day in month t. Additionally, we assume that the low-frequency predictor set must include all monthly price indicators in the dataset at any horizon h, regardless of their publication timing. These choices give rise to missing observations at the end of the sample (ragged-edge problem) when nowcasting on days 8, 15 and 22. As for sg-LASSO, we complete the ragged edges with random-walk nowcasts based on the most recently released information.

Figure 5b points to a considerable worsening of the nowcast precision when the model suffers from the ragged-edge problem. Except for the RF model in the first week of the month, the loss in performance is consistent within both classes of ML methods. Shrinkage methods, particularly at shorter nowcast horizons, exhibit the most pronounced susceptibility to ragged edges, experiencing average losses approaching 60%. These findings confirm the consensus in the MIDAS literature suggesting that ragged edges worsen the forecasting properties of the model, especially in the very short run (see, e.g., Marcellino and Schumacher, 2010; Andreou et al., 2013; Monteforte and Moretti, 2013).

The reasons for this underperformance are twofold. Firstly, during model estimation, substantial weight is assigned to monthly predictors, which can only bring outdated information from $t-1$ to construct the nowcast for π_t, particularly at longer horizons. Secondly, the multicollinearity arising from the inclusion of $p = 3$ high-frequency lags in the predictor set somewhat disorients ML methods, hindering their ability to identify accurate high-frequency signals.

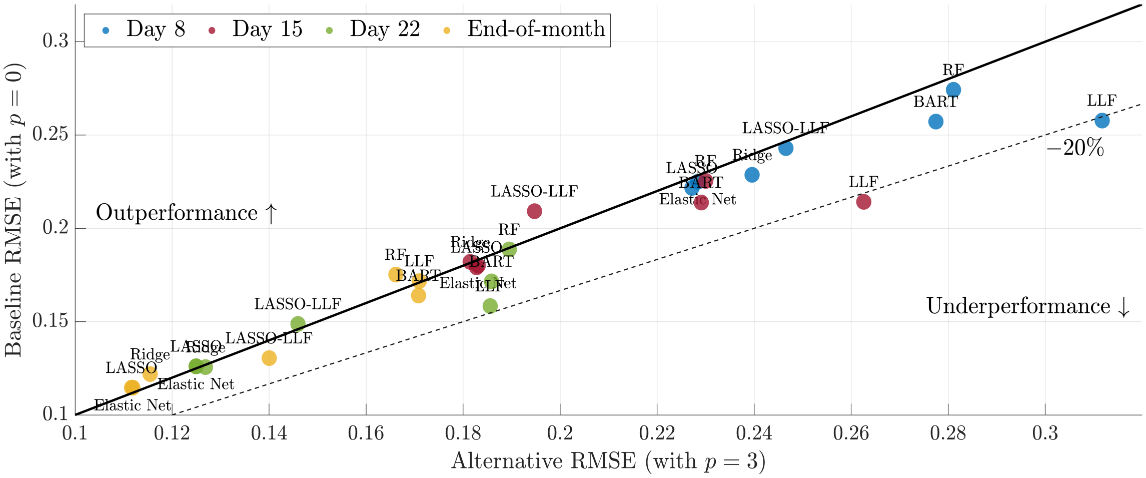
Third, we investigate the benefits of incorporating some degree of informed judgment entailed in SPF median expectations. Professional forecasters do not solely rely on models to form their expectations about short-run inflation dynamics but these can also be attributed to judgment, particularly in challenging times such as the COVID-19 crisis where purely model-based forecasts are adversely affected. Since our baseline strategy includes SPF as a high-frequency predictor, we compare it against the alternative specification that excludes any SPF information from the predictor set. Notably, Figure 5c shows that adding meaningful off-model information from SPF leads to sizeable nowcasting advantages. Specifically, predictive gains are substantially higher across shrinkage methods, averaging from

27% on day 8 to 36% at end-of-month. This indicates that SPF information on inflation expectations can better discipline parametric model structures.

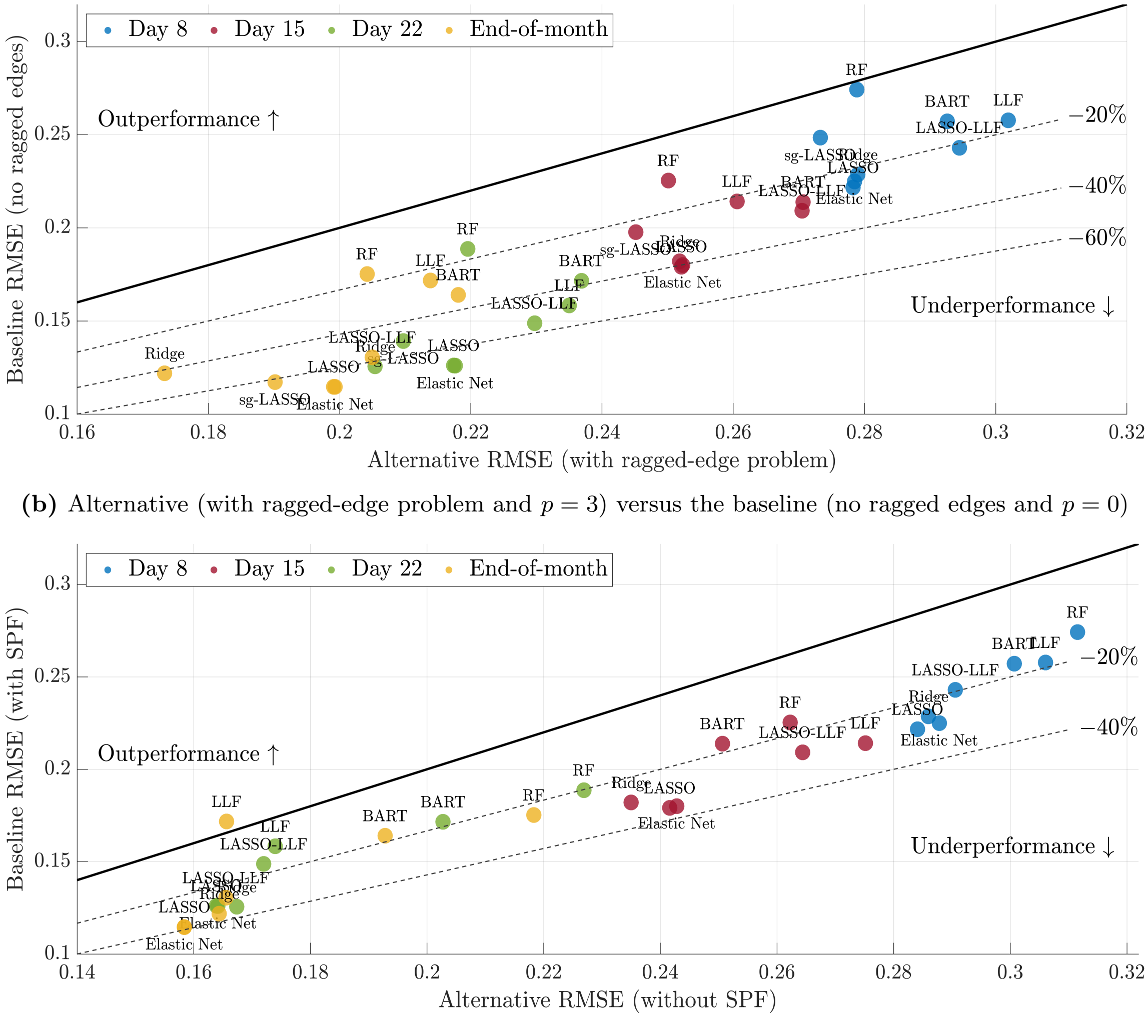
Amidst the turbulence induced by the pandemic, professional forecasters tended to underestimate the inflation surge. Nevertheless, informed judgment in SPF data carried relevant information about rapidly unfolding inflationary trends beyond what was reflected by other predictors in our dataset. In addition, the previous discussion on Figure 4 reveals a growing relevance of SPF in constructing LASSO-based nowcasts, especially post-2021. Consequently, environments marked by highly unpredictable and elevated inflation, like the pandemic, are best suited for enriching model-based inflation nowcasts with SPF expectations. These findings align with most of the previous studies that investigate the value added of SPF expectations into model-based forecasts (see, e.g., Banbura et al., 2021b; Bobeica and Hartwig, 2023).

In summary, superior nowcasting accuracy predominantly stems from the combination of a well-designed mixed-frequency ML structure with carefully selected predictors that include some degree of informed judgment in SPF expectations. Specifically, the prediction model must be free from ragged edges. This is first attained through high-frequency leads, preferably focusing solely on the last available high-frequency signal conveyed by $\boldsymbol{x}_{t-h}^{(w)}$. Furthermore, the inclusion of monthly predictors in the model specification should be guided by their real-time data releases; in particular, only those with available contemporaneous data by the nowcast date.

Figure 5: Absolute RMSE: alternative versus the baseline specification



(a) Alternative (with $p = 3$ high-frequency lags) versus the baseline (with $p = 0$)



(b) Alternative (with ragged-edge problem and $p = 3$) versus the baseline (no ragged edges and $p = 0$)

(c) Alternative (without SPF) versus the baseline (including SPF as predictor)

Notes: This Figure reports the absolute RMSE of the alternative specification versus the baseline model specified in Section 3.1. Points below (above) the 45-degree reference line, in solid black, indicate an underperformance (outperformance) of the alternative specification for a given competing ML method and nowcast horizon.

5 Summary and conclusions

Machine learning methods have recently gained considerable traction as standard tools for macroeconomic nowcasting, offering an effective solution to handle the increasing availability of high-frequency information stemming from all parts of the economy. In the wake of disruptive events like the COVID-19 pandemic, the demand for such nowcasts has intensified. Yet there remains a notable gap in harnessing ML methods to leverage high-frequency signals for real-time inflation nowcasting.

To address this gap, our study compares shrinkage methods with tree-based models in an environment characterized by persistently high inflation. Our empirical exercise underscores the importance of a well-designed mixed-frequency ML framework to construct robust inflation nowcasts that consistently outperform SPF expectations, with major nowcasting benefits during the COVID-19 inflation surge. Moreover, we show that good nowcasts depend on variable selection performed via the LASSO combined with accurate timely signals from price indicators and informed judgment entailed in SPF data. The findings highlight the adaptability of shrinkage methods to produce accurate nowcasts across different horizons while tree-based methods lead to poor performance due to the limited time series sample and the plausible absence of temporal nonlinearities in our setup. Overall, the timely and high-frequency character of the Brazilian real-time dataset offers valuable insights for policymakers and practitioners seeking to refine their inflation forecasting capabilities in uncertain economic landscapes.

Variable importance analysis via the LASSO fitted coefficients shows that model selection heavily depends on the contemporaneous information set available at the time of the nowcast. Specifically, at longer nowcast horizons, a more sparse model delivers higher predictive gains compared to the benchmark, while exploiting early information from weekly price indicators and SPF expectations. At shorter horizons, shrinkage-based models yield a denser structure that also assigns substantial importance to monthly price indicators, which only enter the predictor set when their contemporaneous signal becomes available. In general, financial variables play a minor role but the combination of timely price indicators with SPF judgments proves highly influential in shaping our model-based nowcasts.

Furthermore, our study sheds light on key modeling choices in a mixed-frequency ML framework. The results suggest implementing the following strategies to achieve higher performance: (i) account for expert judgment in the predictor set, (ii) make the prediction model free of ragged edges, (iii) align the model specification with the release calendar of monthly predictors, and (iv) prioritize the most recent high-frequency signal available in the information set. With our framework, we can significantly improve upon SPF expectations, even outperforming the Top 5 SPF institutions, which are widely regarded as the most challenging benchmark for forecasting Brazilian inflation dynamics. As a fruitful avenue for further research, one could expand our analysis to encompass additional classes of ML methods and contrast them with traditional econometric frameworks such as factor models and mixed-frequency Bayesian VARs. Moreover, one could assess the economic value of our nowcasting gains in monetary policy decisions and portfolio allocation strategies.

References

Aliaj, T., Ciganovic, M., and Tancioni, M. (2023). Nowcasting inflation with lasso-regularized vector autoregressions and mixed frequency data. Journal of Forecasting, 42(3):464–480.

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. (2023). Machine learning methods for inflation forecasting in brazil: New contenders versus classical models. Latin American Journal of Central Banking, 4(2):100087.

Arlot, S. and Celisse, A. (2010). A survey of cross-validation procedures for model selection. Statistics Surveys, 4.

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

Atkeson, A. and Ohanian, L. E. (2001). Are Phillips Curves Useful for Forecasting Inflation? Federal Reserve Bank of Minneapolis Quarterly Review, 25(1):2–11.

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.

Banbura, M., Brenna, F., Paredes, J., and Ravazzolo, F. (2021a). Combining bayesian VARs with survey density forecasts: does it pay off? ECB Working Paper.

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

Banbura, M., Leiva-Leon, D., and Menz, J.-O. (2021b). Do inflation expectations improve model-based inflation forecasts? Banco de Espana Working Paper.

Barbaglia, L., Frattarolo, L., Onorante, L., Pericoli, F. M., Ratto, M., and Pezzoli, L. T. (2023). Testing big data in a big crisis: Nowcasting under covid-19. International Journal of Forecasting, 39(4):1548–1563.

Barkan, O., Benchimol, J., Caspi, I., Cohen, E., Hammer, A., and Koenigstein, N. (2023). Forecasting cpi inflation components with hierarchical recurrent neural networks. International Journal of Forecasting, 39(3):1145–1162.

Beck, G., Carstensen, K., Menz, J.-O., Schnorrenberger, R., and Wieland, E. (2023). Nowcasting Consumer Price Inflation Using High-Frequency Scanner Data: Evidence from Germany. Deutsche Bundesbank Discussion Paper, 34.

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.

Bobeica, E. and Hartwig, B. (2023). The covid-19 shock and challenges for inflation modelling. International journal of forecasting, 39(1):519–539.

Borup, D., Rapach, D. E., and Schütte, E. C. M. (2023). Mixed-frequency machine learning: Nowcasting and backcasting weekly initial claims with daily internet search volume data. International Journal of Forecasting, 39(3):1122–1144.

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. (2015). Real-time nowcasting with a bayesian mixed frequency model with stochastic volatility. Journal of the Royal Statistical Society Series A: Statistics in Society, 178(4):837–862.

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).

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

Cascaldi-Garcia, D., Ferreira, T. R., Giannone, D., and Modugno, M. (2023). Back to the present: Learning about the euro area through a now-casting model. International Journal of Forecasting.

Chinn, M. D., Meunier, B., and Stumpner, S. (2023). Nowcasting world trade with machine learning: a three-step approach. Technical report, National Bureau of Economic Research.

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

Cimadomo, J., Giannone, D., Lenza, M., Monti, F., and Sokol, A. (2022). Nowcasting with large bayesian vector autoregressions. Journal of Econometrics, 231(2):500–519.

Clark, T. E., Leonard, S., Marcellino, M., and Wegmüller, P. (2022). Weekly Nowcasting US Inflation with Enhanced Random Forests. Mimeo.

Dahlhaus, T., Guénette, J.-D., and Vasishtha, G. (2017). Nowcasting bric+ m in real time. International Journal of Forecasting, 33(4):915–935.

Diebold, F. X. and Mariano, R. S. (1995). Comparing predictive accuracy. Journal of Business & Economic Statistics, 13(3):253–263.

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

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. (2017). 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.

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.

Harchaoui, T. and Janssen, R. (2018). How Can Big Data Enhance the Timeliness of Official Statistics?: The Case of the U.S. Consumer Price Index. International Journal of Forecasting, 34(2):225–234.

Hauzenberger, N., Huber, F., and Klieber, K. (2023). Real-Time Inflation Forecasting Using Non-Linear Dimension Reduction Techniques. International Journal of Forecasting, 39(2):901–921.

Hindrayanto, I., Koopman, S. J., and de Winter, J. (2016). Forecasting and nowcasting economic growth in the euro area using factor models. International Journal of Forecasting, 32(4):1284–1305.

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

Huber, F., Koop, G., Onorante, L., Pfarrhofer, M., and Schreiner, J. (2023). Nowcasting in a pandemic using non-parametric mixed frequency vars. Journal of Econometrics, 232(1):52–69.

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. Bank of England staff working papers, (915).

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

Knotek II, E. S. and Zaman, S. (2023). Real-time density nowcasts of us inflation: A model combination approach. International Journal of Forecasting, 39(4):1736–1760.

Kohns, D. and Potjagailo, G. (2023). Flexible Bayesian MIDAS: time-variation, group-shrinkage and sparsity. Bank of England staff working papers, 1025.

Krüger, F., Clark, T. E., and Ravazzolo, F. (2017). Using entropic tilting to combine bvar forecasts with external nowcasts. Journal of Business & Economic Statistics, 35(3):470–485.

Macias, P., Stelmasiak, D., and Szafranek, K. (2023). Nowcasting Food Inflation with a Massive Amount of Online Prices. International Journal of Forecasting, 39(2):809–826.

Marcellino, M. and Schumacher, C. (2010). Factor midas for nowcasting and forecasting with ragged-edge data: A model comparison for german gdp. Oxford Bulletin of Economics and Statistics, 72(4):518–550.

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. Banque de France working paper, 520.

McCracken, M. W., Owyang, M., and Sekhposyan, T. (2015). Real-time forecasting and scenario analysis using a large mixed-frequency bayesian var. FRB St. Louis Working Paper, (2015-30).

Medeiros, M. C., Vasconcelos, G., and Freitas, E. (2016). Forecasting brazilian inflation with high-dimensional models. Brazilian Review of Econometrics, 36(2):223–254.

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.

Monteforte, L. and Moretti, G. (2013). Real-Time Forecasts of Inflation: The Role of Financial Variables. Journal of Forecasting, 32(1):51–61.

Powell, B., Nason, G., Elliott, D., Mayhew, M., Davies, J., and Winton, J. (2018). Tracking and modelling prices using web-scraped price microdata: towards automated daily consumer price index forecasting. Journal of the Royal Statistical Society Series A: Statistics in Society, 181(3):737–756.

Richardson, A., van Florenstein Mulder, T., and Vehbi, T. (2021). Nowcasting gdp using machine-learning algorithms: A real-time assessment. International Journal of Forecasting, 37(2):941–948.

Schorfheide, F. and Song, D. (2015). Real-time forecasting with a mixed-frequency var. Journal of Business & Economic Statistics, 33(3):366–380.

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

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

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 Mixed-frequency framework in matrix form

For expositional simplicity, let us reduce the general multiple-predictors specification (1) to the single generic high-frequency predictor $x_t^{(w)}$ and neglect both the low-frequency predictors and seasonal dummies. From there, assume the latest data release for the target variable is associated with a given month t. Based on the high-frequency information set available up to the nowcast point, say $t+1-h$, and pre-sample information $\{\pi_0, x_0^{(w)}, x_{0-1/4}^{(w)}, \ldots, x_{0-p/4}^{(w)}\}$, one can construct the nowcast for π_{t+1} at horizon $h = j/w$, with $j \in \{0,1,2,3\}$, by using the following matrix representation for model estimation:

$$\begin{bmatrix} \pi_1 \\ \pi_2 \\ \vdots \\ \pi_t \end{bmatrix} = \begin{bmatrix} 1 & \pi_0 & x_{1-h}^{(w)} & x_{1-h-\frac{1}{4}}^{(w)} & x_{1-h-\frac{2}{4}}^{(w)} & \cdots & x_{1-h-\frac{p}{4}}^{(w)} \\ 1 & \pi_1 & x_{2-h}^{(w)} & x_{2-h-\frac{1}{4}}^{(w)} & x_{2-h-\frac{2}{4}}^{(w)} & \cdots & x_{2-h-\frac{p}{4}}^{(w)} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & \pi_{t-1} & \underbrace{x_{t-h}^{(w)}}_{\text{nowcast day } (nd)} & \underbrace{x_{t-h-\frac{1}{4}}^{(w)}}_{nd-\frac{1}{4}} & \underbrace{x_{t-h-\frac{2}{4}}^{(w)}}_{nd-\frac{2}{4}} & \cdots & \underbrace{x_{t-h-\frac{p}{4}}^{(w)}}_{nd-\frac{p}{4}} \end{bmatrix} \begin{bmatrix} c \\ \rho_1 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \vdots \\ \beta_{p+1} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_t \end{bmatrix} \tag{A1}$$

For example, suppose we stand at day 15 of December and we want to construct the nowcast for π_{Dec} assuming a general high-frequency lag order p. In this case, the forecast horizon is $h = 2/4$ and we estimate the model using monthly data until November and weekly data until 15 November. To account for the lags p, the last high-frequency observations in (A1) will respectively be associated with 15 November, 8 November, 31 October, 22 October, 15 October, and so on up to the corresponding lag-length p. From there, the nowcast for π_{Dec} is constructed using the estimated coefficients and all the low- and high-frequency information available until 15 December.

Ultimately, note that (A1) makes explicit that the general prediction model is still written at the monthly frequency but accounting for the w high-frequency time increments within each common period t. The nowcast for the inflation rate at periods $t+1, \ldots, T$ can then be updated regularly using the high-frequency data increments that become available after t and well before official releases of the target inflation rate.

Appendix B Supplementary results

Figure B1: CUMSFE: tree-based methods versus the SPF benchmark

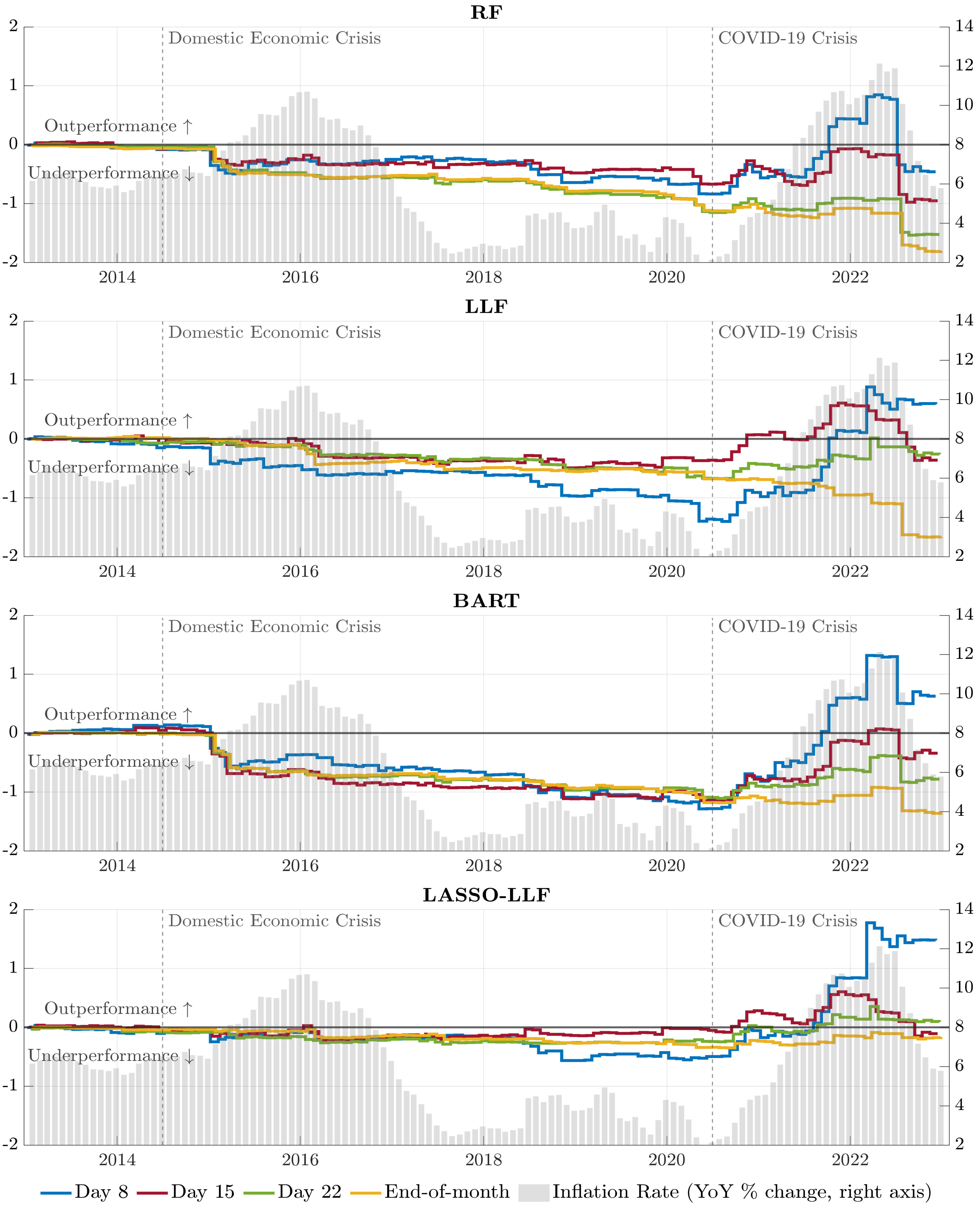
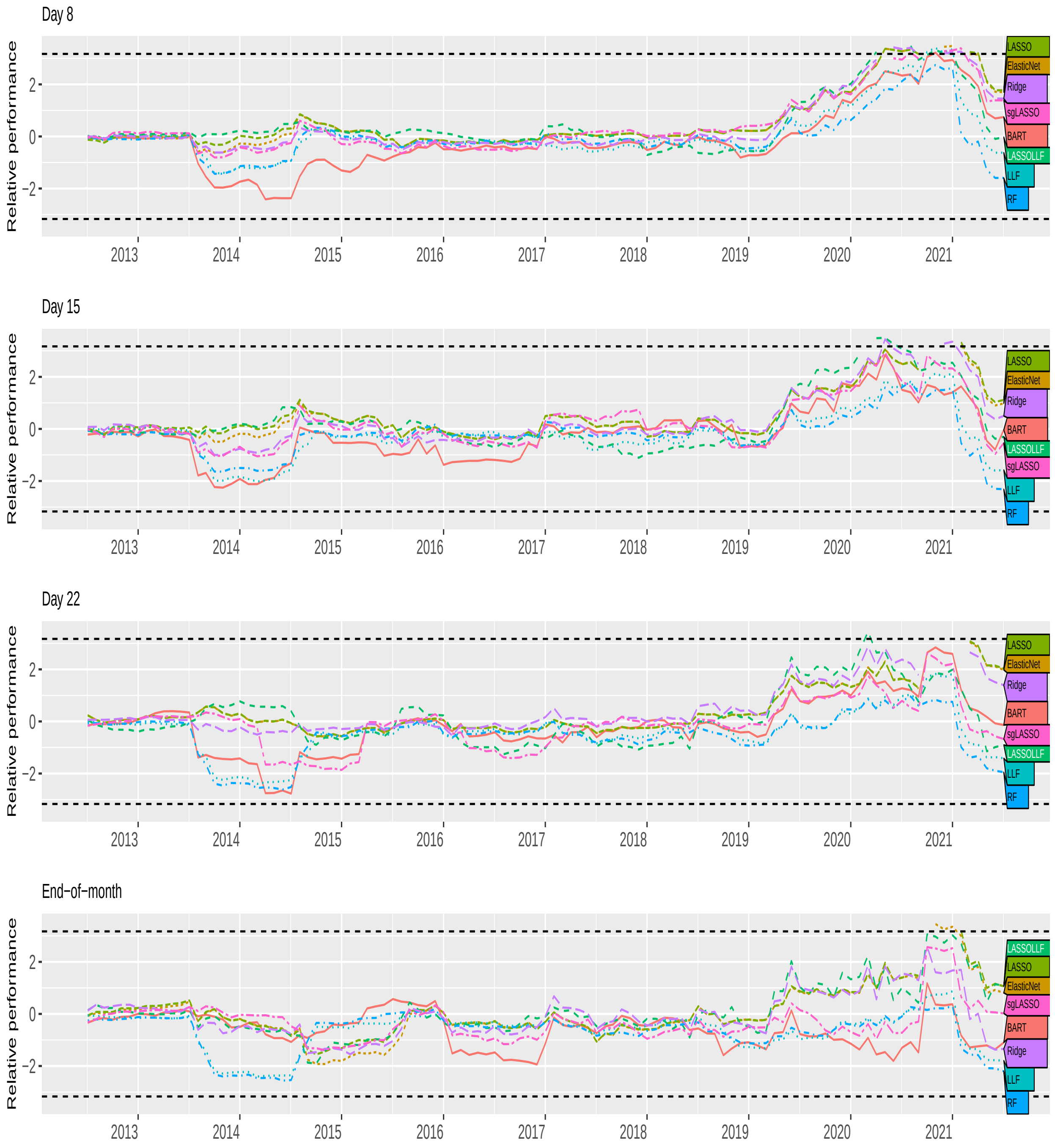
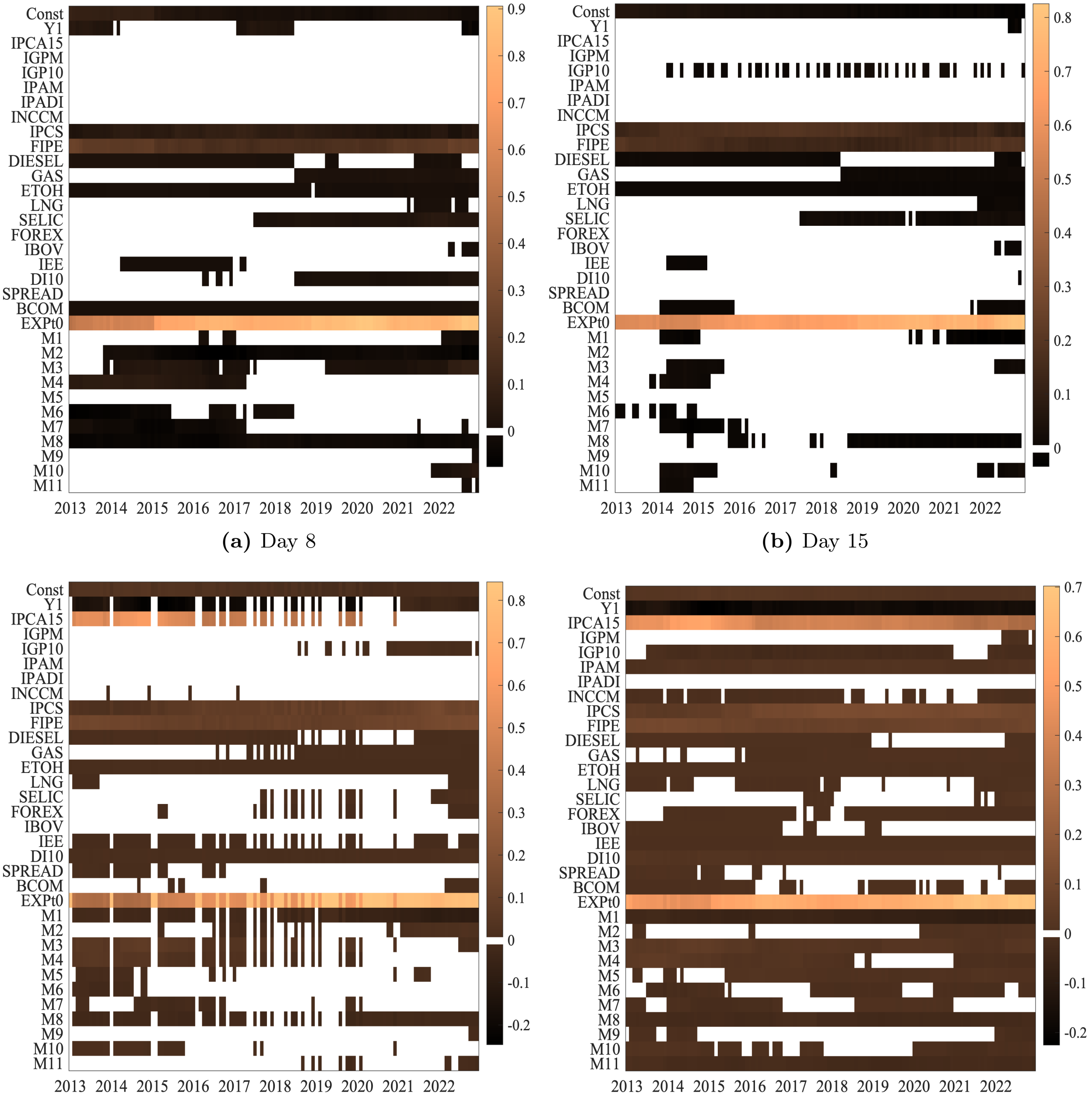


Figure B2: Fluctuation test: ML competing models versus the SPF benchmark



Notes: This Figure reports the fluctuation test from Giacomini and Rossi (2010) based on the squared loss differential between a machine learning method and SPF nowcasts. Areas between the horizontal dashed lines correspond to the 90% confidence interval of the two-sided statistical 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.

Figure B3: Heatmap of coefficient estimates using LASSO



(a) Day 8

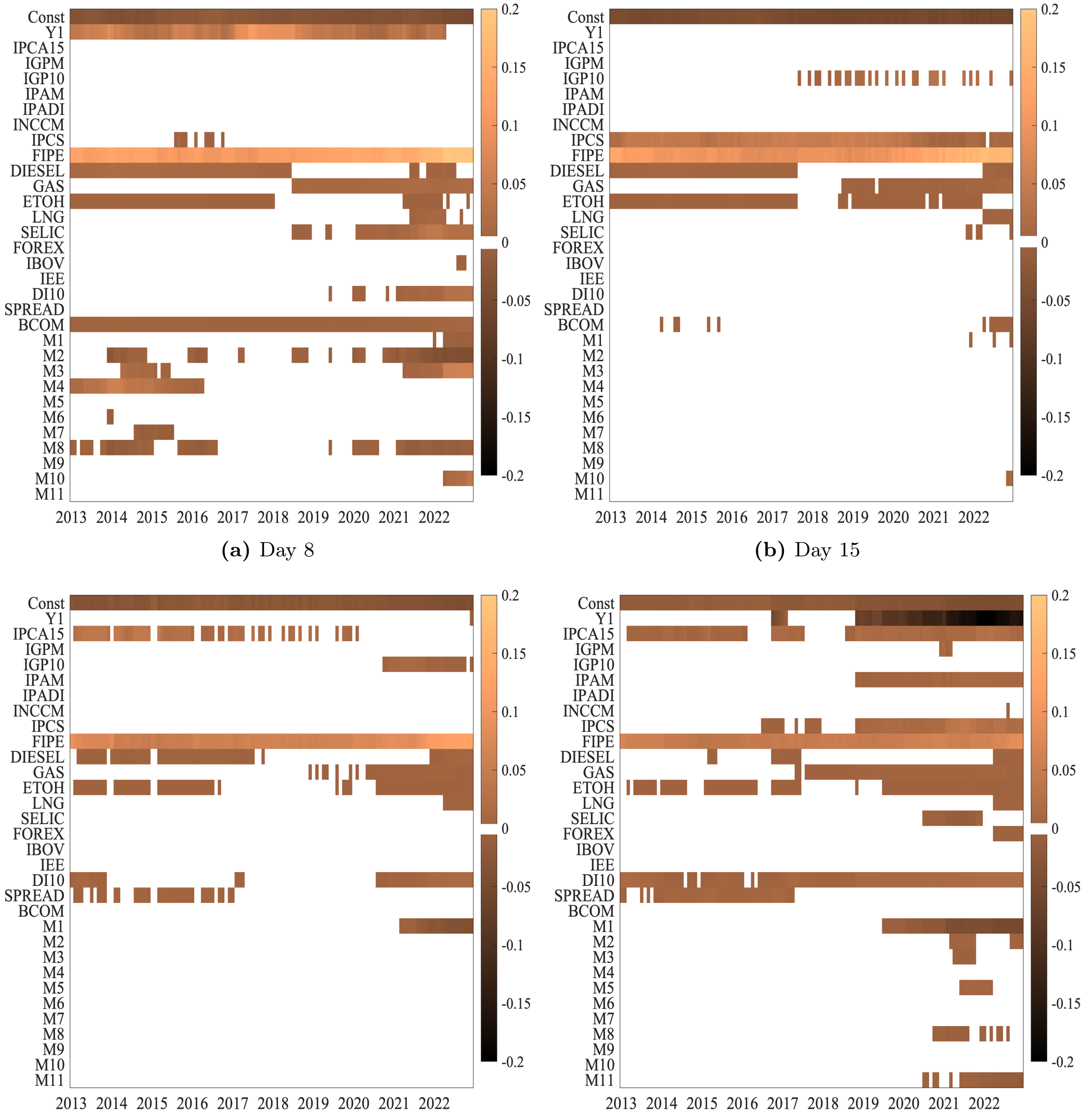
(b) Day 15

(c) Day 22

(d) End-of-month

Notes: This Figure depicts heatmaps of LASSO-fitted coefficients over the evaluation period. Empty cells represent a coefficient estimate equal to zero, and thus a predictor that has not been selected at the estimation round t in the evaluation period.

Figure B4: Heatmap of coefficient estimates using LASSO on the SPF nowcasting errors



(a) Day 8

(b) Day 15

(c) Day 22

(d) End-of-month

Notes: This Figure depicts heatmaps of LASSO-fitted coefficients using SPF nowcasting errors as the dependent variable. Empty cells represent a coefficient estimate equal to zero, and thus a predictor that has not been selected at the estimation round t in the evaluation period.

DeNederlandscheBank
EUROSYSTEEM

De Nederlandsche Bank N.V.
Postbus 98, 1000 AB Amsterdam
020 524 91 11
dnb.nl