Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Nowcasting Mexican Manufacturing Activity with Traditional and Non-Traditional Indicators: A High-Dimensional Regularization Approach

Manuel Torres Favela

manuel.torresf@udem.edu

Universidad de Monterrey https://orcid.org/0000-0002-8689-4767

Edgar Mauricio Luna Domínguez

Universidad Autónoma de Nuevo León https://orcid.org/0000-0002-5245-3357

Research Article

Keywords: Nowcasting, Industrial Production, Regularization Models, Dynamic Factor Models, Google Trends

Posted Date: June 11th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9990146/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Nowcasting Mexican Manufacturing Activity with Traditional and Non-Traditional Indicators: A High-Dimensional Regularization Approach

Manuel Torres Favela[1*] Edgar M. Luna Domínguez[2†]

[1]Universidad de Monterrey

[2]Universidad Autónoma de Nuevo León

Abstract

This paper develops a high-dimensional nowcasting framework for Mexico's Monthly Industrial Production Index (IMAI) using 302 monthly indicators combining traditional and non-traditional high-frequency information. Under a pseudo-real-time design, we compare penalized regressions with ARMA errors (Ridge, LASSO, Elastic Net, Adaptive LASSO, Adaptive Elastic Net, and Multi-Step Adaptive Elastic Net) and DFM-based alternatives.

All models outperform a naïve AR(1) benchmark. The best result is achieved by the Multi-Step Adaptive Elastic Net (MSAENET) with a relative RMSFE of 0.273. Non-traditional indicators improve performance mainly under adaptive penalties, with Diebold–Mariano evidence favoring the richer information set. DFM-based specifications improve with non-traditional information but remain less accurate than the best penalized regressions.

Forecast gains stem from disciplined variable selection. MSAENET selects 10.4 predictors on average, with the Google Trends topic *Drought* as the only non-traditional predictor consistently retained, reflecting Mexico's industrial vulnerability to climatic conditions. A Ridge model estimated on the seven most consistently selected predictors achieves a relative RMSFE of 0.266—statistically indistinguishable from the best full-sample specifications—using only publicly available series, offering a robust and operationally feasible nowcasting implementation.

Keywords: Nowcasting, Industrial Production, Regularization Models, Dynamic Factor Models, Google Trends.

JEL Classification: C32, C53, C55, E27, E37.

*Corresponding author. Email: manuel.torresf@udem.edu

†Email: edgar.luna@uanl.edu.mx

1 Introduction

In economic analysis, timely assessments of current economic activity are crucial for investors and policymakers. While traditional macroeconomic indicators, such as national accounts, labor market surveys, and price indexes, provide valuable insights into an economy's overall health and performance, their inherent publication lags impede timely responses to emerging economic trends. According to Nissilä (2020), this delay can hinder investors' optimal capital allocation and constrain the effectiveness of policy interventions, particularly during periods of rapid economic change, such as recessions.

It is increasingly common to incorporate timely indicators into forecasting frameworks to address the limitations of relying solely on official statistics with time lags. These indicators play a crucial role in nowcasting, enabling the generation of up-to-date estimates of key economic variables. By mitigating the limitations imposed by publication delays, nowcasting facilitates more agile and adaptive decision-making in dynamic economic environments.

In Mexico, the National Institute of Statistics and Geography (INEGI) has implemented several nowcasting initiatives to provide timely estimates of key economic activity indicators, including Gross Domestic Product, private consumption, the Global Indicator of Economic Activity, and manufacturing activity. These initiatives, known as "Timely Estimate of Quarterly GDP" (EOPIBT), "Monthly Indicator of Private Consumption in the Domestic Market" (IMCPMI), "Timely Indicator of Economic Activity" (IOAE), and "Timely Monthly Indicator of Manufacturing Activity" (IMOAM), respectively, leverage a range of timely economic, financial, and non-traditional data, including surveys, interest and exchange rates, stock market indices, and Google search data.

Previous research has highlighted the importance of the Industrial Production Index or Monthly Industrial Activity Index (IMAI) for nowcasting IGAE, as in Corona et al. (2021), and the relevance of IGAE for GDP nowcasting, as in Gálvez-Soriano (2020). However, even IMAI is published with a one-month lag, potentially introducing bias into nowcasts of IGAE and GDP for the most recent month. This paper aims to fill this literature gap by developing a nowcasting framework for Mexico's IMAI that integrates timely traditional indicators with non-traditional high-frequency information in a high-dimensional setting.

Using a dataset of 302 monthly predictors, we evaluate both Dynamic Factor Models and a broad set of regularized estimators—Ridge, LASSO, Elastic Net, Adaptive LASSO, Adaptive Elastic Net, and Multi-Step Adaptive Elastic Net—under a pseudo-real-time design that mimics the information constraints faced by policymakers in practice. The paper makes four main contributions. First, it provides a fully recursive empirical strategy for transformation, selection, and forecasting in real time. Second, it documents that adaptive shrinkage methods

improve predictive accuracy in this environment, with MSAENET standing out as the best-performing specification. Third, among non-traditional indicators, the Google Trends topic *Drought* emerges as the only predictor consistently selected across estimation windows, capturing a structural feature of Mexico's industrial vulnerability to climatic conditions. Fourth, it highlights a practical implementation result: a Ridge model estimated on the seven predictors most consistently selected by MSAENET achieves a relative RMSFE of 0.266—statistically indistinguishable from the best full-sample specifications—using only publicly available series from INEGI, FRED, IMEF, and Google Trends, offering a robust and operationally feasible nowcasting tool.

The remainder of the paper is organized as follows. Section 2 presents the background and motivation. Section 3 describes the data and the pseudo-real-time data-construction strategy. Section 4 details the econometric methodology. Section 5 reports the empirical results, organized into four subsections: out-of-sample forecast accuracy, variable selection dynamics, economic interpretation of selected predictors, and a practical nowcasting implementation. Section 6 concludes with policy implications and avenues for future research.

2 Background

A timely and accurate assessment of current economic conditions is crucial for effective policymaking and informed business decisions. While GDP is the benchmark measure of aggregate economic activity, its inherent publication lags pose a significant challenge, particularly in emerging economies where timely responses to economic shocks are often critical. As documented by Jardet & Meunier (2022), advanced economies such as the United States, Japan, Canada, France, Germany, Italy, and the United Kingdom typically release GDP data within approximately five weeks after the end of the reference quarter. However, for Mexico, an emerging market, the publication lag for GDP is longer, at approximately seven weeks according to Gálvez-Soriano (2020). This extended delay severely constrains policymakers' ability to react promptly to evolving economic circumstances, potentially exacerbating the impact of adverse economic events.

High-frequency indicators are essential for real-time monitoring of economic performance. In Mexico, IGAE provides a timelier perspective than GDP, offering monthly updates on economic activity. While IGAE also suffers from a publication lag of approximately seven weeks, its higher frequency provides earlier signals of cyclical turning points and emerging trends, making it a valuable tool for short-term economic analysis. This early information is particularly important for detecting nascent recessions or recoveries, allowing for more proactive policy responses.

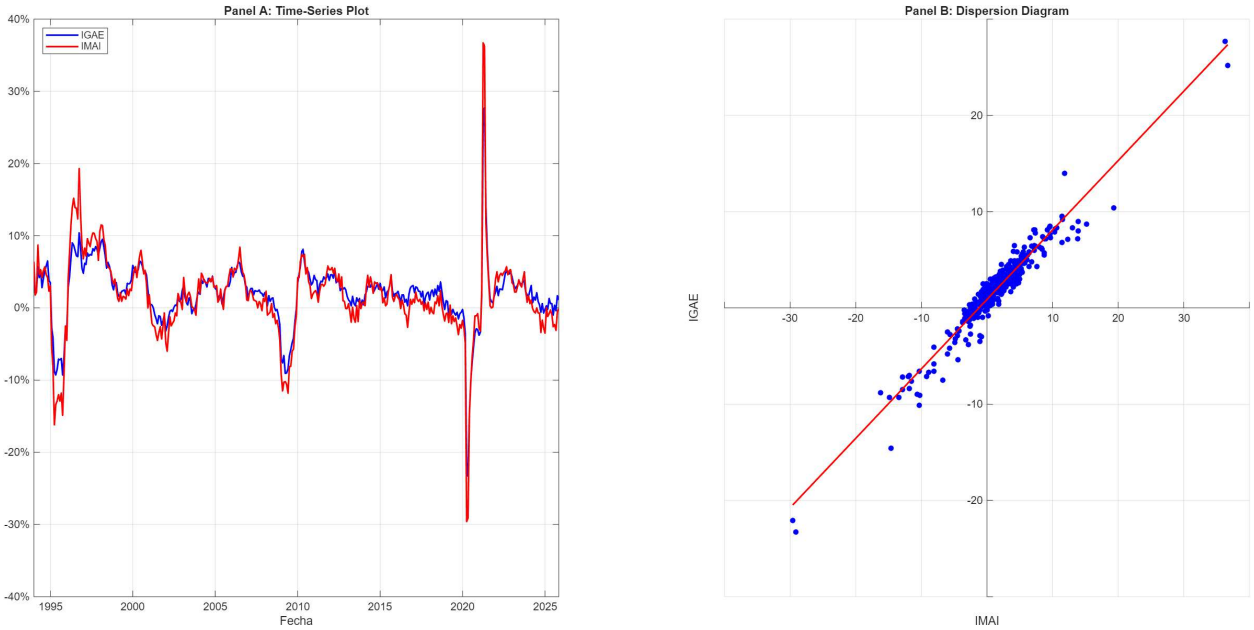


Figure 1: Relationship between IGAE and IMAI (year-over-year % change)

IMAI plays a crucial role in understanding the dynamics of the Mexican economy. As a monthly indicator closely related to the industrial sector, IMAI exhibits a strong empirical linear correlation with IGAE (Figure 1). However, IMAI is subject to a publication delay of approximately four weeks. While it is shorter than GDP and IGAE, this lag still limits its effectiveness for real-time nowcasting of IGAE, especially when considering two-step-ahead forecasts where the IMAI value for the current month is unavailable.

To assess the predictive value of the IMAI for the IGAE, we simulate a scenario in which IMAI data are available without delay (i.e., available at time T). Using this data, we estimate a simple linear regression model with ARMA errors to account for autocorrelation, generating two-step-ahead nowcasts of the IGAE. We then compare the accuracy of these nowcasts with those produced by a naïve AR(1) benchmark. Since the IGAE is a well-established predictor of GDP Gálvez-Soriano (2020), demonstrating the usefulness of the IMAI for forecasting the IGAE underscores the importance of developing timely and accurate nowcasts for the IMAI itself, positioning it as a potential early indicator of broader economic trends. The results of this analysis are presented in Table 1.

Table 1 shows that including IMAI as a predictor of IGAE substantially improves forecast accuracy relative to a naïve AR(1) benchmark. Even when IMAI is observed with a one-period delay, the RMSFE decreases by 75.3% and 33.2% for the one- and two-step-ahead forecasts, respectively. In a counterfactual scenario where IMAI is available without the usual four-week publication lag, the improvement in the two-step-ahead nowcast increases to 79.6%, i.e., 46.4 percentage points above the lagged-IMAI case. These results motivate the development

Table 1: Relative RMSFE for one- and two-step ahead IGAE nowcasts.

Model	Horizon	
	$h = 1$	$h = 2$
SARIMA	0.764	0.862
OLS–ARMA (IMAI available at period $T - 1$)	0.247	0.668
OLS–ARMA (IMAI available at period T)	0.247	0.204

The OLS-ARMA specifications use the year-over-year percentage change in IMAI as the explanatory variable and the year-over-year percentage change in IGAE as the dependent variable. When IMAI is available only with a one-period lag (i.e., at $T - 1$), its value for period T is forecast using an AR(1) model and then used in the OLS-ARMA regression to nowcast IGAE.

of a dedicated IMAI nowcasting framework to provide timely inputs for broader real-time economic monitoring.

The growing availability of high-frequency data ("big data"), together with advances in econometric methods, offers a practical way to reduce the informational constraints created by publication delays. Nowcasting exploits these timely signals to produce current-period estimates of key macroeconomic variables. In this paper, nowcasting IMAI with timely and alternative data sources can improve both the timeliness and the accuracy of IGAE forecasts by supplying an estimate of the current month's IMAI before the official release.

Moreover, combining nowcasted IMAI with additional timely indicators, including financial-market variables, business surveys, and internet search activity, can deliver a more robust and informative real-time monitoring system. A multi-indicator framework is better suited to capture different facets of economic conditions, improving the reliability of short-horizon forecasts for aggregates such as GDP, private consumption, and investment. By providing earlier and more precise assessments of current activity, the proposed approach can support more effective policymaking and better-informed business decisions.

3 Data

3.1 Traditional indicators

For our analysis, we compile a broad set of traditional macroeconomic indicators commonly used to nowcast economic activity in Mexico, focusing on series that are released more promptly than IMAI (see, for example, Caruso (2018); Corona et al. (2021); Gálvez-Soriano (2020)). The dataset includes timely measures closely linked to industrial activity and the business cycle, such as stock market indices, consumer confidence, and business confidence indicators, as well as external indicators (e.g., measures of U.S. economic conditions, industrial

production, and stock markets).

Our baseline dataset comprises 132 traditional indicators observed from January 2003 to December 2025 (when available). All series are collected directly from the sources listed in Table 2. Three series—namely, trade balance forecasts, the 10- versus 3-year fixed-rate bond spread, and the U.S. unemployment rate—contain internal missing values. These gaps are first addressed using Kalman filter imputation based on a structural time-series model.

Regarding seasonal adjustment, the data are already seasonally adjusted whenever the provider supplies such adjustments; otherwise, after imputing missing observations, we apply X-13ARIMA-SEATS to remove seasonality. Table 2 summarizes the composition of the traditional-indicator dataset by category and data source.

Table 2: Traditional indicators: classification and sources

Category	Total	Main sources
Production, sales, and employment indicators	7	ANTAD; IMSS; INEGI
Price indices (consumer and producer prices)	22	INEGI
Financial and market indicators	8	BANXICO
Forecasts, expectations, and perception surveys	10	BANXICO; INEGI
Sector-specific survey indicators	68	BANXICO; IMEF; INEGI
International macroeconomic indicators	17	BANXICO; FRED; Investing

Some indicators enter the dataset at different points in time due to limitations in historical availability. This results in an unbalanced panel with a ragged-edge structure, a common feature in real-time nowcasting applications. While the full dataset covers January 2003 to December 2025, individual indicators may span shorter horizons. Importantly, all indicators are available in the most recent period and are released more promptly than the IMAI, ensuring their relevance for real-time nowcasting. The treatment of this issue is discussed in the Data Treatment section.

3.2 Non-traditional indicators

Recent research has leveraged Google search indices to enhance forecasts of macroeconomic variables (see, e.g., Nagao et al. (2019); Kohns & Bhattacharjee (2023); Bantis et al. (2023)). Google Trends provides near real-time information on the relative intensity of searches for specific terms over a given period and location. Because search activity reflects households' and firms' information needs and economic concerns (e.g., online shopping, job search, credit conditions, or health-related issues), it can serve as a timely complement to traditional "hard" indicators when official statistics are released with delays.

In the Mexican nowcasting literature, Corona et al. (2021) constructs a large set of Google Trends terms tailored to the Mexican economy and uses them to nowcast IGAE. Their results show that COVID-19-related queries (e.g., "facemask" and "quarantine") contain useful predictive information and improve out-of-sample accuracy during the pandemic period.

However, Kohns & Bhattacharjee (2023) choose not to include Google search terms, arguing that they may capture spurious patterns unrelated to the intended economic signal. For instance, Corona et al. (2021) consider the term 'muertos' ("dead" in English) as a proxy for security concerns (e.g., homicides and violence), yet the series displays pronounced seasonality in November due to Mexico's Day of the Dead celebrations, making it difficult to disentangle crime-related searches from cultural effects. In addition, some candidate terms are ambiguous because they have homographs; for example, 'peso' may refer either to the Mexican currency or to "weight," potentially introducing measurement noise.

To mitigate these concerns, we rely primarily on Google search topics and categories rather than individual search terms. Topics aggregate semantically related queries and reduce the risk of homographs or culturally driven seasonal spikes contaminating the economic signal. We construct a set of topic indices that encompasses the terms proposed by Corona et al. (2021) and complement them with additional topics that are particularly relevant for the Mexican economy, such as drought and water scarcity, power outages, hurricanes, organized crime, and drug-related violence. We retain a limited number of individual search terms only when no equivalent topic is available and when the term is unambiguous. In addition, we incorporate the broad economic categories and topics used by Kohns & Bhattacharjee (2023), which are designed to capture general economic search intensity and are therefore comparable across countries.

We also incorporate Economic Policy Uncertainty (EPU) indices developed by Baker et al. (2016), which are constructed from the frequency of selected policy-related terms in major newspapers for each country. These indices have been widely used to nowcast macroeconomic variables such as GDP and private consumption in several economies, including Brazil, Indonesia, Mexico, South Africa, Turkey, and Spain (see Gil et al. (2018); Cepni et al. (2020); Mishra et al. (2021)). In this study, we include the EPU indices for Mexico and the World, as well as selected U.S. subcategory indices to capture potential external spillovers. Furthermore, we augment these measures with complementary uncertainty indicators that reflect emerging sources of macroeconomic risk. Specifically, we consider an Artificial Intelligence Uncertainty index based on newspaper coverage of technologies such as machine learning and computer vision, and the Equity Market Volatility (EMV) index developed by Baker et al. (2019), which measures stock market volatility associated with policy-related news. Together, these indicators allow us to account for both domestic and global dimensions of uncertainty that

may influence economic activity in Mexico.

Accordingly, we augment the baseline dataset with a total of 170 non-traditional indicators (see Table 3). These series are seasonally adjusted using X-13ARIMA-SEATS, with a LOESS filter used as a fallback whenever the former failed to converge, a common issue in non-traditional series with a large number of zero observations (e.g., the Google search topic on Coronavirus, which displayed zero search intensity before 2019).

Table 3: Non-traditional indicators: classification and main sources

Category	Total	Source
Economic policy, artificial intelligence, and uncertainty indices	15	EPU
News-based EMV trackers	44	EPU
Google Trends search categories	31	Google
Google Trends search topics	71	Google
Google Trends search terms	9	Google

Before turning to the pseudo-real-time forecasting exercises, it is worth noting that seasonal adjustment is performed using the full available sample prior to model estimation. In principle, seasonal factors may be revised as new observations become available, which could introduce a look-ahead bias in real-time evaluation. However, empirical inspection indicates that such revisions are quantitatively small for the series considered. Implementing fully recursive seasonal adjustment for more than 300 indicators would substantially increase computational complexity while having a negligible impact on the results. The pseudo-real-time evaluation procedure is described in detail in Section 5.1.

4 Methodology

Our dataset exhibits three main features: (i) the number of potential predictors is large relative to the sample's time dimension; (ii) there is substantial cross-correlation among covariates, particularly within thematic groups (e.g., financial variables, survey indicators, and Google search topics); and (iii) the underlying data-generating process may undergo structural change, especially during periods of economic stress such as the COVID-19 pandemic. Under these conditions, standard unrestricted linear regression is ill-suited, as it can lead to overfitting, parameter instability, and inflated variance. Accordingly, we consider two complementary econometric strategies for nowcasting IMAI in high-dimensional and potentially unstable environments: regularized regression methods and Dynamic Factor Models (DFMs). Both approaches are designed to summarize information in large predictor sets while mitigating multicollinearity and over-parameterization.

4.1 Ridge regression

The Ridge method, introduced by Hoerl & Kennard (1970), is a regularization technique that provides an alternative to Ordinary Least Squares (OLS) estimation when the explanatory variables exhibit multicollinearity. It is also particularly useful in high-dimensional settings where the number of predictors is large relative to the sample size. The key idea is to impose a penalty on the magnitude of the coefficients: in the presence of multicollinearity, OLS estimates tend to be unstable and may take large absolute values, which increases estimator variance and deteriorates out-of-sample predictive performance. In its constrained form, the Ridge estimator is defined as

$$\hat{\boldsymbol{\beta}}^{(\text{Ridge})} = \arg\min_{\boldsymbol{\beta}} \frac{1}{n} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 \quad \text{subject to} \quad \|\boldsymbol{\beta}\|_2^2 \leq \tau, \tag{1}$$

where $\tau > 0$ controls the degree of shrinkage. An equivalent penalized formulation replaces the constraint with a Lagrange multiplier and adds an ℓ_2 penalty to the objective function as follows:

$$\hat{\boldsymbol{\beta}}^{(\text{Ridge})} = \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{n} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_2^2 \right\} \tag{2}$$

where $\lambda \geq 0$ controls the degree of shrinkage. While Ridge regression stabilizes estimation by shrinking coefficients continuously, it does not set any of them exactly to zero. To obtain sparse specifications that simultaneously regularize and perform variable selection, a desirable property in high-dimensional nowcasting applications, we next consider the LASSO estimator.

4.2 LASSO regression

Building on the Ridge framework above, the least absolute shrinkage and selection operator (LASSO), introduced by Tibshirani (1996), replaces the ℓ_2 penalty with an ℓ_1 penalty. This modification is crucial because, unlike Ridge, the ℓ_1 norm can set some coefficients exactly to zero, thereby yielding sparse specifications and performing variable selection, features that are particularly attractive in high-dimensional environments. As in the Ridge case, the LASSO problem can be written either in constrained form (imposing an upper bound on the ℓ_1 norm of the coefficients) or, equivalently, as a Lagrangian (penalized) formulation. In the latter, the estimator solves

$$\hat{\boldsymbol{\beta}}^{(\text{LASSO})} = \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{n} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 \right\}. \tag{3}$$

When $\lambda = 0$, the estimator reduces to OLS; as λ increases, the penalty becomes more stringent and more coefficients are set to zero. This allows LASSO to select a subset of

relevant predictors and helps mitigate overfitting in high-dimensional applications. However, in contrast to Ridge, LASSO does not necessarily provide a "grouping" effect for highly correlated regressors: under multicollinearity it may select one variable from a correlated set and discard the others, leading to less stable model specifications across samples. This limitation motivates the Elastic Net, which combines ℓ_1 and ℓ_2 penalties to retain sparsity while improving the treatment of correlated predictors.

4.3 Elastic Net regression

Elastic Net, introduced by Zou & Hastie (2005), combines the Ridge and LASSO penalties, yielding a compromise between coefficient shrinkage and sparsity. Relative to Ridge, which tends to shrink correlated predictors toward each other and thus helps mitigate multicollinearity, the LASSO component can still set coefficients exactly to zero. At the same time, the additional ℓ_2 term helps address the instability of pure LASSO under multicollinearity, where it may arbitrarily eliminate one variable and retain another within a correlated group. The corresponding penalized (Lagrangian) objective is given in Equation (4).

$$\hat{\boldsymbol{\beta}}^{(\text{ENET})} = \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{n} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \left[\alpha \|\boldsymbol{\beta}\|_1 + (1-\alpha) \|\boldsymbol{\beta}\|_2^2 \right] \right\}. \tag{4}$$

The tuning parameter $\alpha \in [0,1]$ governs the relative contribution of the two penalties: when $\alpha = 1$, Elastic Net collapses to LASSO (pure ℓ_1 penalization), whereas when $\alpha = 0$ it reduces to Ridge regression (pure ℓ_2 penalization). Values $0 < \alpha < 1$ deliver an intermediate specification that combines variable selection with improved handling of correlated predictors.

4.4 Adaptive LASSO regression

Zou (2006) shows that introducing data-driven coefficient-specific penalties in the LASSO framework can achieve the oracle property, i.e., consistent variable selection and asymptotically efficient estimation of the nonzero coefficients. Implementing Adaptive LASSO requires an initial $\sqrt{n}$-consistent estimator to construct the adaptive weights. This can be obtained from an OLS regression when $p < n$; otherwise, in high-dimensional settings, a Ridge estimator is commonly used as the initial estimator. The model specification is given by:

$$\hat{\boldsymbol{\beta}}^{(\text{ALASSO})} = \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{n} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \sum_{j=1}^{p} w_j |\beta_j| \right\}, \tag{5}$$

where $w_j = \left| \hat{\beta}_j^{(\text{init})} \right|^{-\gamma}$ for $\gamma > 0$ and $\hat{\beta}^{(\text{init})}$ denotes an initial $\sqrt{n}$-consistent estimator (typically OLS when $p < n$, or Ridge otherwise). Intuitively, the adaptive weights impose stronger

shrinkage on coefficients that are close to zero in the initial step, while penalizing large coefficients less aggressively, improving variable selection and estimation stability.

Finally, motivated by the same trade-off discussed above between sparsity and stability under multicollinearity, we next consider an Adaptive Elastic Net specification that combines data-driven weights with a mixed ℓ_1–ℓ_2 penalty.

4.5 Adaptive Elastic Net regression

While Elastic Net improves upon LASSO by stabilizing estimation under multicollinearity, it does not generally satisfy the oracle property. To address this limitation, Zou & Zhang (2009) propose the Adaptive Elastic Net, which augments the mixed ℓ_1–ℓ_2 penalty with coefficient-specific adaptive weights. This modification combines the grouping effect and numerical stability of Elastic Net with the selection consistency properties of Adaptive LASSO. The Adaptive Elastic Net estimator is defined as

$$\hat{\boldsymbol{\beta}}^{(\text{AENET})} = \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{n} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \left[\alpha \sum_{j=1}^{p} w_j |\beta_j| + (1-\alpha) \|\boldsymbol{\beta}\|_2^2 \right] \right\}, \tag{6}$$

where w_j is defined as in the Adaptive LASSO regression. In this way, the Adaptive Elastic Net retains the consistency and asymptotic normality properties of Adaptive LASSO, while also being better suited to handle highly correlated covariates and high-dimensional settings.

4.6 Multi-step Adaptive Elastic Net regression

Despite its theoretical advantages, the Adaptive Elastic Net remains a single-step procedure in which the adaptive weights are constructed from a preliminary estimator and then kept fixed throughout the optimization. This feature can be a limitation in practice: forecasting performance and variable selection may depend critically on the quality of the initial estimate, particularly in finite samples and in environments with strong multicollinearity or structural instability. In such cases, the first-step weights may not fully discriminate between weakly and strongly relevant predictors, leading to residual shrinkage bias and/or insufficient sparsity in the selected specification.

To address this weakness, Xiao & Xu (2015) proposed the Multi-Step Adaptive Elastic Net (MSAENET), which extends the Adaptive Elastic Net of Zou & Zhang (2009) by iteratively updating the coefficient-specific weights using the estimates from the previous step. By progressively increasing the penalization of small coefficients while relaxing shrinkage on large ones, this reweighting mechanism can enhance sparsity, reduce false positives, and improve robustness in high-dimensional macroeconomic nowcasting settings.

Formally, let $\hat{\boldsymbol{\beta}}^{(0)}$ denote an initial estimator (e.g., Ridge, Elastic Net, or Adaptive Elastic Net). For iterations $m = 1, \ldots, M$, the MSAENET procedure updates coefficient-specific weights according to

$$w_j^{(m)} = \left|\hat{\beta}_j^{(m-1)}\right|^{-\gamma}, \qquad j = 1, \ldots, p, \tag{7}$$

and then computes

$$\hat{\boldsymbol{\beta}}^{(m)} = \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{n} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \left[\alpha \sum_{j=1}^{p} w_j^{(m)} |\beta_j| + (1-\alpha) \|\boldsymbol{\beta}\|_2^2 \right] \right\}. \tag{8}$$

The final MSAENET estimator is given by $\hat{\boldsymbol{\beta}}^{(\text{MSAENET})} = \hat{\boldsymbol{\beta}}^{(M)}$. Overall, MSAENET provides a flexible compromise between sparsity and stability by allowing the penalty weights—and, thus, the effective amount of shrinkage—to adapt endogenously through iterative reweighting. This feature is particularly appealing in high-dimensional macroeconomic applications with correlated predictors and potential instability.

4.7 Dynamic factor model

The dynamic factor model (DFM) summarizes the information contained in a large set of predictors by means of a small number of latent common factors. The use of DFMs for macroeconomic nowcasting became popular following Giannone et al. (2008), and subsequent work provided a formal statistical justification and feasible estimation methods for such models; see, for instance, Doz et al. (2011). In this framework, each observed series is decomposed into a common component, driven by a few economy-wide factors that capture cross-sectional co-movements, and an idiosyncratic component that is specific to the series. The DFM representation for each predictor can be written as:

$$X_t = \Lambda F_t + \eta_t, \tag{9}$$

where $F_t = (f_{1,t}, \ldots, f_{R,t})^\top$ is an R-dimensional vector of stationary common factors, Λ is an $(N \times R)$ matrix of factor loadings, and η_t is an N-dimensional vector of idiosyncratic components. The common and idiosyncratic components are assumed to be orthogonal. Moreover, the factors are typically modeled as a vector autoregression (VAR) of order P, that is,

$$F_t = \sum_{p=1}^{P} \Psi_p F_{t-p} + \epsilon_t, \qquad t = 1, \ldots, T, \tag{10}$$

where Ψ_p are $(R \times R)$ coefficient matrices and $\epsilon_t \sim \mathcal{N}(0, \Sigma_\epsilon)$ is an R-dimensional white-noise process.

When the database has "ragged edges," the methodology estimates the common factor(s) in two steps. In the first step, the model parameters are calculated using the available information, and the common factors obtained are treated as true ones. The common factors are estimated in the second step via the Kalman smoother.

5 IMAI nowcasting process

In this section, we describe the IMAI nowcasting routine. We first explain the data-treatment process. We then apply the variable-selection methods and use the selected variables in different models to generate IMAI nowcasts. Finally, we use the Root Mean Squared Forecast Error (RMSFE) and the Diebold–Mariano test to compare model performance.

5.1 Data treatment and pseudo-real-time framework

To assess real-time forecasting performance, we construct a pseudo-real-time database starting in January 2015, corresponding to approximately half of the full sample in terms of time span. At each evaluation date t, the model is estimated using an expanding window beginning in January 2004 and ending at $t-1$, and a nowcast is produced for period t. This procedure is repeated recursively on a monthly basis until December 2025.

Prior to transformation selection, variables are pre-processed according to their economic characteristics. Real activity and price indices, which are strictly positive, are transformed using logarithms so that subsequent differences can be interpreted as approximate growth rates and heteroskedasticity is reduced. Variables expressed as rates, such as interest rates, unemployment rates, and spreads, are kept in levels, as logarithmic transformations lack a meaningful economic interpretation in this context. Variables that may take zero or negative values are likewise not log-transformed to avoid arbitrary nonlinear adjustments. Google Trends indicators are normalized following the procedure proposed by Corona et al. (2021), rescaling each series by its in-sample maximum to account for Google's normalization scheme and to ensure comparability over time.

At each evaluation date and for each predictor, we consider a fixed set of six candidate transformations: levels, monthly differences, annual differences, and their respective first lags. To ensure reliable statistical inference, a predictor is considered only if at least 49 non-missing monthly observations are available at date t, guaranteeing sufficient sample size for lag annual differencing and residual-based stationarity testing.

Before assessing predictive relevance, each candidate transformation is subjected to a stability filter. Importantly, we do not require each transformation to be individually stationary. During periods of large macroeconomic disruptions, such as the COVID-19 pandemic, the augmented Dickey-Fuller (ADF) test may fail to reject the null of a unit root due to abrupt level shifts and heightened volatility. Discarding transformations solely on the basis of their univariate non-stationarity could therefore eliminate economically meaningful relationships. Instead, we estimate a bivariate regression between the stationary target variable and the candidate predictor and apply the ADF test to the resulting residuals. Since the dependent variable is stationary, this procedure should not be interpreted as a formal cointegration test. Rather, the residual-based ADF test serves as a joint stability filter, ensuring that the predictor's transformation is dynamically compatible with the stationary behavior of the target variable. Even if both series exhibit pronounced movements during extreme episodes, such as during the pandemic, their co-movement may still yield stationary residuals, implying that forecast errors do not systematically diverge. Only transformations whose residuals are stationary are retained, thereby excluding specifications that may give rise to spurious correlations while preserving relationships that remain stable in relative terms.

Among the admissible transformations, we select the specification that maximizes the average absolute correlation with the target variable computed over a rolling three-year window. This rolling selection procedure serves two purposes. First, it prevents look-ahead bias by ensuring that transformation choice relies exclusively on information available at each evaluation date. Second, it mitigates the disproportionate influence of episodes of extreme volatility, most notably during the 2020 pandemic, which can mechanically inflate correlations for high-frequency (monthly) transformations when correlations are computed over long samples. By restricting the evaluation window, the procedure reduces the dominance of such transitory episodes and allows transformation choice to adapt gradually to evolving macroeconomic conditions.

Importantly, all preprocessing steps—including transformation construction, residual-based stationarity testing, and correlation-based selection—are conducted using only information available up to time $t-1$. The realization of the target variable at time t is used exclusively for out-of-sample evaluation.

Overall, the transformation selection mechanism is fully recursive, statistically filtered, and implemented in pseudo real time, ensuring internal consistency and robustness to structural instability. Figure 2 summarizes the procedure in a compact flowchart. In particular, it highlights the sequence of (i) economic preprocessing of each predictor, (ii) minimum-sample screening to ensure feasibility of annual differencing and residual-based testing, (iii) residual-stationarity filtering via the ADF test, and (iv) rolling correlation selection used to determine

the transformation that is carried forward into the pseudo-real-time unbalanced dataset.

5.2 Variable selection process

At each evaluation date t, after completing the transformation–selection procedure, we obtain two pseudo-real-time datasets. Both are unbalanced because indicators start at different dates, although all variables are observed in the most recent period. To ensure comparability across penalized regressions, we construct a balanced estimation sample by imputing remaining missing values with a Kalman filter; if this fails, we use an ARIMA-based fallback.

For each evaluation date, the estimation sample spans from January 2004 to $t-h-1$, where h is the forecast horizon. All tuning parameters are selected via time-series cross-validation using an expanding window scheme. Concretely, we implement a rolling-origin procedure in which the initial training window excludes the last 36 observations, and one-step-ahead forecast errors are computed recursively. The optimal hyperparameters are chosen according to a user-defined selection criterion based on the out-of-sample RMSFE.

We consider six regularization methods: Ridge, LASSO, Elastic Net, Adaptive LASSO, Adaptive Elastic Net, and MSAENET. For Ridge, LASSO, and Elastic Net, the regularization parameter λ (and, for Elastic Net, the mixing parameter α) is selected by cross-validation over the λ grid produced by `glmnet`. For Elastic Net, we also specify a mixing-parameter grid $\alpha = 0.05a$ for $a = 0, \ldots, 20$. Variable selection is based on predictors with nonzero coefficients at the optimal tuning parameters.

For Adaptive LASSO, coefficient-specific penalty weights are constructed using a $\sqrt{n}$-consistent initial estimator, following Zou (2006). For Adaptive Elastic Net, we use the Elastic Net estimator as the initial estimator, following Zou & Zhang (2009), who also discuss the asymptotic properties of using Elastic Net-type penalization in adaptive procedures. Denoting generic initial coefficients by $\hat{\beta}_j^{(\text{ini})}$, where

$$\hat{\beta}_j^{(\text{ini})} = \begin{cases} \hat{\beta}_j^{(\text{Ridge})}, & \text{for Adaptive LASSO,} \\ \hat{\beta}_j^{(\text{Elastic Net})}, & \text{for Adaptive Elastic Net,} \end{cases}$$

the adaptive weights are defined as

$$w_j = \left| \hat{\beta}_j^{(\text{ini})} \right|^{-\gamma}.$$

where $\gamma = 1$ in the baseline specification. To avoid numerical instability, extremely large weights induced by near-zero initial coefficients are truncated at a large finite constant, and the corresponding variables are excluded from penalization when necessary. Conditional on

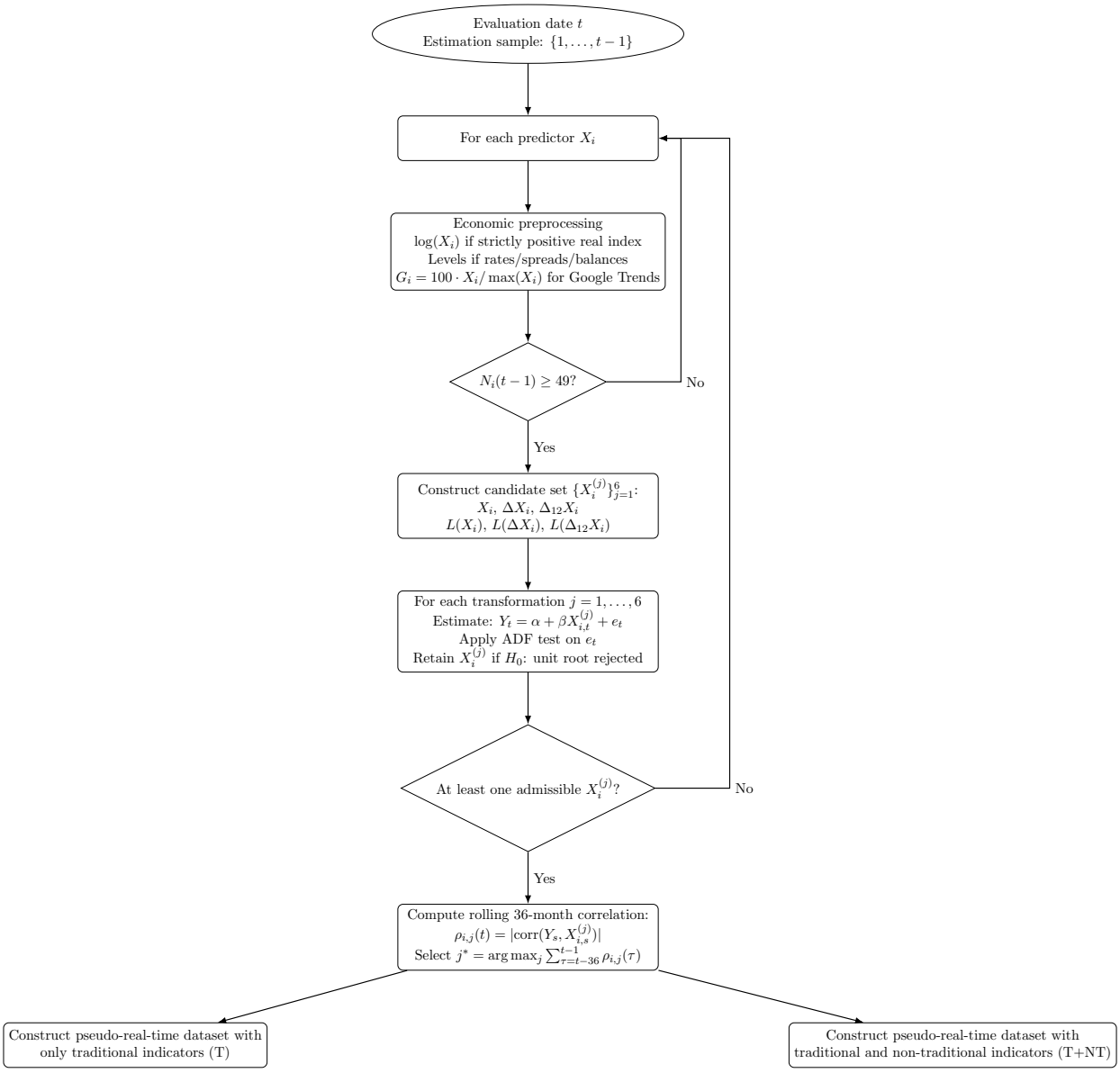


Figure 2: Recursive transformation selection procedure implemented in pseudo real time.

these weights, the optimal λ (and α in the Adaptive Elastic Net case) is again selected by time-series cross-validation.

Finally, we implement MSAENET as an iterative reweighting procedure. The algorithm is initialized with the standard Elastic Net estimator. At each iteration m, adaptive weights are updated as

$$w_j^{(m)} = \left|\hat{\beta}_j^{(m-1)}\right|^{-\gamma},$$

and a weighted Elastic Net regression is re-estimated with tuning parameters selected via cross-validation. We restrict the procedure to three iterations: after the initial Elastic Net estimation (step 1), two successive reweighting steps are performed, and the third-step Adaptive Elastic Net coefficients define the final model. The resulting coefficient vector determines the selected predictor set at date t.

To assess robustness to the information set, the entire procedure is implemented separately for two datasets: (i) the Traditional Only (T) dataset, including only traditional indicators, and (ii) the Full Dataset (T+NT), which additionally incorporates non-traditional indicators. For each method and evaluation date, we record a binary selection indicator for every predictor, allowing us to compute selection frequencies and analyze the stability of variable inclusion across models and information sets.

Overall, the selection process is fully recursive and embedded within the pseudo-real-time framework, ensuring that both hyperparameter tuning and variable selection rely exclusively on information available at each evaluation date.

Table 4: Regularized selection with rolling cross-validation (one-step ahead).

Algorithm 1. Regularized Variable Selection with Adaptive Extensions

Input: Balanced predictor matrix $X \in \{X_T, X_{T+NT}\}$, target vector y, model index $m \in \mathcal{M}$ with $\mathcal{M} = \{\text{Ridge}, \text{LASSO}, \text{ENET}, \text{ALASSO}, \text{AENET}, \text{MSAENET}\}$, grid of $\alpha = 0.05a$ for $a = 0, \ldots, 20$, and tuning parameter $\gamma = 1$.

Step 1. Sample and Validation Scheme

Consider one-step-ahead forecasting ($h = 1$). Restrict sample to $t - 36 - 1$ observations. Define a rolling/expanding window over the last 36 forecasts and select hyperparameters via cross-validation minimizing RMSFE.

Step 2. Model Estimation

For each $m \in \mathcal{M}$:

1. Generate grid of λ (and α when applicable).
2. Select optimal hyperparameters $(\alpha_m^\star, \lambda_m^\star)$ via cross-validation.
3. Estimate the final model to obtain

$$\hat{\beta}_m = \beta_m(\alpha_m^\star, \lambda_m^\star).$$

Step 3. Adaptive Extensions

For $m \in \{\text{ALASSO}, \text{AENET}, \text{MSAENET}\}$:

1. Compute weights $w_j = |\hat{\beta}_j^{ini}|^{-\gamma}$ using Ridge or Elastic Net estimates.
2. Re-estimate the weighted model via cross-validation to obtain $\hat{\beta}_m$.
3. For MSAENET, update weights iteratively using previous adaptive estimates.

Step 4. Variable Selection Set

For each model m, define the selected set as

$$S_m = \{x_i \in X : \hat{\beta}_{i,m} \neq 0\}.$$

5.3 Common factor construction

Following the variable selection process in Section 5.2, which applies multiple methods and yields twelve distinct subsets, we next construct a common factor for each. Because these subsets form an unbalanced panel with uneven coverage, we use a Dynamic Factor Model

(DFM) to extract latent factors. Specifically, we adopt the two-step estimator of Doz et al. (2011), which is well suited to large datasets with missing values because its second step relies on the Kalman filter. This approach provides robust factor estimates despite the non-uniform availability of series. Importantly, the factor estimation uses the variables selected by each model as they appear in the original dataset, without prior Kalman imputation; that is, we take the selected series in their raw, unbalanced form. The analysis is implemented in R using the `dfms` package, which efficiently executes the two-step procedure. For each variable-selection model $m \in \mathcal{M}$, we denote the resulting common factor by f_m and set $R = 1$ and $P = 1$ (see Section 4.7).

5.4 Training and testing models

We developed twelve models for each dataset within the IMAI nowcasting framework. The first six correspond to Ridge, LASSO, Elastic Net, Adaptive Elastic Net, Adaptive LASSO, and MSAENET, each estimated with ARMA errors to address autocorrelation. The remaining six are simple linear regressions that use the common factor constructed from the variable-selection set of each regularization method; these models also include ARMA errors to control autocorrelation.

The procedure is as follows. For each model $m \in \mathcal{M}$, we first compute residuals using the estimation sample up to $t-1$. We then fit an ARMA(p, q) process to these residuals and obtain the one-step-ahead forecast error. The nowcast $\hat{y}_t$ is constructed using the period-t predictors in X and the forecasted ARMA error term. Formally,

$$\hat{y}_t = x_t^\top \hat{\beta}_m + \sum_{i=1}^{\hat{p}} \hat{\phi}_i e_{t-i} + \sum_{j=1}^{\hat{q}} \hat{\theta}_j \varepsilon_{t-j}, \qquad \varepsilon_t \sim \text{i.i.d.}(0, \sigma^2),$$

where $\hat{p}$ and $\hat{q}$ denote the selected ARMA orders, and $\{\hat{\phi}_i\}_{i=1}^{\hat{p}}$ and $\{\hat{\theta}_j\}_{j=1}^{\hat{q}}$ are the corresponding estimated ARMA coefficients.

For DFM-based models, we follow an analogous procedure. For each common factor associated with selection model $m \in \mathcal{M}$, we first estimate the forecasting equation by OLS using data available up to $t-1$. We then compute the nowcast at time t using the contemporaneous factor observation $f_{m,t}$ and the estimated coefficients. Formally,

$$\hat{y}_t = \hat{\alpha}_m + \hat{\beta}_m f_{m,t} + \sum_{i=1}^{\hat{p}} \hat{\phi}_i e_{t-i} + \sum_{j=1}^{\hat{q}} \hat{\theta}_j \varepsilon_{t-j}, \qquad \varepsilon_t \sim \text{i.i.d.}(0, \sigma^2),$$

where $(\hat{\alpha}_m, \hat{\beta}_m)$ are estimated on the sample ending at $t-1$, $\hat{p}$ and $\hat{q}$ denote the selected

ARMA orders, and $\{\hat{\phi}_i\}_{i=1}^{\hat{p}}$ and $\{\hat{\theta}_j\}_{j=1}^{\hat{q}}$ are the corresponding estimated ARMA coefficients.

Finally, we use the Root Mean Squared Forecast Error (RMSFE) to evaluate the accuracy of nowcasts generated by each variable-selection method and by the corresponding common-factor models.

6 Results

This section presents the main results on nowcasting accuracy and variable-selection behavior for the regularized models described in Section 5. Section 6.1 evaluates out-of-sample nowcast performance across models using the relative RMSFE against a naïve AR(1) benchmark. Section 6.2 examines the dynamics of variable selection, emphasizing the role of traditional versus non-traditional predictors and how model complexity evolves under different penalization strategies, with particular attention to differences between standard and adaptive regularization and the added value of incorporating DFM components. Section 6.3 discusses the economic interpretation of the predictors most consistently selected by the best-performing model. Finally, Section 6.4 translates these findings into a practical nowcasting implementation that reduces the information requirements to a minimal, operationally feasible predictor set.

6.1 Out-of-sample forecast accuracy

Out-of-sample forecast performance for the regularized regression models and the DFM-based specifications built from their selected predictor sets is reported in Table 5. Accuracy is evaluated using RMSFE relative to a naïve AR(1) benchmark under two information sets: Traditional Only (T) and Full Dataset (T+NT). Within each model, significance marks indicate the side statistically favored by the Diebold–Mariano test in the comparison between these two information sets.

The results in Table 5 show that all specifications outperform the naïve AR(1) benchmark (all relative RMSFE values are below 1). Within the baseline high-dimensional model comparison, the strongest performance is achieved by MSAENET estimated on the Full Dataset (T+NT) (0.273). Overall, adaptive penalization delivers the strongest outcomes when non-traditional indicators are incorporated: Adaptive Elastic Net improves from 0.322 to 0.280, and MSAENET improves from 0.320 to 0.273, with both gains supported by Diebold–Mariano significance marks. By contrast, Adaptive LASSO shows no gain when expanding the information set (0.319 vs 0.322).

A similar pattern emerges in the DFM-based specifications: performance generally im-

proves when non-traditional information is included, except under LASSO- and Ridge-type penalization, where gains are not observed. Taken together, these results indicate that richer information sets are most valuable when combined with sparse and adaptive regularization rather than with dense shrinkage.

Table 5: Relative RMSFE compared to naïve AR(1) benchmark

Model	Traditional Only (T)	Full Dataset (T+NT)
Ridge	0.461***	0.629
LASSO	0.317	0.339
Elastic Net	0.332	0.339
Adaptive LASSO	0.319	0.322
Adaptive Elastic Net	0.322	0.280**
MSAENET	0.320	0.273**
DFM & Ridge	0.612	0.706
DFM & LASSO	0.556	0.449***
DFM & Elastic Net	0.574	0.449***
DFM & Adaptive LASSO	0.445	0.446
DFM & Adaptive Elastic Net	0.597	0.360***
DFM & MSAENET	0.582	0.322***

Notes: Asterisks indicate Diebold–Mariano test significance for the difference in predictive accuracy between the two information sets within each model (T vs T+NT). They are placed on the side (cell) of the specification statistically favored by the test (i.e., the lower RMSFE with significantly better accuracy). Significance levels: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

6.2 Variable selection dynamics

Table 6 complements the forecast evaluation by reporting the average number of variables selected by each model across the estimation windows, disaggregated by dataset: Traditional Only (T) and Full Dataset (T+NT). This allows us to assess not only how many predictors each penalization strategy retains on average, but also the extent to which non-traditional data sources contribute meaningfully to the selected predictor sets.

A first striking result is the stark contrast in model complexity across penalization strategies. Ridge, by construction, retains all available variables and assigns non-zero coefficients to every predictor, resulting in an average of 116.6 variables under the Traditional Only (T) dataset and 285.2 under the Full Dataset (T+NT), effectively precluding any form of variable selection. In contrast, sparse methods reduce the active predictor set substantially: LASSO and Elastic Net select between 41 and 61 variables on average, while adaptive specifications achieve even greater parsimony. Notably, MSAENET operating on the Full

Dataset (T+NT) selects only 10.4 variables on average, the most parsimonious specification considered.

A second key finding concerns the role of non-traditional indicators. While expanding the information set mechanically increases the pool of candidate predictors, adaptive models are highly selective in which non-traditional variables they retain. MSAENET, for instance, incorporates on average only 1.7 non-traditional indicators alongside 8.7 traditional ones. This suggests that the forecast gains documented in Table 5 are not driven by a broad incorporation of alternative data, but rather by the identification of a small subset of non-traditional predictors with genuine informational content for nowcasting purposes.

Table 6: Average variable selection among datasets

Model	Traditional	Non-Traditional	Total
Ridge (T)	116.6	–	116.6
LASSO (T)	41.1	–	41.1
Elastic Net (T)	48.6	–	48.6
Adaptive LASSO (T)	34.0	–	34.0
Adaptive Elastic Net (T)	38.7	–	38.7
MSAENET (T)	36.5	–	36.5
Ridge (T+NT)	118.7	166.5	285.2
LASSO (T+NT)	25.0	22.0	47.0
Elastic Net (T+NT)	31.2	29.4	60.6
Adaptive LASSO (T+NT)	19.1	8.1	27.2
Adaptive Elastic Net (T+NT)	12.9	4.9	17.8
MSAENET (T+NT)	8.7	1.7	10.4

6.3 Economic interpretation of selected predictors

Table 7 reports the selection frequency of each predictor retained by MSAENET on the Full Dataset (T+NT) in more than half of all estimation windows, together with its data source. The composition of this predictor set offers meaningful economic insights into the drivers of Mexican manufacturing activity.

Among traditional indicators, Automobile and Truck Production are selected in every estimation window (100%), reflecting the tight structural link between vehicle assembly and the broader industrial cycle in Mexico. As one of the world's leading automotive exporters, Mexico's manufacturing sector is heavily integrated into North American supply chains, making vehicle production a near-perfect contemporaneous proxy for IMAI. The U.S. Unemployment Rate (96.2%) and the U.S. Manufacturing Production Index (70.7%) capture external demand conditions: given that roughly 80% of Mexican exports are destined for the

United States, fluctuations in U.S. labor market conditions and industrial output translate rapidly into orders for Mexican manufacturers. The IMEF Manufacturing Production Index (95.5%) provides an early domestic signal, as it is survey-based and released before official INEGI figures, making it particularly valuable in a real-time nowcasting context. Finally, the PPI for Construction (55.6%) likely proxies input cost pressures and investment activity upstream of manufacturing.

Among non-traditional indicators, the Google Trends topic *Drought* stands out as the only predictor selected in more than half of all estimation windows (79.7%), and the second most frequently selected variable in the entire predictor set. This result is economically interpretable in the Mexican context: severe drought episodes directly constrain hydroelectric generation and agricultural supply chains, both of which have significant upstream and downstream linkages with industrial activity. Moreover, water scarcity affects manufacturing operations directly in water-intensive sectors such as food processing, chemicals, and beverages, all components of IMAI. The sustained selection frequency of this topic suggests it captures a structural feature of Mexican industrial vulnerability to climatic conditions rather than a transitory effect. The remaining non-traditional predictors—including *Health Crisis*, *Deportation*, and *PEMEX*—are selected sporadically and below the 10% threshold, reflecting episodic rather than structural informational content.

Table 7: Variable selection frequency: MSAENET (T+NT)

Variable	Frequency	Source
Automobile Production	100.0%	INEGI
Trucks Production	100.0%	INEGI
Unemployment Rate (US)	96.2%	FRED
Manufacturing Production	95.5%	IMEF
Drought (Topic)	79.7%	Google
Manufacturing Production Index (US)	70.7%	FRED
PPI: Construction	55.6%	INEGI

Notes: The table reports only predictors selected in more than 50% of all estimation windows ($n = 132$ recursive windows). Across all windows, MSAENET selected at least one non-zero coefficient for 47 distinct predictors in total—40 traditional and 7 non-traditional. The remaining 40 predictors appear with selection frequencies below the 50% threshold; among non-traditional indicators, these include topics such as *Health Crisis*, *Deportation*, and *PEMEX*, each selected in fewer than 10% of windows, suggesting episodic rather than structural informational content.

6.4 A practical nowcasting implementation

The results in Section 6.1 establish that, within the baseline high-dimensional model set, MSAENET estimated on the Full Dataset (T+NT) delivers the strongest performance, with a

relative RMSFE of 0.273. Notably, this is also the most parsimonious specification, selecting only 10.4 predictors on average. Among the remaining models, LASSO (T), Adaptive Elastic Net (T+NT), and DFM & MSAENET (T+NT) achieve statistically comparable forecast accuracy, suggesting that the marginal gain from expanding model complexity beyond these benchmarks is limited.

However, the computational demands of MSAENET are substantial. At each evaluation date, the procedure requires processing 302 indicators through the full transformation-selection pipeline and running multiple rounds of iterative reweighting with cross-validation. While this design is appropriate for a research setting, it is operationally demanding for a practitioner who needs to produce timely monthly updates.

A practical alternative emerges directly from the variable-selection evidence in Table 7. Because MSAENET consistently retains the same seven predictors across more than half of all estimation windows, these variables capture structural rather than incidental relationships with Mexican manufacturing activity. This stability suggests that a practitioner can bypass the full high-dimensional pipeline and estimate instead a Ridge regression using only these seven indicators. All seven series are publicly available at monthly frequency from INEGI, FRED, IMEF, and Google Trends, and can be downloaded and updated in a matter of minutes.

This approach not only simplifies implementation but also delivers strong forecast performance. The seven-indicator Ridge model achieves a relative RMSFE of 0.266, which is statistically indistinguishable from both DFM & Adaptive Elastic Net (T+NT) and MSAENET (T+NT) based on Diebold–Mariano tests, and marginally lower than the best-performing full model. Figure 3 displays the pseudo-out-of-sample nowcast generated by this model alongside the observed IMAI series and the associated 95% confidence interval, providing visual confirmation of its tracking accuracy across the full evaluation period, including the sharp contraction and rapid rebound associated with the COVID-19 pandemic in 2020. Taken together, the evidence confirms that forecast gains in this environment are driven by disciplined variable selection rather than by the breadth of the information set. For applied nowcasting purposes, the seven-indicator Ridge model thus represents a robust, computationally efficient, and operationally feasible implementation of the broader framework developed in this paper.

7 Conclusion

This paper develops a high-dimensional nowcasting framework for Mexico's IMAI that combines traditional and non-traditional indicators under a pseudo-real-time design. The

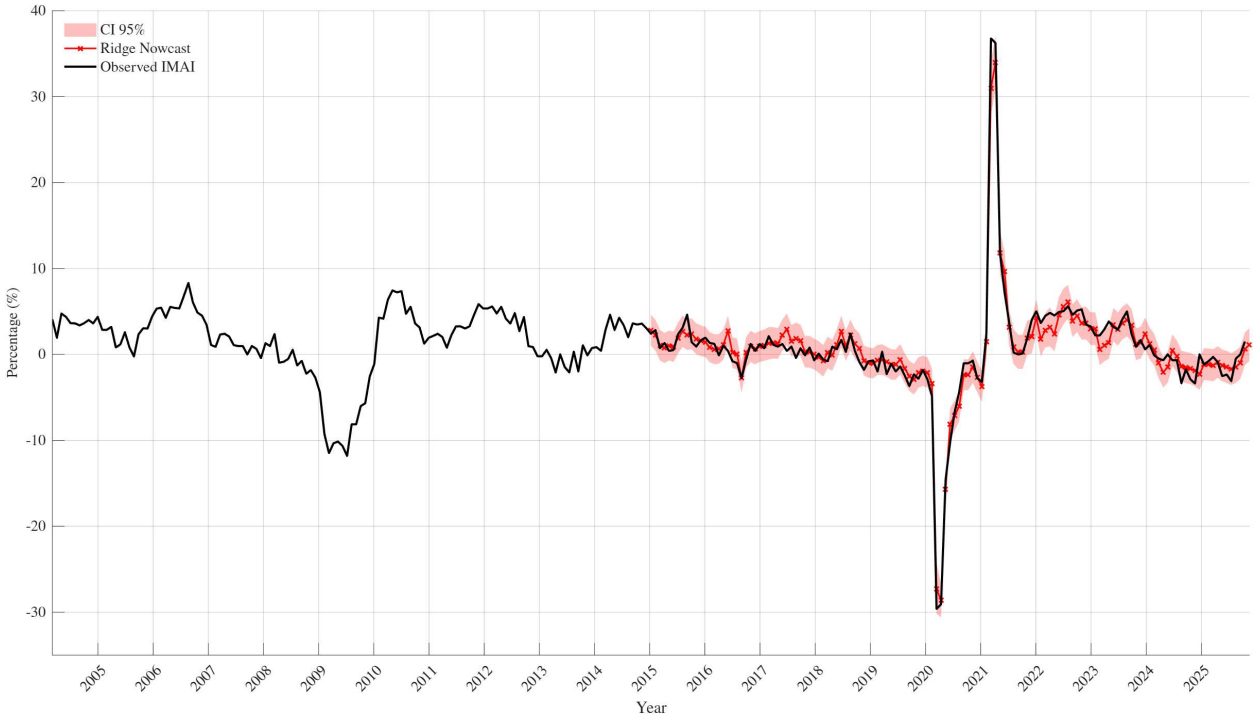


Figure 3: Pseudo-out-of-sample nowcast: seven-indicator Ridge model. The figure displays the nowcast generated by a Ridge regression estimated on the seven predictors selected by MSAENET in more than 50% of evaluation windows (Table 7), alongside the observed IMAI series and the associated 95% confidence interval. The evaluation period runs from January 2015 to December 2025; the observed series extends back to January 2004 to provide historical context.

empirical results show that adaptive penalization methods, especially MSAENET, deliver the most accurate forecasts and maintain strong performance even during periods of elevated macroeconomic volatility. Among non-traditional indicators, the Google Trends topic *Drought* stands out as the only predictor consistently selected across estimation windows, capturing a structural feature of Mexican industrial vulnerability to climatic conditions rather than a transitory effect.

A key practical finding is that forecast gains stem from disciplined variable selection rather than from indiscriminate predictor expansion. MSAENET identifies a stable set of seven predictors that capture structural relationships with Mexican manufacturing activity. A Ridge model estimated on this reduced set achieves a relative RMSFE of 0.266, statistically indistinguishable from the best full-sample specifications, while requiring only publicly available series from INEGI, FRED, IMEF, and Google Trends. This result transforms the proposed framework from a research exercise into an operationally feasible nowcasting tool.

As future work, it would be useful to evaluate non-linear approaches, including neural networks, decision-tree-based models (e.g., random forests and gradient boosting), and hybrid specifications that can capture regime-dependent dynamics. In parallel, Bayesian variable-selection methods such as Stochastic Search Variable Selection (SSVS) and related spike-and-slab frameworks could provide a probabilistic treatment of model uncertainty and improve the robustness of predictor inclusion decisions in high-dimensional environments.

A final research avenue concerns the long-run relevance of Google-search-based indicators. As AI assistants and generative search tools become more prevalent, user behavior may shift away from traditional keyword searches, potentially weakening the informational content of Google Trends for macroeconomic nowcasting. Future nowcasting systems should therefore reassess the stability of search-based signals and broaden the digital-information set to include alternative real-time data sources.

References

Baker, S. R., Bloom, N., & Davis, S. J. (2016). Measuring economic policy uncertainty. *The Quarterly Journal of Economics*, *131*(4), 1593–1636.

Baker, S. R., Bloom, N., Davis, S. J., & Kost, K. J. (2019). *Policy news and stock market volatility* (Tech. Rep.). Cambridge, MA: National Bureau of Economic Research.

Bantis, E., Clements, M. P., & Urquhart, A. (2023). Forecasting gdp growth rates in the united states and brazil using google trends. *International Journal of Forecasting*, *39*(4), 1909–1924.

Caruso, A. (2018). Nowcasting with the help of foreign indicators: The case of mexico. *Economic Modelling*, *69*, 160–168.

Cepni, O., Guney, I. E., & Swanson, N. R. (2020). Forecasting and nowcasting emerging market gdp growth rates: The role of latent global economic policy uncertainty and macroeconomic data surprise factors. *Journal of Forecasting*, *39*(1), 18–36.

Corona, F., González-Farías, G., & López-Pérez, J. (2021). A nowcasting approach to generate timely estimates of mexican economic activity: An application to the period of covid-19. *arXiv preprint arXiv:2101.10383*.

Doz, C., Giannone, D., & Reichlin, L. (2011). A two-step estimator for large approximate dynamic factor models based on kalman filtering. *Journal of Econometrics*, *164*(1), 188–205.

Gálvez-Soriano, O. d. J. (2020). Nowcasting mexico's quarterly gdp using factor models and bridge equations. *Estudios Económicos (México, DF)*, *35*(2), 213–265.

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

Gil, M., Pérez, J. J., Sanchez Fuentes, A. J., & Urtasun, A. (2018). Nowcasting private consumption: traditional indicators, uncertainty measures, credit cards and some internet data. *Banco de España Working Paper*.

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

Jardet, C., & Meunier, B. (2022). Nowcasting world gdp growth with high-frequency data. *Journal of Forecasting*, *41*(6), 1181–1200.

Kohns, D., & Bhattacharjee, A. (2023). Nowcasting growth using google trends data: a bayesian structural time series model. *International Journal of Forecasting*, *39*(3), 1384–1412.

Mishra, P., Alakkari, K., Abotaleb, M., Singh, P. K., Singh, S., Ray, M., ... others (2021). Nowcasting india economic growth using a mixed-data sampling (midas) model (empirical study with economic policy uncertainty–consumer prices index). *Data*, *6*(11), 113.

Nagao, S., Takeda, F., & Tanaka, R. (2019). Nowcasting of the us unemployment rate using google trends. *Finance Research Letters*, *30*, 103–109.

Nissilä, W. (2020). Probit based time series models in recession forecasting: A survey with an empirical illustration for finland. *BoF Economics Review*(7/2020).

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

Xiao, N., & Xu, Q.-S. (2015). Multi-step adaptive elastic-net: reducing false positives in high-dimensional variable selection. *Journal of Statistical Computation and Simulation*, *85*(18), 3755–3765.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, *101*(476), 1418–1429.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *67*(2), 301–320.

Zou, H., & Zhang, H. H. (2009). On the adaptive elastic-net with a diverging number of parameters. *Annals of statistics*, *37*(4), 1733.