Borradores de ECONOMÍA

Nowcasting Colombian Economic Activity: DFM and Factor-MIDAS approaches

By: Franky Juliano Galeano-Ramírez
Nicolás Martínez-Cortés
Carlos D. Rojas-Martínez

No. 1168
2021

tá - Colombia - Bogotá - Colombia - Bogotá - Colombia - Bogotá - Colombia - Bogotá - Colombia - Bogotá - Colombia - Bogotá - Col

Nowcasting Colombian Economic Activity: DFM and Factor-MIDAS approaches

Franky Juliano Galeano-Ramírez
Banco de la República
fgaleara@banrep.gov.co

Nicolás Martínez-Cortés
Banco de la República
nmartico@banrep.gov.co

Carlos D. Rojas-Martínez*
Banco de la República
crojasma@banrep.gov.co

Abstract

Economic policy decision-making requires constantly assessing the state of economic activity. However, this is not an easy task: official figures have significant lags, and the timely information is usually partial and has different frequencies. This paper applies two types of short-term forecasting methodologies (Factor-MIDAS and DFM) for Colombian economic activity involving information with mixed frequencies. We present a heuristic process to select relevant variables, and we evaluate the proposed models' fits by comparing them with traditional forecasting methodologies. Overall, DFM and Factor-MIDAS forecasts are better than those generated by conventional methodologies, especially as the flow of information increases. In times of COVID-19, the model with the best relative fit was the DFM.

Keywords: Colombian economic activity, nowcast, forecast, mixed frequency factor models.

JEL Classification: C53, E27, E52.

*We are thankful for comments from Juan Pablo Cote, Sergio Restrepo, Adolfo Cobo, Carlos Huertas and Celina Gaitán.

Pronósticos de corto plazo para la actividad económica colombiana: aproximaciones DFM y Factor-MIDAS

Franky Juliano Galeano-Ramírez
Banco de la República
fgaleara@banrep.gov.co

Nicolás Martínez-Cortés
Banco de la República
nmartico@banrep.gov.co

Carlos D. Rojas-Martínez*
Banco de la República
crojasma@banrep.gov.co

Resumen

La toma de decisiones de política económica requiere evaluar constantemente el estado de la actividad económica. Sin embargo, ello no es una tarea fácil: las cifras oficiales tienen rezagos importantes y la información más oportuna suele ser parcial y tener frecuencias dispares. Este artículo aplica dos tipos de metodologías de pronóstico de corto plazo (Factor-MIDAS y DFM) para la actividad económica colombiana involucrando información con frecuencias mixtas. Se propone un proceso heurístico para la selección de variables relevantes y se evalúa el ajuste de los modelos comparándolo respecto a metodologías usuales de proyección. En general, los pronósticos de los modelos Factor-MIDAS y del DFM superan los generados por metodologías tradicionales, con resultados más precisos en la medida que aumenta el flujo de información. En tiempos del COVID-19, el modelo con el mejor ajuste relativo fue el DFM.

Palabras clave: actividad económica colombiana, *nowcast*, pronóstico, modelos de frecuencia mixta con factores.

Clasificación JEL: C53, E27, E52.

*Agradecemos los comentarios de Juan Pablo Cote, Sergio Restrepo, Adolfo Cobo, Carlos Huertas y Celina Gaitán.

1 Introduction

Knowing the current state of the economy is key to estimate and forecast correctly the economic cycle and, therefore, for timely policy decision-making. Nevertheless, several challenges arise in the estimation process of the economic activity: 1) Gross Domestic Product (GDP), which is the benchmark aggregate, has about 45 days delay in its quarterly data publication in Colombia, once the respective quarter has ended; 2) each publication usually implies two types of changes in the historical data, the first related to the gradual revision of the GDP due to the inclusion of newly available information, and the second as a result of the seasonal adjustment process; furthermore, 3) the heterogeneity of the potentially helpful indicators poses another issue insofar as they have different frequencies (daily, weekly, monthly, or quarterly) and various publication dates throughout the quarter.

This paper shows some methodologies used by the technical staff of the Colombian Central Bank -Banco de la República (BR)- to estimate and forecast the present and very recent past of the Colombian economic activity (*nowcasting*[1]). In this sense, the paper directly addresses the first and third challenges mentioned above.

Nowcasting processes are not new in the economic literature, so the current state of the art counts with a vast number of related techniques. Some of the most relevant works, which also serve as inspiration for this paper, are Marcellino and Schumacher (2010) and Bok et al. (2017). The first work is pathbreaking as they apply mixed-data sample analysis techniques (MIDAS) (Ghysels et al., 2004) and factor models in a single framework (Factor-MIDAS), achieving a method to short-term forecasts from a broad set of explanatory variables of equal or higher frequency than the target variable, even in the presence of unbalanced samples. The authors find that higher frequent information (monthly) improves the forecast compared to models that include only quarterly information for the German GDP. In turn, Bok et al. (2017) detail a methodology used by the technical staff in the New York Federal Reserve to

[1] The term nowcasting is the contraction of *now* and *forecasting*. In this paper, it is used, as it is usual in economic literature, to refer to the forecast of the present or the very recent periods, while the term *forecasting* is used for future periods forecasts.

forecast the present growth of US GDP. Their technique is a dynamic factor model (DFM) that considers many variables of different frequencies with information from several sectors of the economy. This technique exhibits good forecasting ability and allows them to detail the contributions of new information. Stock and Watson (2011) offer a detailed review of DFM techniques.

For the Colombian case, several studies address some of these challenges. Without aiming to be comprehensive, within the literature, we must mention Cristiano et al. (2012), where the authors use a state-space model for nowcasting and apply an algorithm based on the explained variance of GDP to select relevant variables. In turn, Pérez-Castañeda (2009) and Cárdenas-Cárdenas et al. (2020) apply mixed-data sampling (MIDAS) models to forecast GDP and the changes in food prices in Colombia, respectively. The use of different nowcasting techniques and short-term forecasting of Colombian GDP is broad. For instance, León and Ortega (2018) used a non-linear autoregressive exogenous neural network model to forecast the Colombian Economic Monitoring Indicator (ISE) considering electronic transactions[2]. Their results lead to the conclusion that this information helps to forecast economic activity.

Under this framework, our work contributes to the efforts to nowcast economic variables in Colombia through mixed frequencies data and synthesizing information with factor analysis, mainly using DFM and Factor-MIDAS models. We consider electronic transaction data and other indicators not included in previous documents. Moreover, we use some additional statistical and computational methods to evaluate different possibilities within these methodologies, seeking satisfactory nowcasts and forecasts of GDP.

The development of the paper focuses on 1) the selection of a convenient set (not necessarily optimal) of indicators and 2) the estimation and out-of-sample evaluation of the different models through pseudo-real-time simulations. The latter means that we use the latest revision of each indicator (i.e., the latest data published at the cut-off date of their collection and not those published at each moment in time) to carry out the out-of-sample forecast evaluation exercise. The effects of data revisions, which for some variables might be important, therefore

[2]In Colombia, several commercial banks use the daily electronic transactions of their payment channels to monitor different economic aggregates.

are not considered. However, the impact of these can be negligible in environments with many indicators.

Structure of the paper. Section 2 introduces the econometric framework, both Factor-MIDAS models and DFM. Section 3 presents the database and the procedure to select indicators. Section 4 develops the estimation and evaluates the performance of the models for two quarters ahead (nowcast and forecast made in six different months); section 5 presents the models' forecasts in times of COVID-19. Finally, section 6 concludes.

2 Econometric framework

Why do we use a large number of indicators? The synthesis of national accounts is an exercise of gathering a large amount of information in few aggregates for the general diagnosis of the economy, which could well be classified as a *Big Data* problem. In this regard, the methodologies used in this document synthesize various indicators for nowcasting and forecasting economic activity. Stock and Watson (2017) masterfully summarize the motivation for this type of approach:

> "Twenty years ago, economists who monitored the economy in real time used indexes of economic indicators and regression models for updating expectations of individual releases (such as the monthly employment report), combined with a large dose of judgment based on a narrative of where the economy was headed. While this approach uses data, it is not scientific in the sense of being replicable, using well-understood methods, quantifying uncertainty, or being amenable to later evaluation. Moreover, this method runs the risk of putting too much weight on the most recent but noisy data releases, putting too little weight on other data, and being internally inconsistent because each series is handled separately. Because knowing the current state of the economy in real-time is an ongoing, arguably increasingly important responsibility of policymakers, time series econometricians at central banks and in academia have put considerable effort into improving the

foundations and reliability of real-time macroeconomic monitoring." (p. 72)

The kind of models presented below could incorporate variables with mixed frequencies, so enlightenment in this regard is worthwhile. The traditional and most straightforward approach resides in aggregating the variables with the highest frequency to obtain a balanced set at the lowest frequency. Such is the case of the so-called *bridge models* of sectoral indicators, widely used in forecasting economic activity. However, this type of approach implies loss of information and changes in the data generating process, which entails potentially different dynamics from those present in higher frequency series. The possibility of considering and evaluating the current state of the economy from indicators that preserve their original frequency has made some models, which involve sets of variables with mixed frequencies, become important.

This section presents two kinds of these approaches: Factor-MIDAS and DFM. In both, factor analysis is used to summarize the information of a *large* number of indicators, with potential differences in their publication lags (the so-called *ragged-edge problem*), and mixed frequency techniques for the short-term forecast of economic activity.

2.1 Factor-MIDAS models

The first kind of model consists of establishing MIDAS-type specifications where the high-frequency variables are the factors obtained from the set of monthly indicators, denoted as $\mathbf{X}_t$, using three possible methodologies. This section closely follows Marcellino and Schumacher (2010) and proceeds through two stages. First, it deals with the estimation of factors in unbalanced sets of indicators and, second, with the estimation of MIDAS models using the above factors as high-frequency variables.

2.1.1 Factors

We will assume that the monthly observations have the following structure:

$$\mathbf{X}_t = \mathbf{\Lambda} \mathbf{F}_t + \varepsilon_t, \tag{1}$$

where the vector of r-dimensional factors is denoted as $\mathbf{F}_t = (f'_{1,t}, \ldots, f'_{r,t})$. The matrix of the loadings is $\mathbf{\Lambda}$, and that of the idiosyncratic components is ε_t. Under the assumption that $\mathbf{X}_t$ is balanced, several factor estimation methods have been provided in the literature, with principal components as in Stock and Watson (2002) and dynamic principal components as in Forni et al. (2005) being the most usual. Since $\mathbf{X}_t$ is, typically, an unbalanced set of monthly indicators, three methods are proposed for its treatment.

Vertical realignment and dynamic principal components (VA-DPCA). A convenient way to balance $\mathbf{X}_t$ is to realign each of the series given the last month of data available. Suppose that the variable i is published with a lag of k_i months with respect to the last month of information in the dataset, T_m. The realigned of variable i is:

$$\tilde{x}_{i,t} = x_{i,t-k_i}, \qquad \text{for } t = k_i + 1, \ldots, T_m. \tag{2}$$

The balanced data set results in $\tilde{\mathbf{X}}_t$ for $t = \max(\{k_i\}_{i=1}^N) + 1, \ldots, T_m$. Given this set, the factors are obtained by dynamic principal components.

A disadvantage of this approach is that, due to delays in the data publication, the dynamic correlations change with each new publication, modifying the factors with every estimate. However, in our case, the magnitude of the changes may be relatively small as the publication periods tend to be homogeneous among different types of indicators from the same source.

Principal components and EM algorithm (EM-PCA). This method follows the structure of Stock and Watson (2002), which combines an EM algorithm with a standard estimation of principal components: consider a variable i of $\mathbf{X}_t$ as a column vector $\mathbf{X}_i = (x_{i,1}, \ldots, x_{i,T_m})'$, with final data missing. Vector $\mathbf{X}_i^{obs}$ comprises the available observations of $\mathbf{X}_i$:

$$\mathbf{X}_i^{obs} = \mathbf{A}_i \mathbf{X}_i, \tag{3}$$

where $\mathbf{A}_i$ is a matrix that establishes the relationship between the missing data and the observed figures. The EM algorithm is:

1. Provide naive values for the missing observations for each of the variables. This gives

$\hat{\mathbf{X}}^{(0)}$ and, using a standard estimation of principal components, $\hat{\mathbf{F}}^{(0)}$ and $\hat{\mathbf{\Lambda}}^{(0)}$.

2. Stage E: An update of the missing data estimates for the variable i is provided by the expectation of $\mathbf{X}_i$ conditional on $\mathbf{X}_i^{obs}$, the factors $\hat{\mathbf{F}}^{(j-1)}$ and the loads $\hat{\mathbf{\Lambda}}_i^{(j-1)}$, the iteration continues with:

$$\hat{\mathbf{X}}_{\mathbf{i}}^{(\mathbf{j})} = \hat{\mathbf{F}}^{(j-1)}\hat{\mathbf{\Lambda}}_i^{(j-1)} + \mathbf{A}_i'(\mathbf{A}_i'\mathbf{A}_i)^{-1}\left(\mathbf{X}_i^{obs} - \mathbf{A}_i\hat{\mathbf{F}}^{(j-1)}\hat{\mathbf{\Lambda}}_i^{(j-1)}\right). \tag{4}$$

   The two components of the above equation are: i) the common component of a step back $\hat{\mathbf{F}}^{(j-1)}\hat{\mathbf{\Lambda}}_i^{(j-1)}$, plus the idiosyncratic component $\mathbf{X}_i^{obs} - \mathbf{A}_i\hat{\mathbf{F}}^{(j-1)}\hat{\mathbf{\Lambda}}_i^{(j-1)}$ projected by $\mathbf{A}_i'(\mathbf{A}_i'\mathbf{A}_i)^{-1}$.

3. Stage M: repeat stage E for each all variable i obtaining a balanced set. Get $\hat{\mathbf{F}}^{(j)}$ and $\hat{\mathbf{\Lambda}}^{(j)}$ by standard estimation of principal components. Return to 2 until the algorithm converges.

Once it has converged, the algorithm provides estimates of $\hat{\mathbf{F}}_t$ and missing data for each series.

Space state (KFS-PCA). The model is assumed to have the following form:

$$\mathbf{X}_t = \mathbf{\Lambda}\mathbf{F}_t + \varepsilon_t, \tag{5}$$

$$\Psi(L)\mathbf{F}_t = \mathbf{B}\eta_t. \tag{6}$$

The previous equations are, respectively, the static representation of $\mathbf{X}_t$ and the process of the factors using a lag polynomial $\Psi(L)$, where L is the lag operator. The vector η_t contains the orthogonal dynamic shocks. The model is in the form of a state-space system, where the factors $\mathbf{F}_t$ are the states. Given the potentially large number of indicators, the estimation follows from Doz et al. (2012).

2.1.2 MIDAS models

Once the monthly factors have been estimated, MIDAS models (Ghysels et al., 2004) are used to forecast the quarterly change of GDP. This type of approximation resides in direct forecasts, so they do not consider dynamic relationships between contemporary or lagged factors and the respective values of the lowest frequency variable. MIDAS models are essentially regressions that involve different frequency variables through polynomials with few coefficients what avoids the loss of degrees of freedom in the regression (Ghysels and Marcellino, 2018, p. 459). We will denote by t_m monthly frequency variables and by t_q quarterly frequency variables.

The basic MIDAS model. For the sake of simplicity, suppose a single factor $\hat{f}_{t_m}$. The model for a horizon of h_q quarters ahead, with $h_q = h_m/3$, is:

$$y_{t_q+h_q} = y_{t_m+h_m} = \beta_0 + \beta_1 b(L_m, \theta)\hat{f}^{(3)}_{t_m} + \varepsilon_{t_m+h_m}, \tag{7}$$

where $\hat{f}^{(3)}_{t_m} = \hat{f}_{t_m}$, for all $t_m = \ldots, T_m - 6, T_m - 3, T_m$, y_{t_q} is the quarterly GDP, L_m is the monthly-frequency lag operator, and the polynomial (*exponential Almon lag*, in our case) $b(L_m, \theta)$ is:

$$b(L_m, \theta) = \sum_{k=0}^{K} c(k, \theta)L_m^k, \qquad \text{and} \qquad c(k, \theta) = \frac{\exp(\theta_1 k + \theta_2 k^2)}{\sum_{k=0}^{K} \exp(\theta_1 k + \theta_2 k^2)}. \tag{8}$$

The MIDAS model can be estimated using non-linear least squares (NLS) in a regression of y_{t_m} on $\hat{f}^{(3)}_{t_m-k}$, with $\hat{\theta}_1, \hat{\theta}_2, \hat{\beta}_0$ and $\hat{\beta}_1$ as results. The forecast is the projection of the equation on the estimated factor:

$$y_{T_m+h_m|T_m} = \hat{\beta}_0 + \hat{\beta}_1 b(L_m, \hat{\theta})\hat{f}_{T_m}. \tag{9}$$

The generalization for many factors $r > 1$ follows from equation (7):

$$y_{t_q+h_q} = y_{t_m+h_m} = \beta_0 + \sum_{i=1}^{r} \beta_{1,i} b_i(L_m, \theta_i)\hat{f}^{(3)}_{i,t_m} + \varepsilon_{t_m+h_m}. \tag{10}$$

The AR-MIDAS model. Autoregressive models are, in normal situations, good competitors to static forecasting models. In this sense, it may be convenient to include an autoregressive term in the MIDAS model. Clements and Galvao (2008) propose the following equation with an autoregressive term, estimated by NLS:

$$y_{t_m+h_m} = \beta_0 + \lambda y_{t_m} + \beta_1 b(L_m, \theta)\hat{f}^{(3)}_{t_m} + \varepsilon_{t_m+h_m}. \tag{11}$$

The MIDAS Smooth model. The European economic activity indicator -*The New Eurocoin Index*- (Altissimo et al., 2010) makes it possible to consider mixed frequencies in the projection of low-frequency variables, as follows:

$$y_{T_m+h_m|T_m} = \hat{\mu} + \mathbf{G}\hat{\mathbf{F}}_{T_m}, \tag{12}$$

$$\mathbf{G} = \tilde{\mathbf{\Sigma}}_{y\mathbf{F}}(h_m)\hat{\mathbf{\Sigma}}_{\mathbf{F}}^{-1}, \tag{13}$$

where $\hat{\mu}$ is the sample mean of the quarterly change in GDP, assuming that the factors have zero mean, and $\mathbf{G}$ is the matrix of projection coefficients. $\hat{\mathbf{\Sigma}}_{\mathbf{F}}$ is the factor covariance matrix and $\tilde{\mathbf{\Sigma}}_{y\mathbf{F}}(k)$ is the covariance between the k lags in smoothed GDP and the factors. See Marcellino and Schumacher (2010) for details.

The UMIDAS model. The basic MIDAS model and the AR-MIDAS are based on an exponential function to treat lags in a parsimonious way. An alternative is to consider an unrestricted MIDAS model (Foroni et al., 2015):

$$y_{t_m+h_m} = \beta_0 + \mathbf{D}(L_m)\hat{\mathbf{F}}_{t_m}^{(3)} + \varepsilon_{t_m+h_m}, \tag{14}$$

where $\mathbf{D}(L_m) = \sum_{k=0}^{K} \mathbf{D}_k L_m^k$ is an unrestricted lag polynomial of order K. Equation (14) can be estimated using OLS. The number of lags considered is selected using the Bayesian information criteria (BIC).

2.2 Dynamic Factor Model

The second approach is a DFM (Stock and Watson, 2011).

The dynamics of the monthly series. We will assume that $\mathbf{X}_t$ adopts the following specification:

$$\mathbf{X}_t = \mu + \mathbf{\Lambda}\mathbf{F}_t + \varepsilon_t, \tag{15}$$

$$\mathbf{F}_t = \mathbf{A}_1\mathbf{F}_{t-1} + \cdots + \mathbf{A}_p\mathbf{F}_{t-p} + \mathbf{u}_t, \quad \mathbf{u}_t \sim i.i.d.N(0, \mathbf{Q}), \tag{16}$$

$$\varepsilon_{i,t} = \alpha_i\varepsilon_{i,t-1} + e_{i,t}, \quad e_{i,t} \sim i.i.d.N(0, \sigma_i^2), \tag{17}$$

$$E(e_{i,t}e_{j,s}) = 0, \quad \text{for } i \neq j. \tag{18}$$

In addition, $\mathbf{\Lambda}, \mathbf{A}_1, \ldots, \mathbf{A}_p$ and $\mathbf{Q}$ are restricted assuming that $\mathbf{F}_t$ is composed of three mutually independent factors: a *global* factor (G) common to all variables, a *hard* factor (H) referring to those series that best approximate economic activity[3], and a *soft* factor (S). This type of specification is useful because it allows considering the correlations between *specific* groups of variables and thus obtaining a more precise estimation of the global factor and a better GDP forecast. The matrices mentioned above are:

$$\mathbf{F}_t = \begin{pmatrix} \mathbf{F}_{G,t} \\ \mathbf{F}_{H,t} \\ \mathbf{F}_{S,t} \end{pmatrix} \qquad \mathbf{\Lambda} = \begin{pmatrix} \mathbf{\Lambda}_{H,G} & \mathbf{\Lambda}_{H,H} & 0 \\ \mathbf{\Lambda}_{S,G} & 0 & \mathbf{\Lambda}_{S,S} \end{pmatrix}$$

$$\mathbf{A}_i = \begin{pmatrix} \mathbf{A}_{i,G} & 0 & 0 \\ 0 & \mathbf{A}_{i,H} & 0 \\ 0 & 0 & \mathbf{A}_{i,S} \end{pmatrix} \qquad \mathbf{Q} = \begin{pmatrix} \mathbf{Q}_G & 0 & 0 \\ 0 & \mathbf{Q}_H & 0 \\ 0 & 0 & \mathbf{Q}_S \end{pmatrix}$$

The dynamics of the quarterly series. The quarterly series follows the specification of Mariano and Murasawa (2003), where it is assumed that these are partially observed monthly variables. In this way, quarterly GDP, GDP_t^Q, is observed at $t = 3, 6, 9, \ldots$ and can be expressed in terms of a monthly GDP as:

$$\log(GDP_t^Q) = \frac{1}{3}\left(\log(GDP_t^M) + \log(GDP_{t-1}^M) + \log(GDP_{t-2}^M)\right), \qquad t = 3, 6, 9, \ldots \tag{19}$$

Let $Y_t^Q = \log(GDP_t^Q)$, $Y_t^M = \log(GDP_t^M)$, $y_t^Q = Y_t^Q - Y_{t-3}^Q$ and $y_t = Y_t^M - Y_{t-1}^M$. Mariano and Murasawa (2003) approach is:

$$y_t^Q = \frac{1}{3}\left(Y_t^M + Y_{t-1}^M + Y_{t-2}^M\right) - \frac{1}{3}\left(Y_{t-3}^M + Y_{t-4}^M + Y_{t-5}^M\right) \tag{20}$$

$$= \frac{1}{3}\left(y_t + 2y_{t-1} + 3y_{t-2} + 2y_{t-3} + y_{t-4}\right), \qquad t = 3, 6, 9, \ldots \tag{21}$$

Estimation. The Doz et al. (2012) algorithm, extended by Bańbura and Modugno (2014) for sets with arbitrary patterns of missing information in the series, is used in the estimation of the factor models and consists of the combination of principal components and the Kalman filter: 1) the factors are computed by principal components, 2) estimates of state-space coefficients

[3] In the DFM, GDP will be included in the global and the hard factors.

are obtained with the help of OLS regressions, and 3) run the Kalman filter or smoother to get improved estimates of the factors.

3 Variable selection

We began with a database composed of 148 monthly indicators (see Appendix), with October 15, 2020, as the cut-off date for data collection. It roughly included industrial production, retail sales, electronic transactions, price indices, interest rates, terms of trade, exports, imports, portfolio indicators, labor market series, monetary aggregates, and expectations about activity and prices. The variables involve qualitative information obtained from surveys (*soft data*) and quantitative information from direct measurements of economic activity (*hard data*). All of them were seasonally adjusted[4] and transformed in order to assure stationarity. Altogether, grouping the two types of variables into the same set allows us to involve quantitatively precise information, which has significant lags[5], and qualitative and timely information, which provides convenient signals to guide the forecast.

Given that the main objective of nowcasting methodologies is the accurate and pertinent estimation of the current state of the economy, we must incorporate the corresponding lags of each of the variables and consider the adjustment of each model with different information patterns. Hence, the debugging of the base was carried out using the following heuristic procedure:

Step 1. Simple correlations and Lasso regression. Lasso regression (Tibshirani, 1996) is a shrinkage method defined as:

$$\hat{\beta}^{lasso} = \underset{\beta}{\operatorname{argmin}} \sum_{t=1}^{N} \left(y_t - \beta_0 - \sum_{i=1}^{p} x_{ti}\beta_i \right)^2 \quad \text{subject to } \sum_{i=1}^{p} |\beta_i| \leq \lambda, \tag{22}$$

where λ is the penalty parameter whose value is chosen by minimizing the forecast error. In

[4] X-13 ARIMA-SEATS with holidays and Easter effects adjustments.

[5] Electronic transactions could be a remarkable exception. León and Ortega (2018) study the usefulness of this type of series in short-term forecasts of Colombian economic activity.

our case, the monthly indicators were added in a quarterly frequency[6] and selected using 1) the absolute value of the simple correlation with quarterly GDP growth and 2) a cross-validated Lasso regression.

Step 2. Minimization of RMSFE. Once the initial set of variables was reduced, the fit of the different subsets in each of the models (Factor-MIDAS and DFM) is evaluated in order to select the one that allows us to lower the root mean square forecast error (RMSFE) -out of sample-. However, due to the massive amount of possibilities, not all possible models are considered. Instead, we use the following simple selection algorithm for reducing the RMSFE and choosing a convenient set of variables:

1. Compute the RMSFE for the set with N variables ($RMSFE_N$). Initially, N is the resulting set from step 1.
2. Evaluate the N possible subsets of $N-1$ variables and take the subset with the lowest RMSFE ($RMSFE_{N-1}$).
3. If $RMSFE_{N-1} < RMSFE_N - \delta$, repeat 2 and 3 updating N as $N-1$. If $RMSFE_{N-1} \geq RMSFE_N - \delta$, stop the algorithm and keep the variables. We chose $\delta = 0.001$.

Step 3. Expert judgment. The final step of the series selection process considers the expert judgment of the technical staff[7]; thereby, the set of variables resulting from step 2 is expanded with those essential and which would have been deleted so far, and it is reduced whether it is considered that one or more of the series does not add relevant information. Why do we have to consider the expert judgment in the selection of the variables? Expert judgment is essential for consistency between the resulting variables: statistical criteria can remove some, but they could provide valuable information for forecasting different and varied shocks that the economy could face. Why do we have to consider more than expert judgment in the selection of the variables? Although expert judgment is essential, it could miss series

[6] For the sake of simplicity, we used a simple quarterly average.

[7] Some studies help us to reinforce the judgment. Giannone et al. (2008) conclude that prices and monetary indicators do not improve GDP nowcasts. The evidence about financial indicators relevance is mixed (Bańbura et al., 2013; Forni et al., 2003).

that might prove valuable in the nowcasting process because they could not get frequent or regular attention.

It is worth clarifying: 1) the out-of-sample evaluation period ranges from the first quarter of 2014 to the fourth quarter of 2019, which was selected following the nature of the 2015 Base of national accounts (see section 4); 2) the relevant RMSFE for choosing the best subset of variables was computed with one and two months remaining for the publication of GDP since these are the months when the BR's technical staff presents updated forecasts to the Board of Directors; and 3) since the DFM allows the use of indicators with mixed frequency and arbitrary patterns of missing observations, additional indicators were added with such features: CAMACOL housing series, DANE Building Census variables, the Civil Works investment indicator, BR regional economic pulse, and real GDP from relevant countries. The final sets of selected variables (Table 1) contain macroeconomic indicators which "move markets and make front-page news." (Bok et al., 2017).

Table 1: Selected variables

Indicator	DFM	Factor-MIDAS	DFM-local factor	Transformation	Source
ACH Colombia real value	✓	✓	*Hard*	2	ACH Colombia
CEDEC real value	✓	✓	*Hard*	2	BR
Central Government Investment	-	✓	-	2	Ministry of Finance and Public Credit
Mining exports - quantity index	✓	✓	*Hard*	2	BR
Total imports - quantity index	✓	-	*Hard*	2	BR
DOAM Index	✓	-	*Soft*	1	BR
Regional Economic Pulse	✓	-	*Hard*	1	BR
Economic Monitoring Indicator (ISE)	✓	✓	*Hard*	2	DANE
ISE of tertiary sectors	✓	✓	*Hard*	2	DANE
Retail sales without fuel or vehicles	✓	✓	*Hard*	2	DANE
Semi-durable goods retail sales	-	✓	-	2	DANE
Durable goods retail sales	✓	✓	*Hard*	2	DANE
Manufacturing industrial production	✓	✓	*Hard*	2	DANE
Employed wage-earners in urban areas	-	✓	-	2	DANE
Urban unemployment rate	✓	✓	*Hard*	1	DANE
Building census: completed area (15 areas)	✓	-	*Hard*	2	DANE
Building census: area in process (15 areas)	✓	-	*Hard*	2	DANE
Secondary sector energy demand	✓	✓	*Hard*	2	XM
Industrial business confidence - Q2	✓	-	*Soft*	1	Fedesarrollo
Industrial business confidence - Q5	✓	-	*Soft*	1	Fedesarrollo
Industrial business confidence - Q8	-	✓	*Soft*	1	Fedesarrollo
Commercial business confidence - Q6	✓	-	*Soft*	1	Fedesarrollo
Commercial business confidence - Q9	✓	-	*Soft*	1	Fedesarrollo
Consumer confidence - Economic conditions index	✓	-	*Soft*	1	Fedesarrollo
Consumer confidence - Economic expectations index	✓	-	*Soft*	1	Fedesarrollo
Oil price - Brent	✓	-	*Soft*	2	Bloomberg
House sales	✓	-	*Hard*	2	CAMACOL
House offer	✓	-	*Hard*	2	CAMACOL

Notes: ✓indicates the inclusion of the variable in the model; the transformations are 1 - original variable, and 2 - first diff. of the logarithm.

4 Out-of-sample error

Table 2 presents the model errors for six months prior to the GDP publication and Figure 1 the forecasts for each horizon (nowcast: $m = 1, 2, 3$, and forecast: $m = 4, 5, 6$). These results are obtained from a pseudo-real-time exercise that replicates the patterns of the available information in the first fortnight of each month. It is an iterative out-of-sample process, and it is called pseudo-real-time because despite mimicking the information patterns generated by publication lags, it does not use exactly the information available at each time[8]. Hence there are several series that already include historical revisions.

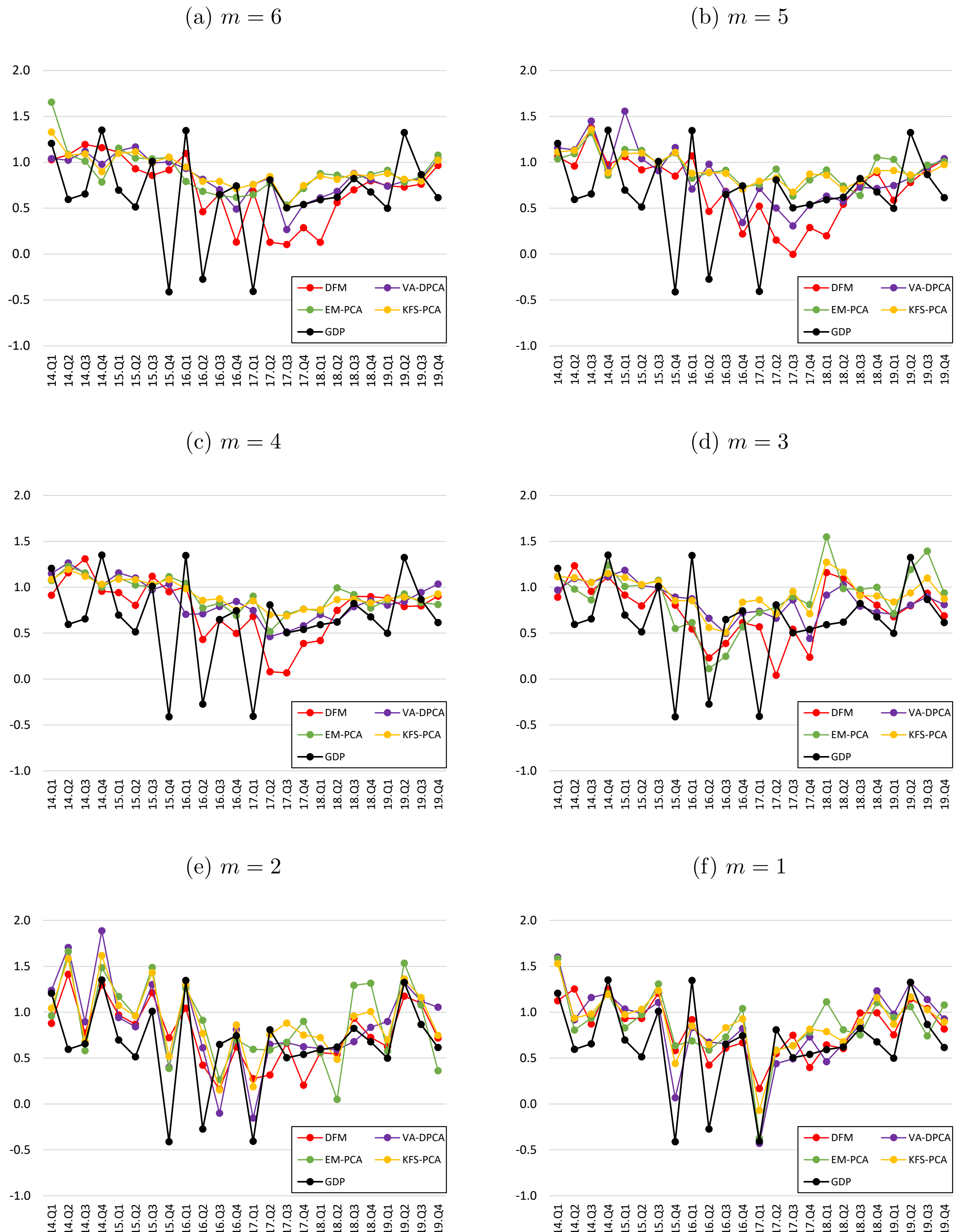
Table 2: RMSFE - quarterly GDP growth

Factor	**Model**	**Nowcast**			**Forecast**		
		$m = 1$	$m = 2$	$m = 3$	$m = 4$	$m = 5$	$m = 6$
-	ARIMA	0.538	0.538	0.538	0.554	0.554	0.554
-	Bridge-ISE	0.375	0.418	0.528	0.518	0.531	0.535
VA-DPCA	MIDAS	0.358	0.433	0.506	0.517	0.555	0.487
	UMIDAS	0.407	0.439	0.550	0.658	0.672	0.588
	MIDAS Smooth	0.538	0.462	0.528	0.571	0.652	0.579
	ARMIDAS	0.369	0.465	0.499	0.525	0.581	0.489
EM-PCA	MIDAS	0.405	0.518	0.495	0.516	0.534	0.498
	UMIDAS	0.396	0.532	0.496	0.635	0.705	0.609
	MIDAS Smooth	0.572	0.600	0.488	0.593	0.600	0.559
	ARMIDAS	0.416	0.522	0.481	0.516	0.549	0.524
KFS-PCA	MIDAS	0.376	0.437	0.528	0.523	0.535	0.495
	UMIDAS	0.389	0.451	0.540	0.611	0.693	0.608
	MIDAS Smooth	0.565	0.583	0.536	0.582	0.578	0.558
	ARMIDAS	0.381	0.436	0.519	0.499	0.557	0.517
-	FAMIDAS Median	0.377	0.443	0.503	0.532	0.578	0.512
-	DFM	0.365	0.414	0.491	0.509	0.496	0.516

Notes: i) m is the number of months remaining for the publication of the reference quarter GDP, ii) the FAMIDAS Median refers to the forecast error resulting from the median of the MIDAS forecast models in each period, it is not the median of the RMSFE.

[8] Although some series have information since 2005 or before, many of these have been published recently for the first time, so it is impossible to find the contemporary version of some of these series in past several years.

Figure 1: Quarterly GDP growth - nowcast and forecast



Notes: i) m is the number of months remaining for the publication of the reference quarter GDP, ii) the figures show the median for each family of Factor-MIDAS models (VA-DPCA, EM-PCA, and KFS-PCA).

The out-of-sample evaluation period was selected from the first quarter of 2014 to the fourth quarter of 2019 due to the features of the national accounts available at the time of data collection. Specifically, the structural year of the national accounts series is 2014, which means that previous information constitutes an *extrapolation* using past statistical operations (DANE, 2018). Therefore, the forecast evaluation is carried out only from the structural year, although we use the series before 2014 to estimate the models under analysis.

Benchmark models. In order to evaluate how accurate the Factor-MIDAS and the DFM adjustments are, two competing models are used: 1) an ARIMA[9] model, and 2) a bridge model using the ISE as a regressor (Bridge-ISE). In particular, bridge models are linear regressions that link high-frequency variables with low-frequency variables through two types of equations: one that allows the high-frequency variable to be projected if there is missing data at the end of the sample, and other that performs the forecast of the low-frequency variable from the aggregated high-frequency variable (e.g., using the quarterly sum or average). In this case, the Bridge-ISE model uses an ARIMA model to forecast the ISE and a linear regression to link the quarterly growth of GDP with the variations of the sum of ISE in each quarter.

Overall, Table 2 shows the DFM and some Factor-MIDAS models outperform ARIMA in the short-term forecasts and compete quite well with the Bridge-ISE. In addition, Figure 1 presents an outstanding performance one month before the GDP publication. The MIDAS-smooth has the worst fit among the Factor-MIDAS models, regardless of the factors used, while the other models have a relatively small error. Note that, as in Marcellino and Schumacher (2010), increasing the nowcast and forecast horizons not always leads to worse RMSFE: it could seem the methods employed cannot always improve the nowcast and forecast with new information, even though it happens because the relatively short period of evaluation could induce high sampling uncertainty in estimates.

[9]We use auto.arima function available in R, set without restrictions, for both ARIMA and Bridge-ISE models.

5 Nowcasting in times of COVID-19

The global economic crisis as a result of the COVID-19 and the public health interventions intended to contain its spread became an unprecedented shock: the scale of the quarterly contraction has no record in recent history for any of the major economies and the speed at which the changes occurred made nowcasting and forecasting economic activity a particularly intricate task. In Colombia, for instance, the most significant contraction of quarterly GDP (since 1994) occurred in 1998 and 1999, with figures close to -6% in annual growth rates and near to -2.5% in quarterly terms. In comparison, the COVID-19 pandemic led to an annual and quarterly contraction in 2020Q2 close to 15%. This section presents the behavior of the models in this harsh context.

Figure 2: Nowcasting results 2020Q1-Q3 (Quarterly GDP growth)

(a) Nowcast: 20Q1 (b) Nowcast: 20Q2 (c) Nowcast: 20Q3

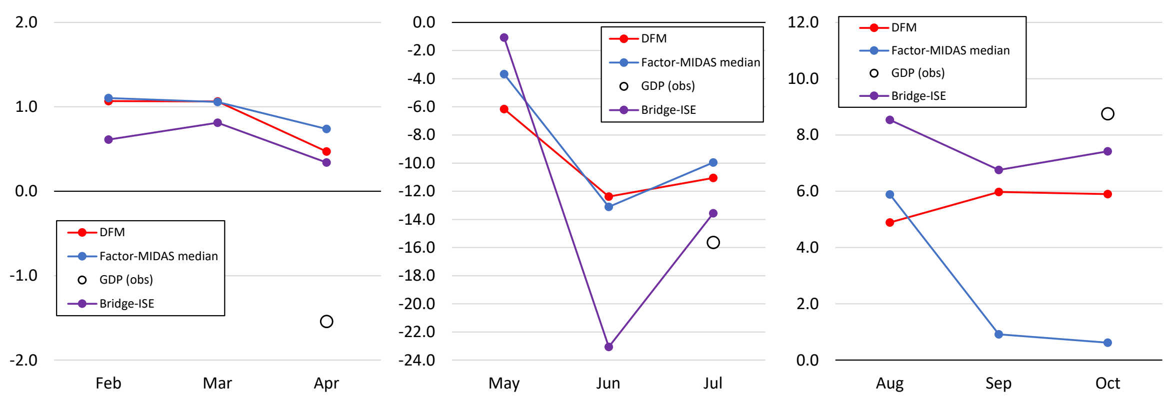


Figure 2 shows the observed GDP and the nowcast for each relevant month according to the different methodologies expounded[10]. For the sake of simplicity, we focus on the median of the Factor-MIDAS models and the DFM's result[11]: although in both cases the forecasts follow the correct direction, they deviated considerably from the observed data, as might be expected. They could not accurately foresee the contraction (2020Q1 and 2020Q2) and

[10]We omit the results of the ARIMA model because they are highly inconvenient (not accurate) in this context.

[11]We show the results of the DFM preserving the parameters estimated in February 2020 owing to *noises* that the inclusion of the never-before-seen values of the COVID-19 period can generate.

the subsequent rebound in quarterly growth (2020Q3). The deficient performance of the Factor-MIDAS models might be because they rely on polynomials estimated with historical information, which could lead to leaving out the rapid and extreme changes that characterized the COVID-19 period. In the case of the DFM, its performance could respond to the lack of indicators to capture the shock[12], mainly in 2020Q1, as well as to its structure from historical information. However, its results are better than those of Factor-MIDAS. As a result, a simpler model, such as the Bridge-ISE, might be better in this context.

A possible strategy to improve the performance of the models in this challenging time is to include variables that allow us to involve suitably the sharp changes of the period (e.g., mobility variables). However, a significant drawback of this framework is how to estimate the models with series of such an extremely short history. Sampi and Jooste (2020) propose to use the Kalman filter along with air pollution data to estimate the history of the series; nonetheless, its reliability is not evident.

6 Conclusion

This paper presents applications of the DFM and Factor-MIDAS models into the Colombian case, widely used techniques. These models allow a correct assessment of macroeconomic conditions in real-time by including many variables of different frequencies in unbalanced samples. In addition, seeking to improve the results, we propose a heuristic algorithm to select variables for the models according to their out-of-sample forecast performance.

The results are suitable for short-term forecasts. Some Factor-MIDAS models and notably the DFM have good nowcasting and forecasting capacity, and they outperform benchmark models, especially in the nowcasting periods and, foremost, with one month remaining before the publication of GDP. Moreover, the DFM seems to outperform the Factor-MIDAS models in this application. However, in times of COVID-19, their results were not accurate. Notwith-

[12]The interventions to contain the health crisis began in March. There were very few indicators for this month when the last nowcast (for 2020Q1) was done.

standing that they follow the correct direction, they deviate significantly from the observed data, as expected given the sharp changes that characterize this period. Indeed, a simpler model such as the Bridge-ISE might present better results.

Despite having applied a wide variety of models, many other alternatives can be tested and might be worth researching. For example, Antolín-Díaz et al. (2020) develop a cutting-edge DFM which can explicitly captures three features of business cycles: low-frequency variation in the mean and variance of the variables, heterogeneous responses to common shocks, and fat-tailed outliers.

References

Altissimo, F., Cristadoro, R., Forni, M., Lippi, M., and Veronese, G. (2010). New eurocoin: Tracking economic growth in real time. *Review of Economics and Statistics*, 92(4):1024–1034.

Antolín-Díaz, J., Drechsel, T., and Petrella, I. (2020). Advances in nowcasting economic activity: Secular trends, large shocks and new data. Working paper.

Bańbura, M., Giannone, D., Modugno, M., and Reichlin, L. (2013). Now-casting and the real-time data flow. *Working Paper Serie ECB 1564*.

Bańbura, M. and Modugno, M. (2014). Maximum likelihood estimation of factor models on datasets with arbitrary pattern of missing data. *Journal of applied econometrics*, 29(1):133–160.

Bok, B., Caratelli, D., Giannone, D., Sbordone, A., and Tambalotti, A. (2017). Macroeconomic nowcasting and forecasting with big data. *Federal Reserve Bank of New York Staff Reports*, 830.

Clements, M. P. and Galvao, A. B. (2008). Macroeconomic forecasting with mixed-frequency data: Forecasting output growth in the united states. *Journal of Business and Economic Statistics*, 26(4):546–554.

Cristiano, D., Hernández, M., and Pulido, J. D. (2012). Pronósticos de corto plazo en tiempo real para la actividad económica colombiana. *Borradores de economía, Banco de la República*, 724.

Cárdenas-Cárdenas, J. A., Caicedo-García, E., and González-Molano, E. R. (2020). Estimación de la variación del precio de los alimentos con modelos de frecuencias mixtas. *Borradores de economía, Banco de la República*, 1109.

DANE (2018). Cuentas nacionales de colombia base 2015. Technical report, DANE.

Doz, C., Giannone, D., and Reichlin, L. (2012). A quasi–maximum likelihood approach for large, approximate dynamic factor models. *Review of Economics and Statistics*, 94(4):1014–1024.

Forni, M., Hallin, M., Lippi, M., and Reichlin, L. (2003). Do financial variables help forecasting inflation and real activity in the euro area? *Journal of monetary economics*, 50(6):1243–1255.

Forni, M., Hallin, M., Lippi, M., and Reichlin, L. (2005). The generalized dynamic factor model: One-sided estimation and forecasting. *Journal of the American Statistical Association*, 100(471):830–840.

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 (Statistics in Society)*, 178(1):57–82.

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

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

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

León, C. and Ortega, F. (2018). Nowcasting economic activity with electronic payments data: A predictive modeling approach. *Revista de Economía del Rosario*, 21(2):381–407.

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.

Mariano, R. S. and Murasawa, Y. (2003). A new coincident index of business cycles based on monthly and quarterly series. *Journal of Applied Econometrics*, 18(4):427–443.

Pérez-Castañeda, G. C. (2009). Using the MIDAS approach for now- and forecasting colombian GDP. Master's thesis, Erasmus Universiteit Rotterdam.

Sampi, J. and Jooste, C. (2020). Nowcasting economic activity in times of covid-19: An approximation from the google community mobility report. *Policy Research Working Paper 9247*.

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

Stock, J. H. and Watson, M. W. (2011). Dynamic factor models. In Clements, M. P. and Hendry, D. F., editors, *The Oxford Handbook of Economic Forecasting*. Oxford University Press.

Stock, J. H. and Watson, M. W. (2017). Twenty years of time series econometrics in ten pictures. *Journal of Economic Perspectives*, 31(2):59–86.

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

Appendix

Table 3: Indicators used

Indicator	Source
ACH Colombia real value	ACH Colombia
ACH CENIT real value	BR
CEDEC real value	BR
Economic Monitoring Indicator (ISE)	DANE
ISE of primary sectors	DANE
ISE of secundary sectors	DANE
ISE of tertiary sectors	DANE
Retail sales without fuels or vehicles	DANE
Non-durable goods retail sales	DANE
Semi-durable goods retail sales	DANE
Durable goods retail sales	DANE
Energy demand	XM
Secondary sector energy demand	XM
Approved area of building licenses	DANE
Coffee production	Colombian Coffee Growers Federation
Cement production	DANE
Consumer confidence index	Fedesarrollo
Consumer confidence - Economic expectations index	Fedesarrollo
Consumer confidence - Economic conditions index	Fedesarrollo
Industrial business confidence	Fedesarrollo
Commercial business confidence	Fedesarrollo
Commercial business confidence - Q1	Fedesarrollo
Commercial business confidence - Q2	Fedesarrollo
Commercial business confidence - Q3	Fedesarrollo
Commercial business confidence - Q4	Fedesarrollo
Commercial business confidence - Q5	Fedesarrollo
Commercial business confidence - Q6	Fedesarrollo
Commercial business confidence - Q8	Fedesarrollo
Commercial business confidence - Q9	Fedesarrollo
Industrial business confidence - Q1	Fedesarrollo
Industrial business confidence - Q2	Fedesarrollo
Industrial business confidence - Q3	Fedesarrollo
Industrial business confidence - Q4	Fedesarrollo
Industrial business confidence - Q5	Fedesarrollo
Industrial business confidence - Q6	Fedesarrollo
Industrial business confidence - Q7	Fedesarrollo
Industrial business confidence - Q8	Fedesarrollo
Industrial business confidence - Q9	Fedesarrollo
Industrial business confidence - Q10	Fedesarrollo
Industrial business confidence - Q11	Fedesarrollo
Break-even inflation (BEI) 1 year forward	BR
BEI 2 year forward	BR
BEI 3 year forward	BR
BEI 4 year forward	BR
BEI 5 year forward	BR
BEI 10 year forward	BR
Forwad Break-even inflation (FBEI) 2-1	BR
FBEI 3-1	BR
FBEI 4-1	BR
Commerce Employment	DANE
Oil production	ANH
National air passengers	Aerocivil
International air passengers	Aerocivil
National air cargo	Aerocivil
International air cargo	Aerocivil

Indicator	Source
Monthly survey of economic expectations (EMEE) - Q1	BR
EMEE - Q2	BR
EMEE - Q3	BR
EMEE - Q4	BR
EMEE - Q5	BR
EMEE - Q7	BR
EMEE - Q8	BR
EMEE - Q10	BR
EMEE - Q11	BR
Manufacturing production index	DANE
Manufacturing employment	DANE
Brent Oil Price	Bloomberg
Credit Default Swaps (CDS) - 5-year	Bloomberg
EMBI Spread Colombia	Bloomberg
USA inflation	Bloomberg
USA unemployment	Bloomberg
Euro Zone inflation	Bloomberg
Euro Zone unemployment	Bloomberg
Colcap index	Bloomberg
Colombian monetary policy rate	BR
Federal funds real rate	Bloomberg
Consumer price index (CPI)	DANE - BR
PPI for total domestic supply items	DANE - BR
CPI excluding food	DANE - BR
CPI excluding food and regulated items	DANE - BR
CPI for foods	DANE - BR
CPI for goods excluding food and regulated items	DANE - BR
CPI for services excluding food and regulated items	DANE - BR
CPI for regulated items	DANE - BR
PPI for produced and consumed items	DANE - BR
PPI for imported items	DANE - BR
PPI for exported items	DANE - BR
PPI for imported capital goods	DANE - BR
Representative market rate of exchange (TRM) average month	BR
Gross international reserves	BR
Net international reserves	BR
New housing price index	BR
Financial stability indicator	BR
DOAM index	BR
Monetary base	BR
Cash	BR
Money supply M1	BR
Money supply M2	BR
Money supply M3	BR
Total savings	BR
CDT (deposits held by the public)	BR
Bank Reserve	BR
Remittances	BR
Real wage index	DANE
Construction cost index	DANE
Installed capacity index	ANDI
Employed wage-earners in urban areas	DANE
Urban unemployment rate of heads of household	DANE
Urban unemployment rate	DANE
Terms of trade (ToT) - PPI	DANE - BR
ToT - international trade (IT)	DANE - BR
Real exchange rate - PPI - IT	BR
Real exchange rate - CPI - IT	BR

Indicator	Source
Total exports - quantity index	DANE - BR
Mining exports - quantity index	DANE - BR
Agricultural exports - quantity index	DANE - BR
Non-mining and agricultural exports - quantity index	DANE - BR
Total imports - quantity index	DANE - BR
Intermediate goods imports - quantity index	DANE - BR
Intermediate goods imports - quantity index	DANE - BR
Capital goods imports - quantity index	DANE - BR
Financial intermediation margin - ex ante	BR
Financial intermediation margin - ex post	BR
Liquid liabilities to liquid assets of the financial sector	BR
Financial system return on assets (ROA)	BR
Banks ROA	BR
Financial system return on earnings (ROE)	BR
Banks ROE	BR
Commercial portfolio quality indicator	BR
Consumption portfolio quality indicator	BR
Mortgage portfolio quality indicator	BR
Commercial default portfolio indicator	BR
Consumption default portfolio indicator	BR
Mortgage default portfolio indicator	BR
DTF interes rate (90-days CDs)	BR
Colombian public debt securities (TES) zero coupon rate - 1 year	BR
TES zero coupon rate - 5 years	BR
TES zero coupon rate - 10 years	BR
Adjusted local currency mortgage portfolio	BR
Local currency consumption portfolio	BR
Local currency commercial portfolio	BR
Benchmark Banking Indicator (IBR) overnight	BR
Central Government total expenses - obligations	Ministry of Finance and Public Credit
Central Government total expenses - payments	Ministry of Finance and Public Credit
Central Government tax income	Ministry of Finance and Public Credit
Central Government investment - obligations	Ministry of Finance and Public Credit
Central Government investment - payments	Ministry of Finance and Public Credit
Taxes to GDP	Ministry of Finance and Public Credit
Regional Economic Pulse	BR
Building census: completed area (15 areas)	DANE
Building census: area in process (15 areas)	DANE
House sales	CAMACOL
House offer	CAMACOL
Trading partners GDP	Bloomberg -BR
Latin America GDP (excluding Colombia)	Bloomberg -BR
US GDP	Bloomberg
Euro Zone GDP	Bloomberg
China GDP	Bloomberg
Used housing prices index	BR
New housing prices index	DANE
New housing prices index	DNP

Note: We took the monthly average for the variables whose source reports in daily frequency. The necessary transformations were carried out to guarantee compliance with the econometric requirements.


BANCO DE LA REPÚBLICA
COLOMBIA