# Colombian Agricultural Sector's Early Estimator of Gross Domestic Production Using Nowcasting and Big Data Methods

*Diego Fernando Bravo Higuera*[1], *León Darío Parra Bernal*[2*], *Milenka Linneth Argote Cusi*[3], *Grace Andrea Torres Pineda*[4]

**Abstract**

Facing challenges like the COVID-19 pandemic, statistical production increasingly relies on non-traditional data sources for timely and accurate information. In this regard, The National Statistical Office of Colombia (DANE, by its acronym in Spanish) initiated a project, supported by the Statistics Advisory Council, to develop an early estimator for the Colombian agricultural sector. This paper presents the results for the implementation of a Ridge model and Zero Shot Classification to estimate the Gross Domestic Product (GDP) of the agricultural sector, leveraging Google News and Google Trends. Results reveal that these alternative sources offer valuable insights into economic trends. Combining machine learning techniques with Nowcasting methods yielded precise projections. The Ridge method demonstrated the lowest estimation error, providing an early GDP indicator for the agricultural sector of 8,188 billion Colombian pesos for 2022 Q2, 30 days ahead of official publication.

**Keywords:** Google News; GoogleTrends; forecasting, agriculture sector; Ridge model, Machine Learning; Dynamic Factor Models.

Submitted: January 25, 2024/ Approved: July 1, 2024

## 1. Introduction and Background

One of the biggest challenges and concerns of economic science has been to predict the behavior of the economy in the coming years and to establish guidelines and public policies in fiscal and monetary decisions that allow strengthening and promoting growth and development. To achieve this aspect, the tools traditionally used are due to the production of macroeconomic indicators framed in the national accounts of a given country. However, to produce these indicators, it is necessary to collect data at an aggregate level by economic sector, which implies that their production and publication must be carried out periodically.

However, the rhythm of generation and production of information in modern society is carried out every minute through social networks, blogs and discussion forums, and internet searches. According to figures from (Big Data-Statista, 2022), every minute about 500 h of content is generated on the web and more than 69 million searches or messages are produced on social networks, which means that by 2030, humanity will produce more than 600 terabytes per year following the same source.

At the same time, the COVID19 pandemic has been one of the most complex events of this century. Societies and governments faced challenges not only to comprehend the nature of the virus but also to comprehend the environmental, economic, political, and social impacts of the COVID19 pandemic, being both peremptory actions to develop adequate strategies and surpass the crisis.

One of the priorities for governments in statistical matters was the so-called macroeconomic fundamentals, like employment, Gross Domestic Product, and prices, for those economic sectors that have an important weight in the economic growth of the countries. According to (Corona et al., 2021), if an economic contraction is foreseeable, businesses can adjust their investment or expansion plans, governments can apply countercyclical policy, and consumers can adjust their spending patterns.

According to the above, and considering the new advances in nowcasting models and their integration with Big Data sources, the National Statistical Office of Colombia (DANE, by its acronym in Spanish) has structured a project with the support of members of the Statistics Advisory Council to develop a methodology for calculating a Colombian agricultural sector´s early estimator using Google News and Google Trends, as an alternative source of information, with the aim of providing policymakers with timely statistics.

In this regard, the objective of this paper was to obtain an index for early measurement of the behavior of the economy for its different economic sectors, which uses Nowcasting techniques to take advantage of information from alternative sources such as Google Trends and Google News. The hypothesis to be tested behind this indicator would be using textual information from the Internet. Economic growth could be modeled for a given sector with less frequency than is required with traditional methods, obtaining a close approximation to the real values of the Gross Domestic Product calculated conventionally.

For the above purpose, a Zero Shot Classification model was used to analyze the text information collected from the two sources described in the last 17 years, with the implementation of a Ridge model to estimate the GDP of the agricultural sector for the case of Colombia.

(1) Universidad Nacional de Colombia, Colombia.
(2) Universidad EAN, Colombia.
(3) Women in Global Health, Colombia.
(4) Departamento Nacional de Estadística DANE, Colombia.
*Corresponding author: ldparra@universidadean.edu.co

ISSN: 0718-2724. (http://jotmi.org)

The results of this study for the agricultural sector in Colombia indicate that the alternative sources of Google Trends and Google News provide invaluable information to predict trends in the economic behavior of a given sector.

In turn, the combination of machine learning techniques, such as neural networks, with Nowcasting techniques yielded a positive result with high levels of adjustment and precision of the projection of the indicators compared with conventional methods. The comparative analysis of the different forecasting methods establishes that the Ridge method presents the lowest estimation error (10% less) based on which it can have an early indicator of the GDP of 8,188 m.m.p (billions of Colombian pesos) for 2022Q2, 30 days before the publication of the official data of 8,125 m.m.p.

## 2. Literature Review

The increasing availability of Big Data has revolutionized the field of economic nowcasting, which is the practice of estimating current economic activity in real time. A recent bibliometric analysis using Scopus data (Huang et al., 2023) highlights the growing interest in this area, particularly from government institutions. This trend aligns with the findings of Bok (2018), who emphasizes the potential of Big Data to improve the accuracy of nowcasting macroeconomic variables.

Big Data methods offer several advantages for nowcasting. They allow researchers to incorporate a wider range of high-frequency indicators, such as web search queries, social media sentiment, and satellite imagery, into their models. This can lead to more comprehensive and timely insights into the current state of the economy compared with traditional methods that rely on limited (Bok, 2018). However, many challenges remain regarding issues such as data quality, model selection, and interpretability when working with Big Data for nowcasting (Huang et al., 2023)

On the other hand, there is a long literature review regarding the use of nowcasting methods to estimate the behavior of the economy. ((Eurostat), 2016) introduced a technical framework for flash estimation of GDP using a pragmatic approach. In the presence of related indicators, the simplest extrapolation of a GDP component can be based on a regression method. Under a direct strategy, commonly used methods are autoregressive distributed lag (ADL) models and factor models, whereas under the indirect approach, temporal disaggregation regressions adopting either an ARIMA structure on the residuals or ADL forms are used.

The National Institute of Statistics and Geography (INEGI by its acronym in Spanish) in México developed an important work related to timely GDP estimation to perform podcasts for the percentage annual variation of the Mexican Global Economic Activity Indicator (IGAE by its acronym in Spanish) using economic and financial time series and real-time variables such as social mobility and significant topics extracted by Google Trends and taking advantage of dynamic factors models (Corona et al., 2021).

This study went beyond including nontraditional information to capture more drastic frictions that occur in the very short run, one or 2 months. The authors identified that previous works focused on traditional information, which limits their capacity to predict the historical declines attributed to COVID19 and the associated economic closures since March 2020. However, this approach could maximize the structural explanation of the already relevant macroeconomic and financial time series with the timeliness of other high-frequency variables commonly used in big data analysis. Broken down by sector and using IGAE, the economy suffered devastating consequences in the secondary and tertiary sectors.

Another studies, such as (Li et al., 2022) focused to disentangle specific influential market-related factors in agricultural futures through a text-based framework (TBF) to forecast soybean future prices in the Chinese market, which takes full advantage of the information in online news to improve forecasting performance. In this way, a topic model (DP-Sent-LDA) was employed to extract influential factors of agricultural futures from online news headlines, and the sentiment analysis method (Bi-LSTM) was used to quantify some important factors that were previously difficult.

From another perspective, some investigations have focused on the construction of statistical models for predicting indicators, or what is currently known as prospective analytics. Among the most widely used models and techniques are proportionality functions, moving averages, regressions, and imputations that allow the forecasting of indicators such as the consumer price index, sectoral economic growth, or even the level of consumption of certain resources (Wang & Cao, 2021, Joseph et al., 2021, and Ashouri et al., 2018).

Other studies have focused on the use of ARIMA models (or integrated moving average autoregressive models) and machine learning (ML) to estimate the behavior of economic activity based on their analysis of time series and historical data that allow project variables or indicators based on their behavior in the past (Masini et al., 2021; Dave et al., 2021 and Menculini et al., 2021). However, the above models are mostly based on parametric estimates that use structured data (mainly variables and indicators) to estimate or forecast the future behavior of a variable.

Less frequently, a few studies use unstructured data from internet search engines and social networks to make estimates. Among these, the most frequent is the use of seasonal or Bayesian models (Choi & Varian, 2012, Levenberg et al., 2013; Chakraborty et al., 2016; Argote and Parra, 2020) and to a lesser extent, the use of Big Data and unstructured data sources (Tuo et al., 2021; Xie et al., 2021; Giannone et al., 2021; Feldmeyer et al., 2021 and Bok et al., 2018). In the same way, high information production about news and the availability of new techniques and algorithms to analyze unstructured information open the way to a set of investigations that abstract this information, model it, emit metrics, and influence the predictive capacity of data for decision-making.

The nowcasting techniques for short-term economic cycle forecasting have become highly relevant in recent years because decision makers require tools for real-time projection of the economic cycle. Dauphin et al. (2022) showed how the use of ML algorithms, including Random Forest, Neural Networks, LASSO models, and Support Vector Machines (SVM), expands the range of available techniques to make economic cycle predictions.

Cepni et al. (2019) compared different machine learning techniques to forecast the behavior of GDP in emerging economies, indicating the importance of the pre-selection and prior filtering process that must be performed on the information before being processed. In turn, Jardet and Meunier (2022) compared the goodness of fit between the LASSO model and the MIDAS model for the GDP forecast of a set of countries with more than 718 data. In their study, they emphasize that although this type of technique produces reasonable results in GDP forecasts in periods of relative macroeconomic stability, they are distorted in periods of crisis such as the one that occurred in 2020 with the COVID-19 pandemic.

Richardson et al. (2021) used ML techniques to structure a real-time evaluation of the economic cycle measured from the GDP of New Zealand. In their study, they used more than 600 predictors to improve the estimation of the country's GDP, finding that ML algorithms that involve a greater number of data had a better level of adjustment in the prediction of the economic cycle compared to simple autoregression models that only consider past indicators data.

In accordance with these approaches, in 2020, DANE-Columbia developed two methodological approaches, firstly following to Martínez et al. (2016), a coincident index (ISE) was developed, which represented the state of nature using variables that move contemporaneously with the state (latent stochastic process) of interest, where non-stationary and common stationary factors were extracted from a multivariate stochastic process and the more appropriate factor was chosen to serve as the index. The challenge was to define a coincident profile to choose a common factor.

Second, in an internal report, DANE, Colombia, designed a nonlinear auto-regressive network with an exogenous inputs neural network (NARX), which was useful for modeling long-term dependencies in time series data due to the extended time delays captured, such as developed by Lewis (2016) who built a NARX model to predict the UK annual unemployment rate, with the objective of creating a model that predicted the actual level within a comfort interval of two standard deviations of the model prediction.

Once DANE published the official ISE data for 2020 Q1, the second approach performed better. The challenge identified was to find and select the best indicators that respond to the situation of the economy, considering that indicators with early publication are highly relevant for research, so the information obtained must be published with greater opportunity than the ISE. Estimates could be obtained before 30 days or, in the best of cases, 1 day (DANE, 2020). Although DANE is not a statistics office with a tradition in forecasting, these studies opened the door to continue to work.

One of the main contributions of this paper to the academic literature was to provide empirical evidence for the utility of combining Nowcasting Techniques with the use of unstructured Big Data Sources for early estimation of Economic Activity. This considering that nowcasting combines current data, including big data sources, to predict economic trends as they occur. It identifies changes in GDP and other critical indicators in real time, making it especially valuable in countries such as Colombia, where higher frequency GDP statistics are not readily available.

On the other hand, to test the study's hypothesis, the authors performed a set of machine learning experiments to determine the relationship between these inputs and the agricultural sector´s GDP, finally choosing the models with the best performance metrics. This approach makes it possible to take advantage of the news to improve the forecast of the GDP of any sector of the economy and obtain estimates up to 30 days before the official results, but at the same time, very close to the real value obtained through traditional field data collection methods. It should be noted that the authors have consolidated a time series for the agricultural, industrial, and tourism sectors; however, this paper shows only the results for the agricultural sector with the aim of explaining in depth the methodological aspects.

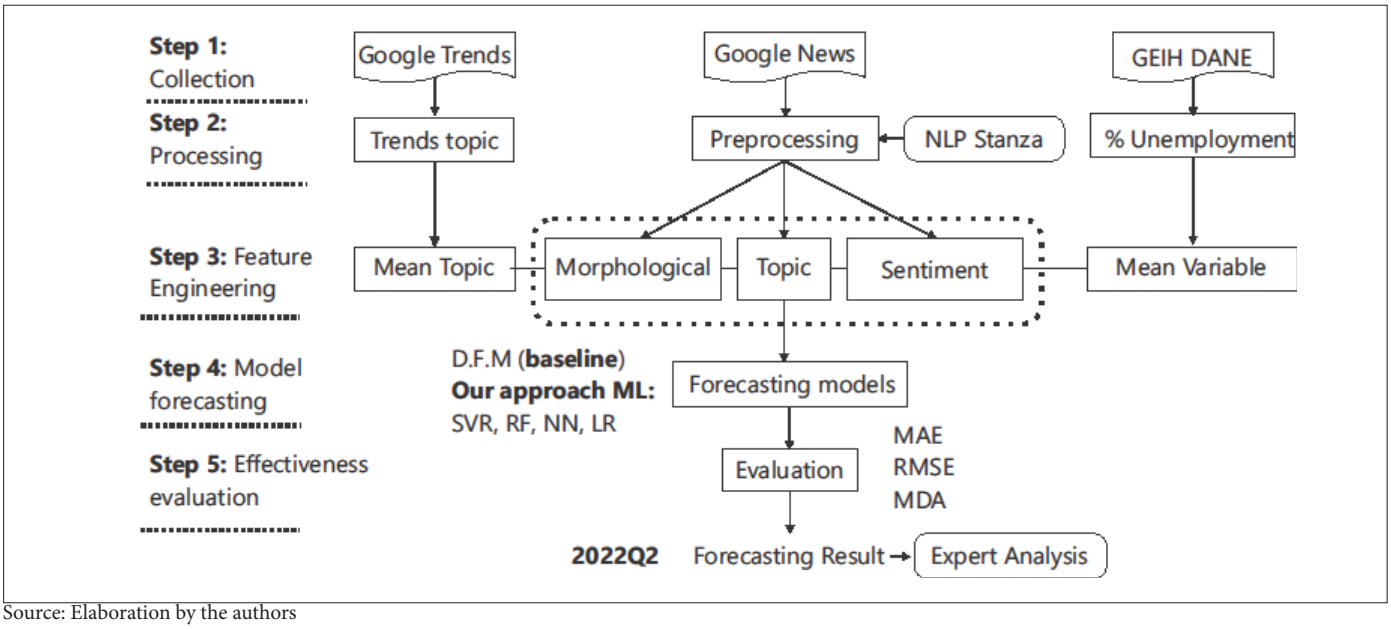
The remainder of the article is organized as follows: Section method and Dataset presents the dataset and preprocessing strategy, section Data and Model Evaluation shows the experimental evaluation of the method for calculating the early estimator, the experimental setup and results, and finally, section Conclusions and future work report the final discussion and upcoming research lines.

## 3. Methodology

The methodology comprises five principal components. First, data collection concentrates on both data sources, traditional/structured and nontraditional/unstructured data. As traditional sources, we use the main Colombian labor force indicators provided by the Labor Survey (GEIH in Spanish). On the other hand, we use Google Trends topics, an up-to-date source of information that provides an index of internet searches or queries by category and geography; however, our analysis goes beyond including Google News.

Second, we preprocess this news database to normalize all the characters and obtain comparable data both in format and date for the title of the news and its content. In addition, those frequent words that do not correspond to the information that is needed are identified and cleaned. Third and fourth, using deep learning techniques, we applied feature characteristics with Google News data to obtain information about sentiment, distribution of occurrences of the text morphology, and syntactic proximity between Google News topics and topics associated with the agricultural sector (in this specific case). Fifth, we compare a baseline using the dynamic factor model against a machine learning approach. Finally, we compute performance metrics for effectiveness evaluation. Figure 1 shows a synthesis of the methodology.

**Figure 1.** Pipeline of the proposed methodology.




Source: Elaboration by the authors

**Step 1: Data collection**

As mentioned previously, this study combines the use of traditional data sources such as labor survey statistics with nontraditional sources to capture relevant events and frictions that occur in a very short time.

To collect Google News, an API was used to extract the URLs of the articles and a second API was used to collect the information from the web, whose input was the list of URLs obtained with the first API. To extract the URLs of the articles, the following terms were defined: headlines, language, the time interval of publication, and page numbers of the search result. At the same time, to collect the Google Trends, we used the API to download Google Trends reports from January 2005 to March 2022; 84 parameters related to the sector were used, for instance: agriculture, forestry, hunting, fishing, electricity, natural gas, water, agricultural, food, bank, poverty, wealth, rural, urban, economic openness, etc.

**Step 2: Data Processing**

The information sources from Google Trends and the GEIH did not require preprocessing; however, the text rendering of Google News articles by its form required a cleanup process to ensure no artifacts were introduced into the feature rendering, which consists of five steps that are briefly described in this section and presented in Figure 2.

The text information of articles was normalized through the following parameters: name of the headline, summary, and content of the news; after that, a cleaning process was then applied to the non-relational database. For this purpose, special characters such as symbols between letters, arithmetic signs, text to lowercase letters, emoticons, blank spaces, and HTML format were removed. Additionally, to have comparable data both in format and date, the dates were normalized to ``DD/MM/YYYY". Subsequently, a linguistic process was performed to standardize, disambiguate, segment, and label the information through morphological analysis. For this, each word was reduced to its lemma to obtain a representative of all the inflected forms of the same word, and the morphological analyzer used for this purpose was the Stanford NLP Group® (see Figure 2-C). Finally, stop words such as articles ("el", "la" in Spanish) and prepositions ("de", "en" in Spanish) were removed because they were the most common words in all the documents and their frequency representation was high, which could add noise in the modeling phase.

**Step 3: Feature Engineering**

The period of the fundamental value of this study is quarterly; however, the information from the alternative data sources used in this study differs. For this reason, the transformation of the data into characteristics was advanced to obtain the inputs for machine learning models.

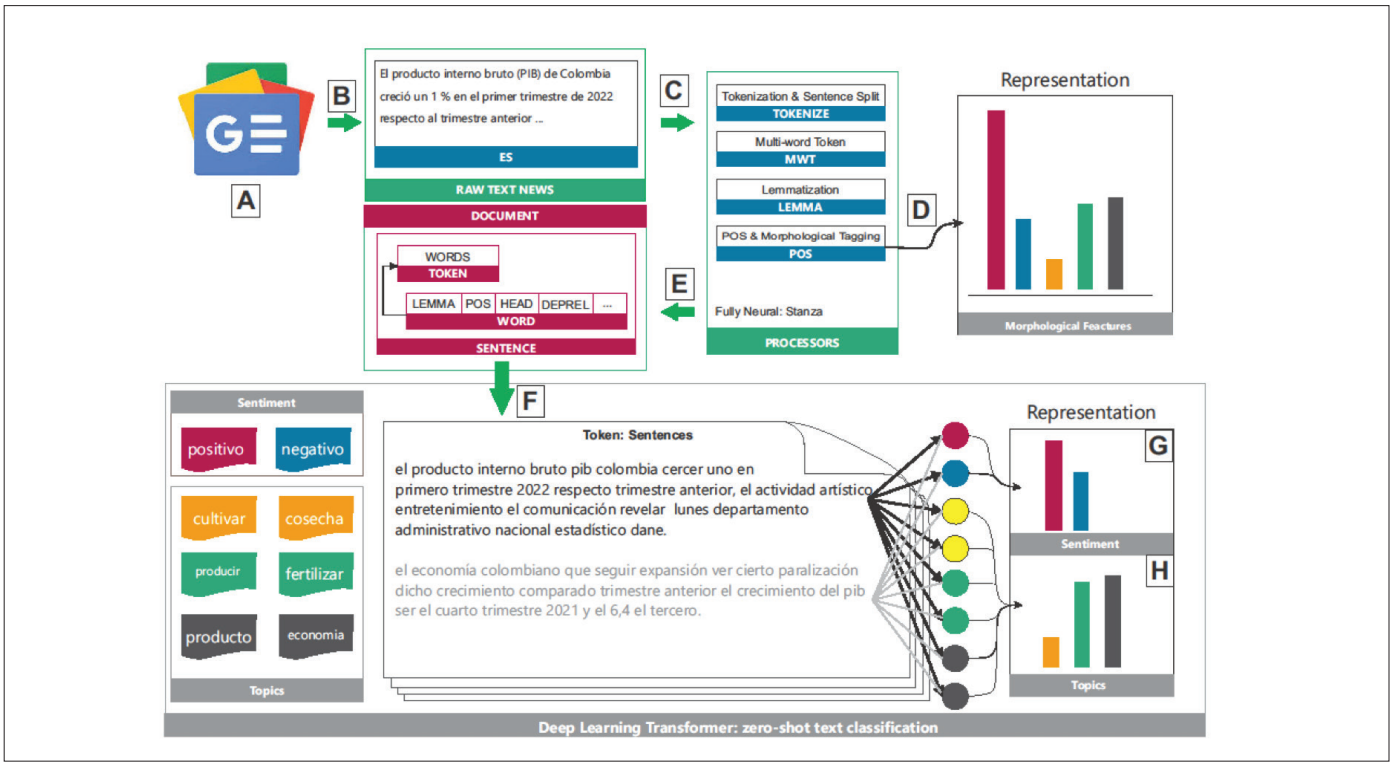
***GEIH and Google Trends***

The statistical information of the GEIH and Google Trends data used in this study have a monthly periodicity. For the feature engineering process, the average was used as a statistic to obtain a representative scalar for each quarter.

***Google News***

The text information from Google News was vectorized to quantify the words and sentences presented in the corpus; and thus obtain a context value for the title, description, and the entire content of the news. Three descriptions were defined for the representation, which are presented in Figure 2.

**Figure 2.** Google News Feature Extraction Process.




Source: Elaboration by the authors

**Morphological Features**

Parts of speech (POS) tags are the properties of words that define their main context, functions, and usage in a sentence. For this work, the distribution of the occurrence of the following parts of speech was used: adjectives, adverbs, auxiliaries, conjunctions, determiners, nouns, numbers, pronouns, proper names, punctuation, and subordinate conjunctions to establish a grouping of labels that represents the information of the corpus, characterized by a histogram of occurrences. Figure 2 (See A-D) summarizes this process.

**Sentiment and Topics**

The underlying model was trained on the Natural Language Inference (NLP) task, which takes two sequences and determines whether they contradict, are related to each other, or neither. The model was implemented for two uses: polarity and multi-label classification. The semantic processing unit comprised the sentence sequences obtained in the morphological analysis applied in the preprocessing stage. Figure 2 (See A-C, E-H) summarizes the process of representing the topics for each document.

The tasks described in this section are based on the principles of unsupervised classification, whose main difference from supervised classification is the development of labeling exercises. In this way, in the experimental exercise developed, the choice of class was essential to determine the predictive power.

In addition, the selection of the set of labels was made in coordination with a team of experts from the agricultural sector of the Direction of Synthesis and National Accounts (DSCN by its acronyms in spanish) of DANE. The following words are a set of labels resulting from this articulated exercise: “cultivate”, “sowing”, “harvest”, “farming”, “produce”, “agricultural”, “fruit”, “farm”, “fertilize”, “invest”, “economy”, “product”.

**Step 4: Baseline Model Forecasting**

Dynamic Factor Models (DFM) posit that some unobserved ``factors” can be used to explain a substantial part of the variation and dynamics in a more significant number of observed variables. A ``large” model typically incorporates hundreds of observed variables, and the estimation of dynamic factors can act as a dimension reduction technique. In addition to producing estimates of unobserved factors, dynamic factor models have many uses in macroeconomic forecasting and monitoring. A popular application for these models is ``Nowcasting”, in which higher frequency data are used to produce “nowcasts” for series that are released only at a lower frequency.

DFMs are based on the fundamental idea that many economic variables show similar developments throughout the business cycle and can therefore be described by some common factors. The statistical model and the EM (Expectation of Maximization) algorithm used

to parameterize estimates are described in (Banbura & Modugno, 2014; Banbura et al., 2011; Bok et al., 2018; and Mariano & Murasawa, 2010). Formally, a DFM assumes that many observed variables $y_t=[y_{1,t},y_2,t),..,y_{n,t}]$ are driven by some unobserved dynamical factors, while features that are specific to individual series, such as measurement errors, are captured by idiosyncratic errors (e1, t, ..., en,t) as it can see in equation 1 and 2:

$$y_t = \wedge f_t + \epsilon_t \quad (1)$$

$$f_t = A_1 f_{t-1} + A_2 f_{t-2} + \cdots + A_p f_{t-p} + u_t \ (2)$$

Where $y_t$ is observed data at time $t$, $\epsilon_t$ is the idiosyncratic disturbance at time $t$, (see details below, including modeling of the serial correlation in this term), $f_t$ is the unobserved factor over time $t, u_t \sim \mathrm{N}(0,\mathrm{Q})$ is the disturbance factor at time $t$, and $\wedge$ is referred to as the matrix of factor loadings, $A_i$ are matrices of autoregression coefficients. Furthermore, we allow idiosyncratic disturbances to be serially correlated. In this regard, the idiosyncratic function is represented accordingly to equation 3.

$$\text{Idiosyncratic\_ar1=True, } \epsilon_t = p_i \epsilon_{t-1} + e_{i,t-1} \sim \mathrm{N}(0, \sigma_2^i). \ (3)$$

If the idiosyncratic Arima is False, instead we have $\epsilon_{t-1} = e_{i,t}$.

**Step 5: Machine Learning Ridge Model**

In the proposed case study, we have a training set $(x_1, y_1), \ldots, (x_T, x_T)$ where $T$ is the number of examples (see Section Dataset), $x_t$ are vectors in $\mathbb{R}^n$ ($n$ is the number of features: Google News + Google Trends + GEIH) and $y_t \in \mathbb{R}, t = 1, \ldots, T$ represents a target variable. Assume a linear model $t = w \cdot x$, where $w \in \mathbb{R}^n$. The Least Squares method recommends computing $w = w_0$ which minimizes equation 5

$$L_T = \sum_{t-1}^{T} (y_t - w \cdot x_t)^2, \ (4)$$

and using for labeling future examples: if a new example has attributes $x$ , the predicted is $w_0 \cdot x$. .

There are three main types of regularization techniques: Ridge Regression, Lasso Regression, and Elastic Net based on the classical linear regression method. (Marquardt & Snee, 1975) presents the Ridge Regression procedure as a slight modification of the least squares method and replaces the objective function of equation 5

$$a\|w\|^2 + \sum_{t-1}^{T} (y_t - w \cdot x_t)^2 \ (5)$$

where (also called hyperparameters) is a fixed positive usually quantified by applying numerical methods (i.e. grid search or random search see Table 1). The parameters are "tuned" using cross-validation resampling methods by minimizing the forecast error in the validation set.

## 4. Dataset Description

The main idea behind forecasting is to take advantage of a diverse set of timely information available before an official estimation is calculated. Our approach maximizes the structural explanation with some of the most relevant indicators to describe the economic dynamic, such as working-age population, total population, and employed and unemployed population, with the timeliness of other high-frequency variables commonly used in big data analysis. The motivation behind such large datasets has been to maximize the information set and thus reduce the risk of bias due to information omission. The dataset also includes variables available from 2005Q1 to the present (2022Q2). The total number of variables used to train the models was 150 characteristics of the economy. The data were organized into the following groups:

- Primary Sector Value-Ground Truth from DANE: 70 stationary variables from 2005Q1 to 2022Q1 for training and validation.
- Google News: 24699 unique news items from different media outlets were used for training and validation of the ML models for the period 2005Q1-2022Q2 and 2392 for the 2022Q2 forecast (Morphological features: 12 + Sentiment value of the title, description, and article: 3 + Quantity News and sentences: 2 + Topics in the title, description, and article: 36 = 53 features).
- Google Trends: Eighty-five Google trending search terms were used from 2005Q1 to 2022Q2 (85 features).
- GEIH: Twelve variables from the seasonally adjusted labor market series were obtained monthly from 2005Q1 to 2022Q2 (12 features).

## 5. Data and Model Evaluation

**Experimental Setup**

The proposed approach requires information processing to train a set of Machine Learning (ML) algorithms. These results were compared with a classic baseline for time series forecasting, such as dynamic factor models. To obtain an estimation of the added value, a regression-supervised classification task was selected for the ML algorithms, where the representation of Google Trends, Google News, and GEIH were established as inputs, and the objective value was defined as the information published by DANE.

Due to the nature of the data, it was necessary to evaluate the proposal using different algorithms to obtain a model that would fit the desired response. For this research, a set of experiments was developed, each corresponding to a case study, and each case study was differentiated based on the defined parameters, such as headlines and news filters out of context (Google News), nature of the models, and variation of the training and test percentages. Based on the above, different configurations were tested for the baseline and ML models.

**Baseline: Dynamic Factor Model**

Table 1 displays the results of the algorithms used, which are subdivided into two types of models:

Dynamic Factor Model (DFM 1): In this case, the predictor variables used were those obtained in the feature engineering process. That is, the input of this model is the same as that used for the Machine Learning models, establishing the same representation (Google News + Google Trends + GEIH DANE) to obtain a comparison of the performance metrics of the models.

Dynamic Factor Model (DFM 2): It is established only for this case that the information used to make the forecast comes from the GEIH-DANE. This source of information comes from a survey carried out by the information custodian in Colombia, and in this sense, it has a methodological rigor establishing the representatives of the information in the country. In this context, the information from Google Trends and Google News was not used because the data obtained is dependent on the headlines and the importance of the words, which may have an independent bias if chosen by experts.

**Table 1.** Machine Learning models used in the experiment (ANN: Artificial Neural Network, RF: Random Forest, LR: Linear Regression, SVR: Support Vector Regression)

| Model | Parameter optimization | Method |
|---|---|---|
| ANN 1 | Optimizer, activation function and loss cost, batch size, neurons, and epochs. | genetic algorithm |
| ANN 2 | Optimizer, activation function and loss cost, batch size, neurons, and epochs. | genetic algorithm |
| RF | Number of trees, maximum number of features, minimum number of data points, and data allowed in a leaf node. | Random hyperparameter search |
| LR | - | - |
| Lasso | Method, alpha values, cross-validation, and iterations | Random hyperparameter search |
| Ridge | Method, alpha values, cross-validation, and iterations | Random hyperparameter search |
| SVR | Kernel, epsilon values, cross-validation, and iterations | genetic algorithm |

Source: Elaboration by the authors

Researchers selected a variation of 70%–99% of the observations to train and test the model. Note that the selected observations consider temporality; that is, their selection was not random because it maintains the notion of the neighborhood.

Based on the experimental configuration, a set of parameters, such as test observations, algorithm, and information representation, was selected that would allow modeling of the problem. The best results were selected on the basis of a set of quantitative and qualitative rules, since the techniques may work better than others in specific circumstances.

Some models can better capture significant changes and therefore better predict turning points, such as crises. Other models can better filter noise and work better in a more stable environment. Not surprisingly, DFM and various ML models can produce results that send mixed and unclear signals, even when applied to an identical dataset. To determine the predictive accuracy of individual models, a backrest was performed, and the accuracy indicators were quantified using the test residuals.

- Backtest: In the first stage, to represent a real-time environment, in each period $t$, data on all explanatory variables are available from the beginning of the sample to the last quarter $t$. However, information on the target variable is available only up to quarter $t-1$. In each quarter , the parameters of each model are estimated using data up to quarter $t-1$. Subsequently, the estimated models are applied to the explanatory variables at time $t$ to produce a one-step-ahead out-of-sample prediction (pseudonowcast) of the target variable. Finally, the forecast errors are quantified by comparing the actual and estimated values of the target variable.

- Quantification: The quantification of the model estimates was obtained using different metrics that represent the error of the prediction; it evaluates how close it is to the real value. Even though each evaluation measure weighs the error from a different point of view, they have in common the measurement of the effectiveness of the methods used.

- Qualitative: A set of experts on economic issues was defined, which comprised three specialists. Based on the quantitative results, the researchers presented a set of models to the experts. As a result, the experts selected a model that adjusted to the forecast established by their experience in accordance with the reality of the economic sector.

Quantitative evaluation was performed using the metrics suggested by (Dauphin et al., 2022). The trained models were evaluated quantitatively using equations 6, 7, and 8 related to the mean absolute error (MAE), Root Mean Square Error (RMSE), and Mean Directional Accuracy (MAD). The information on the metrics obtained by each previously trained model was tabulated, and the behavior of each model was evaluated in both the training and test stages.

$$MAE(y,\hat{y}) = \frac{1}{n}\sum_{i=0}^{n}|y_i - \hat{y}| \quad (6)$$

$$RMSE(y,\hat{y}) = \sqrt{\frac{1}{n}\sum_{i=0}^{n}|y_i - \hat{y}|} \quad (7)$$

$$MDA(y,\hat{y}) = \frac{1}{n}\sum_{i=0}^{n} sign(y_i - y_{i-1}) = (\hat{y}_i - \hat{y}_{i-1}) \quad (8)$$

Where $\hat{y}$ is the predicted value of the sample (value estimated by the model), and $y$ is the real value (corresponds to that published

by DANE). The first two indicators (RMSE and MAE) represent the measures of the average prediction error of the model. By squaring the error terms, RMSE assigns greater weight to large errors, which are useful indicators to compare the predictive power of different models, particularly when there are large errors.

Although MAE is only slightly different in definition from MSE, it has different properties in that it assigns the same weights to large and minor errors. Therefore, unlike the MSE, MAE gives little weight to outliers and provides a generic measure of how well a model is performing. MDA is a measure of predictive accuracy that indicates the probability that a model predicts the true direction of a time series. In other words, MDA compares the predicted direction (up or down) with the actual realized direction.

**Models' performance**

Based on the experimental setup, a set of ML and DMF models was established as the baseline. The objective of the exercise was to forecast the added value of the agricultural sector. Based on this information, the top 10 models with the best performance metrics were chosen, and their behavior over time was observed for each of these models. The predictions of the observations were evaluated quantitatively by established metrics and qualitatively by a group of expert economists. The results obtained for the configuration 70% to train (2005Q1-2016Q4: 48 samples), 30% to test (2017Q1-2022Q2: 21 samples), and the forecast (2022Q2: 1 sample) are shown below.

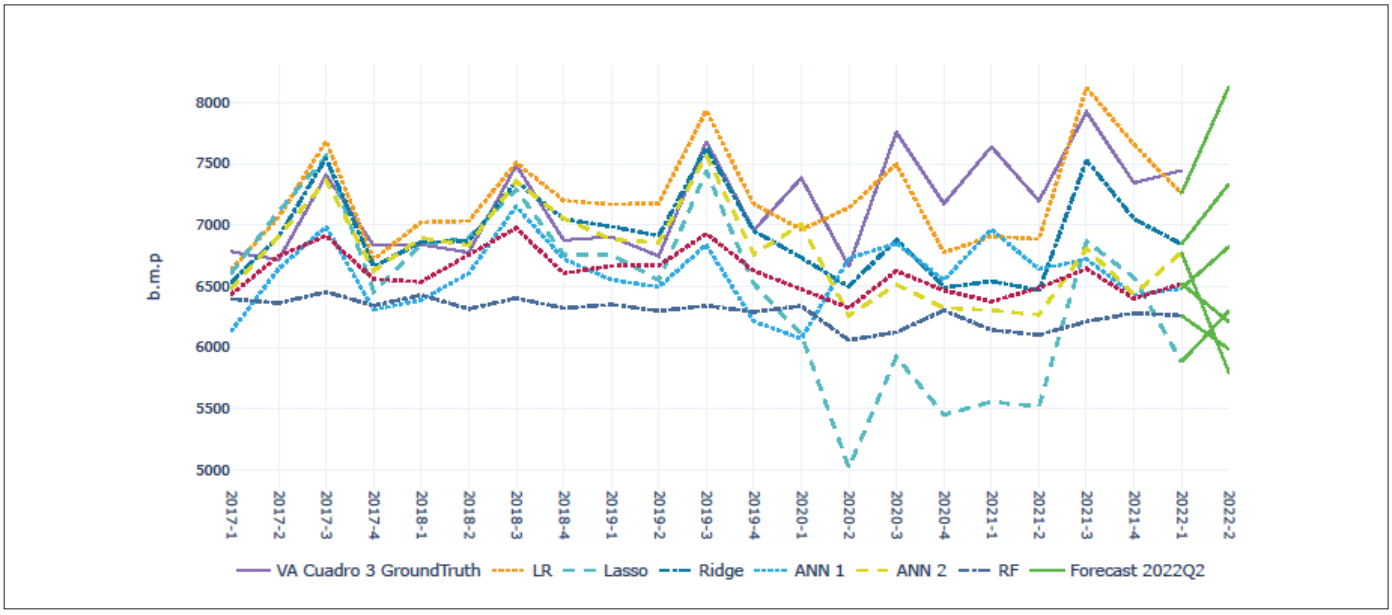
**Table 2.** Comparative results: pipeline vs ML. Note: the black bold words indicate the selected model, whereas the green bold values indicate the model with the lowest error against ML and DFM. The m.m.p. means (billions of Colombian pesos).

| Model | Error Training MAE | Error Test MAE | Error Train RMSE | Error Test RMSE | Error Training MDA | Error Test MDA | Forecasting 2022Q2 b.c.p |
|---|---|---|---|---|---|---|---|
| ANN 1 | 0,02 | 0,18 | 0,03 | 0,21 | 0,50 | 0,50 | 7505.37 |
| ANN 2 | 0,03 | 0,14 | 0,04 | 0,20 | 0,50 | 0,50 | 6133.58 |
| Lasso | 0,01 | 0,24 | 0,01 | 0,33 | 0,94 | 0,75 | 6807.95 |
| RF | 0,05 | 0,28 | 0,05 | 0,31 | 0,94 | 0,80 | 6392.99 |
| SVR | 0,08 | 0,18 | 0,08 | 0,22 | 0,85 | 0,65 | 6687.57 |
| Ridge | 0,01 | 0,10 | 0,01 | 0,14 | 0,91 | 0,75 | 8188.09 |
| LR | 0,00 | 0,09 | 0,00 | 0,10 | 1,00 | 0,65 | 9248.01 |
| DFM 1 | - | 0.39 | - | 0.41 | - | 0.5 | 6882.47 |
| DFM 2 | - | 0.13 | - | 0.17 | - | 0.5 | 5909.48 |

Source: Elaboration by the authors

Table **2** reports the results obtained. As can be seen, the ML models tend to outperform the ARIMA (1) reference model because the MDA is lower. In the forecast for 2022Q2, the trend could present erroneous predictions considering that the added value for the sector in 2022Q1 was 8,333 billion Colombian pesos, and the DFM estimates imply a decrease to the forecast value of 17% and 29%, respectively.

**Figure 3.** Results by ML models to predict GDP (Purple line: Xmax=8965.3 - Xmin=4761.2)




Source: Elaboration by the authors

Figure 3 shows the goodness of fit of the ML models for the sector studied in the different time periods as observed in the LR, Lasso, and Ridge algorithms; in most cases, these trends were adjusted to the real value. However, in Colombia, the situation of COVID19 strongly affected the added value for 2020Q2.

Additionally, the time series consistently reported an upward trend for this period; in this case, the LR model does not fit this trend, and therefore it is not chosen by the economist group as a reference model; for the Lasso and Ridge model, an adjustment close to the trend of the target value is observed and it shows that for 2020Q2, its trend is downward. To select one of these methods, the metrics and as presented in

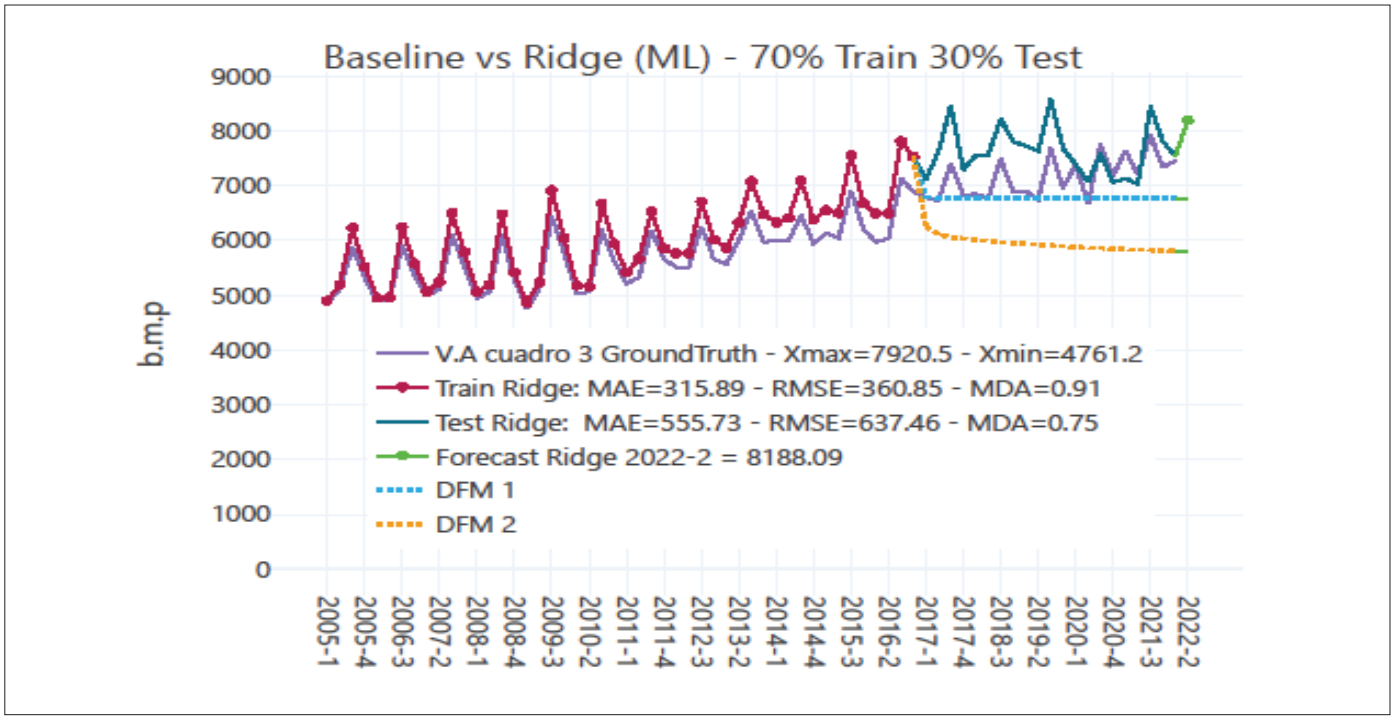
Table 2, the algorithm that is quantitatively and qualitatively adjusted considering the concept of the experts is the Ridge algorithm to obtain the estimate of 2022Q2.

## Results-Forecasting 2022Q2

The findings of this exercise are specific to the experimental setup shown in

Table 2 and should not be generalized. In this sense, Ridge regression and Lasso models could significantly reduce the average forecast RMEs from 10% to 14%. In fact, during the pandemic and specifically in 2020Q2, these models predicted a decrease that was a big challenge for all National Statistical Offices (NSO) in most of the countries. The results are quite different when considering nowcasting performances during isolation by COVID-19. Notably, the ML methods appeared to estimate better than DFM. In other words, the best model that captures GDP agricultural dynamics is the Ridge model across time, as shown in Figure 4. It was presented with the qualitative results of the selected model and baseline.

**Figure 4**. Comparison of Baseline from Ridge model.




Source: Elaboration by the authors

Although DANE has not had a tradition in exploring nowcasting methods, the COVID-19 pandemic prompted not only to explore innovative methods to produce official statistics but also the complexity of this type of phenomenon became an almost impossible historical event to predict. Therefore, the challenge for NSO was twofold. In this sense, the Ridge results presented in Figure 3 are a good fit to the official data.

It is observed that the drop in value added in 2020-Q2 was an abrupt change in accordance with the pattern of the series in the same period. In this sense, the Ridge model predicts that the fall starts in 2020-Q1 and in 2020-Q2 it establishes a downward trend for the first two quarters. At the state-of-the-art level, some models used to perform the same task did not have context information to approach the trend arising from conjunctural issues.

The performance of the individual DFM and ML models can also differ with respect to the experimental configuration. For the DFM, it is better to use more than 90% of the samples to obtain a forecast close to the target value. However, one of the benefits of ML models is to learn a representation in the training stage to predict a value based on the description and the same characterization to predict the

value without having ground truth, which allows establishing a frontier decision based on information that seeks to establish evidence of the quarter under study. In this sense, a good description can establish a good prognosis.

For this study, a model that captures the GDP dynamics of the studied sector is the Ridge because it obtained errors of 0.01, 0.08, and 0.91 for MAE, RMSE, and MAD, respectively, in the training stage, in contrast to the 0.10, 0.14, and 0.75 metrics obtained for the test samples.

Finally, the experimental set-up and evaluation presented previously are the results of a methodological proposal developed from the validations that have been carried out on the GDP of the agricultural sector in different periods of time. In this context, the historic results obtained with the methodology are presented in Table 3, which confirms the benefits of a novel solution using non-traditional information with daily periodicity, and subsequently obtaining a rich representation to train a model and forecast a proxy indicator of the economic sectors.

**Table 3.** Historic results using our methodology ([a]: previous official publication)

| Model | Forecasting b.c.p[a] | Published Data |
|---|---|---|
| ANN | 2021Q2: 8,044 | 2021Q2: 8,037 |
| ANN | 2022Q1: 7.940 | 2022Q1: 8,333 |
| Ridge | 2022Q2: 8,188 | 2022Q2: 8,125 |

Source: Elaboration by the authors

In accordance with the above, the following contributions of this research are highlighted: First, to test non-conventional methods of estimating sectoral economic activity that use information from the Internet such as Google Trends and Google News; second, to combine ML techniques with conventional indicator projection models; and third, to generate a methodology for calculating indicators for early measurement of economic behavior that allows it to be monitored in short periods of time, while they are produced and generate official statistics based on structured data.

At the same time, the investigation sought to contribute to this topic by exposing and implementing an early measurement index of economic activity for the agricultural sector. As part of the findings, it could highlight the fact that the combination of neural networks used for the selection and analysis of information from news and internet searches about the agricultural sector in Colombia, with the LASSO model implemented for the forecasting of GDP of this sector, allow estimating, with a high level of precision, what would be its behavior in short intervals of time, and in turn, the estimates that were obtained converged with the official statistics of the agricultural sector obtained through conventional statistics.

## 6. Conclusions and future work

The use of ML algorithms has gained great relevance in the last decade in the social sciences, largely because the different problems and phenomena that affect society and the economy do not have a linear behavior that can be modeled with the tools and instruments provided by conventional predictive statistics, but are immersed in complex behaviors that require the participation of multiple variables.

The present investigation aims to obtain an index for early measurement of the behavior of the economy for its agricultural sector, which uses Nowcasting techniques to take advantage of information from alternative sources such as Google Trends and Google News. Simultaneously, the hypothesis to be tested behind this indicator would be using textual information from the Internet. Economic growth could be modeled for a given sector with less frequency than is required with traditional methods, obtaining a close approximation to the real values of the gross domestic product calculated conventionally.

According to the above, the present research met the objective and tested the study hypothesis given that, on the one hand, it implemented an early measurement index of GDP for the agricultural sector for the Colombian case, and on the other hand, the estimates obtained from the machine learning and dynamic factor models were very close to the subsequently published official estimates, with an estimation error level below 10%. This means that nowcasting techniques combined with Big Data sources are a excellent alternative for capturing the dramatic economic frictions that occur in the very short run, specifically within a one-to-two-month timeframe in the "very short run". At the same time, another interesting finding that it can highlight is that the combination of the neural network used for the selection and analysis of news information and Internet searches on the agricultural sector in Colombia, with the LASSO model implemented for the projection of the internal production of the said sector, allows the estimation of a GDP of 8,188 MMP for 2022Q2 15 days before the publication of the official data of 8,125 MMP.

With regard to this, the information from Internet searches and news about the agricultural sector in Colombia showed great potential to improve the prediction models of the economic behavior of the said sector. Among the advantages of this type of information is knowing the users' perception of the country's economy, the issues that affect the indicators, and the low periodicity in obtaining said information, which allows for generating an indicator that summarizes and projects the production of the sector in very short periods of time.

Although this research sought to model a specific sector of the economy; the methodology is scalable to other economic sectors, using only news information and internet search trends. In this sense, it was observed that the models used (Ridge and LASSO model) presented an absolute error in the test between 9% and 10% for the indicators of GDP of the sector, which indicates low errors compared with the baseline.

Nevertheless, this study has a few limitations that promote future research directions. First, several studies confirmed that factors affecting analyst quality can be mined from unstructured data (including report text, collaborative networks, analyst appearance, etc.). The potential predictors presented in the proposed methodology could expand the text approach. In the future, deep learning techniques could be employed to extract more enriching predictors from these unstructured data.

Second, the predictors as inputs and the quality of data are closely related to the choice of a forecasting model that includes macroeconomic information from internal and external sources.

Finally, ML models or algorithms could be used to forecast the economic behavior of a sector in periods of strong turbulence to improve the probability of obtaining a precise radiography for all GDP.

## Acknowledgments

We gratefully acknowledge the sponsorship by the National Statistical Office of Colombia (DANE by its acronym in Spanish), in particular the support of the Prospective and Data Analytics, and Synthesis and National Accounts Units, as well as Julieth Solano, Technical Director of the DANE Regulation Unit. We would especially like to recognize the DANE specialists for their participation in carrying out exploratory exercises during the first phases of the research: Santiago Smith, Jennifer Salguero López, Daniel Pérez, Nicolás Chud, Camilo Castro, Nicolas Romero, Samuel Silva, and Nicolas Silva for their English written style review.

## Bibliography

Argote, M., & Parra, L. (2020). *Global Entrepreneurship Analytics, using GEM Data*. Routledge, Taylor and Francis. https://doi.org/10.4324/9780429316715

Ashouri, M., Cai, K., Lin, F., & Shmueli, G. (2018). Assessing the value of an information system for developing predictive analytics: The case of forecasting school-level demand in Taiwan. *Service Science, 10*(1), 58–75. https://doi.org/10.1287/serv.2017.0200

Banbura, M., Giannone, D., & Reichlin, L. (2011). Nowcasting. In L. Oxford Academy (Ed.), *Oxford Handbook on Economic Forecasting* (pp. 193-224). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780195398649.013.0008

Banbura, M., & 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. https://doi.org/10.1002/jae.2306

Big Data-Statista. (2022). Amount of data created, consumed, and stored 2010-2023, with forecasts to 2025. *Global state of big data/AI adoption 2023*. Statista.

Bok, B., Caratelli, D., Giannone, D., Sbordone, A. M., & Tambalotti, A. (2018). Macroeconomic nowcasting and forecasting with big data. *Annual Review of Economics, 10*, 615–643. https://doi.org/10.1146/annurev-economics-080217-053214

CCSA. (2020). *How COVID-19 is changing the world: A statistical perspective, Volume II*. ReliefWeb. Retrieved from https://www.reliefweb.int

Cepni, O., Güney, I. E., & Swanson, N. R. (2019). Nowcasting and forecasting GDP in emerging markets using global financial and macroeconomic diffusion indexes. *International Journal of Forecasting, 35*(2), 555–572. https://doi.org/10.1016/j.ijforecast.2018.10.008

Chakraborty, S., Mengersen, K., Fidge, C., Ma, L., & Lassen, D. (2016). A Bayesian Network-based customer satisfaction model: A tool for management decisions in railway transport. *Decision Analytics, 3*(1), 1–24. https://doi.org/10.1186/s40165-016-0021-2

Choi, H., & Varian, H. (2012). Predicting the present with Google Trends. *Economic Record, 88*, 2–9. https://doi.org/10.1111/j.1475-4932.2012.00809.x

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*. https://doi.org/10.48550/arXiv.2101.10383

DANE. (2020). *Boletín Técnico Producto Interno Bruto (PIB)*. Retrieved from https://www.dane.gov.co/files/investigaciones/boletines/pib/bol_PIB_Itrim20_producion_y_gasto.pdf

Dauphin, M. J.-F., Dybczak, M. K., Maneely, M., Sanjani, M. T., Suphaphiphat, M. N., Wang, Y., & Zhang, H. (2022). Nowcasting GDP-A Scalable Approach Using DFM, Machine Learning and Novel Data, Applied to European Economies. *International Monetary Fund*.

Dave, E., Leonardo, A., Jeanice, M., & Hanafiah, N. (2021). Forecasting Indonesia exports using a hybrid model ARIMA-LSTM. *Procedia Computer Science, 179*, 480–487. https://doi.org/10.1016/j.procs.2021.01.031

DSCN. (2020). Primera estimación rápida del Indicador de Seguimiento a la Economía. Ed. Departamento Nacional de Estadística, DANE, Bogotá, Colombia.

(Eurostat), E. C. (2016). Overview of GDP flash estimation methods: 2016 edition. *Publications Office*. https://doi.org/10.2785/51658

(Eurostat), E. C. (2020). Methodological note GUIDANCE ON QUARTERLY NATIONAL ACCOUNTS ( INCLUDING FLASH ) ESTIMATES IN THE CONTEXT OF THE COVID-19 CRISIS.

Feldmeyer, D., Nowak, W., Jamshed, A., & Birkmann, J. (2021). An open resilience index: Crowdsourced indicators empirically developed from natural hazard and climatic event data. *Science of the Total Environment, 774*, 145-734. https://doi.org/10.1016/j.scitotenv.2021.145734

Giannone, D., Lenza, M., & Primiceri, G. E. (2021). Economic predictions with big data: The illusion of sparsity. *Econometrica, 89*(5), 2409–2437. https://doi.org/10.3982/ECTA17842

Guevara, D. (2021). Crisis del COVID-19: Impactos socioeconómicos y retos para países Latinoamericanos. *Cuadernos de Economía, 40*(85), 1-6.

Jardet, C., & Meunier, B. (2022). Nowcasting world GDP growth with high-frequency data. *Journal of Forecasting, 41*(6), 1181-1200. https://doi.org/10.1002/for.2858

Joseph, A., Kalamara, E., Kapetanios, G., & Potjagailo, G. (2021). WITHDRAWN: Forecasting UK inflation bottom up. *International Journal of Forecasting, 3*(5), 1-15. https://doi.org/10.1016/j.ijforecast.2021.03.005

Levenberg, A., Simpson, E., Roberts, S., & Gottlob, G. (2013). Economic Prediction using heterogeneous data streams from the World Wide Web. In *Scalable Decision Making: Uncertainty, Imperfection, Deliberation (SCALE), Proceedings of ECML/PKDD Workshop*.

Lewis, N. D. (2016). *Deep Time Series Forecasting with Python: An Intuitive Introduction to Deep Learning for Applied Time Series Modeling*. Create Space Independent Publishing Platform.

Li, J., Li, G., Liu, M., Zhu, X., & Wei, L. (2022). A novel text-based framework for forecasting agricultural futures using massive online news headlines. *International Journal of Forecasting, 38*(1), 35–50. https://doi.org/10.1016/j.ijforecast.2020.02.002

Mariano, R. S., & Murasawa, Y. (2010). A coincident index, common factors, and monthly real GDP. *Oxford Bulletin of Economics and Statistics, 72*(1), 27–46. https://doi.org/10.1111/j.1468-0084.2009.00567.x

Marquardt, D. W., & Snee, R. D. (1975). Ridge regression in practice. *The American Statistician, 29*(1), 3–20. https://doi.org/10.2307/2683673

Martínez, W., Nieto, F. H., & Poncela, P. (2016). Choosing a dynamic common factor as a coincident index. *Statistics & Probability Letters, 109*, 89–98. https://doi.org/10.1016/j.spl.2015.11.008

Masini, R. P., Medeiros, M. C., & Mendes, E. F. (2021). Machine learning advances for time series forecasting. *Journal of Economic Surveys, 1*(37), 76-111. https://doi.org/10.1111/joes.12429

Menculini, L., Marini, A., Proietti, M., Garinei, A., Bozza, A., Moretti, C., & Marconi, M. (2021). Comparing prophet and deep learning to ARIMA in forecasting wholesale food prices. *Forecasting, 3*(3), 644–662. https://doi.org/10.3390/forecast3030040

Richardson, A., van Florenstein Mulder, T., & Vehbi, T. (2021). Nowcasting GDP using machine-learning algorithms: A real-time assessment. *International Journal of Forecasting, 37*(2), 941–948. https://doi.org/10.1016/j.ijforecast.2020.10.005

Tuo, S., Chen, T., He, H., Feng, Z., Zhu, Y., Liu, F., & Li, C. (2021). A regional industrial economic forecasting model based on a deep convolutional neural network and big data. *Sustainability, 13*(22), 12789. https://doi.org/10.3390/su132212789

Wang, C., & Cao, Y. (2021). Forecasting Chinese economic growth, energy consumption, and urbanization using two novel grey multivariable forecasting models. *Journal of Cleaner Production, 299*, 126863. https://doi.org/10.1016/j.jclepro.2021.126863

Xie, G., Qian, Y., & Wang, S. (2021). Forecasting Chinese cruise tourism demand with big data: An optimized machine learning approach. *Tourism Management, 82*, 104208. https://doi.org/10.1016/j.tourman.2020.104208

ISSN: 0718-2724. (http://jotmi.org)