GDP Nowcasting: A machine learning and remote sensing data-based approach for Bolivia

Osmar Bolivar*

April 2023

Abstract

The present study seeks to introduce and validate an alternative methodology to perform nowcasting of the monthly behavior of the Bolivian economy. The approach proposed herein represents an innovative strategy that leverages the power of machine learning algorithms and the utility of remote sensing techniques as inputs for economic activity forecasting models. This methodology is applied to Bolivia, where official monthly indicators of economic growth are published with a delay of up to six months from contemporary time. Specifically, the study develops a nowcast metric for monthly economic growth that reduces the gap between real-time estimates and official reports from 6 months to only 2 months, thereby providing more timely metrics that can inform policy-making decisions, aid researchers, and assist economic analysts in their analyses.

Keywords: nowcast; machine learning; remote sensing; economic growth forecast

JEL Codes: C10; C14; C22; C53; C80; E17

*The author of this paper expresses his own opinions and perspectives, which may not necessarily coincide with those of the institutions to which he is associated. E-mail: xxxx@gmail.com

1 Introduction

The Bolivian economy has undergone significant transformations since 2005, characterized by fluctuations in growth and challenges of both economic and political nature that have had a profound impact on its macroeconomic performance. Bolivia experienced a period of remarkable expansion, driven by a surge in domestic demand and a rise in commodity exports, particularly in natural gas and minerals. From 2006 to 2013, the country witnessed an average GDP growth rate of approximately 5%, which was supported by redistributive economic policies such as the nationalization of strategic industries, greater state intervention, higher wages for workers, and increased social spending.

However, the impact of low commodity prices in the global market, climate change on agricultural production, social conflicts, and a decline in public investment have all contributed to reduced growth rates after 2014. The COVID-19 pandemic has further exacerbated the economic downturn in the country, with border closures, a decline in demand, and dwindling trade all contributing to the contraction of the Bolivian economy by -8.7% in 2020.

In 2021, the Bolivian economy demonstrated gradual signs of recovery, although its GDP fluctuated in response to the COVID-19 outbreaks. From 2022 onward, Bolivia's economic metrics stabilized until September of that year. However, the last quarter of 2022 was marked by renewed social conflicts, resulting in a downturn in GDP.

Given the aforementioned challenges, obtaining timely and reliable economic data, particularly with regard to GDP, in Bolivia during periods of fluctuation has long been a concern for policymakers and economists. The absence of such data impedes the capacity to make informed decisions, implement effective policies, and accurately evaluate the effects of economic interventions. Nevertheless, recent developments in machine learning and remote sensing technologies present promising opportunities for overcoming these limitations.

Machine learning algorithms have facilitated the expeditious processing of voluminous economic data, thereby enabling economists and policymakers to promptly discern trends and patterns in economic indicators (Mullainathan & Spiess, 2017). The application of machine learning in economic forecasting has garnered significant scholarly attention in recent years, with studies attesting to the feasibility and efficacy of these methodologies. For instance, machine learning algorithms have been employed to forecast financial crises (L. Liu, Chen, & Wang, 2022), GDP growth volatility (Balcilar, Gabauer, Gupta, & Pierdzioch, 2022), unemployment (Gogas, Papadimitriou, & Sofianos, 2022), inflation (Medeiros, Vasconcelos, Veiga, & Zilberman, 2021), demand (Bajari, Nekipelov, Ryan, & Yang, 2015), sovereign risk (Belly et al., 2023), daily and monthly exchange rates (Plakandaras, Papadimitriou, & Gogas, 2015), agricultural commodity prices (Bonato, Çepni, Gupta, & Pierdzioch, 2022), house prices (Milunovich, 2020), entrepreneurial firm valuation (Zhang, Tian, McCarthy, Wang, & Zhang, 2023), among others.

The introduction of remote-sensing data into economics is a significant contribution to the field. Henderson, Storeygard, and Weil (2012) developed a statistical framework that utilizes satellite data on night lights to estimate GDP growth in countries with poor national income accounts. This method enables growth measurement at sub and supranational geographical levels. Subsequent studies have explored the use of night lights as a proxy for GDP growth, as well as its utility in deriving regional income (Pinkovskiy & Sala-i Martin, 2016), the combination of night lights and land cover to predict economic activity (Keola, Andersson, & Hall, 2015), and its suitability for capturing urban growth

(Storeygard, 2016). Additionally, Donaldson and Storeygard (2016) propose and describe how remote sensing methods can be used to process satellite imagery data to generate metrics about night lights, climate and weather, topography, agricultural land use and crop choice, urban land use, building types, natural resources, and pollution monitoring, which are appropriate predictors of economic activity.

By leveraging these technologies, policymakers in Bolivia could gain a more comprehensive and accurate understanding of the country's economic performance, anticipate changes in economic trends, and make more informed decisions about policy interventions. This, in turn, can help mitigate the adverse effects of economic downturns, support sustainable economic growth, and promote greater social welfare for the population.

Accordingly, the present study introduces an innovative approach for nowcasting monthly economic growth in Bolivia that combines machine learning and remote sensing data. This method represents a significant advancement in the field of economic forecasting, as it allows for the rapid and accurate estimation of GDP growth rates. The incorporation of remote sensing data provides a unique perspective on the country's economic activity, which complements traditional economic indicators. This research contributes to the development of more robust and reliable methods for economic forecasting, with potential applications in other contexts.

It is important to note that the terminology of Giannone, Reichlin, and Small (2008) and Bańbura, Giannone, Modugno, and Reichlin (2013) is adopted to define nowcasting as *"the process of predicting values for a time series that are not officially published for the current or nearest period"*.

2 Methodology

The present study endeavors to nowcast Bolivia's monthly economic growth through the application of machine learning algorithms, which will be trained using data from conventional sources of information, such as administrative records, as well as remote sensing data, such as nightlights satellite imagery and others.

The current research centers on the nowcasting of the Global Index of Economic Activity (IGAE, for its Spanish acronym), a metric that provides an estimation of monthly GDP growth through the combination of sectoral production indices.[1] The IGAE and its related measures are regarded as prompt and proximate gauges of overall economic activity. However, it is worth noting that the release of IGAE data is typically delayed by as much as six months from the present period. Indeed, The present research was completed on the 20th of April 2023; nevertheless, the IGAE —as well as GDP— published by the National Institute of Statistics is available only until September 2022. This implies a temporal gap of six months without a metric for Bolivia's aggregated economic activity.

Subsequent sections expound upon the methodological approach employed to accomplish the research objective in greater detail.

2.1 Identifying appropriate predictors for the IGAE

Generating and having access to data in Bolivia presents several challenges and limitations that impact the country's decision-making processes and development strategies.

[1] Greater detail about the IGAE in: https://www.ine.gob.bo/index.php/estadisticas-economicas/indice-global-de-actividad-economica-igae/

One of the main issues is the lack of availability and frequency of certain key indicators, such as GDP, which is not produced on a monthly basis and only has a limited geographical disaggregation. This hinders the ability to monitor economic activity and respond swiftly to changes and challenges. Furthermore, official statistical data is often subject to significant lags in publication, which can be problematic for timely decision-making. Additionally, data collection and analysis face various resource constraints and technical challenges, such as limited budgets, insufficient staff, and outdated methodologies. All these factors contribute to a situation where data often does not reflect the current reality, thus limiting the usefulness and accuracy of data-driven decision-making in Bolivia.

In light of the aforementioned context, the initial phase of the methodology for nowcasting the IGAE involves assembling a collection of variables that exhibit a robust correlation with the target variable, thereby serving as viable predictors of the IGAE. A noteworthy contribution of the current investigation is the construction of the set of features or predictors of the IGAE, which combines indicators from conventional sources of information with remote sensing data derived from satellite imagery.

2.1.1 Potential predictors from conventional sources

Given that the objective is to produce a nowcast of Bolivia's monthly economic growth, it is imperative that the predictors possess a monthly frequency. Additionally, the chosen predictors or features (X) must reflect the fundamental behaviors that drive the Bolivian economy. It is essential to incorporate production metrics from key sectors such as hydrocarbons, mining and others, as well as variables from the fiscal, monetary, and financial systems and trade indicators. Furthermore, it is worthwhile to consider other indicators that exhibit a statistically significant correlation with the IGAE.

As such, a comprehensive process has been undertaken to gather features possessing the aforementioned attributes. As a foundational step, the Table 1 presents a description of variables obtained from traditional sources, which are potential input features to be incorporated in the matrix X used for nowcasting the IGAE.

A total of 261 potential predictors were gathered for the prediction of the IGAE. However, studies related to prediction using machine learning algorithms indicate that the reduction of features and the presence of a positive and relatively strong degree of correlation between predictor variables and the predicted variable tend to enhance the performance of the prediction (Kotsiantis, 2011). Consequently, a correlation analysis was conducted to examine the relationship between the IGAE and the 261 variables collected, both for their contemporaneous values and lags ranging from the first to the twelfth, given the monthly frequency. Correlation analysis was conducted on a monthly growth rate basis because the target variable is the monthly growth rate of the IGAE.[2]

The analysis revealed that contemporaneously, electricity generation and its specific consumption by small and large industries, the production of liquefied petroleum gas and tin, rail transport, and construction permits exhibited a higher correlation with the IGAE. From an economic perspective, the selection of these variables is consistent with the significance of sectors such as hydrocarbons, mining, industry, transport, and construction in the Bolivian economy. Furthermore, several variables in foreign trade, such as mineral exports (total, zinc, and borates) and agricultural exports (total and chia), exhibited a strong and positive correlation with the contemporaneous growth of the IGAE.

[2]Predictions of the y-o-y growth rate and others transformations of the IGAE were evaluated as well, but the best performance was reached when working with monthly growth rates.

Increases in the importation of durable consumer goods, raw materials, and intermediate products for agriculture and industry were also associated with monthly and contemporaneous increases in the product. Imports of clothing, chemicals and chemical products, manufactured goods, rubber and plastic products, non-metallic mineral products, and food and beverages, classified according to the International Standard Industrial Classification (ISIC), would have a high and positive correlation with the monthly variation of the IGAE.

Table 1: Potential predictors from conventional sources

Sector	Predictors	Source
Production	The generation of electrical power, the manufacture of cement, and the extraction of oil, natural gas, and various minerals including tin, copper, lead, zinc, tungsten, silver, antimony, and gold.	National Institute of Statistics
Services	Electricity consumption across various sectors, including large and small industries, mining, residential, public urban illumination, and rural areas; water and transport consumption indices, as well as construction permits.	National Institute of Statistics
Fiscal	Domestic and customs tax revenues.	Ministry of Economy and Public Finance[1]
Monetary	Monetary aggregates, base and monetary emission, and interest rates.	Central Bank of Bolivia
Financial	Deposits and credits, and reserve requirements.	Central Bank of Bolivia[2]
Trade	Numerous disaggregations of both exports and imports.	National Institute of Statistics
Prices	Consumer Price Index (CPI) in its entirety, as well as its various categories, alongside commodity prices.	National Institute of Statistics

(1): Tax revenue time series are constructed with information from publications and other online sources of Ministry of Economy and Public Finance, National Tax Service, and National Customs.
(2): Financial data is also extracted from publications from the Financial System Supervisory Authority.

Moreover, certain lags of international commodity prices were found to have a positive relationship with the IGAE. For instance, the first lag of oil price exhibited a positive correlation, consistent with the importance of natural gas —its price is indexed to the price of oil— in Bolivia's exports. Gold prices showed a positive degree of association with the IGAE when considered with two lags.

2.1.2 Remote sensing data

Remote sensing is a technique that employs sensors to gather information about the earth's surface and atmosphere from a distance. This method has proven to be a valuable tool in various fields, including agriculture, forestry, geology, and environmental studies, among others. Remote sensing has also been used to generate socioeconomic indicators by processing satellite information, which provides detailed and accurate data about the distribution and characteristics of human settlements, infrastructure, and economic activities. This approach has demonstrated to be effective in monitoring urbanization, pop-

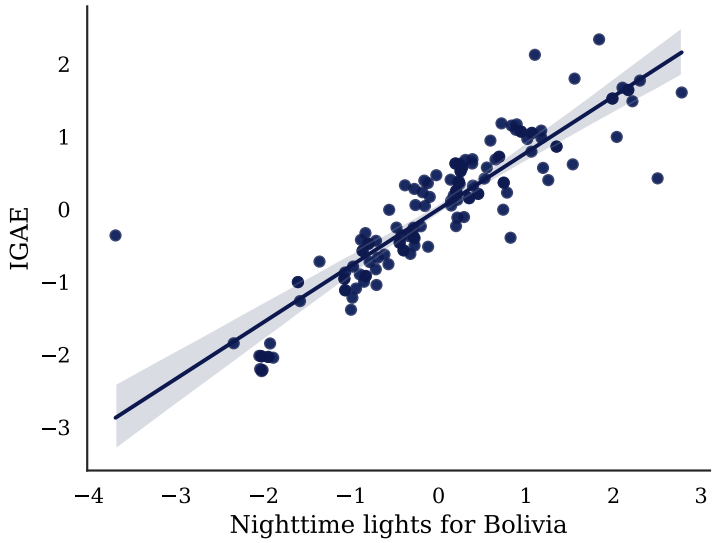
ulation growth, and economic development, among other aspects, enabling policymakers and researchers to make informed decisions and develop sustainable strategies. Therefore, the application of remote sensing techniques in generating socioeconomic indicators could become an essential instrument for understanding and managing the complex interactions between human activities and the environment.

Nighttime lights

Extensive research has shown a robust correlation between nighttime lights and measures of Gross Domestic Product (GDP) at various levels, including national, state, and regional (Chen & Nordhaus, 2011; Ghosh et al., 2010; Henderson et al., 2012), as well as at a more granular resolution. Therefore, nighttime light observations can serve as a reliable proxy for economic activity in areas where data are scarce, or statistical systems are of low quality, or where recent population or economic censuses are not available. Furthermore, economists can utilize changes in luminosity[3] intensity as an additional measure of income growth when other measures are not available.

Within this framework, one of the potential predictors of the IGAE constructed based on remote sensing techniques is nighttime lights captured by satellite images for Bolivia. To achieve this, daily metrics of visible and near-infrared (NIR) nocturnal lights are utilized, which are produced by NASA's Land Processes Distributed Active Archive Center (LP DAAC); more specifically, the image collection is the VIIRS Lunar Gap-Filled BRDF Nighttime Lights Daily L3 Global 500m. This collection encompasses the band *"DNB_BRDF_Corrected_NTL"*, which is an enhanced version of the average radiance band from the conventionally used Visible Infrared Imaging Radiometer Suite (VIIRS) Day/Night Band (DNB). The aforementioned product represents a daily 500-meter-pixel moonlight- and atmosphere-corrected nighttime lights product (Román et al., 2018).[4]

Figure 1: Scatter plot between the monthly growth rates of the IGAE and the median value of night lights in Bolivia



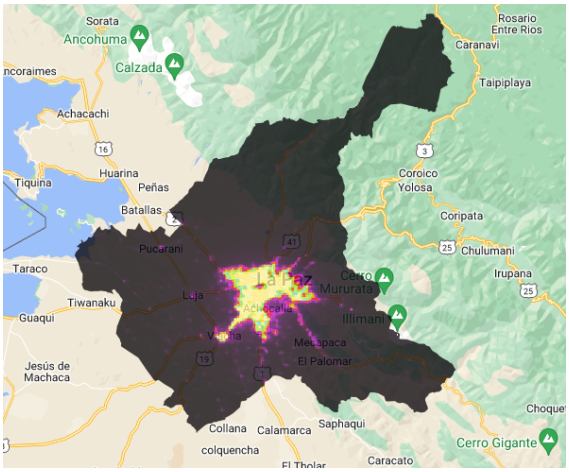
[3]Nighttime lights, light brightness, and luminosity are used interchangeably.

[4]In contrast to the US Defense Meteorological Satellite Program's (DMSP) night lights, this product does not exhibit blurring or overglow effects, nor does it suffer from oversaturation. These technical characteristics are highly relevant for the accurate measurement and mapping of nighttime lights, as they enable the acquisition of clear and reliable data that is free from distortions and inaccuracies.

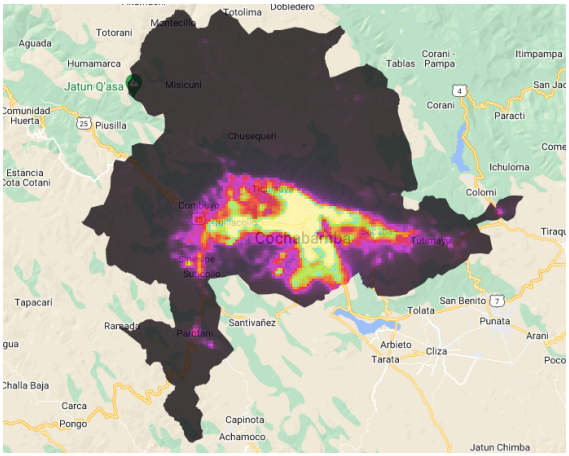
To execute the nowcast of the IGAE, inputs with a monthly frequency are required. Therefore, a time series is assembled, which comprises the monthly median values of brightness in pixels located within the territory of Bolivia. This newly constructed variable records its first observation in January 2012 when nighttime VIIRS images become available. Figure 1 displays a scatter plot —with both variables z-normalized— between the monthly growth rates of the IGAE and the median value of night lights in Bolivia. The scatter plot shows a relatively strong association between the two variables, with a correlation coefficient of 0.87.

Figure 2: Median luminosity intensity for the metropolitan areas of La Paz, Kanata, and Santa Cruz in the year 2022

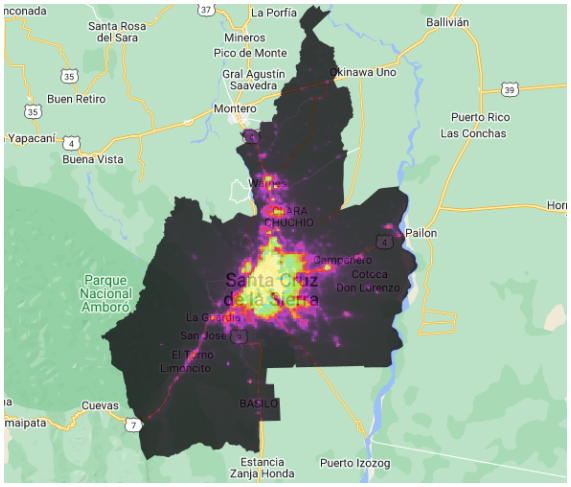
(a) Metropolitan area of La Paz



(b) Metropolitan area of Kanata



(c) Metropolitan area of Santa Cruz



Note: The intensity of luminosity in each pixel is defined by a color palette of black, purple, red, orange, green, and yellow. The lowest levels are represented by black, while the highest levels are represented by yellow.

Furthermore, to assess the suitability of regional time series of nighttime lights as inputs and as part of a sensitivity analysis, prediction models for the IGAE will be executed using variables of the monthly median brightness within the three major metropolitan regions of Bolivia: La Paz, Kanata, and Santa Cruz.[5] These regions are recognized as the

[5]The Metropolitan Area of the City of La Paz is the generic term used to refer to the regional unit composed of the municipalities of Nuestra Señora de La Paz, El Alto, Viacha, Achocalla, Palca, Laja,

primary population and economic centers of Bolivia. Figure 2 illustrates the brightness of nighttime lights in these regions.

Vegetation indices

Remote sensing vegetation indices are directly related to agricultural production, which, in turn, implies greater overall economic activity (Hu & Xia, 2019; Wiegand, Richardson, Escobar, & Gerbermann, 1991). Vegetation indices are mathematical formulations that quantify the status and health of vegetation using satellite data. By analyzing vegetation indices, researchers can estimate crop yields, identify potential areas of crop stress, and optimize crop management strategies. In Bolivia, agricultural production plays a vital role in the economy, providing food and income for millions of people.[6] As such, any increase in agricultural production typically results in an increase in overall economic activity, as more food and resources become available. Therefore, vegetation indices provide a valuable tool for monitoring and predicting agricultural production and its impact on economic activity.

Machine learning models can effectively employ information provided by vegetation indices to make accurate predictions of agricultural output and economic activity (Jung et al., 2021; Sharifi, 2021; Tang, Liu, & Matteson, 2022). By using vegetation indices as inputs, machine learning models can detect patterns and trends in the data, which can be used to develop reliable and precise predictions. In this study, the Normalized Difference Vegetation Index (NDVI) is utilized as input to generate predictors of the economic growth in Bolivia. The NDVI is widely used in vegetation studies due to its sensitivity to vegetation density. NDVI images are obtained from the Suomi National Polar-Orbiting Partnership (Suomi NPP) NASA Visible Infrared Imaging Radiometer Suite (VIIRS) Vegetation Indices (VNP13A1) Version 1 data.

Specifically, for each of the geographical zones spanning Bolivia's crop lands, including rice, sugar cane, soybean, and wheat, time series with monthly average values of the NDVI are calculated; results are fairly similar to those when the median NDVI is applied, but average values exhibit higher correlation with the IGAE.[7] These time series are then used as inputs to machine learning models to enhance the accuracy of Bolivia's monthly economic growth predictions. Figure 3 exemplifies how the NDVI maps soybean cultivation areas in the department of Santa Cruz.

Pucarani, and Mecapaca in the highlands of Bolivia; the 8 municipalities have a population of nearly 2 million inhabitants. The Kanata Metropolitan Region is composed of the municipalities of Cochabamba, Colcapirhua, Quillacollo, Sacaba, Sipe Sipe, Tiquipaya, and Vinto, totaling seven con-urbanized municipalities where approximately 1.5 million inhabitants coexist. The Metropolitan Area of Santa Cruz de la Sierra, currently the most populated urban center in the country (2.4 million inhabitants), is a con-urbanization of six municipalities in the department of Santa Cruz: Santa Cruz de la Sierra, La Guardia, Warnes, Cotoca, El Torno, and Porongo.

[6]Between 1990 and 2021, the economic activity of Agriculture represented an average of 13% of Bolivia's nominal GDP; data from the National Institute of Statistics.

[7]Although not all crop areas in Bolivia were included in the construction of variables with the average value of the Normalized Difference Vegetation Index (NDVI), the most representative ones in terms of being large-scale production crops with export destinations were included. Additionally, vector files with the georeferencing of these crop zones are publicly available.

Figure 3: Satellite and NDVI mapping soybean crop lands

(a) Satellite imagery

(b) NDVI

Note: Regarding the NDVI image, crops are highlighted in a yellowish color, while areas with abundant vegetation tend to be fluorescent green.

Built-up areas

Built-up classification involves identifying and mapping areas of land with high-density human development, such as buildings and infrastructure. These areas are strong proxies for economic activity, as they typically correspond to areas of high population density and economic productivity. Remote sensing techniques can provide valuable insights into the spatial and temporal dynamics of built-up areas (Y. Liu, Zuo, & Dong, 2021;

Zhou et al., 2021). By analyzing built-up classification data, researchers can estimate economic growth, identify areas of economic activity, and monitor urban expansion. This information is crucial for policymakers and city planners who seek to promote economic development and urban sustainability. Remote sensing built-up classification can also be used to assess the impact of economic activity on the environment and identify areas of concern, such as pollution and urban heat island effects. Overall, remote sensing built-up classification provides a valuable tool for understanding and predicting economic activity in urban areas, aiding policymakers and stakeholders in developing effective strategies for sustainable economic growth.

The Global Human Settlement Layer (GHSL) is a widely recognized mapping tool that offers global coverage of built-up areas; this high-resolution dataset was developed by the European Commission and provides detailed information on the spatial distribution of human settlements worldwide (Melchiorri et al., 2018). The GHSL is generated using satellite imagery, census data, and other data sources, and provides information on built-up areas, population density, and other urban characteristics. Unfortunately, the last update of the GHSL is available for 2015. This limitation can pose a challenge for researchers and policymakers who require more frequent updates on urban development. To address this issue, machine learning algorithms can be used to train models to accurately predict built-up areas with monthly frequency (Corbane et al., 2021; Kussul, Lavreniuk, Skakun, & Shelestov, 2017).

To predict economic growth, it is important to possess time series data on the changes in built-up areas in Bolivia. To address this objective, a Random Forest algorithm is trained using GHSL images to classify 1km pixels into built-up or non-built-up areas, using data from the Landsat-8 satellite's blue, green, red, near-infrared, and shortwave infrared 1 and 2 bands and pixel-level light brightness levels.[8] The RF model is chosen due to its effectiveness and widespread use across various domains and problems (Duro, Franklin, & Dubé, 2012).

The GHSL images classify areas according to their degree of urbanization, such as uninhabited, rural, and urban clusters of low and high density. However, as this study focuses on a binary classification of built-up areas, urban clusters of low and high density are defined as built-up areas for algorithm training, while uninhabited and rural areas are defined as non-built-up areas. The images used in the training stage are from the 2015, which is the latest update of GHSL. The pixels within the images are randomly separated into a training set (70%) and a validation set (30%).

To evaluate the classifier algorithm's performance, Accuracy (ACC) and the Matthews Correlation Coefficient (MCC) are used as metrics. The ACC measures the proportion of correct predictions made by the classifier, while the MCC measures the correlation between the predicted and actual labels, taking into account both true positive and true negative rates. The performance of the algorithm on the training sample is usually higher than that on the validation sample, because the algorithm has learned from the training sample and has optimized its parameters to fit the training data. Therefore, the ACC and MCC on the training sample are usually higher than those on the validation sample.

MCC is more appropriate for classifying built-up areas since most GHSL pixels in Bolivia are non-built-up areas; MCC outperforms ACC and other metrics in binary classification evaluation (Chicco & Jurman, 2020). The region of interest for classification encompasses zones around Bolivia's main urban areas, with a buffer extension to ensure

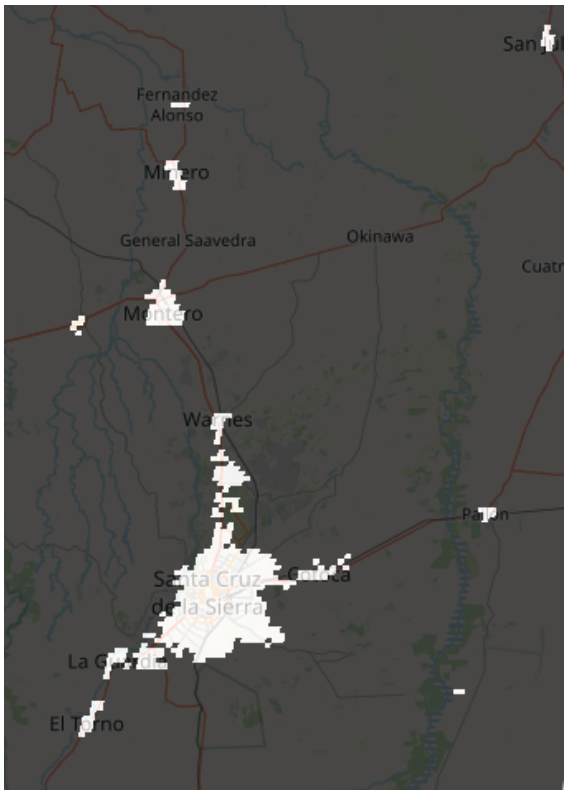
[8]For example, Jiang et al. (2020) detected the dynamics of urban growth in Africa using nighttime lights.

inclusion of representative built-up and non-built-up areas from different regions of the country and varied geographical features. The region of interest contains a total of 5722 1km pixels (i.e., 5722 observations; 4005 in the training set and 1717 in the validation set).

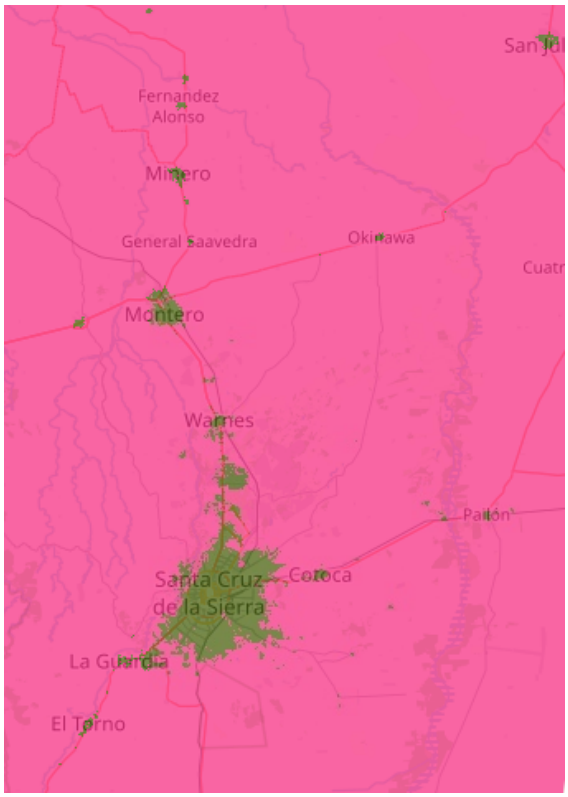
In the training sample, the values of ACC and MCC are considerably high, measuring 0.995 and 0.917, respectively; these results indicate an exceptional performance in terms of classification within this specific subset. Regarding performance on the validation sample, the ACC remains considerably high (0.987); however, the MCC is relatively lower (0.702) compared to that observed in the training sample. This signals that the algorithm may have overfitted to the training data and is not generalizing well to new data. Nonetheless, in general, an MCC score of 0.5 or above is considered a good performance, while a score of 0 indicates a completely random classifier. A comparison between GHSL and Random Forest classified built-up is presented in Figure 4.

Figure 4: Comparison between GHSL and Random Forest classified built-up areas

(a) Original GHSL



(b) Classified with trained RF model



Note: In the image on the left, built-up areas are identified with white color and non-built-up areas with black color. In the image on the right, built-up areas are identified with green color and non-built-up areas with pink color. Due to the spatial resolution of Landsat-8 and VIIRS night lights images, the built-up area classification image is less pixelated and, therefore, can provide more granular inputs regarding the identification of this type of area.

The algorithm that was trained using Landsat-8 and nighttime lights monthly imagery is utilized to predict built-up areas spanning Bolivia's boundaries at a 500-meter-pixel resolution with monthly frequency from April 2013 to January 2023 . The initiation of this period is marked by the availability of Landsat-8 images.

The primary objective is to introduce new inputs to forecast Bolivia's monthly overall economic activity. To attain this goal, the outputs of the built-up image classification are employed to generate monthly time series depicting the proportion of built-up pixels over

total pixels within regions of interest. Specifically, 4 built-up time series are created for Bolivia and the Metropolitan Areas of La Paz, Kanata, and Santa Cruz.

Precipitation

Precipitation, or the amount of water that falls from the atmosphere to the ground in the form of rain, snow, or hail, plays a critical role in the economic growth of regions across the world. Precipitation is essential for agricultural production, which remains a significant source of income and livelihood for many communities, particularly in developing countries such as Bolivia. Rainfall is also important for industrial processes, hydroelectric power generation, and transportation infrastructure. Therefore, the availability of sufficient precipitation is a key determinant of economic growth, particularly in water-dependent sectors.

Remote sensing technology has emerged as an alternative tool for calculating precipitation in a non-intrusive and accurate manner. Remote sensing uses sensors and other instruments mounted on satellites to measure various physical properties of the Earth's surface, such as temperature, reflectance, and moisture content. These measurements can be used to calculate rainfall, which is particularly useful in regions where traditional meteorological data collection is limited or non-existent. Remote sensing can also provide information on precipitation patterns, such as the timing and intensity of rainfall events, which can be used to optimize agricultural practices, manage water resources, and mitigate the impacts of natural disasters.

Hence, an additional metric that inputs processed satellite imagery is a time series featuring the median precipitation within Bolivia's rice, sugar cane, wheat, and soybean cultivation regions —the identical regions utilized in the NDVI computation—. These variables are derived from the Climate Hazards Group InfraRed Precipitation with Station data (CHIRPS), which furnishes precipitation measurements within pixels of 0.05° resolution.

2.2 Training, validation and test sets

After analyzing the relevance and correlation of potential predictors to the IGAE, as described in section 2.1, a list of 35 variables is defined that combines time series from both conventional sources and those constructed based on remote sensing techniques. Additionally, considering that the variable to be predicted is the monthly growth rate of the IGAE, which is characterized by a behavior influenced by time (e.g., seasonality), series with lags 1, 2, 3, 6, and 12 of the IGAE growth are added, as well as variables that signal the monthly and annual patterns of this variable based on sine and cosine functions. There are 44 variables that make up the feature matrix X, input for training and validating IGAE prediction algorithms (the list of these variables is in section 3.2).

Most of the variables from conventional sources have information available since 2010 or earlier. However, for variables generated based on satellite image information, the data availability, including its transformations into lags, is from April 2013. In this context, the total sample of features is for the period from April 2013 to January 2023.

As a result, a subset of features is defined as the training set, starting from April 2013 to December 2018. On the other hand, the validation set corresponds to information from January 2019 to September 2022. It is important to mention that a criterion for this period in the validation set is to analyze the forecast performance during the months of the COVID-19 pandemic, where atypical behavior is observed, and therefore to perceive

if the trained algorithms can respond under those conditions.

Since the IGAE is publicly available only until September 2022, the testing set corresponds to October 2022 to January 2023.[9] Ideally, forecasting the IGAE for months after January 2023 would be optimal, but most of the chosen features do not have information for those months. It is important to note that the nowcast proposed in this research is a conditional forecast; therefore, the forecast horizon is conditioned by the availability of information.

Finally, to guarantee comparability and prevent bias, all variables undergo z-score normalization. In particular, the features are adjusted based on the formula provided.

$$x_j^{(i)} = \frac{x_j^{(i)} - \mu_j}{\sigma_j} \tag{1}$$

Where j represents the selected variable or column in the matrix of features (X). μ_j and σ_j represent the mean and standard deviation, respectively, of the training set values for the selected variable j. The same rationale is applied to z-score normalization of y (i.e., IGAE growth rate).

2.3 Training machine learning algorithms to nowcast IGAE

Given the availability of training, validation, and testing sets, the next step is to train the algorithms for predicting the IGAE. This section provides methodological details about the selected algorithms to achieve the research objective. These algorithms are explained below, and it should be noted that their selection criteria are based on their common usage in the literature for time series prediction, as well as the fact that some of them specialize in modeling linear relationships, while others have more focused abilities in identifying non-linear patterns.

1. **Ridge Regression** is a linear regression technique that incorporates regularization by augmenting the least squares objective function with a penalty term based on the sum of squared weights (L2 penalty). The regularization parameter, a hyperparameter, governs the magnitude of the penalty term and influences the balance between model fit and model complexity. This approach is effective in preventing overfitting by reducing the coefficients' magnitude towards zero, thereby enhancing the model's stability and interpretability (Hoerl & Kennard, 1970).

2. The **Lasso algorithm**, short for least absolute shrinkage and selection operator, is a linear regression technique that employs regularization through an L1 penalty term instead of an L2 penalty. By promoting certain coefficients to be precisely zero, this penalty encourages sparsity and performs feature selection, thereby reducing the model's complexity. Lasso proves particularly beneficial in scenarios where a limited number of predictors exert a significant impact on the response variable, or when the number of predictors far exceeds the number of observations (Tibshirani, 1996).

3. The **ElasticNet Regression (EN)** technique is a hybrid of the Ridge and Lasso regression methods, incorporating both L1 and L2 penalties. This approach achieves

[9] IGAE time series is available monthly since January 1990.

a balance between the two methods by simultaneously promoting sparsity and coefficient shrinkage. This technique is especially valuable when dealing with data that contains a large number of features (Zou & Hastie, 2005).

4. The **Decision Tree Regressor (DT)** is a non-parametric regression algorithm that employs decision rules to recursively partition the predictor space into subregions. The algorithm aims to maximize the reduction in variance of the response variable. The resulting tree can be represented as a series of if-else statements that enable the mapping of a new observation to a predicted value. This visualization facilitates interpretation and analysis of the model. Decision Tree Regressor is particularly useful when the relationship between the predictors and response variable is complex and non-linear (Loh, 2011).

5. The **AdaBoost Regressor (ADA)** is an algorithm that utilizes a boosting technique to combine weak learners in a weighted sum to generate a robust learner capable of predicting the response variable for new observations. The algorithm places emphasis on the data points that were misclassified by the prior weak learner at each iteration, assigning them higher weights to compel the next weak learner to pay more attention to them. AdaBoost Regressor is particularly useful in situations where the data is noisy or complex and where there are interactions and non-additive effects (Freund & Schapire, 1997).

6. The **Gradient Boosting Regressor (GBR)** is a boosting algorithm that constructs an additive model of weak learners in a step-by-step manner. In each step, the algorithm fits a weak learner to the negative gradient of the loss function, relative to the current model's prediction, thereby modifying the residual errors of the prior model. The Gradient Boosting Regressor can be particularly advantageous when the data is complex or noisy, and when there are interactions and non-additive effects to consider (Friedman, 2001).

7. The **Random Forest Regression (RF)** algorithm is a popular machine learning technique that enhances prediction accuracy and robustness by amalgamating multiple decision trees. The algorithm operates by generating a multitude of decision trees, each trained on a random subset of the data and with a restricted number of features. The predictions of all trees are subsequently merged to yield a final outcome. Random Forest Regression is especially advantageous when dealing with intricate, non-linear relationships between variables and when preventing overfitting. This method is extensively employed in diverse domains, such as finance, healthcare, and environmental science (Breiman, 2001).

8. The **Extra Trees Regressor (ET)** is an ensemble learning technique that involves constructing numerous decision trees using random subsets of both data and features. However, this method incorporates additional randomness during the tree building process, where the splitting thresholds for each feature are selected randomly rather than being based on information gain or Gini impurity. This added randomness serves to mitigate overfitting and enhance generalization performance (Geurts, Ernst, & Wehenkel, 2006).

Additional information regarding the methodological characteristics and execution of the aforementioned algorithms can be found in Appendix A.

Furthermore, the process of hyperparameter tuning is a crucial step in attaining optimal performance for a machine learning model. It involves the selection of values that control the algorithm's behavior, known as hyperparameters. The procedure entails the identification of a range of values for these hyperparameters, followed by an iterative selection of a subset of them to optimize the model's performance.

As a result, the algorithms mentioned earlier undergo a rigorous process of hyperparameter tuning, primarily based on a time-aware cross-validation method. This methodology guarantees that the validation set always follows the training set in time, preventing the model from receiving information from the future during training and generating excessively optimistic outcomes. The following paragraphs detail the applied methodology for hyperparameter tuning for each algorithm.

1. **Ridge** regression is a powerful tool for balancing the fit of a model to training data with the need to keep model coefficients small in order to avoid overfitting. To achieve this balance, the regularization strength parameter λ must be carefully tuned. One common approach to tuning this hyperparameter is to use k-fold cross-validation to evaluate the performance of the model over a range of λ values. To search for an appropriate range of λ values, it is often useful to explore a wide range of values, from very small to very large, spanning several orders of magnitude. In this study, a 5-fold cross-validation process —recall it is a time-aware cross-validation— was conducted with an array of 1000 λ values, evenly spaced on a logarithmic scale from 10^{-5} to 10^{5}, as the impact of λ on the model tends to be multiplicative. The optimal value of λ was then selected based on the one that minimized the Mean Square Error (MSE).

2. The **Lasso** algorithm shares identical hyperparameters with Ridge regression. The method for fine-tuning the hyperparameters resembles that of Ridge. Specifically, a 5-fold cross-validation technique is employed to assess the model's efficacy across a spectrum of 1000 λ values, evenly distributed on a logarithmic scale ranging from 10^{-5} to 10^{5}. The optimal value is determined by identifying the one that minimizes the MSE.

3. The **ElasticNet** regression features two hyperparameters that play a critical role in the regularization process. The first hyperparameter, denoted as λ, governs the strength of the regularization, while the second hyperparameter, denoted as α, controls the mixing parameter between L1 and L2 regularization. In order to optimize these hyperparameters, a rigorous 5-fold cross-validation is performed using a range of 1000 λ values, similar to the approach taken in Ridge and Lasso regression. Additionally, values for α ranging from 0.05 to 0.95 in 0.01 increments are considered to ensure optimal model performance. The optimal values are selected based on the ones that minimize the MSE.

4. The **Decision Tree Regressor** hyperparameters include the maximum depth of the decision tree (d), the minimum number of samples required to split an internal node (mss), and the minimum number of samples required to be at a leaf node (msl). In this study, a 5-fold cross-validation was executed over a range of values for each of the hyperparameters, which was selected based on prior research and machine learning best practices (Mantovani et al., 2018).

5. The **AdaBoost Regressor** hyperparameters encompass a variety of factors, including the maximum depth of the tree (d), the maximum number of estimators utilized during boosting (T), the learning rate (α_t), and the loss function employed to update weights after each boosting iteration. The selection of hyperparameter ranges for tuning is contingent upon several considerations, such as model complexity and available training resources. To determine the optimal value for α_t, a broad range of values is typically assessed to strike a balance between convergence speed and overfitting risk; a starting range of 0.01 to 3 with an increment of 0.01 was used for tuning. For T the range is chosen based on the trade-off between model complexity and performance, with a large number of estimators potentially leading to overfitting and a small number resulting in underfitting; thus, a range of 50 to 200 with an increment of 1 was tested. Additionally, linear, square, and exponential loss functions were evaluated. d was evaluated in a range from 3 to 10. Optimal values are selected based on minimizing the MSE through 5-fold cross-validation.

6. The hyperparameters utilized in **Gradient Boosting Regressor** are the learning rate (γ_m), maximum depth of the tree (d), the maximum number of estimators at which boosting is terminated (T), and the minimum number of samples required to split an internal node (mss). To determine the optimal values for each of these hyperparameters, 5-fold cross-validation is performed over a range of values. For T hyperparameter, a range of 100 to 500 with a step of 5 is chosen; this range is deemed reasonable as it encompasses a wide range of values and includes some values that have been known to work well in practice. Similarly, for the d hyperparameter, a range of 3 to 9 with a step of 1 is chosen, because it covers a range of depths that have been known to work well in practice, while exceeding a depth of 9 can result in overfitting. In the context of the mss hyperparameter, it is worth noting that the range of admissible values spans from 2 to 20, with a step size of 1; a choice of 2 would result in a greater number of splits and a heightened risk of overfitting, whereas a higher value of 20 would lead to fewer splits and a potential for underfitting. For the learning rate hyperparameter, a range of 0.01 to 0.1 with a step of 0.01 is chosen. Additionally, Stochastic Gradient Boosting is tested because it can help reduce overfitting by introducing randomness into the training process; the parameter in sklearn library to assess this option is *subsample*. The optimal values are selected based on the ones that minimize the MSE.

7. The hyperparameters chosen for tuning in the **Random Forest Regression** model were selected based on prior knowledge and common practice in the literature. The number of estimators (T) was tested in the range of 100 to 300 with increments of 5, as a larger number of trees can improve the model's performance, but at a certain point, the improvement may be negligible. The criterion for splitting nodes ($criterion$) was set to test both the mean squared error and the mean absolute error, as both are common loss functions in regression problems. The minimum number of samples required to split an internal node (mss) was tested in the range of 2 to 10 with increments of 1; this range was selected to avoid overfitting, as having too few samples required to split a node may lead to over-complex trees, while having too many may lead to underfitting. A 5-fold cross-validation process is conducted to identify the optimal hyperparameters that minimize the MSE.

8. For tuning hyperparameters in the **Extra Trees Regressor**, a wide range of values

were evaluated to identify the combination that yields the best performance. For instance, a range of values from 100 to 500, with a step size of 5 were evaluated for the number of trees in the forest (T). Besides tuning other common parameters such as d, mss, and msl, the algorithm was executed allowing the option to whether or not bootstrap samples are used when building trees. If bootstrap is enable, there is a parameter to be tuned which is the number of samples to draw from X to train each base estimator ($MaxSample$); a range of values from 0.1 to 1.0, with a step size of 0.01, are tested. The final hyperparameters are selected based on the ones that minimize the MSE over a 5-fold cross-validation strategy.

Finally, after training the algorithms, the mean squared error is calculated for the validation set —the root mean square error and the mean absolute error are calculated as well—. Based on these results, the algorithms with the best performance in predicting the IGAE are identified. The prediction of the algorithm with the best performance or a combination (e.g. average) of the forecasts of a select group of algorithms with the highest predictive power are used to nowcast monthly growth rates of the IGAE between October 2022 and January 2023, which corresponds to the testing set (see section 3.3).

3 Results

3.1 Fine-tuning hyperparameters

The Table 2 showcases the performance of the selected machine learning algorithms in predicting the monthly IGAE. The first column depicts the name of each algorithm, while the second to fourth columns report the mean squared error (MSE), root mean squared error (RMSE), and mean absolute error (MAE) acquired by the algorithm without tuning its hyperparameters. The fifth to eighth columns display the same metrics obtained after hyperparameter tuning through a 5-fold cross-validation strategy. MSE and RMSE are commonly used evaluation metrics for regression problems, preferred when larger errors need to be penalized more severely. On the other hand, MAE is a metric that treats all errors equally, regardless of their magnitude.

Upon examining the results, it becomes evident that all algorithms' performance improved after hyperparameter tuning. This implies that the initial hyperparameters selected for each algorithm were suboptimal, and fine-tuning was crucial to enhancing the models' performance. It is important to note that the parameters in non-tuned models are default values utilized in the sckit-learn library. Cross-validation during the tuning process prevented overfitting and improved the models' generalization ability.

Prior to hyperparameter tuning, both the Lasso and ElasticNet algorithms displayed the highest MSE and RMSE values, while Ridge regression exhibited superior performance. However, after fine-tuning of hyperparameters, Ridge regression was found to significantly outperform the other algorithms in predicting the IGAE in the validation set, displaying a reduction in both metrics. Additionally, ElasticNet showed a substantial reduction in MSE and RMSE values, primarily due to its L2 regularization which corresponds to Ridge regression.[10]

Of note, the Gradient Boosting algorithm achieved the lowest MAE and relatively low MSE and RMSE values among all algorithms, indicating its potential suitability for IGAE

[10]In Appendix B, R^2 scores are presented for all the evaluated algorithms as an additional metric to analyze prediction capacity.

Table 2: Fine-tuned hyperparameters by algorithm

Algorithm	Without tuning			After tuning (5-fold cross-validation)			
	MSE	RMSE	MAE	MSE	RMSE	MAE	hyperparameters
Ridge	0.317	0.563	0.350	0.294	0.542	0.354	$\lambda = 14.508$
Lasso	1.687	1.299	1.034	0.432	0.657	0.383	$\lambda = 0.011$
ElasticNet	0.987	0.993	0.733	0.351	0.592	0.377	$\lambda = 0.087$ $\alpha = 0.11$
Decision Tree	0.721	0.849	0.403	0.679	0.824	0.393	$d = 4$
AdaBoost	0.557	0.746	0.371	0.539	0.734	0.323	$d = 5$ $\alpha_t = 2.1$ $Loss = Exponential$ $T = 60$
Gradient Boosting	0.505	0.711	0.304	0.389	0.624	0.283	$d = 7$ $T = 135$ $mss = 10$ $Subsample = 0.4$
Random Forest	0.560	0.749	0.368	0.540	0.735	0.362	$mss = 3$ $T = 215$
Extra Trees	0.501	0.707	0.311	0.487	0.698	0.357	$d = 13$ $MaxSamples = 0.98$

Note: In the process of training and evaluating the algorithms, the scikit-learn library was employed. It should be noted that in comparison to Section 2.3, certain hyperparameters are not included in the table above. However, during the fine-tuning process, it was determined that the default values in the scikit-learn functions were the most suitable, hence, these default values have been omitted from the table.

nowcasting.

Ultimately, the Ridge algorithm was determined to be the best-performing algorithm, both before and after hyperparameter tuning. This finding underscores the potential of Ridge regression as the most appropriate algorithm for IGAE nowcasting in Bolivia —conditioned upon the data used in this study—, particularly in cases where a linear relationship exists between variables. Notably, Ridge regression does not exclude any predictors when making predictions, unlike Lasso which sets some coefficients to zero. This may be one reason for Ridge regression's superior performance in predicting IGAE, as evidenced by the strong linear correlation between the IGAE variable and its predictors.

It is important to note, however, that non-linear relationship models such as Decision Trees and Tree ensembles may be better suited for capturing complex and intricate patterns in the data. While these models may be advantageous when non-linear relationships exist between variables, their effectiveness may be diminished in cases where the relationship between variables is predominantly linear, as is the case with IGAE prediction.

3.2 Feature importance

So far, the algorithms with the highest capacity to predict the growth of the IGAE have been identified. However, it is important to know the importance of each of the predictors (features) with respect to their contribution or weight in the trained algorithms. Thus, Table 3 shows the weights that the predictors receive in each of the trained algorithms to predict the IGAE; originally, these values are reported on a scale of 0 to 1 (-1 to 1 in Ridge, Lasso and ElasticNet), but for visualization purposes, they were multiplied by 100.

In the cases of Ridge, Lasso, and ElasticNet, the values reported in the table are the coefficients of the resulting fitted regression model from the training. The value of a coefficient indicates the magnitude and direction of the influence of that feature on the target variable; positive coefficients indicate a direct relationship, while negative coefficients indicate an inverse relationship. For the rest of the algorithms, the values in the table show the relative importance of each feature, so the sum of the individual weights equals 100.

First, it is observed that the lag variables of the growth of the IGAE and the variables of signaling temporal patterns are very important for prediction; for example, the 12-lag of the target variable is the most important in comparison to the individual weights of the rest of the input variables.

Second, it is worth noting that the variable of the median value of nighttime lights for Bolivia —built with remote sensing techniques— would position itself as the second most important among the 44 included features. Furthermore, variations in luminosity would be the most important predictor among the 35 available features that do not include temporal variables. These results not only demonstrate the validity of the inclusion of this variable to predict the percentage change of the IGAE, but also that it would be a fundamental and innovative input for forecasting economic activity in Bolivia, with potential for granular analysis (e.g., regional, municipality, community, or other).

If we continue to analyze the importance of variables generated with remote sensing as inputs to predict the growth of the IGAE with machine learning algorithms, the 6-lag of the average NDVI value in sugarcane crops in the department of Santa Cruz would position itself as the fourth or fifth most important variable —excluding time variables— in Ridge and ElasticNet algorithms, which are among the most predictive.

Sugarcane production in Bolivia is an economically important activity for the country, especially for the region of Santa Cruz, due to its capacity to generate income and employment, as well as its potential for the production of ethanol, sugar, and other by-products. In terms of production, sugarcane is the third most important crop in Bolivia after soybeans and rice; Santa Cruz is the main sugarcane-producing region in Bolivia, representing more than 70% of national production.[11] It should be mentioned that the 6-lag of the average NDVI value in these crops was studied because it showed the highest correlation with the IGAE, which is also consistent with the agricultural cycle of this product whose planting and harvesting are separated by a period of almost a year, and in the process, the sugarcane plants reach their maturity stage, accumulating the highest amount of sugars, resulting in higher vegetation quality.

In a similar vein, though with lower relative importance, the lag 9 of the average value of NDVI in wheat crops in the Santa Cruz region is another predictor to be taken into consideration. In Ridge regression, it would be ranked 12th out of 35 input variables, not including time variables. On the other hand, the lag 9 of the average value of NDVI in soybean crops in the Santa Cruz region, a significant export product, also exhibits a positive relationship with Bolivia's economic growth; however, its contribution to prediction is comparatively lower. In summary, variables constructed with NDVI data are also significant inputs for forecasting the growth of IGAE.

Another indicator constructed using remote sensing is the proportion of built-up areas in Bolivia. For the Ridge algorithm, the importance of this variable would be ranked 13th out of 35 input variables, without including time variables. Moreover, in other algorithms such as Decision Tree, Random Forest, and AdaBoost, its importance is higher, ranking it in the 2nd, 3rd, and 6th places among 35 variables, respectively. Once again, concerning the variable with built-up area data —classified with a Random Forest model and satellite image information—, there is evidence of its relevance as a predictor of the growth of IGAE.

The final case of the variables generated from satellite image information consists of the time series of precipitation. In the Ridge regression, the lags of the median precipitation in sugarcane and soybean crops in Santa Cruz, as well as in crops in the rural area of Cochabamba, would have a negative relationship, of significant magnitude, with contemporaneous variations in IGAE. In all of these areas, the crops have a significant degree of irrigation, and thus greater amounts of precipitation could potentially have adverse effects on agricultural production, resulting in reduced crop yields.

In terms of conventional information features, electricity consumption indices (for small and large industry) and railway transportation would be the main predictors in this category. Additionally, tin production and aggregate electricity generation variables would constitute predictors of high importance. Furthermore, variations in the value of imports of manufactured products such as clothing, plastics, and durable consumer goods prove to be useful predictors for predicting the growth of IGAE within the framework of the methodology employed in this research. It is worth mentioning that these variables are used in practice for the calculation of IGAE by the National Institute of Statistics of Bolivia; however, despite having access to this data, the institution still maintains a prolonged lag in the publication of the monthly indicator of aggregate economic activity.

[11] It is noteworthy that in the sugarcane cultivation areas in Santa Cruz, there is also a crop rotation system in which, depending on market demands and prices, soybean or other agro-industrial products are cultivated.

Table 3: Feature importance by algorithm

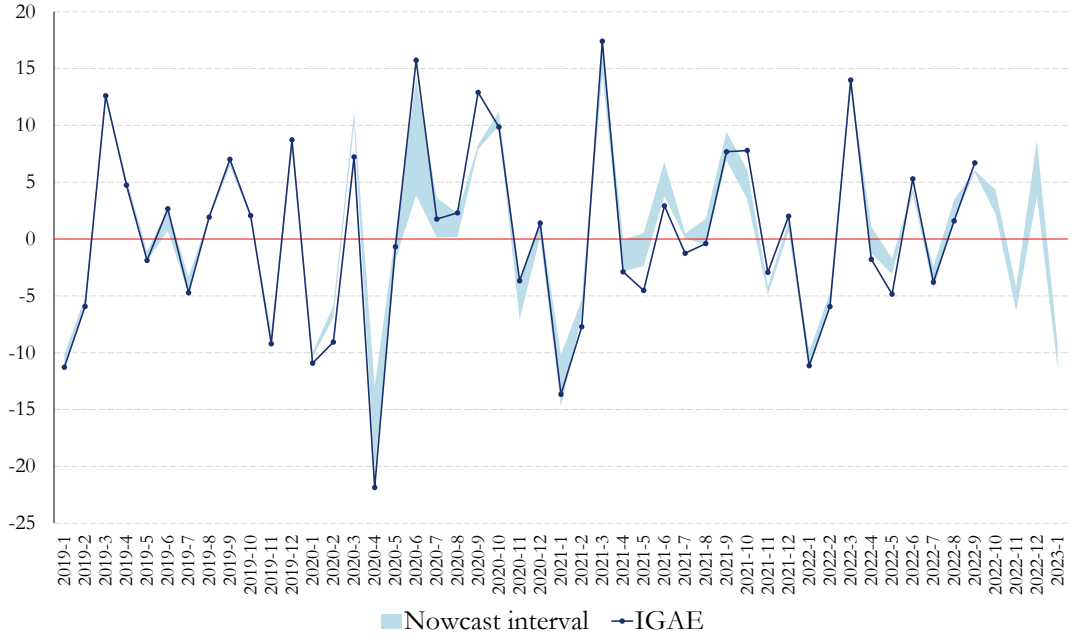
Feature	Ridge	Lasso	EN	DT	ADA	GBR	RF	ET
IGAE (lag 12)	26.70	63.50	36.60	65.30	46.30	43.00	60.00	39.90
Bolivia: Median Nighttime Lights	22.70	23.60	26.30	0.00	29.20	25.90	13.90	27.30
IGAE (lag 6)	17.80	5.10	14.40	0.00	0.40	0.60	0.10	3.50
Sine monthly-time signaling variable	10.00	1.90	8.30	0.00	0.50	1.10	0.90	4.10
Small industries: Electricity consumption	8.90	0.90	7.60	0.00	0.70	1.80	1.10	1.80
IGAE (lag 9)	8.10	4.30	6.00	0.00	0.50	1.80	0.40	0.50
Large industries: Electricity consumption	7.30	5.40	6.60	0.00	0.10	0.10	0.10	0.20
Rail transport consumption index	7.10	0.10	5.10	0.00	0.40	1.00	0.00	0.40
Santa Cruz: Average NDVI in sugar cane crops (lag 6)	6.50	2.40	5.80	0.00	0.10	0.30	0.20	0.50
Imports: Manufacture of clothing	5.20	1.30	4.50	0.00	1.70	0.90	4.30	1.20
Imports: Manufacture of rubber and plastic products	5.20	10.10	4.90	0.00	0.10	0.10	0.10	0.30
Imports: Durable Consumer Goods	4.10	0.00	1.10	0.70	0.20	0.20	0.20	0.20
Production: Tin	4.00	3.20	3.00	0.70	0.40	0.70	0.20	0.10
Generation of electrical power	4.00	0.00	2.40	12.60	1.30	5.00	4.70	2.10
Commodity price: Gold (lag 2)	3.40	0.00	1.70	0.00	0.20	0.20	0.10	0.10
Santa Cruz: Median NDVI in wheat crops (lag 9)	3.20	0.00	0.30	0.00	0.20	0.30	0.20	0.40
Bolivia: Share of built-up pixels	2.50	0.00	0.40	5.10	1.00	0.20	1.40	0.30
Exports: Wood and wood products	2.40	2.30	2.80	0.00	0.40	1.90	0.30	0.20
Exports: Agriculture, livestock, forestry and fishing	2.30	2.60	2.00	0.00	0.10	0.80	0.10	0.30
Cosine yearly-time signaling variable	1.90	0.00	0.20	0.00	0.20	0.10	0.10	0.20
Production: Liquefied petroleum gas	1.90	0.00	0.00	0.00	0.30	0.80	0.10	0.50
Imports: Raw and intermediate products for agriculture	1.70	0.00	0.00	0.00	0.00	0.60	0.10	0.40
Export: Chia	1.70	0.00	2.30	0.00	1.50	1.60	1.20	0.50
Imports: Manufacturing industries	1.70	0.00	0.00	0.00	0.10	0.10	0.10	0.20
Imports: Suitcases, handbags, footwear, etc.	1.40	0.00	0.00	0.00	0.80	0.90	0.40	0.40
Export: Mineral extraction	0.80	0.00	0.00	0.00	0.30	0.10	0.10	0.20
Exports: Zinc ore	0.80	0.20	0.00	0.00	1.40	0.00	0.10	0.30
Commodity price: Oil (lag 1)	0.50	0.20	0.00	0.00	0.10	0.00	0.00	0.20
Cosine monthly-time signaling variable	0.50	0.00	0.00	0.70	1.00	0.10	1.00	0.40
Santa Cruz: Median NDVI in soybean crops (lag 9)	0.50	0.00	0.00	0.00	0.30	0.30	0.10	0.20
Imports: Non-metallic mineral manufactured products	0.50	3.70	0.00	0.00	0.10	0.90	0.20	0.20
Construction permits	0.40	0.00	0.00	0.00	0.20	0.10	0.10	0.10
Sine yearly-time signaling variable	-1.10	0.00	0.00	0.00	1.20	3.20	1.70	0.60
IGAE (lag 3)	-1.50	0.00	0.00	0.00	0.20	0.20	0.10	0.50
Rural Cochabamba: Median precipitation (lag 2)	-1.60	0.00	-0.70	0.00	0.50	0.30	0.40	0.70
IGAE (lag 1)	-1.80	0.00	0.00	0.60	2.40	2.00	1.80	0.90
Imports: Manufacture of food and beverage products	-2.30	-2.90	0.00	0.00	0.10	0.20	0.00	0.10
Imports; Chemical substances and products	-2.40	0.00	0.00	0.00	0.00	0.50	0.00	0.10
Rural Cochabamba: Median precipitation (lag 1)	-2.70	-2.80	-2.60	0.00	0.40	0.10	0.00	0.30
Generation of electrical power (lag 12)	-3.00	-6.50	-2.00	0.00	0.40	0.50	0.30	1.90
Imports: Raw and intermediate products for industry	-3.30	0.00	0.00	0.00	0.10	0.50	0.00	0.30
SantaCruz soybean crops: Median precipitation (lag 2)	-3.50	-1.60	-3.20	0.00	0.30	0.10	0.20	0.10
Tarija sugar cane crops: Median precipitation (lag 3)	-4.40	0.00	-1.00	0.00	1.20	0.20	0.60	1.70
IGAE (lag 2)	-6.30	-3.50	-4.30	14.30	3.20	0.50	2.60	5.50

3.3 Nowcasting of the Bolivian monthly economic activity

So far, the focus was on identifying the algorithms with the highest capacity to predict the monthly growth of the IGAE in Bolivia, based on the training and validation sets. Since these algorithms were identified, it is necessary to establish criteria to generate an aggregated metric for the monthly nowcasting of Bolivia's economic activity. The algorithms with the lowest mean squared error in the validation set were Ridge, ElasticNet (EN), and Gradient Boosting Regressor (GBR), and their predictions were also among those with the lowest mean absolute error values.

To this end, the study propose, first, the development of a nowcast interval, which shows the range of values predicted by the Ridge, EN, and GBR algorithms. In other words, the upper and lower limits of the interval are defined based on the maximum and minimum values of the forecasts generated by these three algorithms. Figure 5 overlays the series of monthly IGAE growth with the nowcast interval for the period analyzed in the validation and test sets.

Figure 5: Observed and nowcast interval of the IGAE monthly growth rate



Note: The interval for nowcasting is delineated by the range between the upper and lower bounds of the most optimal predictions made by the Ridge, ElasticNet, and Gradient Boosting Regressor algorithms.

It is noteworthy that, in most months of this analysis period, the nowcast interval includes the observed growth of the IGAE within its range. This result adds validity and confidence to the assumption that the performance of the Bolivian economy is close to what is proposed by the nowcasting approach of this study. Additionally, the seasonal behavior and direction of the sign in the variation of the IGAE are accurately captured by the nowcast interval. Furthermore, it can be observed that even in atypical moments, such as the outbreak of the COVID-19 pandemic, the nowcast interval is in the vicinity or coincides with the observed performance of the IGAE growth.

On the other hand, the growth of the IGAE shows an exceptionally marked seasonal behavior with respect to the direction —sign— of the percentage variation of this indicator between months. In Appendix C, this seasonal behavior is exemplified in box plots with the distribution of economic growth in each month since 1990. Consequently, in pre-pandemic and post-pandemic months when the Bolivian economy did not face atypical conditions and instead performed in line with historical seasonal patterns, the performance of the nowcast is even superior (i.e. a narrower nowcast interval that is closer to the observed values).

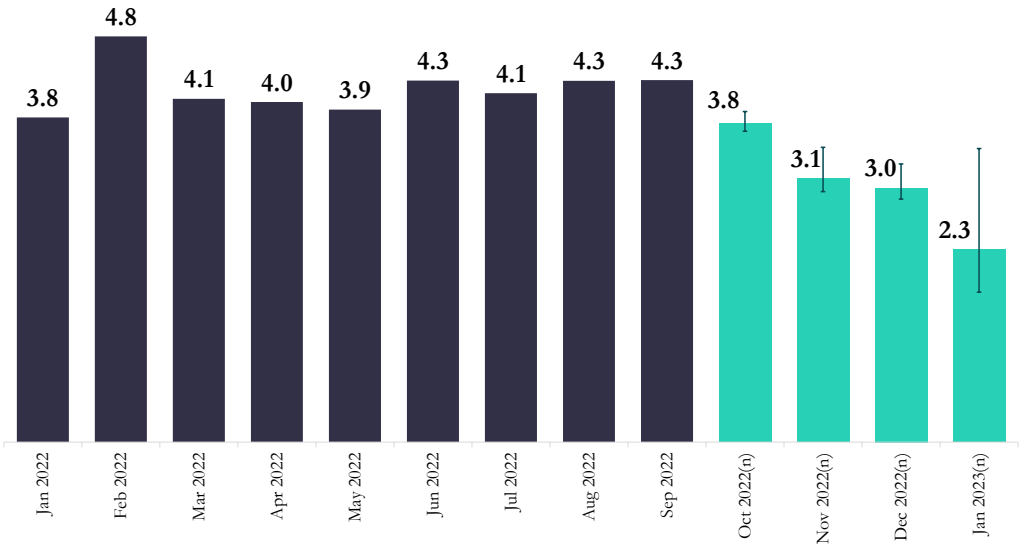
Certainly, while the nowcast interval serves as a useful indicator for the purposes outlined in the preceding paragraphs, in many instances, policymakers, researchers, and economic analysts require a specific number regarding economic growth estimates. For this reason, the IGAE nowcast ($\bar{\hat{y}}$) is defined as the weighted geometric mean of the

predictions generated by the Ridge, EN, and GBR algorithms, with the weights (α_i) determined as the inverse of the mean squared errors of each algorithm i.[12]

$$\bar{\hat{y}} = \left(\prod_{i=1}^{3} \hat{y}_i^{\alpha_i} \right)^{\frac{1}{\sum_i \alpha_i}} \tag{2}$$

Although the algorithms were trained to forecast the monthly growth rates of the IGAE, the Figure 6 illustrates the series of accumulated growth in the Bolivian economy from January 2022 to January 2023, with the last four months corresponding to the nowcast. This presentation of information was chosen as it contributes to the interpretation and analysis of Bolivia's economic situation during these months, as explained in the following paragraphs.

Figure 6: Cumulative growth rate of the IGAE, observed and nowcast



(n): Nowcasted.
Note: During the period from October 2022 to January 2023, the nowcast interval was included alongside the cumulative economic growth nowcast.

Historically, the last quarter has a significant impact on the annual growth of the Bolivian economy (see Appendix C). This is partly because the majority of public resources, particularly public investment, are executed during those months. In the Bolivian economy, public investment is a fundamental determinant of economic growth. Further, the seasonal effect of year-end festivities generates additional dynamism in private economic activities. However, the nowcast for the last quarter of 2022 shows a slowdown of the Bolivian economy.

Conveniently, the coincidence of conducting the nowcast for these months is of great utility to test the predictive capacity of the methodology assumed in this study, given that during that period, atypical and adverse events occurred for the Bolivian economy. Specifically, from October 22, 2022, for 36 consecutive days, a civic strike was carried out in the city of Santa Cruz de la Sierra and other neighboring municipalities, in which

[12] I used the geometric mean instead of the arithmetic mean because is less sensitive to outliers.

an economic activity stoppage was applied due to the request for the realization of the population and housing census for 2023.[13]

Regarding the adverse effects of the civic strike on the economy of the Santa Cruz region and the country, the stoppage of activities generated significant economic losses, especially in sectors such as commerce and transportation.[14] The department of Santa Cruz represents approximately 30% of Bolivia's GDP and constitutes the most dynamic region of the Bolivian economy. This region specializes in agricultural and agroindustry production —with an export focus—, about 35% of its GDP is explained by these two economic activities.[15]

In turn, the impact was not only on Santa Cruz, but also on other economic centers in Bolivia. For example, poultry production in Cochabamba was affected by the shortage of balanced feed produced in Santa Cruz. Other regions of the country were affected by price increases in food products, given that Santa Cruz is one of the main suppliers of these products for consumption by the population throughout the territory of Bolivia. Additionally, exports were limited by mobility restrictions on trunk roads.

While Bolivia had recorded a cumulative economic growth of 4.3% until September 2022, under the restrictive conditions caused by social conflicts, the Bolivian economy decelerated in October (3.8%) and November 2022 (3.1%), according to the IGAE nowcast. This would imply that the year-on-year variation of the IGAE in the months of October and November is negative, with contractions of -0.5% and -2.6%, respectively.

If we examine the last month of 2022 in light of the results of the IGAE nowcast, it can be infered that the Bolivian economy may conclude that period with a growth rate of approximately 3%. In terms of year-on-year growth, the month of December would experience a recovery of economic activity, in the sense that the production of that month would be approximately 2% higher than that recorded in 2021.

Nonetheless, after the contractions registered in October and November of 2022, it is reasonable to expect that economic recovery will be progressive for the Bolivian economy. One potential reason for this difficulty is that the halt in activities would have negatively impacted the country's productive supply, especially in the Santa Cruz region, where important economic sectors such as agriculture, livestock, and industry are located. The shortage of labor and the cessation of operations in companies could have led to a reduction in the production of goods and services, which in turn would have affected the country's ability to meet domestic and external demand. On the other hand, the loss of income for companies and workers affected by the conflicts would also have decreased demand in the country; the lower purchasing power of workers would have reduced their consumption of goods and services, which in turn would have affected the companies that depend on domestic demand.

Finally, based on the January 2023 nowcast indicator, the Bolivian economy is estimated to have grown by 2.3% compared to the same period in 2021. This outcome suggests that the Bolivian economy is sustaining its product expansion, even at a slightly higher rate than the estimated year-on-year growth in December 2022. Notably, the month of January witnessed adverse shocks, albeit with a lesser effect than the events in October

[13]Civic representatives of that region complained that the last time a census was carried out was in 2012 and that since then, a complete update of the data had not been made. This lack of updated information had important implications for decision-making in public policy, as well as for planning and resource allocation in the region.

[14]Statements from government and private sector representatives indicate a loss ranging from USD 500 to USD 1000 million, which in percentage of GDP would represent 1.25% and 2.5%, respectively.

[15]Data for Santa Cruz was obtained from the National Institute of Statistics.

and November. The social conflicts in Peru restricted the flow of goods between Bolivia and Peru, which is a significant factor for the exports and imports of Bolivia.

Furthermore, the nowcast range for January 2023 shows greater variability in the prediction, ranging from a minimum growth of 1.8% (Ridge algorithm) to a maximum growth of 3.5% (GBR algorithm). As these algorithms weigh the behavior of features differently, the greater variability in the prediction may indicate that some predictors exhibit more favorable performance in their growth compared to other features in the overall group.

3.4 Sensitivity to variations in remote sensing indicators

The previous results considered features such as the median value of nighttime lights and the proportion of built-up areas for the entire territory of Bolivia. However, similar time series were also constructed for more specific geographic areas, such as the metropolitan areas of La Paz, Kanata, and Santa Cruz, which are clusters of municipalities where Bolivia's population and income are concentrated.

In Table 4, it is observed a slight improvement in the predictive capacity of the algorithms when variables on median luminosity are included independently for the three metropolitan regions under analysis. Both the mean squared error (MSE) and mean absolute error (MAE) decrease slightly compared to the values of these metrics when the algorithms were trained with a single nighttime lights variable for the entire territory of Bolivia.

Table 4: Forecast evaluation metrics by type of remote sensing indicators

Algorithm	Bolivia		3 luminosity variables		3 built-up variables	
	MSE	MAE	MSE	MAE	MSE	MAE
Ridge	0.294	0.354	0.290	0.344	0.298	0.355
Lasso	0.432	0.383	0.426	0.378	0.433	0.384
ElasticNet	0.351	0.377	0.338	0.367	0.354	0.379
Decision Tree	0.679	0.393	0.604	0.335	0.707	0.427
AdaBoost	0.539	0.323	0.530	0.335	0.542	0.335
Gradient Boosting	0.389	0.283	0.390	0.286	0.381	0.279
Random Forest	0.540	0.362	0.539	0.357	0.540	0.361
Extra Trees	0.487	0.357	0.455	0.340	0.461	0.345

On the other hand, when the variable of the proportion of built-up areas is disaggregated for each metropolitan region, the prediction capacity is equal or slightly lower compared to the case of a single variable for Bolivia.

Given that an improvement —although minor— is achieved with independent variables on the median luminosity by metropolitan area, an analysis of feature importance for the trained algorithms was performed. All three variables would be important to predict the IGAE; however, the corresponding variables for the metropolitan regions of La Paz and Santa Cruz tend to have a greater weight.

In conclusion, for parsimony and given that the improvements in predictive capacity are minimal, it is appropriate to maintain the remote sensing variables for the national territory.

3.5 Does the present nowcasting technique outperform classical econometric methods?

The results presented in the previous sections demonstrate that the proposed methodological approach is well-suited for forecasting monthly economic growth in Bolivia using machine learning algorithms that incorporate information from both conventional sources and satellite imagery. However, it is important to compare the predictive power of this approach with that of econometric models that are conventionally used for this type of forecasting.

In this regard, this section presents a comparison of the predictive performance of the nowcast indicator proposed in this research with two econometric models conventionally used for time series forecasting: a univariate model and a multivariate model.

- The **univariate model** is a Seasonal Autoregressive Integrated Moving Average (SARIMA) model, which is a useful tool for forecasting time series with complex seasonal patterns, but is intensive in terms of the amount of data required to estimate the model coefficients accurately. The optimal specification of the model that minimizes the value of the information criteria was obtained using the *"auto.arima"* function in the pmdarima library in Python.[16] Four SARIMA models are developed, each with a 1-year projection horizon, to compare the predictive capacity of the models over the validation set period (January 2019 to September 2022). For instance, the first model is trained with data from 1990M2 to 2018M12 and its projection covers the period 2019M1-2019M12. The second model is trained with data from 1990M2 to 2019M12 and its projections cover the period 2020M1 to 2020M12, and so on.

- The **multivariate model** follows Arias, Rubio-Ramírez, and Waggoner (2018) and estimates a Bayesian Structural Vector Autoregressive (BSVAR) model, which has the ability to impose restrictions on the identification of zeros, signs, and magnitudes for contemporary and subsequent periods associated with structural shocks. In this model, the endogenous variables are the year-on-year growth of IGAE, inflation, legal reserve requirements surplus, and the price of oil, whose dynamics are explained by supply, demand, monetary policy, and commodity price shocks, as shown in the identification strategy exemplified in Table 5.

  The forecasts of this econometric model —transformed into monthly growth rates— are included in the comparison with the nowcast indicator because it is an advanced model for time series forecasting that captures the dynamics (posterior probability distributions) of the main shocks that determine the behavior of the Bolivian economy; ergo, a good predictive performance is expected. The BSVAR model is trained on a sample of monthly data from January 2007 to December 2018, and the growth of IGAE is forecasted for the months of the validation set conditional on the observed values of the other endogenous variables.

[16] The specification is SARIMAX(1, 0, 2)x(2, 0, [1], 12).

Table 5: Structural shocks identification strategy

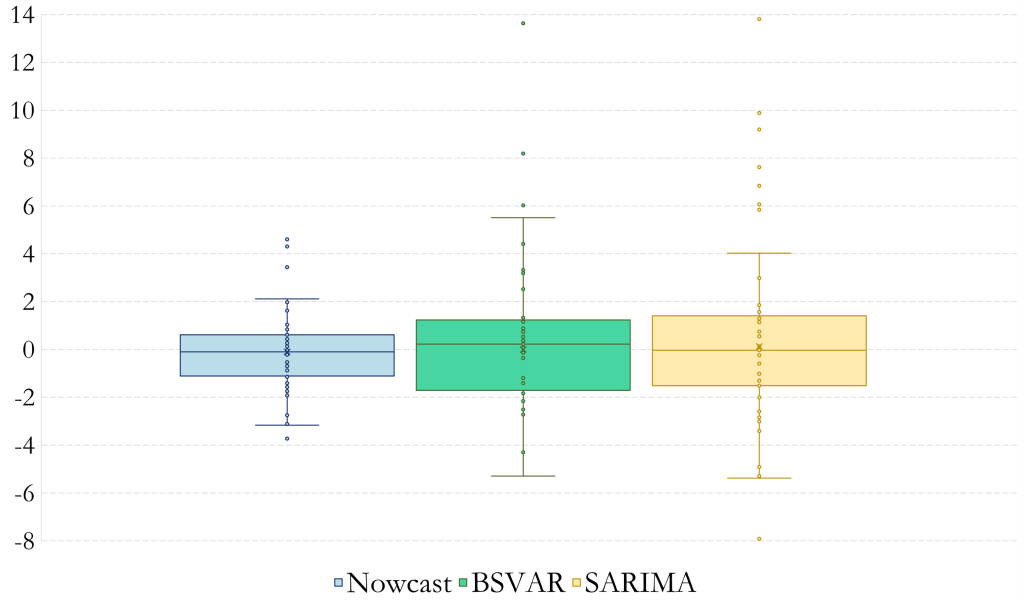
Shocks	IGAE growth	Inflation	Reserve requirements	Oil price
Supply	+	-	.	0
Demand	+	+	.	0
Monetary Policy	.	.	+	0
Commodity price	.	.	.	+

Note: Supply shocks are distinguished from demand shocks as they generate reductions in prices, because they are associated with improvements in productivity (i.e., lower production costs). The prices of commodities are exogenous to internal conditions (i.e., zero effect of domestic supply and demand on international prices of commodities), which is consistent with the characteristics of the Bolivian economy. Expansionary monetary policy shocks are those that increase the liquidity of the financial system (i.e., through a larger surplus of legal reserves), but allowing the model solution to determine whether there are contemporary effects on output and prices. (.) means that no condition is imposed.

The mean squared errors of the SARIMA and BSVAR models, for the validation set, are 1.27 and 0.97, respectively. These results suggest that the nowcast indicator developed in this research would have better predictive performance compared to the two evaluated econometric methods. The mean squared error of the nowcast approach is 0.34; it should be noted that as this indicator is a combination of forecasts from three algorithms, this MSE metric may not necessarily be equal to those reported in Table 2.[17]

Additionally, the Figure 7 exhibits the distribution of forecast errors (observed monthly growth of the IGAE minus the predicted) for the three models: Nowcast, BSVAR, and SARIMA. In the case of the nowcast of the IGAE, it can be observed that the forecast errors are distributed relatively symmetrically, with an average and median close to zero. More importantly, the interquartile range is relatively narrow; that is, 50% of the forecasts deviate at most ±1 percentage point from the observed growth.

Figure 7: Distribution of forecast errors by model



Note: In the distributions of the forecast errors of the BSVAR and SARIMA models, there are atypical errors close to -20 percentage points (during months of strict pandemic quarantine). However, for the purpose of simplifying the visualization, they are not included in the graph.

[17]Recall that MSE is calculated with the observed and predicted z-normalized values of the IGAE monthly growth rates.

In contrast, the BSVAR and SARIMA models show greater variability in the distribution of forecast errors. In these cases, 50% of the forecasts from these models deviate by more than 1.5 percentage points from the observed records. Moreover, the maximum errors, both in positive and negative deviations, are considerably higher in the BSVAR and SARIMA models relative to the distribution of forecast errors in the proposed nowcast with machine learning algorithms.

Among the BSVAR and SARIMA models, it is evident that their average predictive performance is similar, but the SARIMA model has more atypical errors, which is consistent with the weakness of univariate models that are not suitable for time series with seasonal patterns that change over time, as occurred with the COVID-19 pandemic.

In conclusion, the methodology proposed in this paper for nowcasting with machine learning algorithms that incorporate data from conventional sources and satellite images would be superior to the evaluated univariate and multivariate econometric models in terms of precision in forecasting the growth of the IGAE.

4 Concluding remarks

This study is considered to have fulfilled its objective of providing and validating a methodological alternative for nowcasting the monthly behavior of the Bolivian economy. One noteworthy aspect is that this alternative is an innovative proposal that not only takes advantage of the power of supervised machine learning algorithms, but also demonstrates the usefulness of information generated with remote sensing techniques as input for economic activity forecasting models.

Thus, in the face of a six-month lag in the official publication of the Global Index of Economic Activity (IGAE) —a proxy for monthly GDP in Bolivia— by the competent authorities in Bolivia, this study was able to reduce the time gap to only two months relative to the contemporary period, allowing policy makers, researchers, and economic analysts to have a more timely metric of economic growth. It is important to note that the proposed IGAE nowcast is technically a conditional forecast, so exact forecasts for the contemporary period cannot be generated due to the unavailability of most predictor variables.

Among some specific findings, the IGAE nowcast suggests that the Bolivian economy entered an economic slowdown since October 2022, associated with social conflicts that broke out in the last quarter. Moreover, it is estimated that the year-on-year variation of the IGAE in the months of October and November 2022 would have been negative, at -0.5% and -2.6%, respectively; during these months, strict economic activity paralysis occurred in the department of Santa Cruz —a key economic center for Bolivia—. However, despite this economic contraction, activity in December 2022 would show a slight rebound, with a year-on-year growth of approximately 2%. In summary, it is estimated that Bolivia's annual economic growth would be around 3% for 2022.

Considering that official IGAE data is only public up to September 2022, nowcast estimates of economic growth for the last three months of 2022 are of great value, especially for analyzing the degree of impact that social conflicts could have had in those months on Bolivia's real economic activity. In the literature review and other sources, no estimates are found on GDP or IGAE growth for the last quarter of 2022 in Bolivia, which adds greater importance to the results obtained in this document.

It should also be noted that the IGAE nowcast extends to January 2023. The results show a year-on-year growth of 2.3%, which, while relatively low compared to historical

year-on-year growth rates, corresponds to an improvement compared to December 2022 growth, exhibiting a gradual recovery of the Bolivian economy after the contractions of October and November 2022.

Finally, it is important to emphasize that the use of information from unconventional sources, such as satellite images, holds great potential and advantages in economic analysis and research. Satellite images can provide a wide range of geospatial information, such as the location of infrastructure, crops, urban and rural areas, among others. This information, when processed using spatial and statistical analysis techniques, allows economists to identify patterns and trends, enabling them to gain a more detailed and precise understanding of the economy of a region or country.

As part of this study, unprecedented time series were constructed for the Bolivian economy, such as monthly levels of light brightness, vegetation indices and precipitation in crop lands, and the proportion of built-up areas. These series were used to train the algorithms for nowcasting the IGAE. Furthermore, in the analysis of the importance of predictors, the time series generated by remote sensing were found to be of great importance in most of the machine learning algorithms that were evaluated. For example, the most important predictor in the majority of the algorithms was the intensity of nighttime lights in Bolivia. Therefore, it is recommended to make more intensive use of this type of information, not only for the exclusive analysis of the GDP but also of other economic indicators in Bolivia and other countries that face timely information constraints.

References

Arias, J. E., Rubio-Ramírez, J. F., & Waggoner, D. F. (2018). Inference based on structural vector autoregressions identified with sign and zero restrictions: Theory and applications. *Econometrica*, *86*(2), 685–720.

Bajari, P., Nekipelov, D., Ryan, S. P., & Yang, M. (2015). Machine learning methods for demand estimation. *American Economic Review*, *105*(5), 481–485.

Balcilar, M., Gabauer, D., Gupta, R., & Pierdzioch, C. (2022). Uncertainty and forecastability of regional output growth in the uk: Evidence from machine learning. *Journal of Forecasting*, *41*(6), 1049–1064.

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

Belly, G., Boeckelmann, L., Caicedo Graciano, C. M., Di Iorio, A., Istrefi, K., Siakoulis, V., & Stalla-Bourdillon, A. (2023). Forecasting sovereign risk in the euro area via machine learning. *Journal of Forecasting*, *42*(3), 657–684.

Bonato, M., Çepni, O., Gupta, R., & Pierdzioch, C. (2022). El niño, la niña, and forecastability of the realized variance of agricultural commodity prices: Evidence from a machine learning approach. *Journal of Forecasting*.

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

Chen, X., & Nordhaus, W. D. (2011). Using luminosity data as a proxy for economic statistics. *Proceedings of the National Academy of Sciences*, *108*(21), 8589–8594.

Chicco, D., & Jurman, G. (2020). The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics*, *21*, 1–13.

Corbane, C., Syrris, V., Sabo, F., Politis, P., Melchiorri, M., Pesaresi, M., ... Kemper, T. (2021). Convolutional neural networks for global human settlements mapping from sentinel-2 satellite imagery. *Neural Computing and Applications*, *33*, 6697–6720.

Donaldson, D., & Storeygard, A. (2016). The view from above: Applications of satellite data in economics. *Journal of Economic Perspectives*, *30*(4), 171–198.

Duro, D. C., Franklin, S. E., & Dubé, M. G. (2012). A comparison of pixel-based and object-based image analysis with selected machine learning algorithms for the classification of agricultural landscapes using spot-5 hrg imagery. *Remote sensing of environment*, *118*, 259–272.

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, *55*(1), 119–139.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189–1232.

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine learning*, *63*, 3–42.

Ghosh, T., L Powell, R., D Elvidge, C., E Baugh, K., C Sutton, P., & Anderson, S. (2010). Shedding light on the global distribution of economic activity. *The Open Geography Journal*, *3*(1).

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

Gogas, P., Papadimitriou, T., & Sofianos, E. (2022). Forecasting unemployment in the euro area with machine learning. *Journal of Forecasting*, *41*(3), 551–566.

Henderson, J. V., Storeygard, A., & Weil, D. N. (2012). Measuring economic growth from outer space. *American economic review*, *102*(2), 994–1028.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55–67.

Hu, M., & Xia, B. (2019). A significant increase in the normalized difference vegetation index during the rapid economic development in the pearl river delta of china. *Land Degradation & Development*, *30*(4), 359–370.

Jiang, S., Wei, G., Zhang, Z., Wang, Y., Xu, M., Wang, Q., ... Liu, B. (2020). Detecting the dynamics of urban growth in africa using dmsp/ols nighttime light data. *Land*, *10*(1), 13.

Jung, J., Maeda, M., Chang, A., Bhandari, M., Ashapure, A., & Landivar-Bowles, J. (2021). The potential of remote sensing and artificial intelligence as tools to improve the resilience of agriculture production systems. *Current Opinion in Biotechnology*, *70*, 15–22.

Keola, S., Andersson, M., & Hall, O. (2015). Monitoring economic development from space: using nighttime light and land cover data to measure economic growth. *World Development*, *66*, 322–334.

Kotsiantis, S. (2011). Feature selection for machine learning classification problems: a recent overview. *Artificial Intelligence Review*, *42*(1), 157–176.

Kussul, N., Lavreniuk, M., Skakun, S., & Shelestov, A. (2017). Deep learning classification of land cover and crop types using remote sensing data. *IEEE Geoscience and Remote Sensing Letters*, *14*(5), 778–782.

Liu, L., Chen, C., & Wang, B. (2022). Predicting financial crises with machine learning methods. *Journal of Forecasting*, *41*(5), 871–910.

Liu, Y., Zuo, R., & Dong, Y. (2021). Analysis of temporal and spatial characteristics of urban expansion in xiaonan district from 1990 to 2020 using time series landsat imagery. *Remote Sensing*, *13*(21), 4299.

Loh, W.-Y. (2011). Classification and regression trees. *Wiley interdisciplinary reviews: data mining and knowledge discovery*, *1*(1), 14–23.

Mantovani, R. G., Horváth, T., Cerri, R., Junior, S. B., Vanschoren, J., & de Carvalho, A. C. P. d. L. F. (2018). An empirical study on hyperparameter tuning of decision trees. *arXiv preprint arXiv:1812.02207*.

Medeiros, M. C., Vasconcelos, G. F., Veiga, Á., & Zilberman, E. (2021). Forecasting inflation in a data-rich environment: the benefits of machine learning methods. *Journal of Business & Economic Statistics*, *39*(1), 98–119.

Melchiorri, M., Florczyk, A. J., Freire, S., Schiavina, M., Pesaresi, M., & Kemper, T. (2018). Unveiling 25 years of planetary urbanization with remote sensing: Perspectives from the global human settlement layer. *Remote Sensing*, *10*(5), 768.

Milunovich, G. (2020). Forecasting australia's real house price index: A comparison of time series and machine learning methods. *Journal of Forecasting*, *39*(7), 1098–1118.

Mullainathan, S., & Spiess, J. (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, *31*, 87-106.

Pinkovskiy, M., & Sala-i Martin, X. (2016). Lights, camera... income! illuminating the national accounts-household surveys debate. *The Quarterly Journal of Economics*, *131*(2), 579–631.

Plakandaras, V., Papadimitriou, T., & Gogas, P. (2015). Forecasting daily and monthly exchange rates with machine learning techniques. *Journal of Forecasting*, *34*(7),

560–573.

Román, M. O., Wang, Z., Sun, Q., Kalb, V., Miller, S. D., Molthan, A., ... others (2018). Nasa's black marble nighttime lights product suite. *Remote Sensing of Environment*, *210*, 113–143.

Sharifi, A. (2021). Yield prediction with machine learning algorithms and satellite images. *Journal of the Science of Food and Agriculture*, *101*(3), 891–896.

Storeygard, A. (2016). Farther on down the road: transport costs, trade and urban growth in sub-saharan africa. *The Review of economic studies*, *83*(3), 1263–1295.

Tang, B., Liu, Y., & Matteson, D. S. (2022). Predicting poverty with vegetation index. *Applied Economic Perspectives and Policy*, *44*(2), 930–945.

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

Wiegand, C., Richardson, A., Escobar, D., & Gerbermann, A. (1991). Vegetation indices in crop assessments. *Remote sensing of Environment*, *35*(2-3), 105–119.

Zhang, R., Tian, Z., McCarthy, K. J., Wang, X., & Zhang, K. (2023). Application of machine learning techniques to predict entrepreneurial firm valuation. *Journal of Forecasting*, *42*(2), 402–417.

Zhou, M., Lu, L., Guo, H., Weng, Q., Cao, S., Zhang, S., & Li, Q. (2021). Urban sprawl and changes in land-use efficiency in the beijing–tianjin–hebei region, china from 2000 to 2020: A spatiotemporal analysis using earth observation data. *Remote Sensing*, *13*(15), 2850.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, *67*(2), 301–320.

A Methodological characteristics and execution of the machine learning algorithms

A.1 Ridge, Lasso and ElasticNet

Ridge, Lasso, and ElasticNet are three popular regularization techniques used in linear regression models to prevent overfitting. These techniques add a penalty term to the objective function, which controls the complexity of the model and reduces the impact of irrelevant features.

Ridge regression, also known as L2 regularization, adds a penalty term to the sum of squared errors (SSE) objective function. The penalty term is proportional to the square of the L2 norm of the coefficients vector. The objective function for Ridge regression is given by:

$$minimize\ J(\beta) = SSE + \lambda \sum_{j=1}^{p} \beta_j^2$$

where β is the vector of coefficients, p is the number of features, and λ is the regularization parameter that controls the strength of the penalty term.

Lasso regression, also known as L1 regularization, adds a penalty term to the SSE objective function. The penalty term is proportional to the L1 norm of the coefficients vector. The objective function for Lasso regression is given by:

$$minimize\ J(\beta) = SSE + \lambda \sum_{j=1}^{p} |\beta_j|$$

ElasticNet is a combination of Ridge and Lasso regression techniques. It adds both L1 and L2 penalty terms to the SSE objective function. The objective function for ElasticNet regression is given by:

$$minimize\ J(\beta) = SSE + \lambda \left(\alpha \sum_{j=1}^{p} |\beta_j| + (1-\alpha) \sum_{j=1}^{p} \beta_j^2 \right)$$

The hyperparameter λ controls the strength of the regularization, with larger values of λ leading to stronger regularization. The hyperparameter α controls the mixing parameter between L1 and L2 regularization, with values between 0 and 1. A value of 1 corresponds to L1 regularization, while a value of 0 corresponds to L2 regularization. Intermediate values correspond to a combination of both L1 and L2 regularization. In summary, Ridge regression shrinks the coefficients towards zero, Lasso regression sets some coefficients to zero, and ElasticNet combines both techniques to overcome their limitations.

A.2 Decision Tree Regressor

The Decision Tree Regressor works by recursively partitioning the feature space into smaller and smaller regions based on the values of the input features. Let X be the input feature matrix of size $n \times m$, where n is the number of observations and m is the number of features. Let y be the corresponding target variable vector of size n. The algorithm works as follows:

1. Choose a feature j and a threshold value t that best splits the data into two subsets R_1 and R_2 based on the mean squared error (MSE) loss function:

$$MSE(R) = \frac{1}{|R|} \sum_{i \in R} (y_i - \bar{y}_R)^2$$

where $|R|$ is the number of observations in region R and $\bar{y}_R$ is the mean target value in region R.

2. Repeat step 1 recursively on each subset R_1 and R_2 until a stopping criterion is met, such as a maximum depth ($MaxDepth$) or minimum number of observations in a leaf node ($MinSamplesLeaf$).

3. Assign the mean target value of the training observations in each leaf node as the predicted value for new observations that fall within that region.

The algorithm can be represented as a binary tree, where each internal node represents a split on a feature and threshold value, and each leaf node represents a predicted target value. The tree can be trained using an iterative algorithm such as CART (Classification and Regression Trees).

A.3 AdaBoost Regressor

AdaBoost Regressor is a boosting algorithm used for regression tasks. It trains a series of weak learners on weighted versions of the training data, with the weights updated at each iteration to focus on the data points that were poorly predicted in previous iterations. The final model is an ensemble of the weak learners, weighted by their performance.

Let $D = (x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ be the training data, where x_i is a vector of input features and y_i is the corresponding output label. Let $h(x)$ be a weak learner that takes an input x and outputs a prediction $\hat{y}$.

The algorithm proceeds as follows:

1. Initialize the sample weights $w_i = 1/n$ for $i = 1, 2, ..., n$.

2. For $t = 1, 2, ..., T$ do:

   a. Train a weak learner $h_t(x)$ on the training data D with weights w_i.

   b. Compute the weighted error ϵ_t of the weak learner:

$$\epsilon_t = \frac{\sum_{i=1}^{n} w_i \left| y_i - h_t\left(x_i\right) \right|}{\sum_{i=1}^{n} w_i} \tag{3}$$

   c. Compute the learning rate α_t:

$$\alpha_t = \frac{1}{2} \ln\left(\frac{1 - \epsilon_t}{\epsilon_t}\right) \tag{4}$$

   d. Update the sample weights w_i:

$$w_i \leftarrow w_i \cdot \exp\left(-\alpha_t \left(y_i - h_t\left(x_i\right)\right)\right) \tag{5}$$

e. Normalize the weights:

$$w_i \leftarrow \frac{w_j}{\sum_{j=1}^{n} w_j} \tag{6}$$

3. Output the final model:

$$f(x) = \sum_{t=1}^{T} \alpha_t h_t(x) \tag{7}$$

In this context, T denotes the number of weak learners, also referred to as base estimators, while ϵ_t measures the error of the weak learner $h_t(x)$ on the training data. The learning rate α_t governs the contribution of each weak learner to the final ensemble model, with a small value chosen in the case of high error rates, and a large value chosen when the error rate is low. The loss function utilized in AdaBoost Regressor is typically the exponential loss function, which places greater emphasis on data points that are difficult to predict.

A.4 Gradient Boosting Regressor

Gradient Boosting Regressor (GBR) is an ensemble method that combines multiple weak learners to create a strong predictor. Let y_i be the actual value of the target variable for the i-th observation, and let $\hat{y}_i$ be the predicted value of the target variable. Then, the MSE between the predicted and actual values is given by:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

The GBR algorithm works by sequentially adding new regression trees to the ensemble, with each tree attempting to correct the errors of the previous trees. Let $f_m(x)$ be the prediction of the m-th tree for the input vector x. Then, the prediction of the ensemble after M trees is given by:

$$\hat{y}_i^{(M)} = \sum_{m=1}^{M} \gamma_m f_m(x_i)$$

where γ_m is the learning rate for the m-th tree, which controls the contribution of each tree to the final prediction. The learning rate is typically set to a small value, such as 0.1, to prevent overfitting.

To train the GBR, it starts with an initial prediction $\hat{y}_i^{(0)}$, which is typically set to the mean value of the target variable. Then iteratively add new trees to the ensemble by minimizing the residual errors of the previous trees. Specifically, at each iteration m, fit a new regression tree $f_m(x)$ to the negative gradients of the MSE loss function:

$$r_{im} = -\frac{\partial}{\partial \hat{y}_i^{(m-1)}} L(y_i, \hat{y}_i^{(m-1)})$$

where $L(y_i, \hat{y}_i^{(m-1)})$ is the loss function used to measure the error between the predicted and actual values. The negative gradients are then used as the target values for the new tree, and the learning rate is applied to control the contribution of the new tree to the ensemble. The final prediction of the GBR is obtained by summing the predictions of all the trees in the ensemble:

$$\hat{y}_i = \sum_{m=1}^{M} \gamma_m f_m(x_i)$$

A.5 Random Forest Regression

Random Forest Regression is an ensemble learning method that combines multiple decision trees to make a more accurate prediction. In this algorithm, a random subset of features is selected at each node of the decision tree, and the best split is made based on the selected features. The final prediction is made by averaging the predictions of all the decision trees in the forest.

Let X be the input data with n samples and m features, and y be the target variable. The Random Forest Regression algorithm can be described as follows:

1. Initialize the number of decision trees T, the number of features to be selected at each node m', and the maximum depth of each tree d.
2. For $t = 1$ to T:
    a. Randomly select m' features from the m features.
    b. Construct a decision tree using the selected features with a maximum depth of d.
3. For a new input x', predict the target variable y' as:

$$y' = \frac{1}{T} \sum_{t=1}^{T} f_t(x')$$

where $f_t(x')$ is the prediction of the t-th decision tree.

A.6 Extra Trees Regressor

The Extra Trees Regressor algorithm is similar to the Random Forest algorithm, but with some key differences in the way the trees are constructed. Let X be the input features and y be the target variable. The Extra Trees Regressor algorithm works as follows:

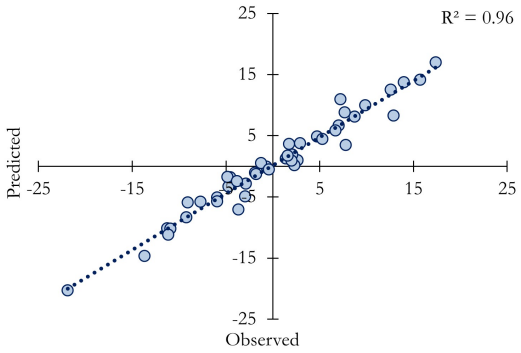
1. Randomly select a subset of the input features for each tree to use.
2. For each tree, randomly select a subset of the training data to use for training.
3. For each split in each tree, randomly select a subset of features to use for determining the best split.
4. Split the data based on the selected feature and threshold value that minimizes the mean squared error (MSE).
5. Repeat steps 3-4 until the tree reaches a maximum depth or the number of samples in a leaf node is below a certain threshold.

The predicted output value for a new input sample x can then be obtained by averaging the output values of all the trees in the ensemble:

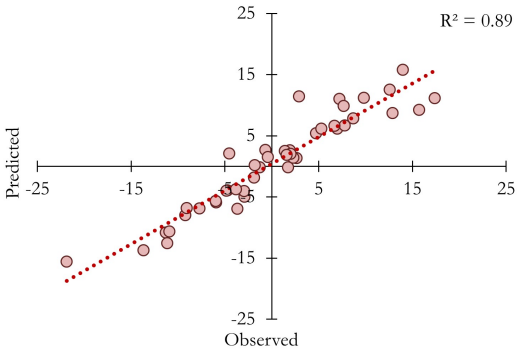
$$\hat{y} = \frac{1}{N} \sum_{i=1}^{N} f_i(x) \tag{8}$$

where N is the number of trees in the ensemble and $f_i(x)$ is the output value of the i-th tree for the input sample x.

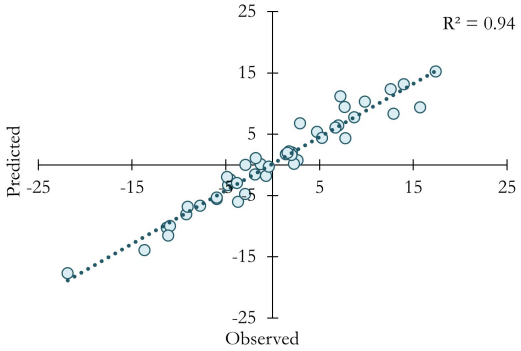
B R^2 between observed and predicted IGAE by algorithm



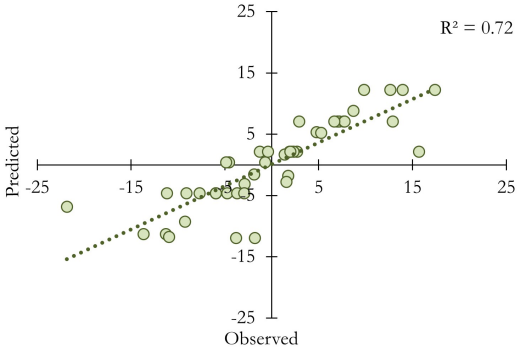
(a) Ridge



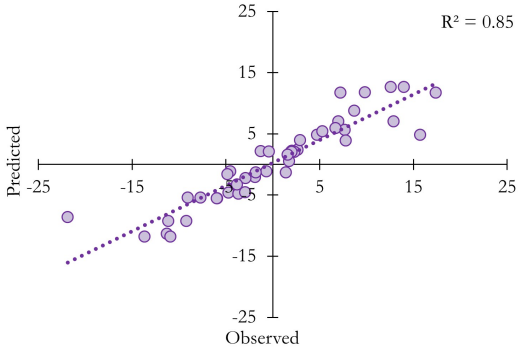
(b) Lasso



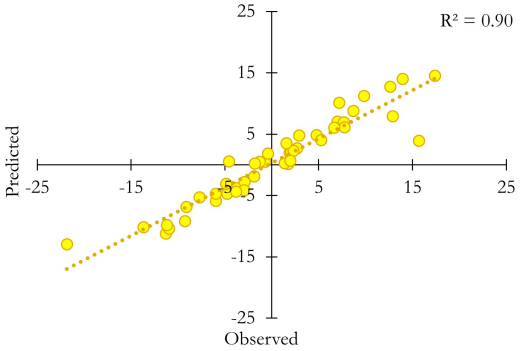
(c) ElasticNet



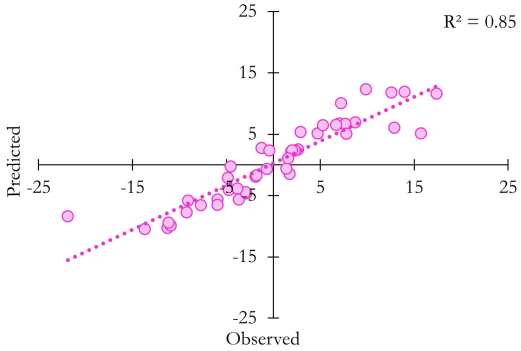
(d) Decision Tree



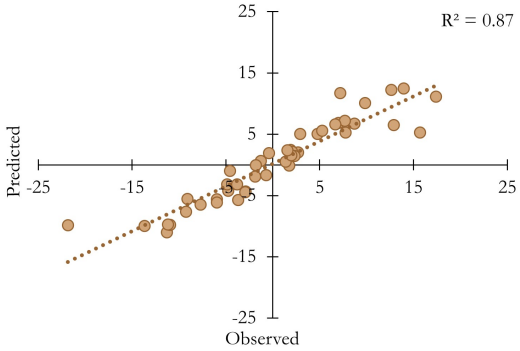
(e) AdaBoost



(f) Gradient Boosting

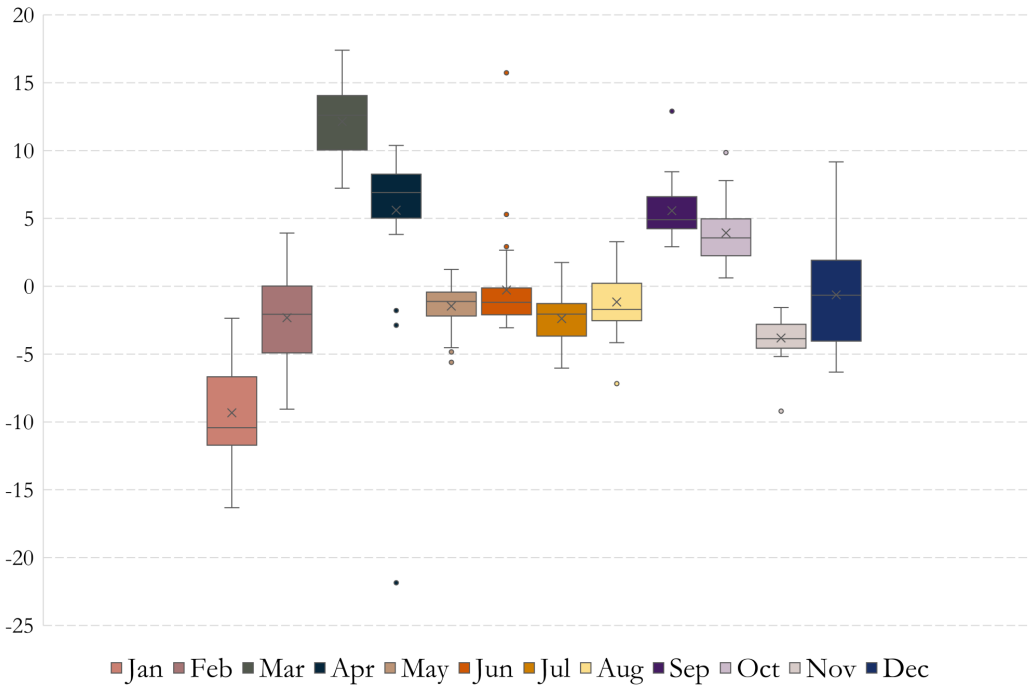


(g) Random Forest



(h) Extra Trees

C Distribution of the IGAE monthly growth rate in each month since 1990



Source: National Institute of Statistics.