INTERNATIONAL MONETARY FUND

Nowcasting Economic Growth with Machine Learning and Satellite Data

Eurydice Fotopoulou, Iyke Maduako, Prachi Srivastava, M. Belen Sbrancia

WP/26/20

2026
JAN

WORKING PAPER

 WP/26/20

IMF Working Paper

Statistics Department

Nowcasting Economic Growth with Machine Learning and Satellite Data

Prepared by Eurydice Fotopoulou, Iyke Maduako, Prachi Srivastava, M. Belen Sbrancia*

Authorized for distribution by Marco Marini
January 2026

ABSTRACT: The absence of reliable data on fundamental economic indicators (e.g. real GDP), combined with structural shifts in the economy, can severely constrain the ability to conduct accurate macroeconomic analysis and forecasting. This paper explores alternatives to address data limitations by integrating machine learning and satellite data to estimate real GDP. Specifically, it finds that incorporating satellite-based nightlight data into a random forest model significantly improves the accuracy of quarterly GDP growth estimates compared with models relying solely on traditional indicators.This empirical application contributes to the emerging nowcasting field to enhance economic forecasting in economies with significant data gaps.

RECOMMENDED CITATION: Fotopoulou, E., Maduako, I., Srivastava, P. and Sbrancia, M. B. (2026). *Nowcasting Economic Growth with Machine Learning and Satellite Data*. Working Paper. International Monetary Fund WP/26/20

JEL Classification Numbers:	C53, C44, C52
Keywords:	Macroeconomic forecast, Machine learning, Nowcasting, GDP, Satellite data, Random Forest
Authors E-Mail Address:	efotopoulou@imf.org, imaduako@imf.org, BSbrancia@imf.org, prachi.jmc13@gmail.com

* The authors would like to thank Jean Francois Clevy Aquilar, El Boukherouaa, Sonali Das, Chris Evans, Aquiles Farias, Pierre Guerin, Beata Jajko, Dora Metodieva Iakova, Jason Lu, Marco Marini, Clément Marsilli, Pau Rabanal, Ivy Sabuga, Giovanni Ugazio, and Kevin Wiseman for their feedback, Roberta Guarnieri for data support, and Amira Archer-Davies for excellent editing support. All errors remain authors' own.

WORKING PAPER

Nowcasting Economic Growth with Machine Learning and Satellite Data

Prepared by: Eurydice Fotopoulou, Iyke Maduako, Prachi Srivastava, M. Belen Sbrancia[1]

[1] The authors would like to thank Jean Francois Clevy Aquilar, El Boukherouaa, Sonali Das, Chris Evans, Aquiles Farias, Pierre Guerin, Beata Jajko, Dora Metodieva Iakova, Jason Lu, Marco Marini, Clément Marsilli, Pau Rabanal, Ivy Sabuga, Giovanni Ugazio, and Kevin Wiseman for their feedback, Roberta Guarnieri for data support, and Amira Archer-Davies for excellent editing support. All errors remain the authors' own. The views expressed in IMF Working Papers are those of the authors and do not necessarily represent the views of the IMF, its Executive Board, or IMF management.

Affiliations: Eurydice Fotopoulou, Iyke Maduako, M. Belen Sbrancia, all IMF. Prachi Srivastava: Heidelberg University.

Contents

FIGURES

TABLES

Glossary

DNB	Day/Night Band
EVI	Enhanced Vegetation Index
GDP	Gross Domestic Product
ML	Machine Learning
NDVI	Normalized Difference Vegetation Index
NTL	Nightlight
USD	United States Dollar
VIIDS	Visible Infrared Imaging Radiometer Suite

Executive Summary

Economic forecasting in Venezuela[2] presents significant challenges due to the lack of timely, reliable, publicly available key macroeconomic data by the Banco Central de Venezuela since Q1 2019. The economy has experienced a sharp and prolonged contraction, particularly between 2013 and 2020, with nominal GDP in 2013 shrinking to a quarter of its 2012 size, and real GDP starting to contract in 2014. The absence of regularly published data on fundamental economic indicators—such as real GDP, trade flows, and manufacturing activity—combined with structural shifts in the economy, has severely constrained the ability to conduct accurate macroeconomic analysis and forecasting.

This paper seeks to close this data gap by integrating traditional and non-traditional data sources with machine learning and econometric techniques to nowcast real GDP. Specifically, it evaluates the applicability of Random Forest (RF) and the Dynamic Factor Model (DFM) in the case of Venezuela, up to 2024. In this work, we use a combination of satellite data (especially nightlight and vegetation data) to proxy majorly for non-oil economic activity, and calibrated and harmonized light datasets, which are then combined with macroeconomic data, to produce a quarterly estimate of GDP. We find that using an RF model with satellite data produces more accurate results than without satellite data or a DFM. The introduction of satellite data improved the RF model by 13.85% while the use of a non-linear model (in this case RF) outperformed the popular DFM model by reducing the RMSE by 32.85%.

The analysis assesses the advantages and limitations of these methodologies, highlighting their potential to bridge critical data gaps and enhance economic forecasting in environments where direct data collection is limited. These approaches may offer broader applicability for forecasting in data-scarce economies, where economic activity may be subject to significant fluctuations, informing policy design and economic decision-making in similarly constrained contexts.

[2] As of the time of publication, the International Monetary Fund's engagement with the Bolivarian Republic of Venezuela remained paused due to the lack of government recognition by the membership. The analysis in this paper is solely technical and is intended to test the model and methodology in a data-constrained environment.

1. Introduction

The complex and shifting economic landscape requires flexibility and stronger analysis of real-time vulnerabilities to ensure responsive and responsible policy. To this end, the role of granular and timely data is paramount, as they allow realistic economic modelling and forecasting to measure and monitor the progress of an economy. Real Gross Domestic Product (GDP) is one of the most commonly used variables to measure an economy. It is an essential input in economic analysis, policy planning, and a key indicator of the evolution over time of an economy and a society. Yet, in some cases reliable real GDP series in a consistent and timely manner are not available either due to institutional capacity or more complex considerations. Under these conditions, the absence of data can significantly hinder the effectiveness of policy measures and obscure the true trajectory of growth in an economy.

The paper focuses on closing this gap for Venezuela[3], a country where the regular publication of statistics has stopped. The motivation for this paper was a desire to experiment with potential methods to close a data gap in a data-constrained environment and test nowcasting in case where the economy is not diversified and reliant heavily on a single commodity.

To understand the challenges in estimating quarterly real GDP growth in the context of Venezuela it is important to first discuss recent economic developments in the country. Due to a prolonged crisis, Venezuela turned into one of the most fragile countries in the world; its economy in 2020 had contracted to 75 percent in real GDP relative to 2013. Since then and as of 2024, real GDP has rebounded, growing on average at 4 percent. However, the economy in 2024 was still 70 percent below its 2013 level. Second, Venezuela experienced one of the most protracted hyperinflation episodes on record between 2017-2021. Third, social and humanitarian conditions are dire, infrastructure and public services have collapsed, and about 25 percent of the population has left the country (mostly to the region) as migrants and refugees. And last, but not least, there is a debt crisis with approximately US$ 140-190 billion of external debt in default.

Economic nowcasting is challenging when there is lack of timely, reliable, and publicly available macroeconomic data. As an alternative, satellite-based indicators can provide valuable insights into economic activity through indirect observation. Although these datasets were not originally designed for economic forecasting, their application in economic analysis and extraction of additional insights has grown significantly. By leveraging multiple satellite-based indicators, we aim to construct a more comprehensive proxy for real GDP, complementing traditional macroeconomic data sources. This constitutes one of the key contributions of this paper, with a proposed application that can be an essential tool for countries that are severely data constrained.

The existence of several structural breaks in the series does not allow for traditional techniques to be used to forecast economic growth. However, the novel techniques used in this paper capture the numerous structural breaks in the series, by combining machine learning and econometric techniques. As we will discuss in more detail in section 5, the model performs better when it combines satellite data and high-frequency data from other sources. We also consider other machine learning regularization methods like LASSO, Ridge and Elastic Net, which produce results inferior to that of random forest. Overall, random forest and dynamic factor models (DFM) tend to perform better when there is no data disruption, which is not the case for Venezuela. In the period we examine there is a severe lack of publication lags and other vignettes (small recurring data releases) which are used under DFM. To address this, we use a combination of satellite data (especially nightlight and

[3] As of the time of publication, the International Monetary Fund's engagement with the Bolivarian Republic of Venezuela remained paused due to the lack of government recognition by the membership. The analysis in this paper is solely technical and is intended to test the model and methodology in a data-constrained environment.

vegetation data) to proxy majorly for non-oil economic activity for Venezuela and calibrated and harmonized satellite data on night light to generate a time series spanning 2000–2024 for the purposes of testing this work. We combine this with self-curated data on macroeconomic variables from private various sources. We use matching statistics for crude oil imports, third party sources for gas consumption, revenue from VAT, monetary aggregates. This helps to understand the other parts of non-oil activity for Venezuela and give us more accurate analysis.

The applicability of the techniques used here is broader though and can serve to improve the quality of existing statistics by making evident gaps or inconsistencies across different series. Newer techniques, such the combination of machine learning and traditional econometrics tested in this paper, have the power to close data gaps and enhance monitoring of economic conditions, while providing more flexibility to understand changing trends over time, with publicly available non-traditional and traditional data.

The remainder of the paper is organized in the following way: the next section covers briefly the relevant literature. We then discuss the availability and treatment of the data required, in section 3, including some stylized facts. Section 4 reviews the methodology used, presenting the building blocks of our approach, while section 5 discusses the results. Finally, section 6 concludes, with further suggestions on the applicability and improvement of our technique.

2. Literature Review

The basic principle of nowcasting is to exploit information, which is published earlier and often at higher frequencies than the variable of interest, to obtain its early estimate before the official figure of an economic indicator is published. Nowcasting of GDP specifically, one of the key metrics for any economy, has been in practice for more than 20 years, due to its significance for policy design and decision making, especially at times of uncertainty or when facing large shocks. Early efforts to nowcast GDP aimed to close a relatively short lag, usually of 1-2 quarters, building on financial variables, for instance with the foundational models by Andreou, Ghysels and Kourtellos (2008) and a project for the Board of Governors of the Federal Reserve since 2003, pioneered by Giannone, Reichlin and Small (2008). Since then, nowcasts of GDP have become ubiquitous, using mostly traditional data (financial and economic variables). More recent work such as Akbar et al. (2023), which highlights the importance of GDP nowcasting in low-income developing countries in sub-Saharan Africa, and particularly, fragile and conflict-affected states, has motivated us to look into data-scarce cases. This paper builds on and contributes to the latest nowcasting developments using machine learning (ML) and satellite data, aiming to overcome the complete lack of any available regular or high-frequency macroeconomic data.

We build on the growing literature of time series forecasting using machine learning methods. An important benchmark for our analysis is the dynamic factor model (DFM) as this model has been widely used to nowcast GDP (Giannone et al., 2008; Bańbura and Rünstler, 2011; Jansen et al., 2016; Hindrayanto et al., 2016; Bok et al. 2018, Dauphin at al., 2022). Another popular class of models that we also consider are MIDAS based models (Marcellino and Schumacher, 2010; Kuzin et al., 2011; Foroni and Marcellino, 2014). We compare this model to machine learning model using the random forest (RF) technique, which is an ensemble learning algorithm that builds multiple decision trees to improve the accuracy and stability of predictions. It combines the outputs of individual trees to produce a final, more reliable result. Additionally, we consider comparisons to LASSO and ridge regressions, following Cashin et Al. (2025) for nowcasting. Shapley values as discussed by Štrumbelj and Kononenko (2014) and Lundberg and Lee (2017) are also relevant in our analysis, especially on the importance of the predictors in the random forest combined with partial dependency plots which shows the marginal effect one or two features have on the predicted outcome of a machine learning model.

Additionally, we draw from (and contribute to) the emerging literature that is utilizing satellite data (Arslanalp et al., 2025; Arslanalp et al., 2021). The literature highlights the advantage of satellite data in cases where obtaining data in traditional ways is particularly challenging, such as in the case of South Sudan (McSharry and Mawejje, 2024) and Afghanistan (Abdel-Latif, Badr, and Maduako, forthcoming). Satellite data, such as nighttime light (luminosity), nitrogen dioxide (NO_2), and vegetation cover indices are being used in the literature as proxies for economic activity, with increasing success recently, as more data become available. For instance, research using NO_2 pollution finds that there is a positive relationship between the two (Bichler and Bittner, 2022; Dauphin at al., 2022), particularly where environmental laws may be flexible (Bichler, Schönebeck, and Bittner, 2023).

The use of nightlight data has been prominent in economic literature, with applications that range from the geographical mapping of economic activity (Sutton and Costanza, 2002; Doll et al., 2006; Ghosh et al., 2010; Tiffin, 2016), regional development analysis (Michalopoulos and Papaioannou, 2013), to the evaluation of the accuracy of economic statistics (Chen and Nordhaus, 2011; Henderson et al., 2012; Nordhaus and Chen, 2015; Pinkovskiy and Sala-i Martin, 2016). Of particular interest is the sigmoid relationship suggested by Zhao et al. (2019) and Li et al. (2020) between the nightlight inputs and income, which we also derive. Moreover, there is a growing literature on the use of Visible Infrared Imaging Radiometer Suite (VIIRS) nighttime light as a subnational indicator of economic activity at an intra-annual timeframe, sufficiently closing data gaps for policy purposes. The VIIRS data have been used to assess the impact of COVID-19 in India (Beyer et al., 2020; Ghosh et al., 2020; Beyer et al., 2021), China (Elvidge et al., 2020), and Morocco (Roberts, 2021). While many of these studies have struggled to convert changes in VIIRS nighttime lights to changes in economic activity, recent modelling efforts have made substantial progress and demonstrated the ability of nighttime lights to predict economic activity. Within this realm, Beyer et al. (2022) exploit the full potential of VIIRS data for economic analysis by estimating an elasticity of 1.55, which is needed to convert changes in nighttime lights into changes in economic activity. Their estimated elasticity is consistent for emerging market economies and low income economies, with only small deviations across country groups and different model specifications. This is a key empirical result which provides support for using nighttime light in such countries, including in countries affected by conflict and fragility, such as the present study.

3. Data

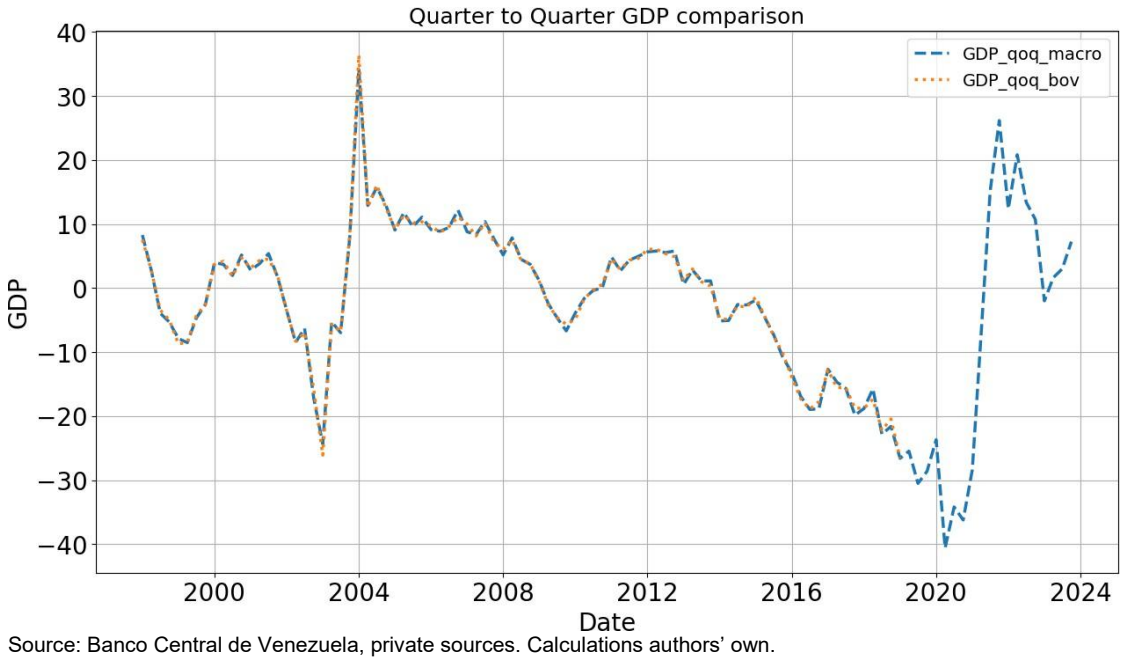
In this section we outline and evaluate the different types of data used in our analysis. These include: (i) real GDP, which serves as the target variable, (ii) alternative indicators derived from non-traditional data sources, such as satellite data, and (iii) macroeconomic variables, some of which are particularly relevant for nowcasting, most of which come from official sources. These variables have been selected due to their timeliness and ability to map different aspects of the state of the economy of Venezuela accurately.

3.1 Real GDP

Given the discontinuation of official GDP data publication after 2019 Q1, we construct a unified GDP series by using BCV data up to 2019 Q1 and private source estimates from 2019 Q2 onward. For the former period, we utilize seasonally adjusted real GDP from the Banco Central de Venezuela (BCV), covering the period 2006 Q1 to 2019 Q1, and for the latter, a real GDP series obtained from private sources, available from 1997 Q1 to 2023 Q4. To ensure consistency, we first apply seasonal adjustment to the private GDP series. We then compute both year-on-year (Y/Y) and quarter-on-quarter (Q/Q) GDP growth rates to validate the comparability of the two sources. Figure 1 presents the Y/Y quarterly growth rate of real GDP from both official and private sources,

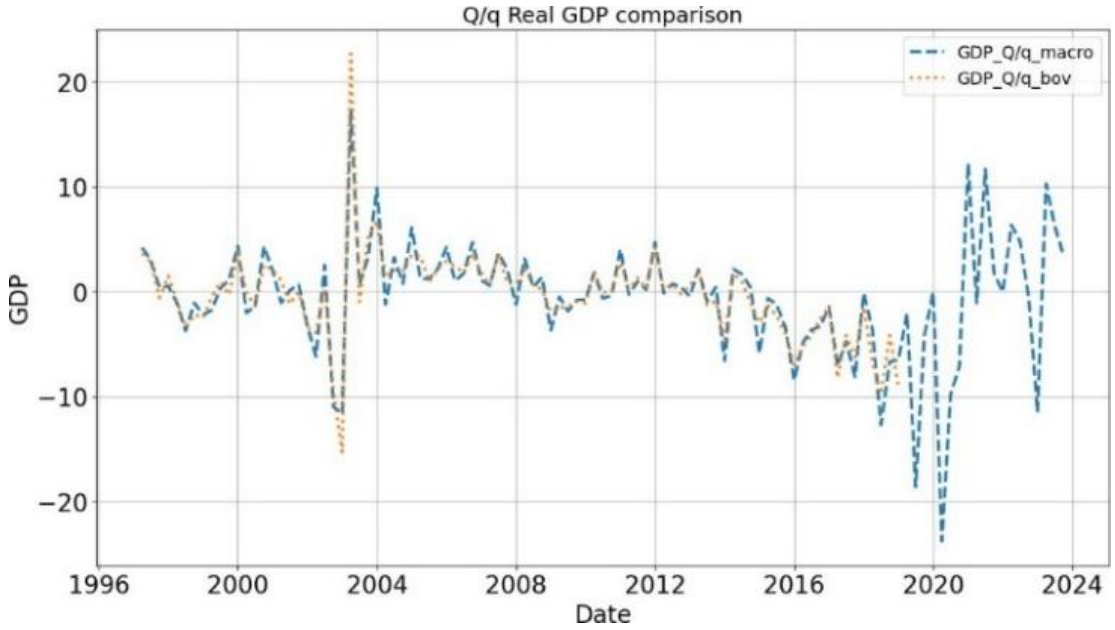
showing a high degree of similarity between the two series, which supports the reliability of private estimates. Figure 2 plots the Q/Q GDP growth rates, further reinforcing the viability of private data as a substitute. Finally, Figure 3 displays the final composite Q/Q real GDP growth rate series, integrating both sources, which serves as the basis for the nowcasting model. This approach allows us to mitigate the challenges posed by data discontinuities and enhance the robustness of our forecasting exercise.

Figure 1. Year-on-Year Real GDP Comparison



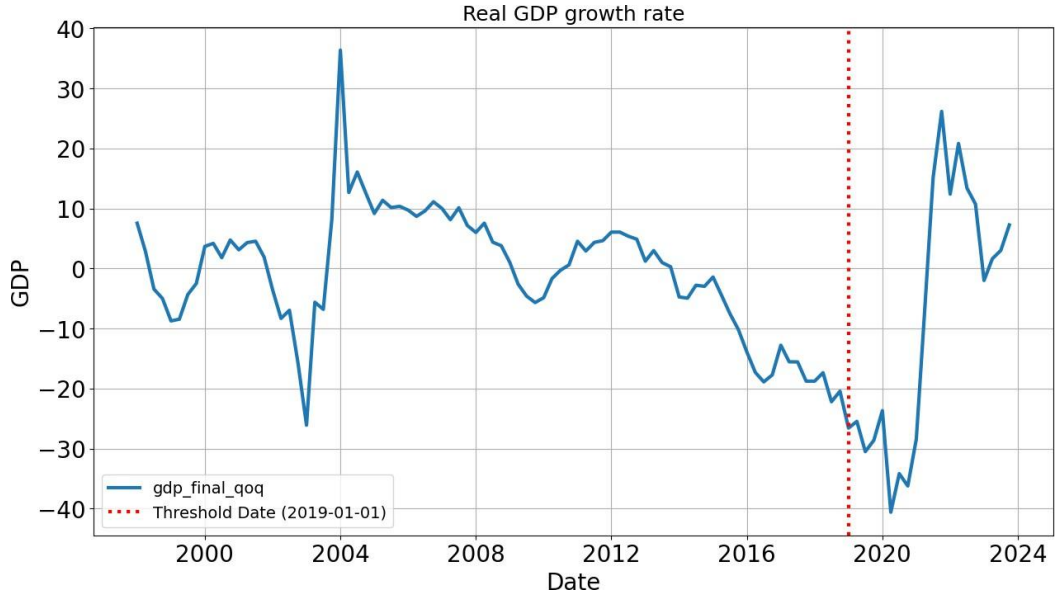
Source: Banco Central de Venezuela, private sources. Calculations authors' own.

Figure 2. Quarter-on-Quarter Real GDP Comparison



Source: Banco Central de Venezuela, private sources. Calculations authors' own.

Figure 3. Composite Quarter-on-Quarter Real GDP Growth Rate



Source: Banco Central de Venezuela, private sources. Calculations authors' own.

3.2.1 Satellite data

The objectivity and real-time nature of satellite-based indicators make them a valuable tool for macroeconomic surveillance, nowcasting/ forecasting, and policy analysis in data-scarce environments. Additionally, these tend to be publicly available data, meaning that researchers and analysts do not have to dedicate financial resources to obtain the data usually. To enhance forecasting accuracy, for this project we have chosen four distinct satellite data sources: (i) Nightlight data (NTL), (ii) Vegetation Indices (NDVI and EVI), and (iii) Nitrogen Dioxide (NO_2) Emissions. Each dataset captures different aspects of economic activity, and their combination provides a more complete and unbiased representation of macroeconomic conditions. A key advantage of using satellite data is that it is not subject to measurement challenges faced by traditional economic indicators, such as survey uncertainty or data revision lags. The following subsections detail the properties and economic relevance of each satellite-based indicator, emphasizing their suitability in cases where regular collection of traditional data is not always possible.

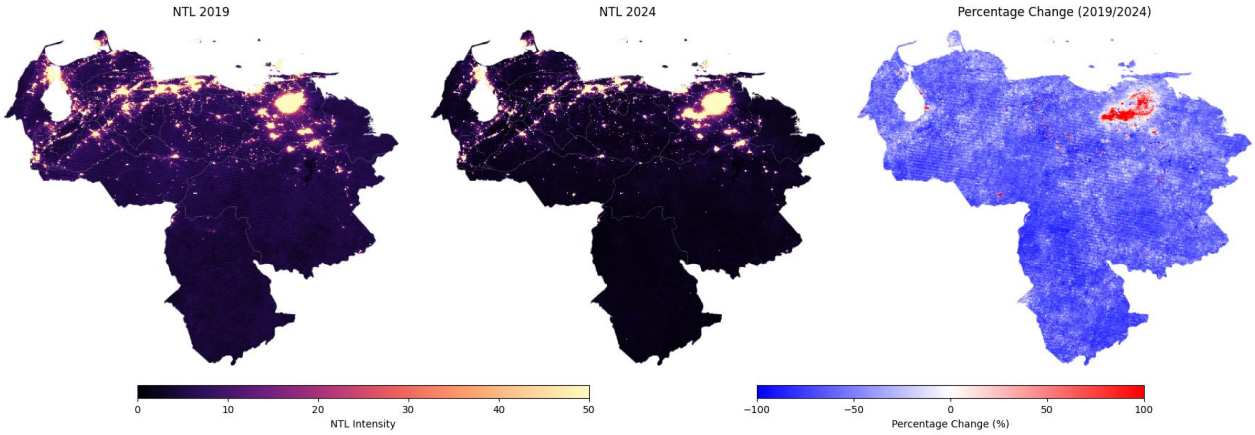
3.2.2 Nightlight Data (NTL)

One of the most widely used satellite-derived economic indicators is Nightlight data (NTL). Nightlight data measures illumination levels from artificial lights at night, which serves as a proxy for urbanization, economic development, and industrial activity. Studies such as Abdel-Latif, Badr and Maduako (2025) have shown that nightlight intensity is strongly correlated with economic activity and serves as a reliable predictor of GDP growth. As mentioned in the previous section, the relevant literature suggests that there is a sigmoid-like curve which reveals a light intensity threshold that separates unlit areas or cells with low light intensity from brighter cells in the process of constructing the luminosity map. Nightlight data is collected from satellites that capture light emissions from urban centres, roads, and industrial zones. We use NTL data primarily collected from the Visible Infrared Imaging Radiometer Suite (VIIRS-DNB). The VIIRS Day/Night Band (DNB), part of NASA's Suomi National Polar-Orbiting Partnership (Suomi NPP) satellite, has been operational since 2012. This offers higher spatial resolution (≈500 m) and improved radiometric sensitivity, making it more suitable for analysing economic activity at a finer spatial scale, monitoring power outages, disaster impact assessment, and tracking

changes in nighttime economic activity. As a result, it becomes possible to study urban expansion, industrialization, and infrastructure development.

Figure 4 presents the evolution of nightlight intensity in Venezuela from 2019 to 2024, highlighting key economic hubs. The concentration of nightlight in Monagas region, which contains Venezuela's largest gas reserves, where is significant gas flaring activity, as well as areas from and around Caracas towards Merida, illustrate the spatial distribution of industrial activity and infrastructure development. The most recent data (2024) indicate a significant decrease in activity compared to earlier data, across the country except in the major cities.

Figure 4. Average Change in Nightlight Map, 2019-2024 (percentage)



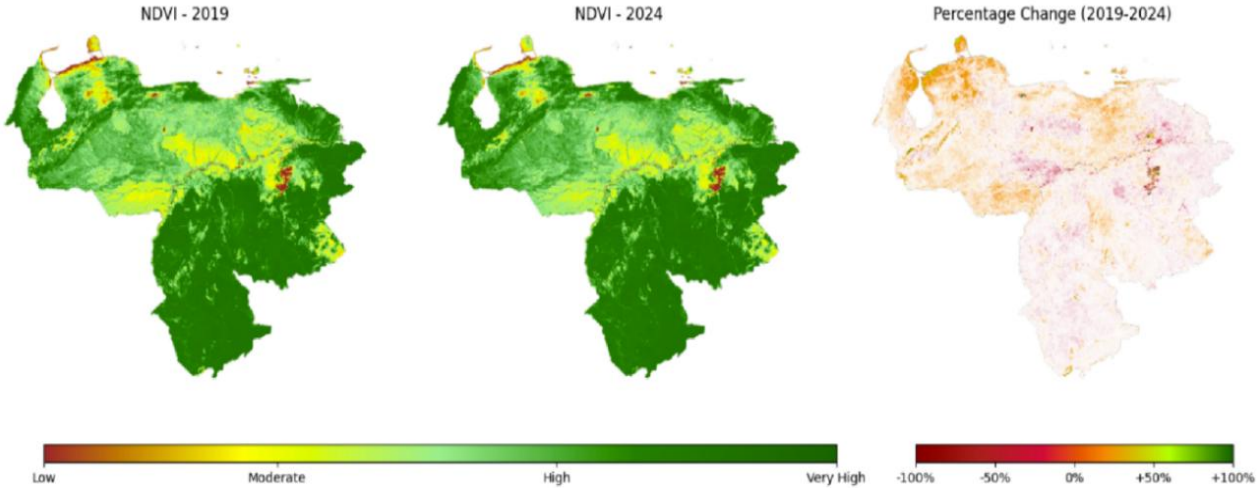
Source: Suomi NPP, authors' calculations.

3.2.3 Vegetation Indices: NDVI and EVI

Another crucial satellite-based indicator is vegetation coverage monitoring, which captures changes in land use, urbanization, and agricultural activity. We utilize two widely recognized vegetation indices, the Normalized Difference Vegetation Index (NDVI) and Enhanced Vegetation Index (EVI). The use of NDVI and EVI in the nowcasting model allows the tracking of fluctuations in agricultural output, deforestation, and land-use transformation, all of which are closely linked to economic activity.

The Normalized Difference Vegetation Index (NDVI) is the most common satellite-derived index for assessing vegetation health and density. NDVI measures the difference between near-infrared (NIR) and red light reflectance, as vegetation absorbs red light while reflecting near-infrared. NDVI is particularly useful for tracking agricultural production and deforestation trends. However, it is moderately sensitive to soil and atmospheric conditions and can become saturated in areas with dense vegetation, making it less effective in regions with high Leaf Area Index (LAI). Despite these limitations, NDVI remains one of the most widely used indicators for monitoring crop cycles, drought conditions, and land degradation, and as such could be used for prediction of crop volume and quality, as well as food sufficiency, in fragile countries where obtaining this data is crucial. NDVI for Venezuela (Figure 5), shows evidence of reduction of vegetation coverage, as gradually the spectrum of colour becomes progressively darker over time, particularly in the northern half of the country.

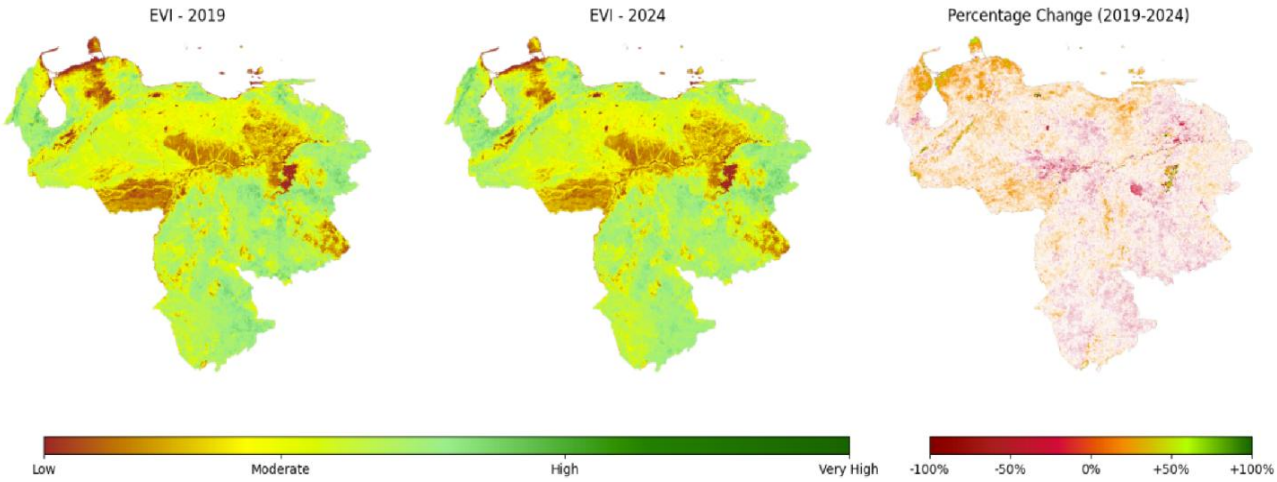
Figure 5. Average Annual Change in NDVI Map for Venezuela, 2019-2024 (percentage)



Source: Calculations authors' own.

The Enhanced Vegetation Index (EVI) is an improvement over NDVI, designed to enhance vegetation signals in areas with high biomass, while reducing sensitivity to atmospheric distortions and soil background effects. EVI has several advantages over NDVI, particularly in regions with dense vegetation, where NDVI tends to saturate. Unlike NDVI, EVI incorporates the blue spectral band, which helps correct for atmospheric distortions. EVI is widely used for monitoring forest health, agricultural productivity, and land-use changes. The EVI results for Venezuela confirm and complement the findings of NDVI on reduction of vegetation, with progressively darker colour areas in the EVI map areas across Venezuela, implying a change in the quality of land and infrastructure development, particularly in Guárico and Anzoátegui regions from 2019 to 2024 (Figure 6). Depending on specific local conditions, a decrease in vegetation could reflect expansion of urbanization, a shift away from agriculture, a poorer harvest, or deforestation. As this is not always to disentangle, EVI should be combined with other intelligence for more informed conclusions.

Figure 6. Average Annual Change in EVI Map for Venezuela, 2019-2024 (percentage)

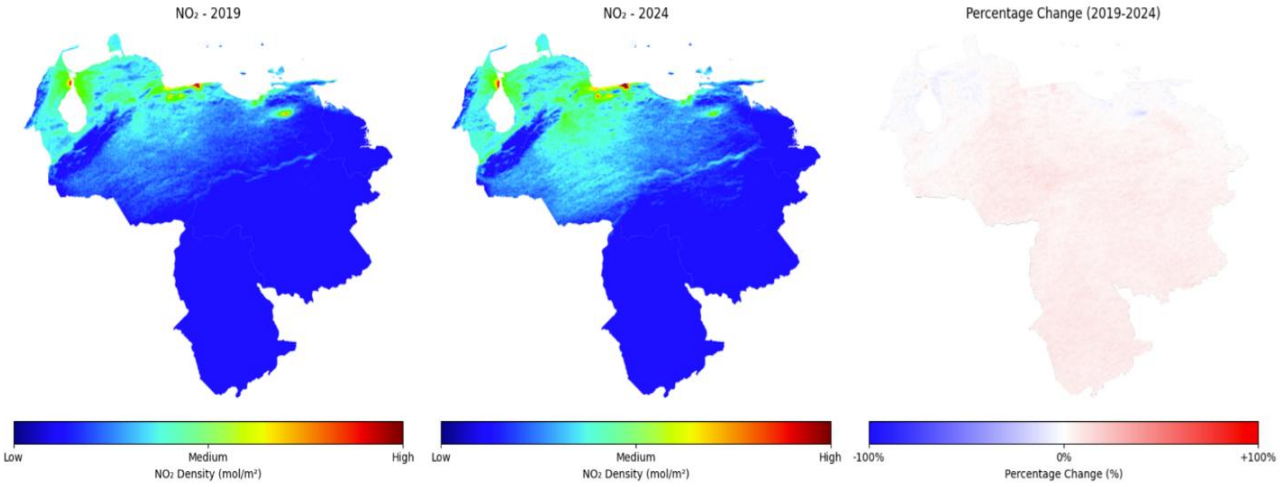


Source: Calculations authors' own.

3.2.4 Nitrogen Dioxide (NO_2) Emissions

Nitrogen dioxide (NO_2) emissions serve as a real-time proxy for industrial production, transportation activity, and energy consumption. NO_2 is primarily released from fossil fuel combustion, including emissions from power plants, vehicle exhaust, and industrial processes. Unlike nightlight, which capture economic activity at night, NO_2 data provides a daytime measure of economic intensity, making it a perfectly complementary source of information. Parubets and Naito (2025) suggest that NO_2 levels correlate well with short-term fluctuations in economic activity, whereas nightlight data are more indicative of long-term economic performance. A decline in NO_2 emissions typically reflects economic slowdowns, reduced manufacturing output, and lower energy consumption. Conversely, rising NO_2 emissions are associated with increased industrial production and economic expansion, due to an increase in emissions of pollutants, where there is lax atmospheric emissions regulation. Figure 7 depicts the evolving dynamics and NO_2 emissions in Venezuela from 2019 to 2024, which could be reflection of emission-generated human activities on the ground. The data in this case show an increase in emissions in the northern part of the country, specifically in Guárico region.

Figure 7. Average Annual Change in NO_2 Emissions Map for Venezuela, 2019-2024 (percentage)

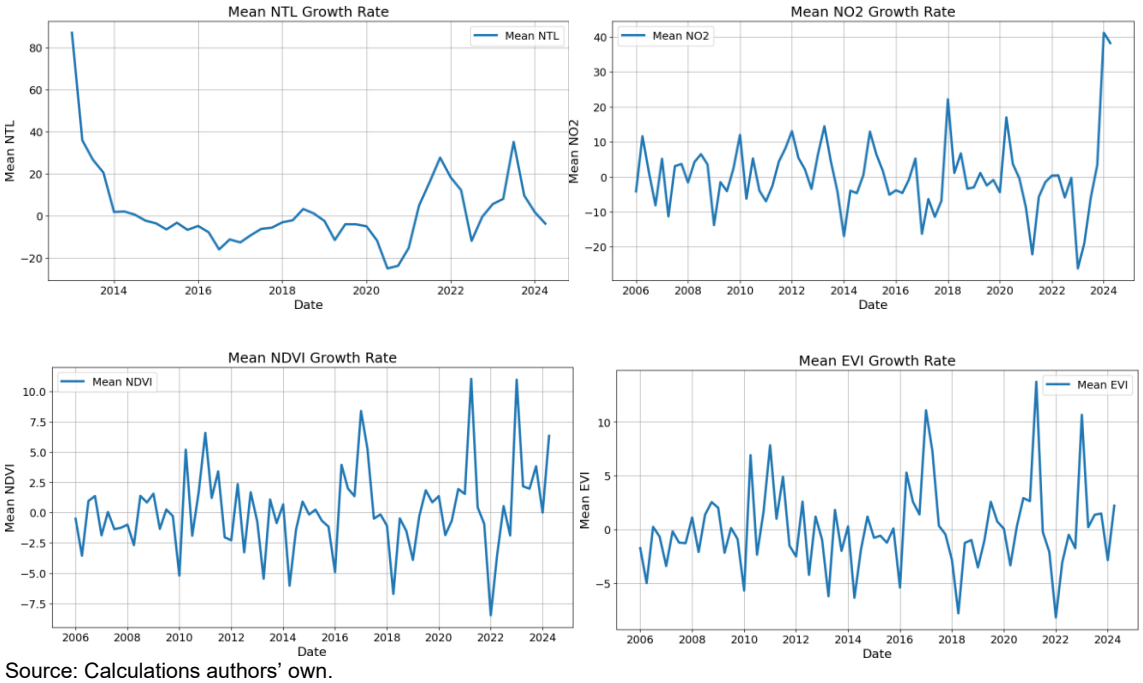


Source: Calculations authors' own.

3.2.5 Aggregation and Application in Nowcasting Models

We aggregate the monthly series to quarterly frequency and compute quarter-on-quarter (Q/Q) growth rates. The combined use of NTL, vegetation indices, and NO_2 emissions allows us to construct a comprehensive, multi-dimensional proxy for economic activity, mitigating the limitations of relying on a single data source. It is important to note how the images were converted into numerical series through a two-step process initially temporal and then spatial reduction. First, daily images were combined into monthly mosaics by creating monthly images which have the sum of the daily pixel values at each pixel. Next, each monthly image was spatially reduced to a single mean and sum by averaging and summing up all pixel values across the country respectively. We exclusively utilized level 3 corrected datasets from NASA and Google Earth Engine. These datasets have already been adjusted for factors such as cloud cover, gap-filling, forest fires, stray light, moonlight, and similar influences. Additionally, calculating the growth rate addresses the inherent seasonality present in the data series.

Figure 8. Mean Growth Rates of Night Light Data (2012-01 to 2024-03), NDVI and EVI (2003-01 to 2024-03), and NO_2 Emissions (2004-10 to 2024-03)



Source: Calculations authors' own.

3.2.6 Macroeconomic Predictors

Incorporating conventional macroeconomic series is essential for capturing different dimensions of economic activity, including demand, supply, energy production, financial sector conditions, industrial capacity, trade flows, household consumption and income dynamics, directly or by the closest proxy. The selection of these, most relevant to the case study, indicators provide critical insights into Venezuela's economic structure (resource-dependent economy) and serve as key inputs for nowcasting.

In general, in economies that are highly dependent on specific commodities/resources (as is the case of many emerging market economies and low-income economies) it is essential for an accurate nowcast to include data on these dominant commodities, at the best quality available, as illustrated in the results section. A limitation may be that "best quality available" statistics may come from non-official sources for the modelling, and as such may not always be compatible with official statistics, unless the data is produced following the same methodology; otherwise, there may be unpredicted disruptions, affecting the quality of the estimates. Including these indicators further provides a comprehensive view of economic conditions and helps improve predictive accuracy when combined with non-traditional data sources, such as satellite imagery, improving timeliness and quality of estimates.

Indicators in Table 1 are available at a monthly frequency, except for total capacity utilization, and are aggregated to quarterly frequency by computing the mean values. We then calculate seasonally adjusted quarter-on-quarter (Q/Q) growth rates for all indicators to facilitate nowcasting. The time series for the macroeconomic indicators used starts from Jan 2005 and ends in April 2024.

Table 1. Key Macroeconomic Indicators for Venezuela Nowcasting

Indicator	Unit	Description
Crude Oil Production	Millions of Barrels	Crude oil production is the single most important predictor of Venezuela's real GDP, given the country's dependence on oil exports. During the study period, Venezuela's oil production has seen a sharp decline, according to OPEC data.
Total Capacity Utilization (TCU)	Percentage	Capacity utilization measures the percentage of an economy's productive capacity that is actively used, sourced from CONINDUSTRIA. It serves as a proxy for industrial activity and manufacturing performance. In Venezuela, capacity utilization remains consistently low across most sectors, reflecting structural inefficiencies, mismanagement, and insufficient investment in infrastructure. This indicator is particularly relevant for assessing the operational status of oil refineries and the broader manufacturing sector.
Credit to the Private Sector	USD, Parallel Exchange Rate	Credit to the private sector reflects the level of domestic financial intermediation and access to financing for businesses and households. This indicator captures bank lending, trade credits, and other financial instruments. Credit availability in Venezuela has been significantly affected by hyperinflation and financial instability, though recent trends suggest a modest recovery in private-sector lending.
Monetary Aggregates	USD	Monetary aggregates are key proxies for liquidity conditions and private consumption dynamics. We use M2 and M3 (monetary base), which include currency in circulation, central bank reserves, and liquid deposits. Changes in money supply provide insights into inflationary pressures and domestic demand conditions. Given Venezuela's hyperinflationary past, shifts in monetary aggregates are crucial for assessing macroeconomic stability.
Crude Oil Exports	Millions of Barrels	Crude oil exports represent a critical driver of foreign exchange earnings and fiscal revenue. Historically, Venezuela was among the largest oil exporters, with daily crude exports exceeding 2.5 million barrels per day (mbpd). However, since 2017, exports have declined sharply, reaching a low of 300,000 barrels per day due to sanctions, declining production, and operational disruptions. Tracking export trends helps gauge Venezuela's external sector resilience and exchange rate pressures.
Government Revenue	Monthly VAT Receipts, USD	We use monthly Value-Added Tax (VAT) revenue data as a proxy for domestic economic activity and consumption trends. VAT collections provide insights into household spending patterns and informal sector activity, which remains significant in Venezuela's economy.
Gas Consumption	Cubic Feet per Capita, USD	Gas consumption serves as an indicator of economic and industrial activity. Venezuela consumes approximately 27,961 cubic feet of natural gas per capita per year (based on 2017 estimates). Gas usage reflects both industrial demand (power generation, manufacturing) and household consumption.

A correlation matrix for the variables used (traditional and non-traditional) is available in Annex II, illustrating the relationship between these two groups of variables.

4. Methodology

Machine learning (ML) techniques have increasingly been applied in macroeconomic forecasting to address data limitations and improve predictive accuracy. In this section, we present two methodologies used in our nowcasting model, namely, Random Forest (RF) and the Dynamic Factor Model (DFM). These approaches allow us to leverage both traditional and non-traditional data sources while mitigating overfitting risks and handling high-dimensional datasets. One important detail here to notice that in case of Venezuela there are severe economic episodes like hyperinflation, sanctions, and a sharp decline of GDP which can complicate nowcasting.

4.1 Random Forest

Random Forest (RF) is a supervised machine learning algorithm that operates as an ensemble learning method, building multiple decision trees and combines their outputs to enhance predictive accuracy. RF is particularly effective in economic nowcasting as it can handle nonlinear relationships, high-dimensional datasets, and missing data, making it well-suited for Venezuela's complex economic environment. RF operates by constructing multiple decision trees during training, each based on a randomly selected subset of the data (both observations and features). The final prediction is obtained by averaging the outputs of all trees, which helps to reduce variance and improve robustness. The process is the following:

1. **Bootstrap Aggregation (Bagging):** The training dataset is randomly sampled with replacement to create multiple sub-datasets for different trees.
2. **Feature Randomization:** At each split in a decision tree, a random subset of features is selected rather than considering all available features. This prevents trees from becoming overly reliant on a few dominant predictors, reducing the risk of overfitting.
3. **Tree Construction:** Decision trees are grown by recursively splitting the data at nodes, optimizing a criterion such as the Mean Squared Error (MSE) for regression problems.
4. **Ensemble Averaging:** The final prediction is obtained by taking the mean (for regression) or the majority vote (for classification) across all decision trees.

Mathematically, the RF prediction for a given input is:

$$\hat{\boldsymbol{Y}} = \frac{\mathbf{1}}{\boldsymbol{T}} \sum_{\boldsymbol{t}=\mathbf{1}}^{\boldsymbol{T}} f_t(X)$$

where $f_t(X)$ is the prediction of the *t*-th decision tree, and *T* is the total number of trees.

To optimize model performance, we tune hyperparameters, such as i) the number of trees (T), where higher values improve stability but increase computation time, ii) the maximum tree depth, which limits the complexity of each tree to prevent overfitting, iii) the minimum samples per leaf, which ensures that splits occur only when there are enough observations. Additionally, to reduce overfitting, we divide the dataset into training (90%) and testing (10%) sets, a slightly different split from the standard 80:20 approach due to the limited available data. We further apply 5-fold cross-validation within the training set to refine the model's hyperparameters.

4.2 Dynamic Factor Model

The Dynamic Factor Model (DFM) is widely used in macroeconomic nowcasting applications, where real-time data updates are incorporated into nowcasts. DFMs are particularly useful when working with high-dimensional datasets, as they reduce the dimensionality by summarizing information into a few unobserved latent factors, enabling efficient modeling of large datasets by extracting common factors rather than relying on a large number of predictors. Additionally, DFMs effectively incorporate both monthly and quarterly data, making them suitable for mixed-frequency nowcasting/forecasting tasks. Furthermore, by focusing on a small number of common factors, DFM help avoid overfitting issues that often arise when using extensive predictor sets.

It is worth noting that for Venezuela the number of available indicators is smaller than it would be regularly required, and the indicators listed above are the maximum that could be employed. The economy's heavy dependence on oil for revenue, however, allows the use of a smaller number for the model (see section 5), indicating that if a very strong predictor is identified among the indicators, it may be possible to avoid bias as critical information is not omitted. Arguably, this observation may be particularly useful in nowcasting using DFM in countries whose revenue relies heavily on specific commodities (e.g., natural resources, etc.). Formally, the DFM represents a high-dimensional time series Y_t as a function of a low-dimensional set of unobserved factors F_t, with some idiosyncratic noise ε_t:

$$Y_t = \Lambda F_t + \varepsilon_t, \quad \varepsilon_t \sim N(0, \sigma^2)$$

where:

Y_t is an N x1 vector of observed economic variables at time

F_t is an r x1 vector of latent factors.

Λ is an N x r matrix of factor loadings, capturing the influence of each factor on economic variables.

ε_t is an N x 1vector of idiosyncratic errors.

The factors F_t are estimated using Principal Component Analysis (PCA) or Kalman Filtering, allowing the model to extract the most informative signals from several macroeconomic and satellite indicators. Once factors are estimated, they are modelled dynamically using a Vector Autoregression (VAR) model:

$$F_t = \phi_1 F_{t-1} + \phi_2 F_{t-2} + \ldots + \phi_p F_{t-p} + u_t$$

where ϕ_i are autoregressive coefficients and u_t is a white noise error term.

Application of a DFM to Venezuela faces certain limitations. The DFM assumes stable economic relationships, but Venezuela's hyperinflation, policy shocks, and external sanctions introduce structural breaks that can distort factor estimates. The DFM also struggles with irregular and ragged edge missing data which is a common issue given Venezuela's lack of publicly available, regularly produced official statistics. Additionally, mixed-frequency dynamics can sometimes lead to inconsistencies when integrating monthly and quarterly data. We standardized all features to a single frequency and performed smoothed imputation. Using the DFM result as our baseline, we show how machine learning methods enhance predictive robustness.

5. Results

Preliminary results suggest that RF outperforms DFM in terms of predictive accuracy, especially when satellite data are included, most likely due to its ability to capture nonlinearities and interactions among features. However, the DFM remains useful for nowcasting when integrating mixed-frequency economic data. We also compare the forecasting performance of RF and DFM in predicting Venezuela's real GDP by evaluating out-of-sample prediction errors.

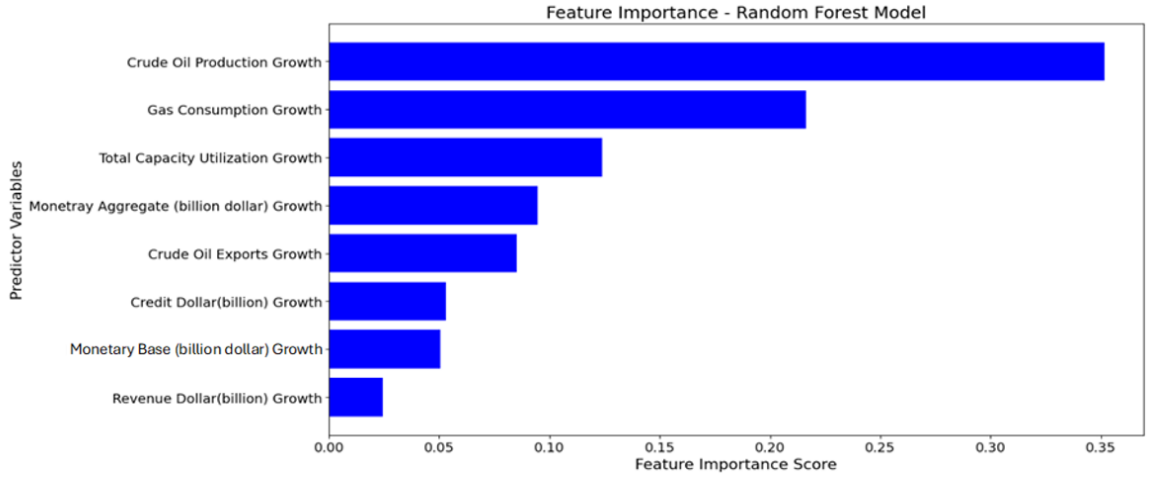
A comparison of our approach with LASSO, Ridge Regression and Elastic Net Regression, is available in Annex I. Validating machine learning models on time series data presents unique challenges due to the sequential and temporal nature of the data. Standard cross-validation approaches, which randomly shuffle and split data, can result in unrealistic performance estimates and data leakage. We used the Time Series Split technique to address these issues by structuring training and test splits in a way that respects the temporal order. We implement a rolling-origin expanding-window cross-validation scheme, in which each fold uses only past observations for training and a forward block for validation, thereby eliminating any look-ahead bias. Hyperparameter tuning is nested within this rolling scheme to avoid using validation information in model selection. Hyperparameters were selected using a greedy grid search, choosing those that minimized the error functions without causing overfitting. A function was used to monitor the convergence of error metrics (such as RMSE and MAE) across different folds (specifically, 5 folds in this case). Since real-time publication lags are not available for Venezuela, all predictors are used as observed in the dataset, and no future information enters either the training or validation windows. The models are assessed using standard performance metrics, such as the Root Mean Squared Error (RMSE). In this section, we discuss in more detail the predictive power and reliability of each approach.

5.1 Random Forest Performance

The standard process for running any ML technique (training, testing and cross validation) is a challenge in the context of large fluctuations, as in the case of the Venezuelan economy, which went into hyperinflation with a steep decline in the economy, followed by a steep recovery. For this reason, the training and testing sample is restricted to a 90:10 ratio. We first use RF over the dataset without satellite data, and then we compare the improvement of RMSE when satellite data are used.

We start with the feature importance from the RF, a measure of how much a feature/ predictor variable contributes to the accuracy of the model. Features that are ranked highly have a significant influence on the model's decision making and improving its importance**.** Unsurprisingly, crude oil production and gas consumption are the top features which can help predict the real GDP for Venezuela (Figure 9), followed by total capacity utilization (TCU) and monetary aggregates.

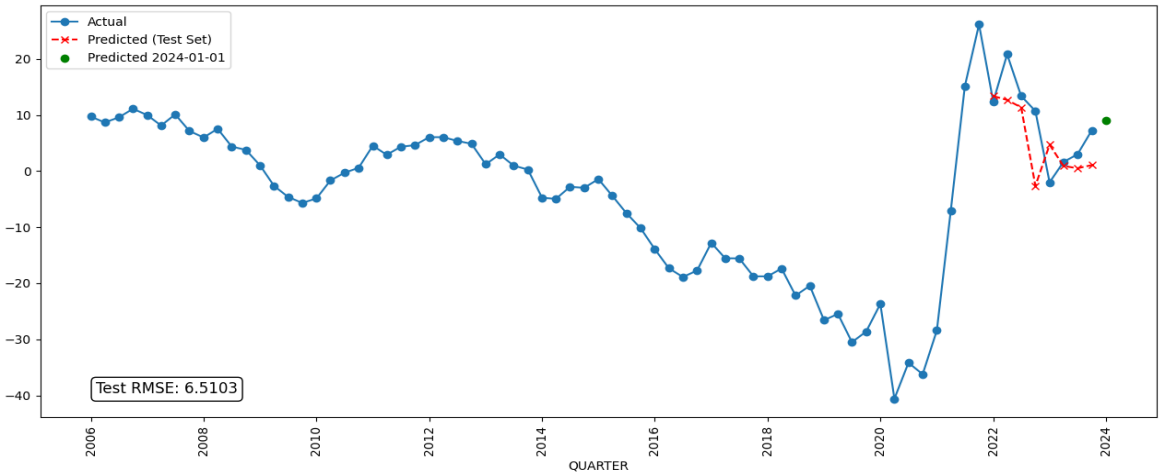
Figure 9. Feature Selection without Satellite Data



Source: Calculations authors' own.

With this in mind, we train and test the sample at a 90:10 split, where RMSE is 6.51 and predicted one-step- ahead quarterly real GDP growth forecast (2024 Q1) is 8.2 percent (Figure 10).

Figure 10. Actual vs Predicted GDP Growth Rate, Random Forest without Satellite Data

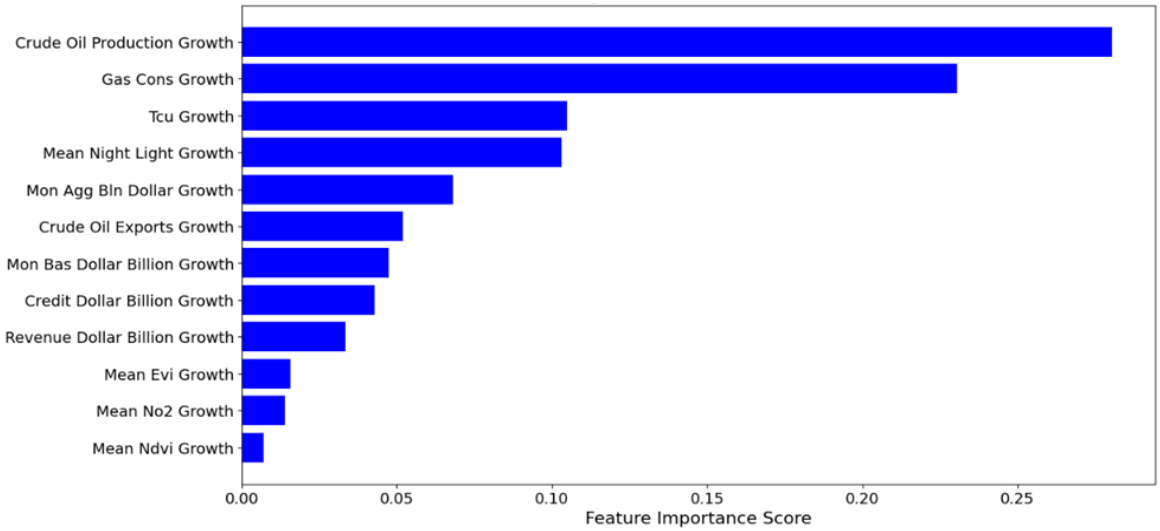


Source: Calculations authors' own.

The process is repeated with the inclusion of the satellite variables, namely, nightlight, NDVI, EVI, and NO_2 emissions. When these are included, nightlight is one of the top 4 features selected for the quarterly real GDP growth prediction (Figure 11), and the most important out of these four alternative indicators.

This finding signifies the importance and relevance of nightlight data as a good proxy for economic activity, corroborating the findings of the relevant literature, and providing evidence on its validity in emerging and developing economies. Additionally, it provides some counterweight, in a way that moderates the over-dominance of crude oil production and gas consumption growth in the estimates, despite the possible bias due to flaring that can affect the quality of some nightlight data, if there is no correction for it. The latter can be much more volatile as indicators compared to nightlight, which captures broader economic activity.

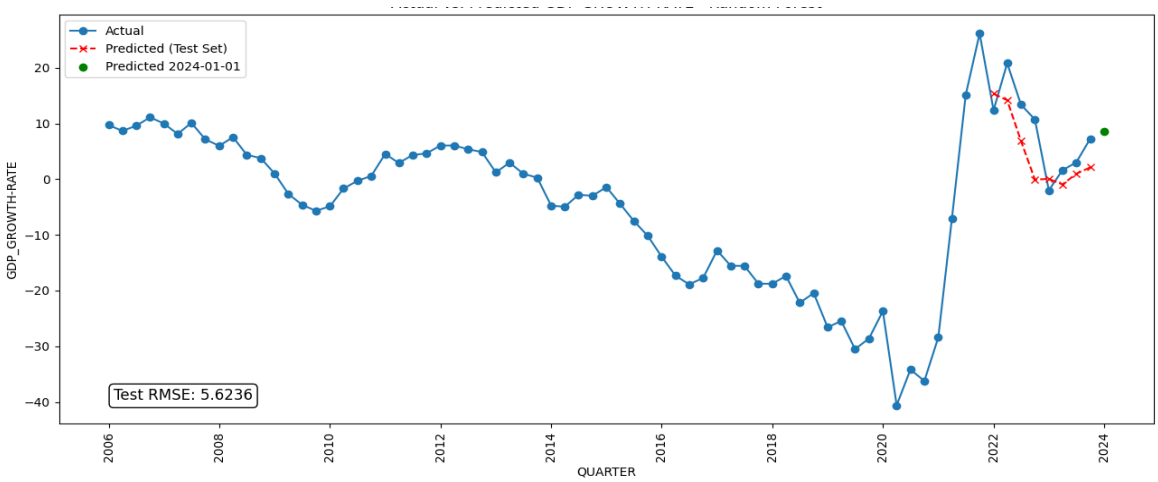
Figure 11. Feature Selection with Satellite Data



Source: Calculations authors' own.

The next step is to examine the predictions using RF with satellite data. Upon the addition of satellite data, RMSE is lower, to 5.6, but a prediction of slightly higher quarterly real GDP, at 8.5 percent of one-step-ahead real GDP growth for 2024 Q1. This result is not perfect; rather it is the lowest possible, given two limitations to test and train the model, namely that the length of the time series is limited and that there are episodes of hyperinflation during the relevant period.

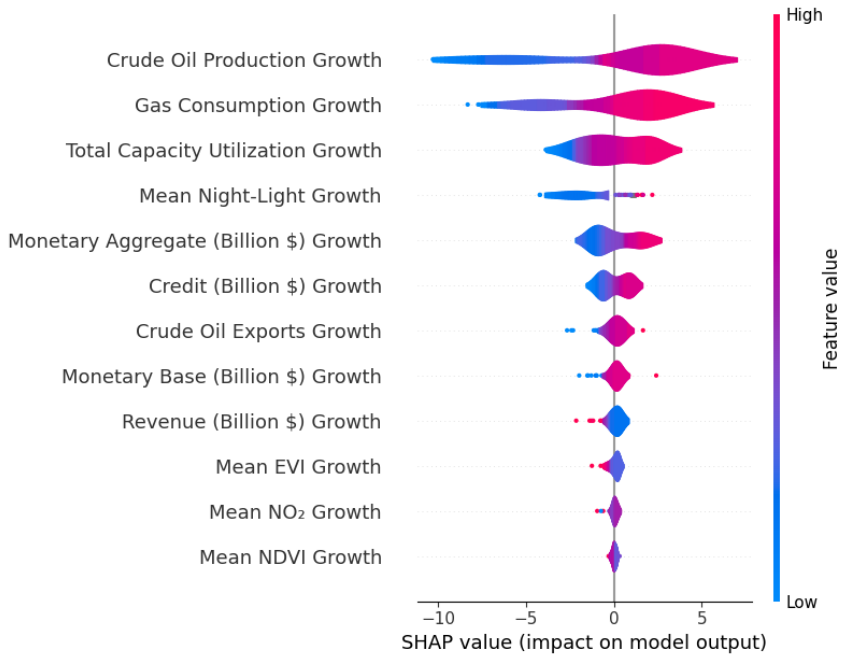
Figure 12. Actual vs Predicted GDP Growth Rate, Random Forest with Satellite Data



Source: Calculations authors' own.

To further evaluate the ML metrics, as customary, we present Shapley and Partial Dependence Plots. The Shapley plot (Figure 13) strengthens the findings that nightlight data can be a suitable proxy, though EVI, NDVI and NO_2 emissions appear to have a lesser explanatory power in this case.

Figure 13. Shapley Plot

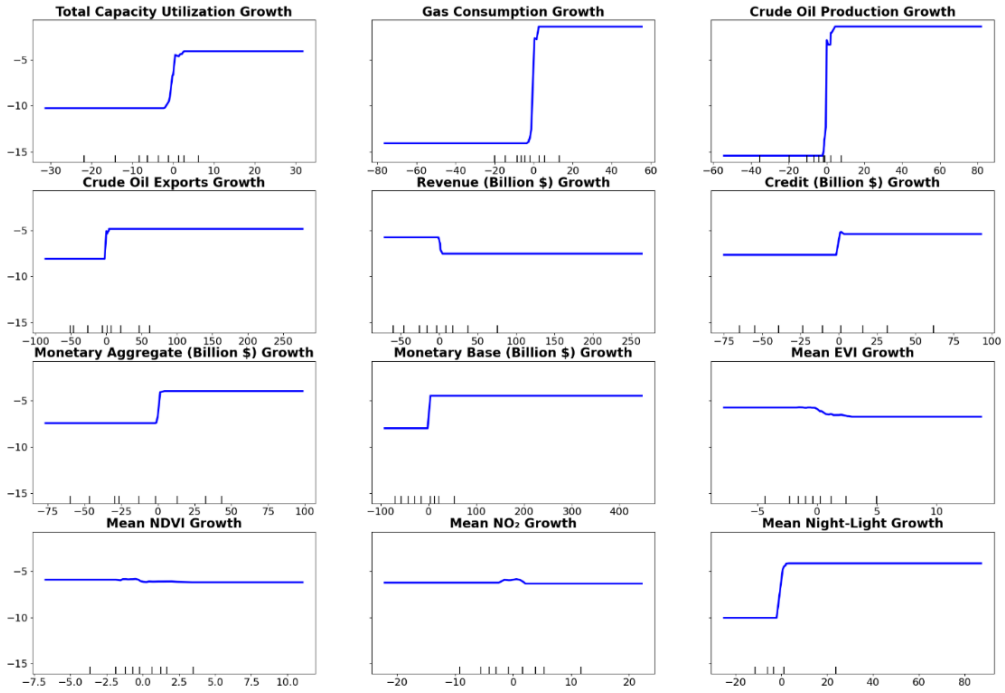


Source: Calculations authors' own.

The Partial Dependence Plot (Figure 14) also shows that nightlight data are the alternative indicator with the highest explanatory power out of the additional features from satellite sources. This finding is somewhat common in emerging market economies and developing economies for which nowcasting using satellite data has been attempted, as it generally fits expectations of a typical developmental path.

At the same time, we note that use of nightlight in cases where oil and gas production are the economic activities with the highest weight in the economy, may bias findings (albeit positively) due to flaring as part of the production, making it more evident. For these reasons, using multiple satellite indicators could help reduce the bias. The plots also clearly show signs of nonlinearity, particularly threshold effects, regime shifts, and structural breaks. This helps explain why a non-linear model like RF performs better than linear models such as DFM. Non-linear models capture complex non-linear relationships, whereas linear models generally assume constant marginal effects, smooth monotonic patterns, and no regime changes or thresholds.

Figure 14. Partial Dependence Plot for all Features

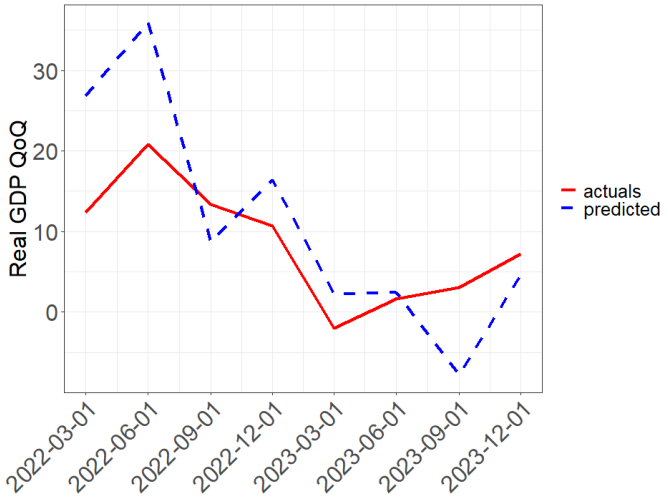


Source: Calculations authors' own.

5.2 Dynamic Factor Model

For DFM, we use as the training dataset the period from 2006 Q1 to 2021 Q4 and use the period from 2022 Q1 to 2023 Q4 as the testing dataset. For the testing dataset we get a RMSE of 8.34, which is higher than for both iterations of the RF (with and without satellite data), but lower than other ML techniques presented in Annex I (LASSO, Ridge, and Elastic Net Regression). We find a one-step-ahead forecast of quarterly real GDP growth rate of 13.2 percent (Figure 15), which is significantly higher than the results of the two RF iterations.

Figure 15. Testing Dataset for DFM, Quarterly Real GDP Growth Rate



Source: Calculations authors' own.

6. Conclusion

Closing data gaps in an economy is essential in order to understand economic activity, and, more significantly, to use for better policy design, with very tangible application for the wellbeing of people and macroeconomic stability. This is why, in this paper, we have undertaken an experiment, using data for Venezuela, a country that does not publish accurate GDP data regularly, to show how non-traditional and traditional data can be combined to produce estimates. With this endeavour we add to the corpus of work using a few different variables sourced from satellite data to improve economic activity measurement in a way that could help practically decision making, when there is data scarcity. The contribution of the paper is twofold: i) we experiment with the usage of satellite data and machine learning in a data-scarce environment, and ii) we demonstrate for the first time an application for a resource-dependent economy with data scarcity in Latin America, to deal with a significant data challenge- the regular publication of real GDP on a quarterly basis. As such, the contributions may prove useful and replicable to a degree in similar economies.

By employing satellite data (vegetation cover, night/day light and nitrogen dioxide) and machine learning techniques (random forest and dynamic factor modelling) for Venezuela up to 2024, we obtain performance improvements over traditional methods and indicators. As a result, we are confident that these alternative methods and data sources can help predict the movement of key economic variables (in our case real quarterly GDP), and consequently close data gaps, even in cases where economic activity is subject to large fluctuations, and thus less suited to traditional forecasting modelling. The application of these techniques, under specific conditions, on countries that face severe data availability challenges can be a powerful alternative, especially understanding the direction of growth of economic activity, using publicly available satellite data. The introduction of satellite data improved the RF model by 13,85% while the use of a non-linear model (in this case RF) outperformed the popular DFM model by reducing the RMSE by 32.85%, which underscores the significance of this contribution.

This does not mean that these data sources and techniques do not have limitations; caution needs to be exercised when interpreting results, and these need to be complemented with case-specific knowledge and regular ground truthing. Additionally, when the benchmark includes traditional, non-official data that are not always available, such as those from private sources, nowcasting using satellite data will also suffer from similar malaise as nowcasting using traditional official data – namely, not being able to update the model with the necessary frequency for economic decision making, e.g. on a monthly or quarterly basis. Similarly, when non-traditional data are not accessible due to changes in technology, for example, the model may not be performing at potential. For these reasons, these experimental data sources and techniques cannot replace official statistics, and at the current stage these are largely used primarily to provide insights for policy.

A potential extension researchers could consider is exploring the integration of additional satellite indicators in such models, such as the Agricultural Stress Index (ASI) and Vegetation Health Index (VHI), to better proxy the agricultural sector beyond what NDVI and EVI provide. Moreover, improving the model with additional feature engineering, such as including date information linked to business cycles, may enhance its ability to capture trends and seasonal variations, and possibly deal better with episodes such hyperinflation. This, in turn, could help the model differentiate between years and identify specific patterns or anomalies, leading to more precise predictions. Finally, fine-tuning hyperparameters using different time-series cross-validation methods, compering fixed, rolling and expanding windows may be necessary for further accuracy.

Bibliography

Abdel-Latif, H., Badr, and Maduako, I. (forthcoming). Nowcasting Economic Activity in Afghanistan. IMF Working Paper.

Akbal, O. F., Choi, S. M., Narita, F., and Yao, J. (2023). *Panel nowcasting for countries whose quarterly GDPs are unavailable* (IMF Working Paper No. 23/158)

Andreou, E., Ghysels, E., and Kourtellos, A. (2008). Regression models with mixed sampling frequencies. *Journal of Econometrics, 158*(1), 204-219. https://doi.org/10.1016/j.jeconom.2010.01.004

Arslanalp, S., Choi, S. M., Kamali, P., Koepke, R., McKetty, M., Ruta, M., Saraiva, M., Sozzi, A., and Verschuur, J. (2025). *Nowcasting global trade from space*. IMF Working Paper 25/93

Arslanalp, S., Koepke, K. and Verschuur, J. (2021). Tracking Trade from Space: An Application to Pacific Island Countries. IMF Working Paper 21/225.

Bańbura, M., and Rünstler, G. (2011). A look into the factor model black box: Publication lags and the role of hard and soft data in forecasting GDP. *International Journal of Forecasting, 27*(2), 333-346. https://doi.org/10.1016/j.ijforecast.2010.01.011

Banbura, M., Giannone, D., Reichlin, L., Clements, M., and Hendry, D. (2011). Oxford handbook on economic forecasting.

Barnett, W. A., Chauvet, M., and Leiva-Leon, D. (2016). Real-time nowcasting of nominal GDP with structural breaks. Journal of Econometrics, 191(2):312–324.

Beyer, R. C., Franco-Bedoya, S., and Galdo, V. (2021). Examining the economic impact of COVID-19 in India through daily electricity consumption and nighttime light intensity. *World Development*, *140*, 105287.

Bichler, R., and Bittner, M. (2022). Comparison between economic growth and satellite-based measurements of NO2 pollution over northern Italy. *Atmospheric Environment*, *272*, 118948.

Bichler, R., Schönebeck, S. S., and Bittner, M. (2023). Observing decoupling processes of NO2 pollution and GDP growth based on satellite observations for Los Angeles and Tokyo. *Atmospheric Environment*, *310*, 119968.

Bok, B., Caratelli, D., Giannone, D., Sbordone, A., and Tambalotti, A. (2018). Macroeconomic nowcasting and forecasting with big data. *Federal Reserve Bank of New York Staff Report No. 830*.

Bolivar, O. (2024). GDP nowcasting: A machine learning and remote sensing data-based approach for Bolivia. Latin American Journal of Central Banking, 5(3):100126.

Cashin, P., Han, F., Sabuga, I., Xie, J., and Zhang, F. (2025). *Parameter proliferation in nowcasting: Issues and approaches—An application to nowcasting China's real GDP* (IMF Working Paper No. 2025/217). International Monetary Fund. https://doi.org/10.5089/9798229027212.001

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

Dauphin, J.-F., Dybczak, K., Maneely, M., Taheri Sanjani, M., Suphaphiphat, N., Wang, Y., and Zhang, H. (2022). *Nowcasting GDP: A scalable approach using DFM, machine learning and novel data, applied to European economies* (IMF Working Paper No. 2022/052). International Monetary Fund. https://doi.org/10.5089/9798400204425.001

De Valk, S., de Mattos, D., and Ferreira, P. (2019). Nowcasting: An r package for predicting economic variables using dynamic factor models. The R Journal, 11(1):230– 244.

Doll, C. N., Muller, J. P., and Morley, J. G. (2006). Mapping regional economic activity from night-time light satellite imagery. *Ecological Economics*, *57*(1), 75-92.

Dutta, P., Paul, S., and Kumar, A. (2021). Comparative analysis of various supervised machine learning techniques for diagnosis of covid-19. In Electronic devices, circuits, and systems for biomedical applications, pages 521–540. Elsevier.

Elvidge, C. D., Ghosh, T., Hsu, F. C., Zhizhin, M., and Bazilian, M. (2020). The dimming of lights in China during the COVID-19 pandemic. *Remote Sensing*, *12*(17), 2851.

Ezran, I., Morris, S. D., Rama, M., and Riera-Crichton, D. (2023). Measuring global economic activity using air pollution. World Bank.

Foroni, C., and Marcellino, M. (2014). Unrestricted mixed data sampling (MIDAS): MIDAS regressions with unrestricted lag polynomials. *Journal of the Royal Statistical Society: Series A, 178*(1), 57-82.

Garcia, M. G., Medeiros, M. C., and Vasconcelos, G. F. (2017). Real-time inflation forecasting with high-dimensional models: The case of Brazil. International Journal of Forecasting, 33(3):679–693.

Gawthorpe, K. (2021). Random forest as a model for Czech forecasting. Prague Economic Papers, 30(3):336–357.

Ghosh, T., Powell, R. L., Elvidge, C. D., Baugh, K. E., Sutton, P. C., and Anderson, S. (2010). Shedding light on the global distribution of economic activity. *The Open Geography Journal*, *3*(1), 147-160.

Giannone, D., Reichlin, L., and Small, D. (2008). Nowcasting: The real-time informational content of macroeconomic data. *Journal of Monetary Economics, 55*(4), 665-676. https://doi.org/10.1016/j.jmoneco.2008.05.010

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

Hindrayanto, A., Swanson, N. R., and van Dijk, D. (2016). Nowcasting GDP using machine learning methods. *AStA Advances in Statistical Analysis, 109*(1), 1-24.

Jansen, W., Pick, A. and de Winter, J. (2016). Nowcasting GDP using machine learning methods. *De Nederlandsche Bank Working Paper No. 754*.

Kuzin, V., Marcellino, M., and Schumacher, C. (2011). MIDAS vs. mixed-frequency VAR: Nowcasting GDP in the euro area. *International Journal of Forecasting, 27*(4), 529-542.

Li, X., Zhou, Y., Zhao, M., and Zhao, X. (2020). A harmonized global nighttime light dataset 1992–2018. *Scientific data*, *7*(1), 168.

Lundberg, S., and Lee, S. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems, 30*. https://arxiv.org/abs/1705.07874.

Manuelito, S. (2017). The use of high-frequency indicators in short-term forecasting models: The case of Latin American and Caribbean countries.

Marcellino, M., and Schumacher, C. (2010). Factor MIDAS for nowcasting and forecasting with ragged-edge data: A model comparison for German GDP. *Oxford Bulletin of Economics and Statistics, 72*(4), 518-550. https://doi.org/10.1111/j.1468-0084.2010.00591.x.

Michalopoulos, S., and Papaioannou, E. (2013). Pre-colonial ethnic institutions and contemporary African development. *Econometrica*, *81*(1), 113-152.

McSharry, P. and Mawejje, J. (2024). Estimating urban GDP growth using nighttime lights and machine learning techniques in data poor environments: The case of south sudan. Technological Forecasting and Social Change, 203:123399.

McSharry, P., and Mawejje, J. (2024). Estimating urban GDP growth using nighttime lights and machine learning techniques in data-poor environments: The case of South Sudan. *Technological Forecasting & Social Change, 203*, 123399. https://doi.org/10.1016/j.techfore.2024.123399

Nordhaus, W., and Chen, X. (2015). A sharper image? Estimates of the precision of nighttime lights as a proxy for economic statistics. *Journal of Economic Geography*, *15*(1), 217-246.

Parubets, S., and Naito, H. (2025). *Predicting Economic Activity Using Atmospheric NO2 Satellite Data: Evidence from Local Economic Indicators in Japan* (No. 2025-002). Faculty of Humanities and Social Sciences, University of Tsukuba.

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

Puttanapong, N., Prasertsoong, N., and Peechapat, W. (2023). Predicting provincial gross domestic product using satellite data and machine learning methods: A case study of Thailand. Asian Development Review, 40(02):39–85.

Roberts, M. (2021). Tracking economic activity in response to the COVID-19 crisis using nighttime lights–The case of Morocco. *Development Engineering*, *6*, 100067.

Schnorrenberger, R., Schmidt, A., and Moura, G. V. Inflation now- casting in persistently high inflation environments.

Sneiderman, R. (2020). From linear regression to ridge regression, the lasso, and the elastic net. en. In: Medium (Nov. 2020). url: https://towardsdatascience. com/from- linear-regression-to-ridge-regression-the-lasso-andthe-elastic-net-4eaecaf5f7e6 (visited on 02/26/2022).

Štrumbelj, E., and Kononenko, I. (2014). Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems, 41*(3), 647-665. https://doi.org/10.1007/s10115-013-0679-x.

Sutton, P. C., and Costanza, R. (2002). Global estimates of market and non-market values derived from nighttime satellite imagery, land cover, and ecosystem service valuation. *Ecological economics*, *41*(3), 509-527.

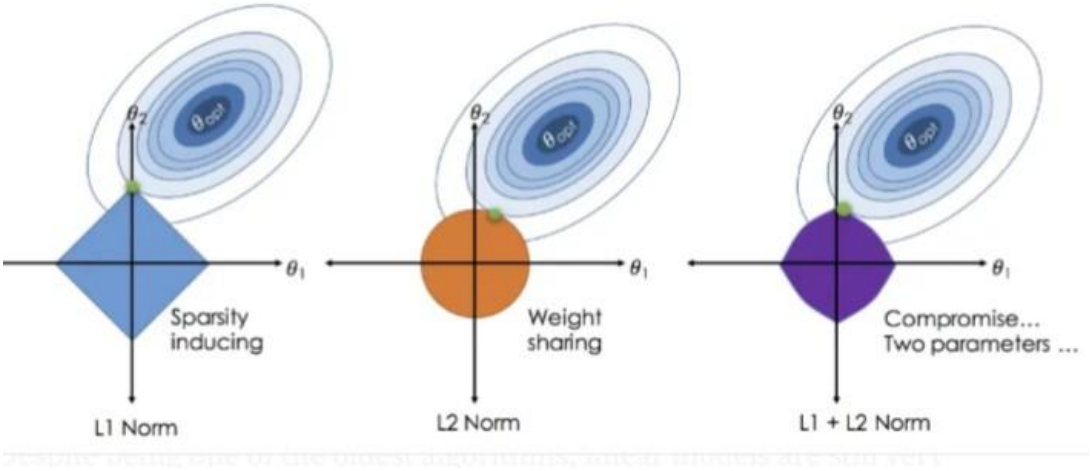
Tiffin, M. A. (2016). Seeing in the dark: A machine-learning approach to nowcasting in Lebanon. (IMF Working Paper No. 2016/056). International Monetary Fund. https://doi.org/10.5089/9781513568089.001

Zhao, M., Cheng, C., Zhou, Y., Li, X., Shen, S., and Song, C. (2021). A global dataset of annual urban extents (1992–2020) from harmonized nighttime lights. *Earth System Science Data Discussions*, *2021*, 1-25.

Annex I. Comparison of Regression Techniques

In Annex I we provide an overview of three widely used regularized regression techniques: LASSO, Ridge, and Elastic Net Regression. These methods introduce penalty terms to traditional Ordinary Least Squares (OLS) regression, addressing issues of overfitting, multicollinearity, and high-dimensionality. We further present the results of our out-of-sample performance evaluation, comparing the Root Mean Squared Error (RMSE) across different models, where lower RMSE indicates better predictive accuracy.

Figure 16. From Linear Regression to Ridge Regression, the Lasso, and the Elastic Net (14)



Source: Calculations authors' own.

A.1 LASSO Regression (L1 Regularization)

The Least Absolute Shrinkage and Selection Operator (LASSO) is a regularized regression technique that applies L1 regularization to a linear model. The primary objective of LASSO is to balance model simplicity and accuracy by introducing a penalty term that encourages sparsity, meaning that some regression coefficients are forced to exactly zero. This makes LASSO particularly useful for feature selection, as it retains only the most relevant predictors.

LASSO modifies the traditional OLS regression by adding an L1 penalty, which constrains the sum of the absolute values of regression coefficients. The optimization problem is given by:

$$min_{\{\beta_0,\beta_1\}} \sum_{i=1}^{n} (y_i - \beta_0 - X_i\beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

where:

y_i is the dependent variable.

X_i is the vector of independent variables.

β_0 is the intercept.

β_j are the regression coefficients.

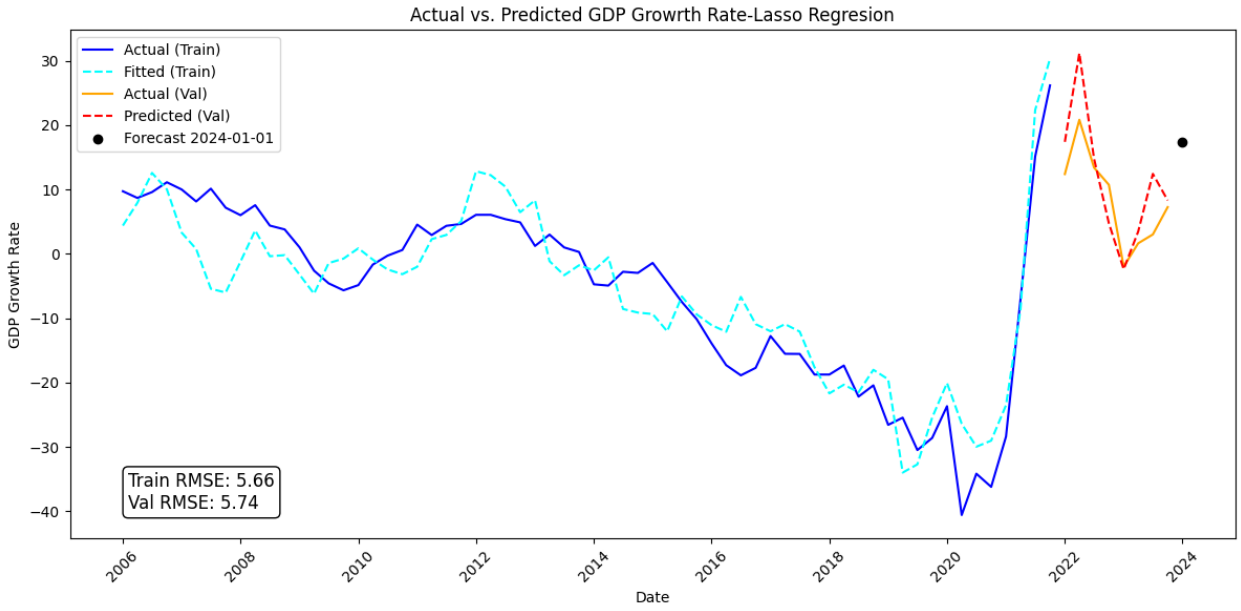
λ is the tuning parameter, controlling the degree of shrinkage.

As λ increases, more coefficients are shrunk to zero, effectively removing non-informative predictors.

LASSO selects the most relevant variables by shrinking irrelevant coefficients to zero, which helps in feature selection. It penalizes large coefficients, thereby reducing overfitting and improving model generalization. By eliminating redundant predictors, LASSO also improves the interpretability of the model. Additionally, it is suitable for handling high-dimensional data, especially when the number of predictors exceeds the number of observations.

Performance Evaluation (Out-of-Sample RMSE)

Figure 17. LASSO Regression: actual vs predicted real GDP growth rate



Source: Calculations authors' own.

A.2 Ridge Regression

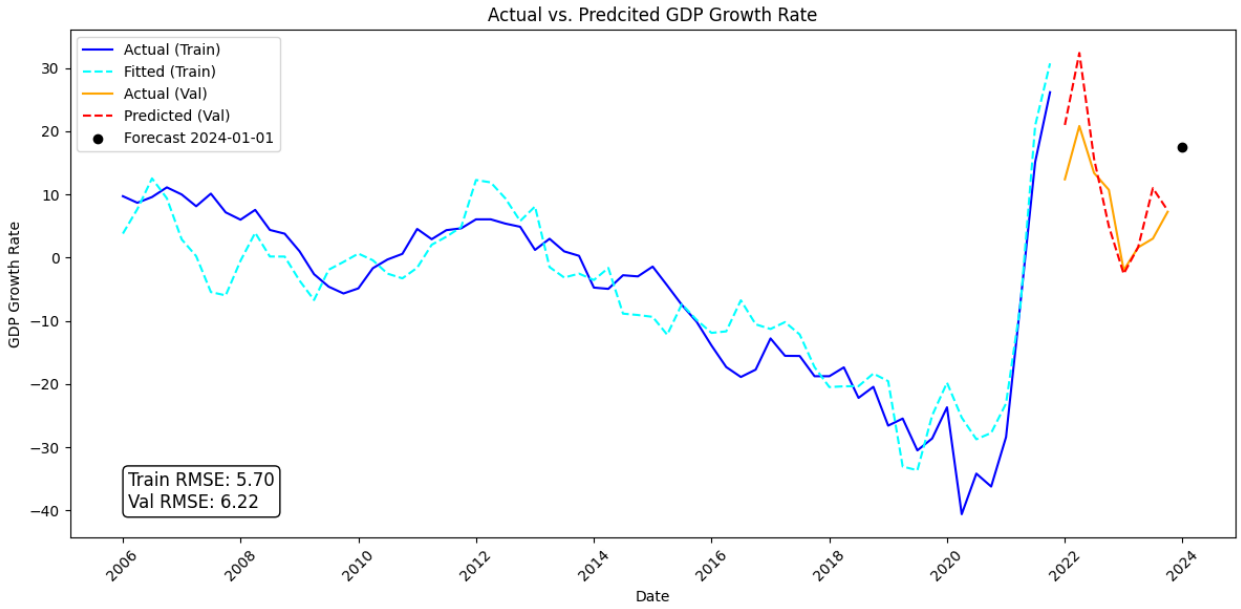
Ridge regression is an L2-regularized regression model designed to address issues of overfitting and multicollinearity in high-dimensional datasets. Unlike LASSO, which enforces sparsity, Ridge shrinks all regression coefficients toward zero without eliminating any variables entirely. Ridge regression modifies OLS by introducing an L2 penalty, which constrains the sum of squared coefficients.
The optimization problem is:

$$min_{\{\beta_0,\beta_1\}} \sum_{i=1}^{n} (y_i - \beta_0 - X_i\beta)^2 + \lambda \sum_{j=1}^{p} {\beta_j}^2$$

where λ controls the extent of regularization. Unlike LASSO, Ridge regression does not shrink coefficients to exactly zero but instead reduces their magnitude, ensuring that all variables contribute to the model. Ridge regression offers several key benefits. It helps in reducing overfitting, thereby preventing excessive variance in model predictions. Additionally, it effectively handles multicollinearity, making it useful when predictors are highly correlated. Unlike LASSO, Ridge regression retains all predictors, ensuring that no variables are entirely eliminated from the model.

Performance Evaluation (Out-of-Sample RMSE)

Figure 18. Ridge Regression: actual vs predicted real GDP growth rate



Source: Calculations authors' own.

A3. Elastic Net Regression

Elastic Net Regression is a hybrid approach that combines both LASSO (L1) and Ridge (L2) regularization. It balances variable selection and coefficient shrinkage, making it particularly effective when dealing with high-dimensional and highly correlated predictors. Elastic Net introduces two penalty terms: L1 (absolute sum of coefficients) and L2 (squared sum of coefficients).

The optimization function is:

$$min_{\{\beta_0,\beta_1\}} \sum_{i=1}^{n}(y_i - \beta_0 - X_i\beta)^2 + \lambda_1\sum_{j=1}^{p}|\beta_j| + \lambda_2\sum_{j=1}^{p}{\beta_j}^2$$

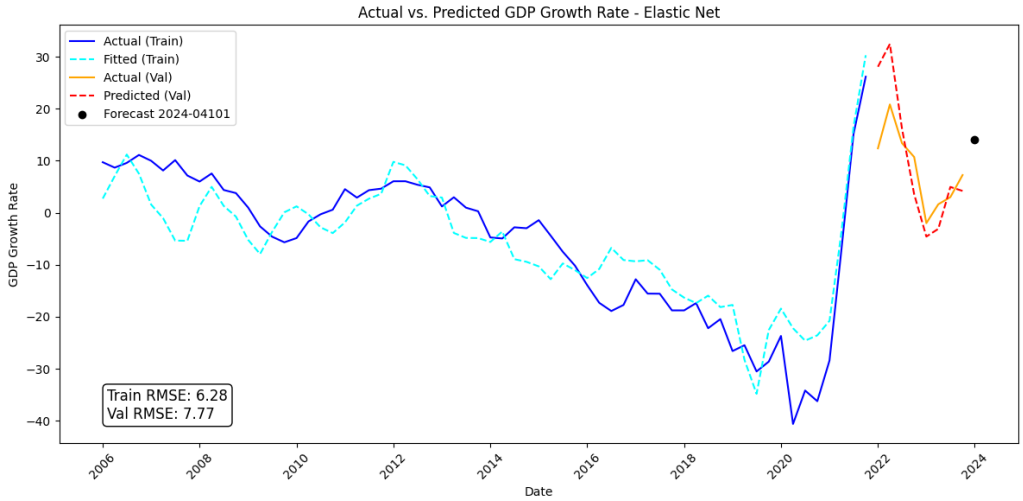
where λ_1 controls L1 regularization (LASSO) and λ_2 controls L2 regularization (Ridge).

Elastic Net has the flexibility to adjust between LASSO and Ridge:
• If $\lambda_1 = 0$, the model simplifies to Ridge Regression.
• If $\lambda_2 = 0$, the model simplifies to LASSO Regression.

Elastic Net Regression has key properties that make it a valuable tool in statistical modeling. It balances feature selection and shrinkage, retaining the benefits of LASSO's feature selection while reducing Ridge's over-penalization. The method handles multicollinearity effectively, performing better than LASSO when predictors are highly correlated. Additionally, it is stable in high-dimensional data, making it particularly useful when the number of predictors exceeds the number of observations.

Performance Evaluation (Out-of-Sample RMSE)

Figure 19. Elastic Net Regression: actual vs predicted real GDP growth rate



Source: Calculations authors' own.

A.4 Summary and Model Comparison

The table below summarizes the **out-of-sample performance** of LASSO, Ridge, and Elastic Net regression models based on **Root Mean Squared Error (RMSE)**, where lower values indicate better predictive performance.

Method	Test RMSE
Lasso Regression	5.74
Ridge Regression	6.22
Elastic Net	7.7

The model comparison reveals key insights into the predictive performance of LASSO, Ridge, and Elastic Net regression models. LASSO Regression, with the lowest RMSE, indicates superior predictive capabilities while selecting the most relevant variables. On the other hand, Ridge Regression performs the worst, suggesting that coefficient shrinkage without feature selection leads to poorer generalization. Elastic Net strikes a balance between Ridge and LASSO, but its RMSE is slightly higher than LASSO, indicating that pure feature selection is more effective in this case.

Annex II. Additional Figures

In this section, we present all the aggregated series used in the final analysis. For each macroeconomic indicator, we aggregate monthly data to quarterly frequency by computing the mean values. The resulting series are transformed into quarter-on-quarter (Q/Q) growth rates to ensure consistency in forecasting models. The selected macroeconomic variables provide a comprehensive view of Venezuela's economic conditions and help improve predictive accuracy when combined with non-traditional data sources, such as satellite imagery. These variables capture key dimensions of economic activity, including energy production, financial sector conditions, industrial capacity, trade flows, and household consumption.

Figure 20. Macroeconomic Series

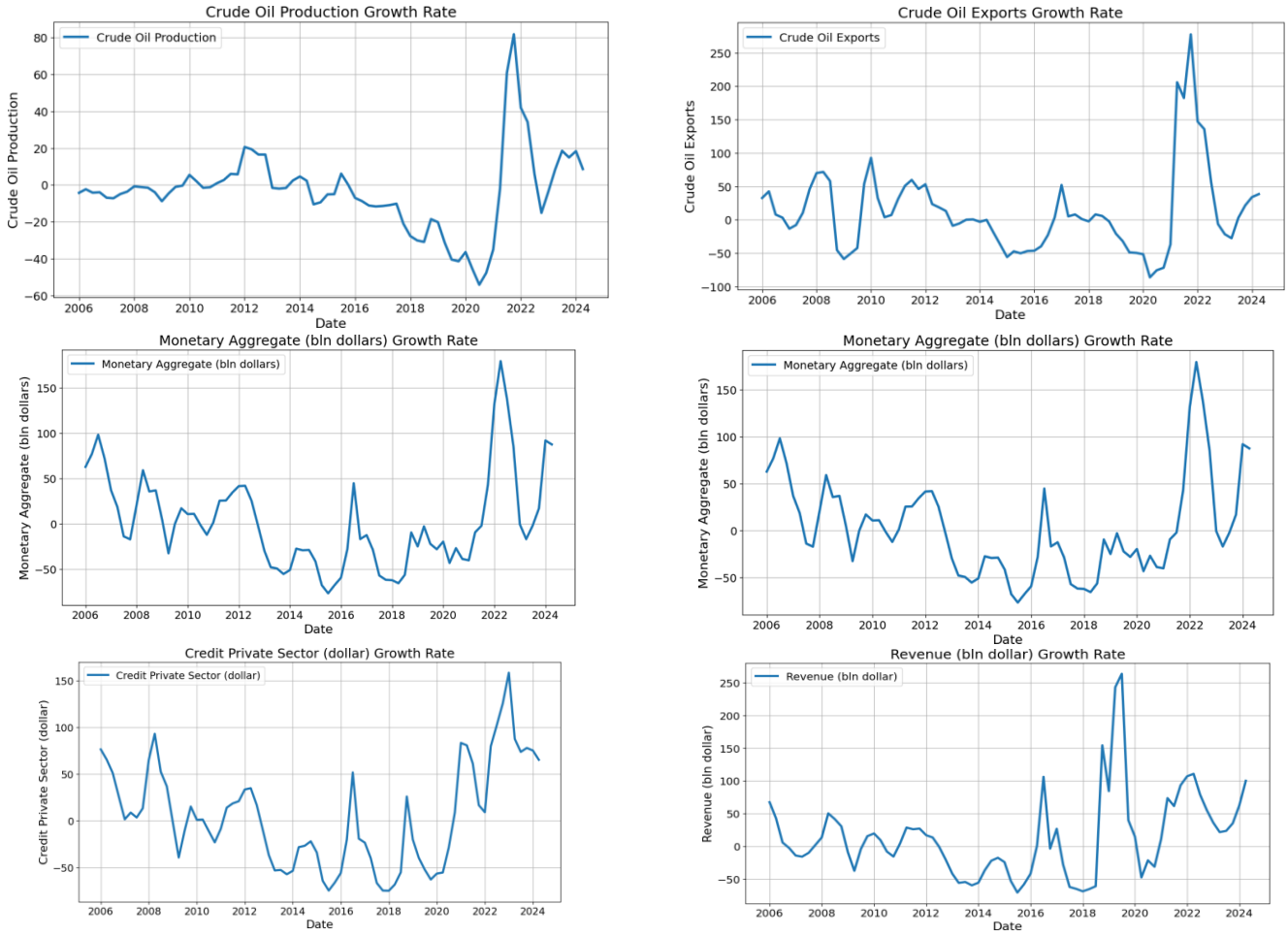


Figure 21. Satellite Data

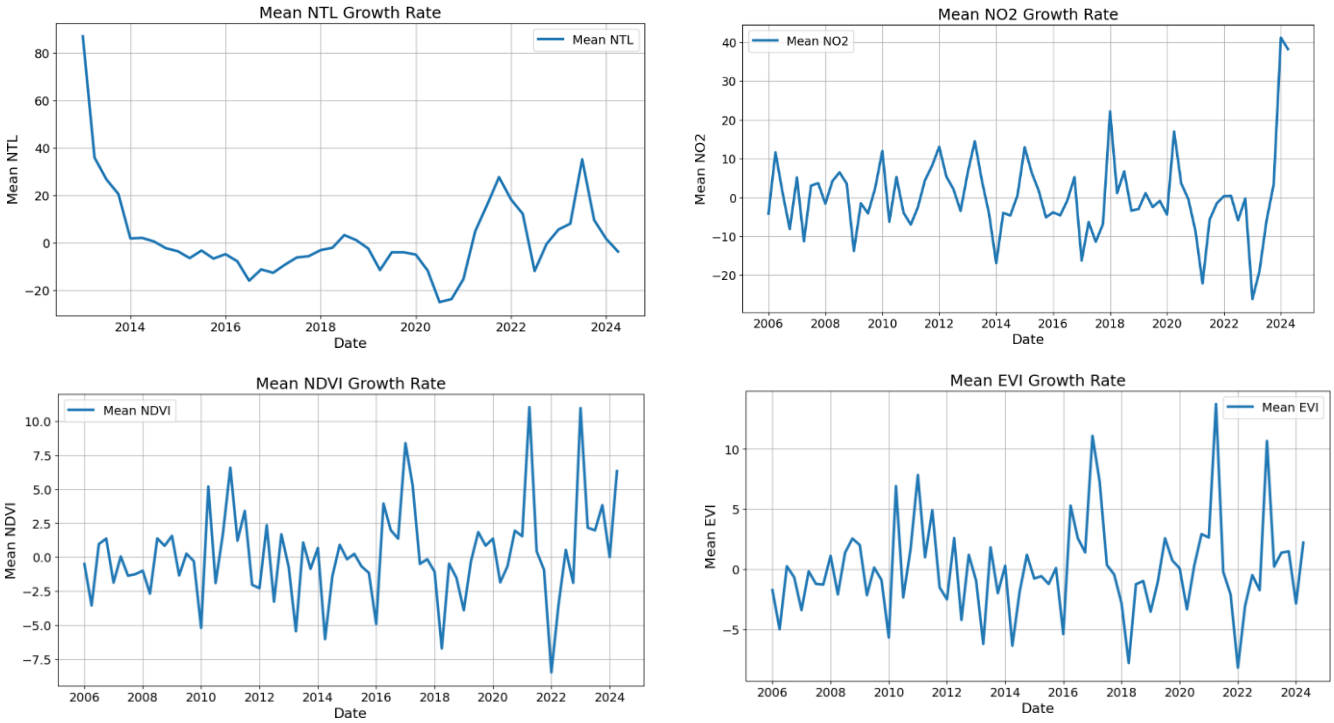
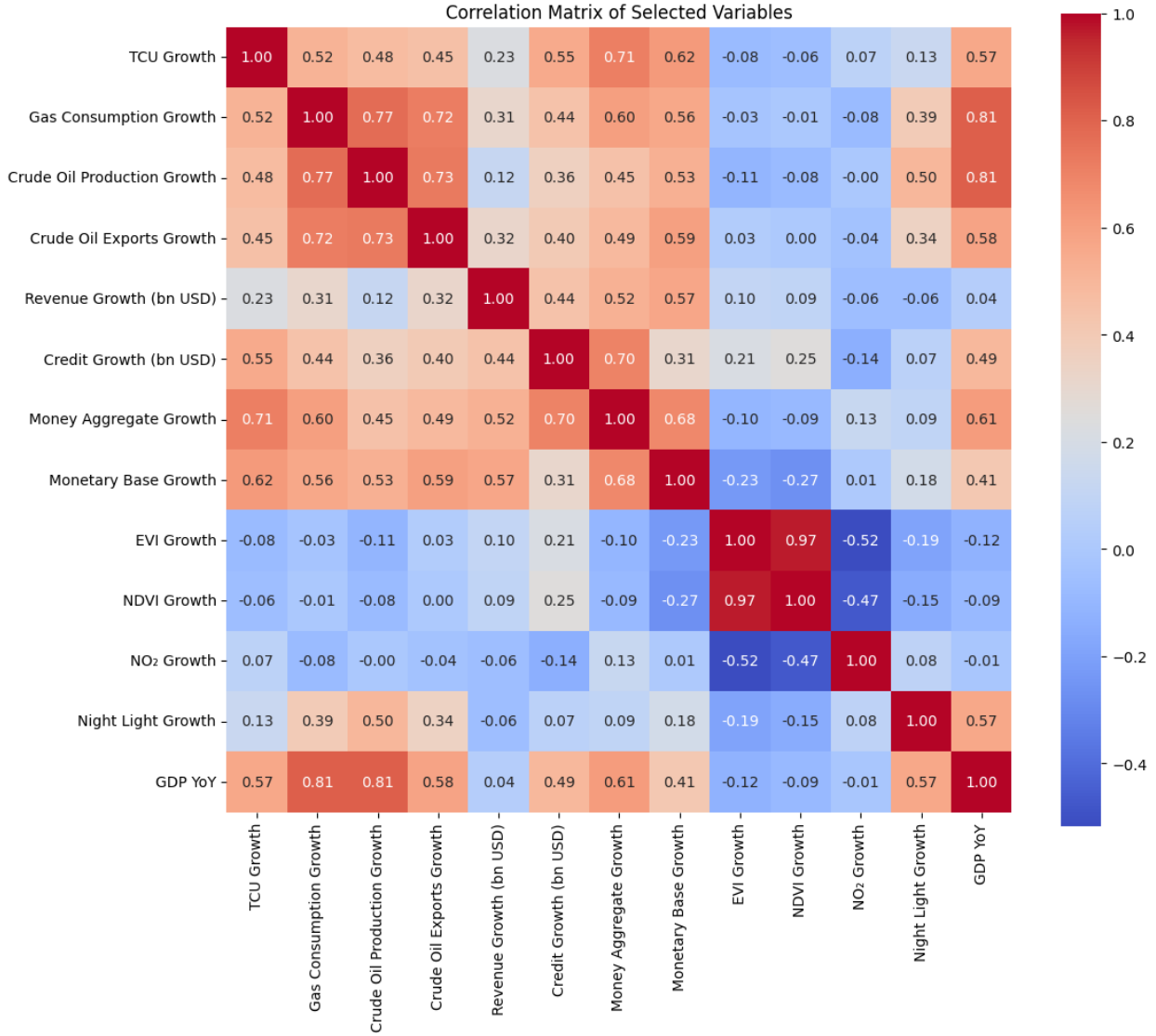


Figure 22. Correlation Matrix of Selected Variables



PUBLICATIONS

Nowcasting Economic Growth with Machine Learning and Satellite Data

Working Paper No. [WP/26/20]