Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Using satellite images of nighttime lights to predict the economic impact of COVID-19 in India

Nataraj Dasgupta

Imperial College Business School, Imperial College London, London SW7 2BU, UK

Received 21 March 2021; received in revised form 9 May 2022; accepted 18 May 2022

Available online 24 May 2022

Abstract

The outbreak of COVID-19 in early 2020 heralded a deep global recession not seen since the Second World War. With entire nations in lockdown, burgeoning economies of countries like India plunged into a downward spiral. The conventional instruments of estimating the short-term economic impact of a pandemic is limited, and as a result, it is challenging to implement timely monetary policies to mitigate the financial impact of such unforeseen events. This study investigates the promise of using nighttime images of lights on Earth, also known as nightlight (NTL), captured by the Visible Infrared Imaging Radiometer Suite (VIIRS) instrumentation onboard the Suomi National Polar-Orbiting Partnership (Suomi NPP) satellite mission to measure the economic cost of the pandemic in India. First, a novel data processing framework was developed for a recently released radiance dataset known as VNP46A1, part of NASA’s Black Marble suite of NTL products. Second, the elasticity of nightlight to India’s National Gross Domestic Product (GDP) was estimated using panel regression followed by machine learning to predict the Year-over-Year (YoY) change in GDP during Fiscal Year (FY) 2020Q1 (Apr-Jun, 2020). Electricity consumption, known to closely track economic output and precipitation were included as additional features to improve model performance. A strong relationship between both electricity usage and nightlight to GDP was observed. The model predicted a YoY contraction of 24% in FY2020Q1, almost identical to the official GDP decline of 23.9% later announced by the Indian Government. Based on the findings, the study concludes that nightlight along with electricity usage can be invaluable proxies for estimating the cost of short-term supply–demand shocks such as COVID-19, and should be explored further.

© 2022 COSPAR. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Keywords: Night lights; Machine learning; COVID-19; Satellite data

1. Introduction

The outbreak of the novel coronavirus disease (COVID-19) in 2020, claimed millions of lives and plunged parts of the world into the worst economic recession since the Second World War. While economies rebounded, understanding the extent of the economic impact was challenging. Estimating the cost of an unprecedented supply–demand shock in the short-term is a non-trivial, data intensive exercise. Oftentimes, it requires data on high-frequency economic indicators, in particular, sector-level production data (del Rio-Chanona et al., 2020), card transactions (Carvalho et al., 2021), unemployment claims (Coibion et al., 2020) and energy consumption such as electricity usage (Fezzi and Fanghella, 2020).

However, in some countries, the presence of a “grey economy,” characterised by the absence of systematic record-keeping protocols, presents a key limitation. In India, this informal sector, according to the International Monetary Fund, accounts for 90% of the labour force and approximately 45% (Murthy, 2019) of the nation’s Gross Domestic Product (GDP). Hence, estimating the near-term impact of COVID-19 using broad sector-level production data in India is infeasible. Transaction data...
is not available in the public domain and requires purchasing expensive datasets. It is also challenging to measure the impact using unemployment data, as the publication of employment reports were officially suspended in 2016 (Basu, 2019). Finally, electricity consumption data is published with a quarterly lag. That is, data for a current month is not available until the end of the calendar quarter. Due to these constraints, the use of alternative instruments to estimate the short-term economic impacts of COVID-19 in a country like India is worth exploring.

This study uses nighttime images of lights on the surface of Earth, i.e., nightlight (NTL), taken by satellites to estimate the economic impact of COVID-19 in India during Fiscal Year (FY) 2020Q1 (Apr-Jun, 2020), the study period. The NASA nighttime radiance dataset, VNP46A1, was used as the NTL datasource for the study.

A few nuances of nightlight makes it uniquely suited to measuring short-term supply-demand shocks such as the COVID-19 crisis. Nightlight data is freely and readily accessible in near real-time from several sources in the US, China and the International Space Station (Zhao et al., 2019). NTL data also cannot be manipulated in the same sense as how governments might be inclined to misrepresent economic data for political purposes (Martinez, 2017). The most important aspect of nightlight that makes it a good fit for India is that the process of estimation is completely decoupled and independent of the idiosyncrasies of the informal sector.

Due to these advantages, as a leading indicator of economic output, nightlight can provide an immense benefit in countries such as India where data collection is limited and it can take years before Real Gross Domestic Product (GDP) figures are published, thus delaying key Government spending and policy decisions.

The impact of COVID-19 in India during the study period, was analysed using a three-step process. First, radiance values (expressed in units of nW.cm⁻².sr⁻¹, i.e., nanoWatts per sq. cm per steradian), were extracted from VNP46A1 using a custom-designed, cloud-based architecture. Second, elasticity, i.e., the rate of change of nightlight with respect to national quarterly GDP was estimated using Panel Regression. Finally, machine learning (ML) algorithms were used to predict Year-over-Year (YoY) change in GDP during FY2020Q1.

The paper is structured as follows: Section 2 (Background) traces the use of Panel Regression and Machine Learning for nightlight-related research and discusses some of the challenges with using NTL data. Section 3 (Data) outlines the datasets used in the study. In Section 4 (Methodology), the process of extracting radiance values from VNP46A1 has been explained along with the specifications for Panel Regression and Machine Learning. Section 5 (Results) presents the estimates of elasticity and predicted economic impact of COVID-19 at the national and sector-level. Section 6 (Analysis and Discussion) examines the results and their practical implications. Section 7 (Conclusion) concludes with a brief discussion on limitations and further research.

2. Background

The use of nightlight for socio-economic research first started in the mid-70’s (Croft, 1973; Croft, 1978; Welch, 1980; Foster, 1983) for a range of research topics from estimating energy consumption to industrial activities (Elvidge et al., 1999). However, use of the data was limited as they were only available on physical film strips (Zhao et al., 2019). The digitalisation of the nightlight data archive by National Oceanic and Atmospheric Administration (NOAA) in 1992 made them accessible to the wider research community.

2.1. Panel regression

In the early 2010’s, using NTL data from 1992-2003 for a cohort of 188 countries, Henderson et al. (2012), economists at Brown University, showed that national income data could be complemented with nightlight to empirically estimate true income growth. Estimates from Henderson’s model, especially for countries with low-quality economic data were more realistic and aligned with the public consensus even though the results sometimes disagreed with official figures published by state governments that may have been inflated or deflated to serve political purposes. Following the findings in this landmark study, nightlight came to be regarded as a viable and reliable proxy for serious economic research. Henderson’s paper also marked the first time that a panel regression framework had been used to study nightlight data and subsequently, the method gained broad acceptance for nightlight-related research.

Several panel-based studies have been undertaken since then. Panel studies with Indian economic data have found that the relation between nightlight and economic metrics varies based on population (Prakash et al., 2019), extent of urbanisation (Bhandari and Roychowdhury, 2011) and income levels (Chakravarty and Dehejia, 2017) among other factors.

One of the earliest study on estimating the elasticity of nightlight to GDP was conducted by Elvidge et al. (1997). The researchers used the logarithm of total area lit, i.e., the number of pixels with a non-zero radiance value within a country’s boundaries to estimate the elasticity of nightlight to GDP for 21 countries. They observed that the relationship between nightlight and GDP was almost linear (correlation ≈ 0.97) with the outliers being poorly developed countries with less lighting. Between 1997-late 2000’s, several other similar studies were conducted that further established the theory that nightlight was strongly correlated with economic measures such as GDP and Gross Value Added (GVA) (Doll et al., 2000; Ebener et al., 2005).

A limitation of the models however, was the use of annual temporality. Research was done using yearly esti-
mates even though economic data was available at sub-annual level. This was partly due to the limitations of the Defense Meteorological Program (DMSP) Operational Line-Scan System (OLS) satellite imaging system used prior to 2012 for taking nighttime images. Due to lower sensitivity, spillover effect (Bansal et al., 2020) and sensor degradation (Chanda and Kabiraj, 2020) over time, it was hard to distinguish gradual Quarter-over-Quarter (QoQ) changes in radiance. There were also operational challenges. Due to differences in the time of day/night when images are captured, multi-temporal series of images require calibration. Hence, before being used for research, NTL had to be calibrated by identifying time-invariant features such as stable surface features or by using histograms (Zhang et al., 2016). Furthermore, pixel saturation (Wu et al., 2013) and Blooming effects (Shen et al., 2019) cause diffusion of brightness, particularly in densely populated urban centers, producing anomalous results. Additional processing is required to remove noise resulting from such artifacts.

The launch of a special imaging instrument for taking nighttime images, the Visible Infrared Imaging Radiometer Suite (VIIRS) aboard the Suomi National Polar Partnership (NPP) (Hille, 2015) satellite in late 2011, with twice the sensitivity compared to DMSP-OLS to distinguish minor variations in light, vastly improved the reliability and adoption of nightlight for research. Apart from explanatory variables such as log of area lit used in regression studies at the time, the higher resolution now allowed researchers to create and apply new features such as log of sum of lights for measuring elasticity.

By leveraging the new capabilities of VIIRS, in one of the first studies to utilise sub-annual data, Prakash et al. (2019), used a combined DMSP & VIIRS dataset from 1992–2017 to explore the elasticity between nightlight and state-level economic metrics in India. Using aggregated quarterly economic data, the researchers observed a robust correlation between NTL and GDP as well as other macroeconomic indicators such as Index of Industrial Production (IIP).

Beyer et al. (2020) also used sub-annual, quarterly data similar to Prakash et al. (2019) to estimate QoQ effects, but included electricity usage as an additional regressor. In the panel model, Beyer et al. (2020), found that electricity consumption was significant at the 1% level, underscoring the high importance of the predictor in measuring economic metrics. The current study bears semblance to research by Prakash et al. (2019) and Beyer et al. (2020) and models quarterly national GDP using panel regression to examine the elasticity of GDP to nightlight.

2.2. Machine learning framework

In recent years, new ways of measuring economic activity using neural networks (Head et al., 2017; Jean et al., 2016; Basihos, 2016; Subash et al., 2018) and more generally machine learning algorithms, such as Gradient Boosting (Bansal et al., 2020), Random Forests (Otchia and Asongu, 2019), Support Vector Machines (Pandey et al., 2013) and others with nightlight data have become commonplace.

In this study, a machine learning (ML) framework using raw nighttime radiance values similar to the approach of Otchia and Asongu (2019) was used to estimate the economic cost of COVID-19 during FY2020Q1. Several ML algorithms were benchmarked using a range of hyper-parameters and the algorithm with the best overall performance was used to build the final model. Following the example of Beyer et al. (2020), electricity consumption, was also used for training the ML models.

2.3. Use of ancillary datasets

The NTL products, VNP46Ax, were scheduled for release between 2020–2021. These datasets, produced using the Black Marble Algorithm, took advantage of the high low-light sensitivity of VIIRS and represented a major improvement in publicly-available NTL data quality (Román et al., 2018). However, despite the improvements, VNP46A1 was suboptimal compared to VNP46A2 which had not yet been released during the course of this study. Specifically, VNP46A1 did not contain BRDF-corrected radiance layers and being top-of-atmosphere (TOA) NTL, was affected by noise due to moon illumination, clouds and other artifacts (Xu and Qiang, 2021). Although this study has attempted to perform robust data cleansing, data on precipitation and population were included as additional features in order to improve ML accuracy.

3. Data

Three major sources of data were used in this study: Indian economic, nightlight and electricity usage data.

3.1. Indian economic data

Quarterly, National-Level GVA and Real GDP at 2011–12 constant prices were downloaded from the website of Indian Ministry of Statistics and Programme Implementation (MOSPI) for FY2012Q1–FY2019Q4.

All figures were quoted in units of Indian Rupee (₹) crores (a numeric quantity in the Indian Number System representing 10 million). The Indian fiscal year starts in April of each year, i.e., FY Q1 corresponds to calendar year Q2 (Apr-Jun) and so on. Hence, quarterly nightlight, electricity usage and other ancillary sources used for analysis were re-mapped to fiscal-year based quarters.

3.2. Nightlight data

The VNP46A1 nightlight data for India between Feb 2012 and June 2020, were downloaded from NASA’s Level 1 Atmosphere Archive and Distribution System Distributed Active Archive Center Data Center (LAADS DAAC).
The dataset is available in the form of Hierarchical Data Format (HDF5) files, where each file corresponds to a square-shaped area on the Earth’s surface, known as a tile. The image in Fig. 9, shows a single tile and the complete set of 8 tiles, required to cover the Indian mainland (excluding Andaman and Nicobar Islands). Further details have been provided in Table 1.

At a high level, each HDF5 file can be thought of as a filesystem with files organised in folders. Each such file within a VNP46A1 HDF5, is a geocoded dataset called GeoTIFF, a type of matrix where each cell is associated with a latitude-longitude co-ordinate (geo-coordinate) in addition to a numeric value. Each GeoTIFF, in this case is of size 2,400 rows × 2,400 columns.

The daily VNP46A1 HDF5 files contain 26 GeoTIFF files and 4 of them were relevant to this analysis - DNB Radiance, Cloud Mask, Solar Zenith, Moon Illumination Fraction layers. (See Table 6.). The layer that contains the geo-coded nightlight radiance values is the ‘DNB Radiance’ layer. Due to extraneous artifacts, the data in the layer requires further processing as discussed in Section 4. A visual inspection of the NTL recorded in May of 2020 shows a subtle change in nightlight compared to March of the same year (See Fig. 1).

3.3. Electricity usage data

Data on state-level electricity consumption was obtained from monthly reports published by Power Systems Corporation of India (POSOCO) from April 2012 to June 2020 and then aggregated to compute quarterly, national-level usage. In this dataset, electricity consumption is recorded as actual drawal (as opposed to projected usage) measured in units of MU, i.e., 1 Million KWh.

3.4. Ancillary datasets

In order to improve model accuracy, data on precipitation and population were included in the ML model. Precipitation data was obtained from the website of Indian Agricultural Research Institute (ARI, 2020). Monthly rainfall, measured in mm, was aggregated by quarters for FY2012Q1 - FY2020Q1. Population estimates were obtained from Trading Economics (tradingeconomics.com, 2019) also for the same time horizon.

4. Methodology

4.1. Overall methodology

The overall study involved 3 steps as shown below:

- Extracting nightlight radiance values to calculate national-level NTL metrics;
- Panel Regression to estimate the elasticity of nightlight; and
- Machine Learning to estimate economic impact of COVID-19

4.2. Nightlight radiance

As noted earlier, radiance values in the VNP46A1 DNB Radiance Layer cannot be used as-is since extraneous artifacts such as moon illumination, cloud cover, snow reflec-
The cloud layer required special handling as it contained 16-bit values with encoded information at different bit positions (Fig. 8). Details of bit array processing and Boolean matrix operations is very involved and has been shown in Section A.4 in the Appendix. Quarterly sum of the radiance values, known as “Sum of Lights” was thereafter obtained by taking a sum of radiance in the pixels within the national boundary.

4.2.1. Tools used

Parallel processing tools were used extensively due to the scale of the data. In particular, the library, multiprocessing in Python (v3.6.6), packages foreach – doParallel in R (v4.0.0) and GNU Parallel in Ubuntu Linux v18.04.5 (Tange, 2011) were used for parallel processing. Geospatial Data Abstraction Library (GDAL) (GDAL/OGR contributors, 2020) was used for geo-spatial analytics and the R package, exactextractr, was used to extract radiance values. A custom tool was also developed using Adobe Java Libraries for MS Office and pattern matching libraries in R to automatically extract monthly state-level electricity consumption data from images in the POSOCO PDF reports. A custom Amazon Web Services (AWS) cloud-based hardware architecture using high-performance and high I/O servers was implemented for faster data processing (Fig. 11).

4.3. Panel regression

A common statistical approach to modelling cross-sectional time-series data is to use multi-level models such as Panel Regression that control for time effects and time-invariant group effects.

Following the approach proposed by Henderson et al. (2012), the elasticity of nightlight and electricity consumption with respect to GDP was estimated using a Panel Regression Framework. A log–log model was used where the coefficients for each variable represented the percentage change in the outcome variable for 1% change in the corresponding regressor.

First, a pooling model was estimated assuming $\beta_0 = \beta$ for all $\{i,t\}$ where $i$ is the group effect and $t$ the time effect. Next, Fixed Effect Panels, also known as “no pooling” (Gelman and Hill, 2009), were estimated to model the heterogeneity across quarters. Finally, Random Effects, i.e., “partial pooling” (Gelman and Hill, 2009), was used to model time-invariant individual effects as well as factors that change with time.

The baseline functional form for a Fixed Effect Model (Wooldridge, 2016) with $k$ explanatory variables for group $i$ at time $t$, is shown in Eq. 1.

$$ y_{it} = \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + z_i + u_{it}, \quad t = 1, 2, \ldots, T $$

where,
\(z(i = 1, \ldots, n)\) is the intercept for the group; 
\(y_{it}\) is the dependent variable; 
\(x_{itk}\) is the k-th independent variable for entity \(i\) at time \(t\); 
\(\beta_i\) is the coefficient of the k-th independent variable; and 
\(u_{it}\) is the error term.

For the Random Effects model, an intercept-term, \(\beta_0\), is added to Eqn. 1 (Wooldridge, 2016). It is assumed that the fixed effect, \(z_i\), is independent of all regressors across all time periods.

\[ y_{it} = \beta_0 + \beta_1 x_{it1} + \beta_2 x_{it2} + \ldots + \beta_k x_{itk} + z_i + u_{it}, \quad t = 1, 2, \ldots, T \]  

(2)

Panel diagnostics for serial correlation in idiosyncratic errors (Breusch - Godfrey, Wooldridge Test), cross-sectional dependence (Parasaran CD Test), heteroskedasticity (Breusch - Pagan test) and stationarity (Levin-Lin-Chu Test) were performed for each model. For models with serial correlation or cross sectional dependency, the spatial correlation consistent standard error (SCC) method proposed by Driscoll and Kraay (1998) with the HC3 weighting scheme (Long and Ervin, 2000) for small sample size was used to estimate Robust Standard Errors (SE). (See Table 8)

4.4. Functional specifications

4.4.1. National

All explanatory variables were aggregated to the Fiscal Year Quarter (t) level. Functional specifications for GDP and Sectoral GVA have been shown in Eqn. 3 and Eqn. 4 respectively.

\[ \ln(GDP) = \beta_1 \ln(SumOfLights) + \beta_2 \ln(SumElectricity) + \beta_3 \ln(SumOfLightsSq) + \beta_4 \ln(Population) + z_i + u_{it} \]  

(3)

\[ \ln(GVA_{Sector}) = \beta_1 \ln(SumOfLights) + \beta_2 \ln(SumElectricity) + \beta_3 \ln(SumOfLightsSq) + \beta_4 \ln(Population) + z_i + u_{it} \]  

(4)

4.5. Predictive modelling

Generally, in NTL research, the elasticity of nightlight is used for predictive analysis. In this study, instead of elasticity, a machine learning approach was implemented.

Several ML algorithms were evaluated with National GDP as the outcome variable. Specifically, Support Vector Machine (SVM) (Cortes and Vapnik, 1995) with Radial Basis Function (RBF) Kernel, K-Nearest Neighbours (Altman, 1992), Random Forest (Breiman, 2001), eXtreme Gradient Boosting (XGBoost) which implements a Gradient Boosted Model (Chen and Guestrin, 2016) and Lasso/ Ridge Regression (Tibshirani, 1996) were compared to assess individual model performance.

The predictors shown in Table 2 were used for feature engineering. First, the features were scaled and centered to have mean 0 and a range between (0,1). Next the dataset was split into 80/20 Training/Test set samples. For each algorithm, a 5-fold repeated cross-validation was performed over a set of hyper-parameters. The best model was selected based on the least Root Mean Square Error (RMSE) on the training set. The model was then retrained using a more narrow and targeted range of hyperparameters.

5. Results

5.1. Economic Impact of COVID-19

At the national level, Lasso Regression outperformed all other machine learning models and had the lowest RMSE (Table 3). After retraining, the final model with the least RMSE had parameters \(z = 1, \lambda = 4201.\) Fig. 7 shows the results of model cross-validation.

Lasso uses L1 regularisation that penalises the absolute value of the coefficients which in the case of Lasso means coefficients can become 0 or drop off resulting in a simpler model. The implementation in the R package, glmnet (Friedman et al., 2010), uses cyclical co-ordinate descent which minimises the objective function shown in Eqn. 5 over a range of \(\lambda\) values.

\[
\text{Objective Function} = \min_{\beta_0, \beta} \sum_{i=1}^{N} w_i l(y_{i}, \beta_0 + \beta^T x_i) + \\
\lambda \left[ (1 - z) ||\beta||_2^2 / 2 + z ||\beta||_1 \right]
\]

(5)

The model estimated that the National GDP for FY2020Q1 would be ₹2,686,859, i.e., 24% lower on a YoY basis and had a Confidence Interval of (-18%,-30%).

The quarterly trend of nightlight (Fig. 2), electricity usage (Fig. 3) and and GDP with the predicted value (Fig. 4), showed a precipitous drop of GDP accompanied by a corresponding drop in both in electricity output and nightlight in FY2020Q1.

| Variable learning features | Variable Description |
|----------------------------|----------------------|
| \(y\)                      | gdp (Gross Domestic Product) |
| \(x_1\)                    | vnpyr_sol (sum of lights)   |
| \(x_2\)                    | vnpareayr_sol (sum of pixels > 0) |
| \(x_3\)                    | sumelec (sum of electricity usage) |
| \(x_4\)                    | meanelec (mean of electricity usage) |
| \(x_5\)                    | pop (population)           |
| \(x_6\)                    | precip (precipitation)     |
### Table 3
ML Performance (Units: ₹ crores).

| Model                  | Parameter | Best Fit                          | Train RMSE | Test RMSE |
|------------------------|-----------|-----------------------------------|------------|-----------|
| Linear Regression      | -         | -                                 | ₹144,625   | ₹422,920  |
| Lasso Ridge Regr.      | $x, \lambda$ | $x = 1, \lambda = 4201$         | ₹108,958   | ₹176,989  |
| Random GLM             | maxIntOrder | maxIntOrder = 1                   | ₹128,633   | ₹150,941  |
| SVM RBF Kernel         | $C, \Sigma$ | $\sigma = 1.2692, C = 2$         | ₹136,189   | ₹620,528  |
| KNN                    | $k$       | $k = 5$                           | ₹300,175   | ₹721,176  |
| RandomForest           | mtry      | mtry = 4                          | ₹147,257   | ₹380,779  |
| XGBoost                | misc*     | nrounds 100, max depth 6          | ₹159,039   | ₹303,559  |

![Fig. 2. National sum of lights.](image1)

![Fig. 3. National electricity usage.](image2)

![Fig. 4. National GDP with FY2020Q1 prediction.](image3)
The sector-level predictions shown in Table 4 indicated that the hardest hit sectors for the quarter would be Mining, Real Estate, Utilities and Agriculture, all experiencing more than a 6% YoY decline. The only sector that registered a YoY increase was Public (Government) Services.

5.2. Elasticity of nightlight

5.2.1. Panel regression

The results of Panel Regression with different independent variables have been shown in Table 5. Diagnostics for the National Panel shown in Table 8 indicated presence of Cross-Sectional dependency and accordingly, Robust Standard Error (SE) were estimated and reported in the output.

As seen in the Pooled OLS model, the coefficient of Sum of Lights was significant across all models. In the Pooled OLS (1) model, a 1% change in Sum of Lights increases GDP by 0.189%. When Sum of Electricity is added to the model, the premium drops to 0.113 although it still remains significant. The variable is significant at the 1% level and the coefficient suggests that a 1% increase in electricity usage, increases GDP by 0.313%.

In Fixed Effect models (3–6) with Quarterly FE, the coefficient of Sum Of Lights is 0.189 when it is the only independent variable. The premium drops to 0.023 when population is added. Population is significant across all models where the variable was included, at the 1% level with an elasticity over 5.5. The magnitude is very high but not surprising. Increase in population size increases the labour pool and hence GDP, the national output. The magnitude of Sum Of Lights reduces by almost 50% to 0.10 when Sum of Electricity is added. It is well known that electricity usage is closely correlated with GDP (Lin and Shi, 2020) and the results are consistent with findings in earlier studies.

6. Analysis and discussion

6.1. VNP46A1 & DNB composite comparison

In order to assess the quality of the processed data, a comparison of nightlight radiance between VNP46A1 and DNB Composite, the gold standard, was performed. The values were found to be strongly correlated (0.85). However, the results, while similar, were not identical. Under ideal conditions with low moon illumination, VNP46A1 can achieve a significantly higher nightlight detection limit with a narrower confidence band of 0.5 nW.cm$^{-2}$.sr$^{-1}$ ($\pm$ 0.10 nW.cm$^{-2}$.sr$^{-1}$) compared to the minimum threshold of 3.0 nW.cm$^{-2}$.sr$^{-1}$, for NTL data to be usable (Román et al., 2018). Hence, VNP46A1 data-

Table 4

| Industrial Sector | RMSE | 2019 Q1 | 2020 Q1 | YoY Diff (%) | ORF Diff (%) |
|-------------------|------|--------|--------|-------------|-------------|
| Mining            | ₹ 10,741.52 | ₹ 92,777.00 | ₹ 83,357.54 | -10.15 | -14.70 |
| Real Estate       | ₹ 77,366.22 | ₹ 810,500.00 | ₹ 757,553.63 | -6.53 | -17.30 |
| Utilities         | ₹ 2,403.44 | ₹ 81,628.00 | ₹ 76,564.31 | -6.20 | -13.90 |
| Agriculture       | ₹ 89,712.19 | ₹ 439,248.00 | ₹ 412,726.08 | -6.04 | -1.30 |
| Communication     | ₹ 31,680.37 | ₹ 644,224.00 | ₹ 618,586.72 | -3.98 | -9.70 |
| Construction      | ₹ 8,424.91 | ₹ 263,653.00 | ₹ 254,386.05 | -3.51 | -13.30 |
| Manufacturing     | ₹ 24,793.88 | ₹ 574,411.00 | ₹ 574,037.46 | -0.07 | -6.30 |
| Public Service    | ₹ 18,261.17 | ₹ 421,191.00 | ₹ 449,404.97 | 6.70 | -0.40 |

Table 5

| Dependent variable: | log(Gross Domestic Product) |
|---------------------|-----------------------------|
|                     | Pooled OLS (1) | Pooled OLS (2) | Qtr FE (3) | Qtr FE (4) | Qtr FE (5) | Qtr FE (6) |
| log(Sum of Lights)  | 0.189*** (0.011) | 0.113*** (0.012) | 0.189*** (0.030) | 0.023*** (0.007) | 0.103* (0.053) | 0.018*** (0.006) |
| log(Sum of Electricity) | 0.313*** (0.078) |               |             |             | 0.346 (0.259) | 0.010 (0.042) |
| log(Sum of Lights$^2$) |              |               |             |             |               | -0.001 (0.001) |
| log(Population)     |               |               |             |             |               | 5.547*** (0.281) | 5.569*** (0.342) |
| Constant            | 11.297*** (0.202) | 9.130*** (0.758) |             |             |             |             |

Observations: 26
R$^2$: 0.852 0.903 0.866 0.996 0.918 0.996
Adjusted R$^2$: 0.846 0.894 0.840 0.995 0.898 0.995

Note: *p<0.1; **p<0.05; ***p<0.01
set can capture minute radiance values that other NTL products such as the DNB Composites may be unable to detect. Conversely, under high moon illumination the performance of VNP46A1 can be suboptimal, failing to even meet the minimum threshold \((\text{Román et al., 2018})\). The inconsistency of VNP46A1 is a major limitation and as a result DNB Composite may produce more reliable, accurate and stable estimates. Such differences, in addition to different methodologies used to produce the datasets such as a 90-day averaging window used in the study compared to 30-day for DNB Monthly Composite, could account for the variation observed between the two data sources.

Despite the differences, a common trend of reduced nightlight during COVID-19 was observed across all major NTL products \((\text{Deep and Gupta, 2021; Xu et al., 2021; Alahmadi et al., 2021; Ghosh et al., 2020})\). Similar to the findings with VNP46A1 in this study, DNB Composites \((\text{Elvidge et al., 2013})\) also showed that radiance decreased significantly from March to May of 2020 relative to the same time frame in 2019 (Fig. 5).

6.2. COVID-19 impact prediction

6.2.1. National level

The projected 24% contraction of National GDP in FY2020Q1 was a significant change from the average YoY GDP increase of 6% in years prior to 2020 \((\text{Plecher, 2018})\). Based on official figures, initially, the ML model appeared to have overestimated the economic cost. Figures published by leading investment banks had predicted a less severe impact. For instance, in August, 2020, State Bank of India Research had projected a YoY FY2020Q1 decrease of 16.5% \((\text{SBI, 2020})\).

On August 31, 2020, while the results of the current study was being reviewed, the Indian Government announced official GDP figures for FY2020Q1 which estimated the YoY decline to be 23.9% \((\text{Mundle, 2020})\). It turned out that the predicted GDP decline of −24% obtained in the study was a much more accurate estimate, almost identical to the actual results.

The industrial sector-wise YoY estimates shown in Table 4, was lower compared to estimates published by the Observer Research Foundation (ORF), \((\text{Mukhopadhyay, 2020})\) and McKinsey \((\text{Gupta et al., 2020})\). For instance, the impact in the real estate sector was estimated to be approx. −17.3% according to ORF, whereas the nightlight model suggested a more modest figure of −6.53%. There are a few factors that may have likely contributed to the difference. The distribution of labour in the informal sector is not uniform across all industries. Consequently, while the impact of a lockdown can be more accurately measured say, in the Public Service sector, it may be less accurate in a sector such as construction or agriculture with no controls for centralised record-keeping.

Moreover, NTL might not be representative of changes across sectors uniformly. A change in nightlight may be a better proxy for changes in manufacturing than for changes in say, the agricultural sector where a better harvest does not lead to more nightlight the same way as an expansion of manufacturing would add more lights to a production plant.

![Radiance in March compared to Radiance in May of 2019](https://example.com/radiance-march-may-2019)

![Radiance in March compared to Radiance in May of 2020](https://example.com/radiance-march-may-2020)

Fig. 5. Difference in NTL between March - May in 2019 and 2020 (covid lockdown) using Average DNB Monthly Composite \((\text{Elvidge et al., 2013})\). Yellow and Red points indicate NTL Radiance was higher in March compared to May.
7. Conclusion and recommendations

7.1. Main findings

The results obtained from the econometric and ML analysis showed that nightlight could be used to obtain reasonably accurate estimates of the economic impact of the COVID-19 pandemic.

A few findings were key to this study. The post-processed VNP46A1 NTL data was closely correlated with DNB Composites, commonly used in NTL research. This demonstrates that high-quality nightlight dataset can be developed efficiently at scale with parallel processing tools and publicly available high-performance servers in the cloud.

Second, based on low RMSE values, the study found that nightlight and electricity usage could predict the short term impact of a supply–demand shock with a high degree of accuracy at the national level in India. Electricity consumption had a high predictive value and since it is easier to obtain data on electricity usage, it should be explored further, similar to studies that have been already conducted elsewhere (Fezzi and Fanghella, 2020). Precipitation data was also found to improve model performance and in general linear ML models performed better at the national level.

In countries such as India, due to lack of data, changes in the output of the informal economy cannot be estimated in the short-run. This means that Real GDP data is not available to institutions immediately and it can sometimes take years before the numbers are known (Bhandari and Roychowdhury, 2011). Hence, nightlight and electricity usage could be invaluable proxies and leading indicators of economic changes.

7.2. Counter arguments

Nightlight might not always be a suitable proxy for estimating the economic impact of a pandemic. During COVID-19, nearly all companies in the services sector adopted work-from-home policies. A study by Lenovo Research India claimed that remote working had improved productivity by 83% (Lenovo, 2020). In urban areas dominated by the services sector, even though nighttime lights had gone down significantly due to reduced highway traffic, business closures and other factors, it did not lead to a proportional decrease in the output of the services sector. Hence, using nightlight for estimating impact in the services sector in this environment could produce misleading results. Similarly changes in manufacturing patterns due to lower resources or shifting of activities to daytime may lead to biased results.

7.3. Next steps

At the time of writing, in August, 2020, NASA had started publishing the next-generation of Black Marble nightlight images called VNP46A2 with BRDF-Corrected Radiance layers. This product may eventually complement DNB Composites used for nightlight research. Given the promising results, researchers from other domains such as finance, should also consider incorporating VNP46A2 and more generally nightlight for short-term supply–demand shock and other economic studies.

The use of machine learning with nightlight radiance data is becoming more prevalent, and as seen in this study, it resulted in a much better predictive performance relative to linear regression models. Pictures taken by satellites are objective, unaffected by disruptive events on Earth, and can show day-to-day changes with extremely precise geolocation information. This makes it an unrivaled instrument of research with a vast array of possibilities for future research.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This study was conducted as part of a final research report at Imperial College London. I’d like to thank my advisor, Prof. Edward Anderson, for his erudite guidance during initial research and Prof. Kalyan Talluri, Programme Director, Business Analytics who encouraged me to pursue a new topic. I’d also like to thank my parents, sister and my wife, Suraiya, whose endless love, caring & patience over many weekends and nights, for the past two years made it all possible. Finally, I’d like to convey my deep gratitude to my friends, Tok, for the thorough and constructive criticism of the manuscript and to Vijaysree, for always being there to lend a helping hand - with her exceptional scholarship and support, we produced the best papers in so many different courses.

The data that support the findings of this study are available from the following sources: VNP46A1 data from NASA LAAD Repository for VNP46A1, VIIRS DNB Composites from Payne Institute, financial data from Indian Ministry of Statistics and Programme Implementation, electricity consumption data from Power System Operation Corporation, precipitation data from Indian Agricultural Research Institute and COVID-19 datasets from the Ministry of Health and Family Welfare.

Ethical Standards. The research meets all ethical guidelines, including adherence to the legal requirements of the study country.
Appendix A. A.1. Tables

See Tables 6 and 7.

Table 6
Description of layers selected for this study.

| SDS HDF5 Layer     | Units                  | Description                | Bit Types | Fill   | Range  | Scale |
|--------------------|------------------------|----------------------------|-----------|--------|--------|-------|
| DNB Radiance       | nW.cm^{-2}.sr^{-1}     | At-Sensor Radiance         | 16-bit uint | 65535  | 0–65534| 0.10  |
| Solar Zenith       | Degrees                | Solar Zenith Angle         | 16-bit int | -32768| 0–18000| 0.01  |
| Moon Illum. Frac.  | Percentage             | Moon Illumination %        | 16-bit int | -32768| 0–10000| 0.01  |
| QF Cloud Mask      | NA                     | Cloud Mask Status          | 16-bit uint | 65535  | 0–65534| NA    |

Table 7
Index of industrial production (IIP) India national level.

| Index of Industrial Production (IIP) | Jan-20 | Feb-20 | Mar-20 | Apr-20 | May-20 | Jun-20 |
|-------------------------------------|--------|--------|--------|--------|--------|--------|
| Primary goods                       | 1.80   | 8.20   | -4.00  | -26.60 | -19.70 | -14.60 |
| Capital goods                       | -4.40  | -9.60  | -38.80 | -92.60 | -65.20 | -36.90 |
| Intermediate goods                  | 15.60  | 23.00  | -18.60 | -65.40 | -40.60 | -25.10 |
| Infrastructure/ construction goods   | -0.30  | 2.80   | -24.30 | -84.70 | -40.70 | -21.30 |
| Consumer durables                   | -3.70  | -6.20  | -36.80 | -96.00 | -69.40 | -35.50 |
| Consumer non-durables               | -0.60  | -0.30  | -22.30 | -48.70 | -11.10 | 14.00  |

A.2. Panel regression diagnostics

Panel Regression Diagnostic Framework

A diagnostic testing framework to select the appropriate model and to address issues such as serial correlation and cross-sectional dependence was also adopted for each set of Panel Regressions. In particular, the following tests were performed,

Hausman Test (phtest): A test for endogeneity to select between the Fixed-Effects and Random-Effects Estimators. The Null Hypothesis states that the disturbance term is uncorrelated with the regressors;

Stationarity (adf.test): Augmented Dickey-Fuller Test to test for non-stationarity (Null Hypothesis). If Unit Root is present the series can be made stationary using first differences;

Cross-Sectional Dependence (pcdtest): The Parasaran CD Tests to determine whether the residuals are correlated across individual groups. The Null Hypothesis states that they are uncorrelated;

Serial Correlation (pbgtest): Breusch-Godfrey/Wooldridge test for serial correlation in idiosyncratic errors. The Null Hypothesis is that there is no correlation. The implementation of pbgtest in the plm package in R allows for testing higher-order serial correlation (Croissant and Millo, 2018), which is particularly helpful for the national quarterly level GDP data; and.

Heteroskedasticity (bptest): Breusch-Pagan test of heteroskedasticity to decide whether a robust covariance matrix should be estimated (see Table 8).

Table 8
Panel regression diagnostics.

| Scope   | Test               | Context                        | p-value | Comment                   |
|---------|--------------------|--------------------------------|---------|---------------------------|
| National| Hausman            | Select Fixed vs Random          | 0.94    | Random-Effects selected   |
|         | Levin-Lin-Chu      | Unit Root                      | <0.01   | Time-Series is stationary |
|         | Breusch-Godfrey    | Serial Auto-Correlation        | 0.08    | No serial auto-corr.      |
|         | Pasaran CD         | Cross-Sectional Dep.           | <0.01   | Cross-Sec. Dep. exists    |
|         | Breusch-Pagan test | Heteroskedasticity             | 0.062   | No heteroskedasticity     |
| State   | Hausman            | Select Fixed vs Random          | 0.29    | Random-Effects selected   |
|         | Levin-Lin-Chu      | Unit Root                      | <0.01   | Time-Series is stationary |
|         | Breusch-Godfrey    | Serial Auto-Correlation        | <0.01   | Serial auto-corr. present |
|         | Pasaran CD         | Cross-Sectional Dep.           | <0.01   | Cross-Sec. Dep. exists    |
|         | Breusch-Pagan test | Heteroskedasticity             | 0.01    | Heteroskedasticity present|
A.3. Figures

See Fig. 6 and 7.

Fig. 6. Correlation of variables.

Fig. 7. National Lasso regression CV results.
A.4. NTL Image Processing

Daily Night-Light data requires extensive image processing to correct for cloud cover, moon illumination, stray light and various other artifacts. To make the data suitable for research, monthly and annual composite images are produced by time-averaging daily data using enterprise-scale computational resources (Elvidge et al., 2017). These datasets - Geocoded TIF (Tagged Image Format) files, known as DNB (Day/Night Band) Composites are available for download from NOAA and the Earth Observation Group (EOG) at the University of Colorado.

A.4.1. DNB Composite images

1. Extract QF Cloud Mask Layer (M_C). Each cell in this layer contains a 16-bit value with the binary values at different bit positions indicating different flags (Fig. 8).
   1..1 If Bits 1–3 <> 000 (Land & Desert) or 001 (Land), set mask pixel value to 0
   1..2 If Bits 6–7 <> 00 (Confident Clear) or 01 (Probably Clear), set mask pixel value to 0
   1..3 If Bit 8 = 1 (Shadow), set mask pixel value to 0
   1..4 If Bit 10 = 1 (Snow/Ice), set mask pixel value to 0
   1..5 If none of the above conditions exist, set mask pixel value to 1

2. Extract Moon Illumination Layer (M_M): For each pixel, if value 20(%) set mask pixel value to 0 or 1 otherwise

3. Extract Solar Zenith Layer (M_S): For each pixel, if value < 101 (degrees), set mask pixel value to 0 or 1 otherwise

4. Extract Radiance Layer (M_R): For each pixel, if value > 150 nW.cm⁻².sr⁻¹, set mask pixel value to 0 (see Table 9)

A.4.2. VNP46A1 masking steps

Note: The Cloud Layer is a 2,400 Geocoded matrix (GeoTIFF). Each cell of the matrix contains 16 binary values. The binary values in different positions shown herein is used to determine presence of Land, Cloud, etc. (see Table 10).

![Cloud Layer Encoding](M_C).](image)

Table 9
Cloud layer (M_C) bitmask.

| Layer          | Bit Pos. | Attribute                        | Values                      | Selected |
|----------------|----------|----------------------------------|-----------------------------|----------|
| Cloud          | 1–3      | Land/Water Presence              | 000 = Land & Desert         | 000      |
|                |          |                                  | 001 = Land no Desert        | 001      |
|                |          |                                  | 010 = Inland Water          |          |
|                |          |                                  | 011 = Sea Water             |          |
|                |          |                                  | 101 = Coastal               | 101      |
| Cloud          | 6–7      | Cloud Detection Confidence       | 00 = Confident Clear         | 00       |
|                |          |                                  | 01 = Probably Clear          |          |
|                |          |                                  | 10 = Probably Cloudy         |          |
|                |          |                                  | 11 = Confident Cloudy        | 01       |
| Cloud          | 8        | Shadow Detected                  | 1 = Yes                     | 0        |
|                |          |                                  | 0 = No                      | 0        |
| Cloud          | 10       | Snow/ice surface                 | 1 = snow/ice                | 0        |
|                |          |                                  | 0 = no snow/ice              | 0        |

Table 10
HDF5 layers used in this study.

| Step | Layer                  | Operation                                                                 |
|------|------------------------|---------------------------------------------------------------------------|
| 1    | Cloud Quality Flag (M_C) | Used to remove pixels that represent a) non-land surfaces; b) Snow c) Shadow; and d) Probably/Confident Cloudy |
| 2    | Moon Illumination (M_M)  | Used to remove pixels where the fraction of lunar illumination is greater than 20% |
| 3    | Solar Zenith (M_S)       | Used to remove pixels where the Angle of the Solar Zenith is less than 101 Degrees (to remove direct sunlight) |
| 4    | DNB Radiance (M_R)       | Pixels where the amount of Radiance is greater than 150 nW.cm⁻².sr⁻¹ is excluded |

875
See Figs. 9–11

\[
C_i = \left( \left( (M_{Si} \land 14) \gg 1 \right) \in \{2, 3\} \right)^{\text{Land}} \lor \left( \left( (M_{Ci} \land 192) \gg 6 \right) > 1 \right)^{\text{CloudConfidence}} \\
\left( \left( (M_{Ci} \land 256) \gg 8 \right) = 1 \right)^{\text{Shadow}} \lor \left( \left( (M_{Ci} \land 1024) \gg 10 \right) = 1 \right)^{\text{Snow}}
\]

\[
L_i = M_{Si} > 20\% \\
S_i = M_{Si} > 101^\circ \\
R_i = M_{Ri} > 150\text{ nW} \\
Z_i = \neg (C_i \lor L_i \lor S_i \lor R_i)
\]

GeoTIFFMatrixOperations : BooleanMaskswithBitShifts (\rightarrow) for finalRadianceLayer, \(Z_i\)

Fig. 9. Eight tiles are required to cover the extent of India (not including Andaman & Nicobar Islands).
Fig. 10. VNP46A1 processing workflow.

VNP46A1 NTL Data Processing Pipeline

The NTL Data Processing Pipeline uses a load-balanced autoscaling EC2 Cluster to process HDFS Data at scale. Data is first stored in S3 as it is downloaded from the NASA LAADS DAAC. EC2 servers access the data and using GNU Parallel and GO Libraries process files simultaneously, distributing load as necessary. Python is used to extract layers from the HDFS datasets. It is used to assign NTL values to geo-coordinates. The extracted NTL data can then be stored in a DB (eg. Redshift) or downloaded as CSV files.

Fig. 11. VNP46A1 processing AWS architecture.
