Overall Methodology:

The overall study involved 3 distinct steps as shown below:

• Geocoding nightlights radiance values to calculate sum of lights (radiance) within each national/state geo coordinates.

Panel Regression: to estimate the elasticity of nightlights

Machine Learning:

Temporal Data Collection and Preprocessing

Data augmentation for generation of composite for the required period from Daily/Monthly composi tes to generate Quarterly data sets.

1.formulating an algorithm computing quarterly state-level contributions to the national GDP.

2. Timeseries analysis of any socio-economic parameters using the NTL data.

3.Study the distribution and density of nighttime lights in rural areas to understand the population di stribution, settlement patterns, and expansion dynamics.

Geocoding Nightlight Radiance Values:

Layer: Cloud Quality Flag (MC).

Used to remove pixels that represent a) non-land surfaces; b) Snow c) Shadow; and d) Probably/Confident Cloudy.

Layer: Moon Illumination (MM).

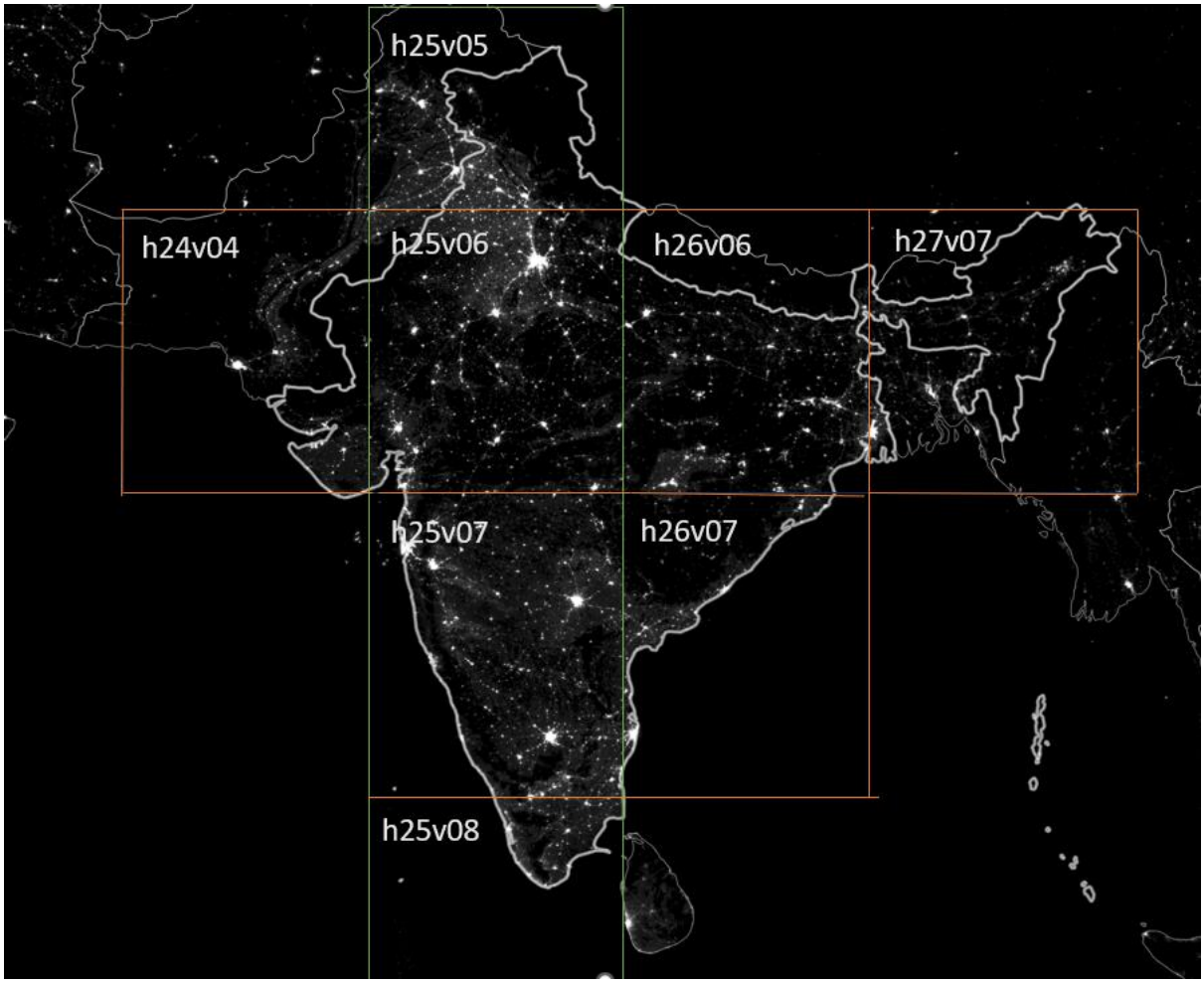
Used to remove pixels where the fraction of lunar illumination is greater than 20%.

Layer: Solar Zenith (MS).

Used to remove pixels where the Angle of the Solar Zenith is less than 101 Degrees (to remove direct sunlight)

Layer: DNB Radiance (MR).

Pixels where the amount of Radiance is greater than 150 nW.cm-2.sr-1 is excluded



Tools Used:

The computations on GeoTIFF matrices are extremely resource intensive. In total, more than 98,000 matrices, each of size 2,400 × 2,400, i.e., 560 billion Values were needed to prepare the quarterly Night light Data.

Layer:

Cloud Quality Flag (MC): Used to remove pixels that represent

a) non-land surfaces;

b) Snow

c) Shadow; and d) Probably/Confident Cloudy

Moon Illumination (MM):

Used to remove pixels where the fraction of lunar illumination is greater than 20%

Solar Zenith (MS):

Used to remove pixels where the Angle of the Solar Zenith is less than 101 Degrees (to remove direct sunlight)

DNB Radiance (MR):

Pixels where the amount of Radiance is greater than 150 nW.cm-2.sr-1 is excluded

Parallel processing tools were used extensively due to the scale of the data. In particular, multiprocessing in Python (v3.6.6), foreach - do Parallel in R (v4.0.0) and GNU Parallel in Ubuntu Linux (v18.04.5) (Tange, 2020) were used for parallel processing, the Geospatial Data Abstraction Library 262 (GDAL) was used for geo-spatial analytics and the R package, exacerbate, was used for assigning radiance values at geo-coordinate

Panel Regression:

The standard 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 groups.

The elasticity of nightlights and electricity consumption with respect to GDP at the national level and GSVA and at the state-level was estimated using a Panel Regression Framework.

A log-log model is used where the coefficients for each variable represents the percentage change in the outcome variable for 1% change in the corresponding regressor.

The gradient of Sum of Electricity vs GSVA ((rightmost image) is steeper relative to the gradient of Sum of Lights vs GSVA (middle image). This shows a stronger linear trend for electricity compared to nightlights and also highlights the motivation of using electricity as an additional regressor.

First, a pooling model is estimated assuming βit = β for all {I, t} where I is the group effect and t, 277 the time effect. Next, Fixed Effect Panels, also known as "no pooling" (Gelman and Hill, 2009), was 278 estimated to model the heterogeneity across quarters (National) and States (Sub-National). Finally, Random Effects, i.e., "partial pooling" (Gelman and Hill, 2009), has been used to model for both time-invariant individual effects as well as factors that change with time.

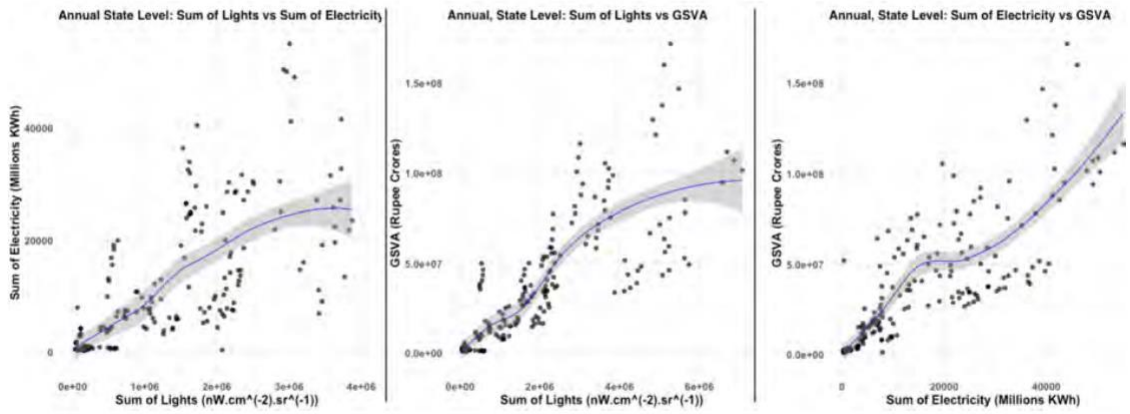


Figure 4: Nightlights, Electricity Usage, GDP

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 284 cross sectional dependency, the 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).

Equation:

$y_{It} = \beta_1 x_{It1} + \beta_2 x_{It2} + \ldots + \beta_k x_{Itk} + \alpha_I + u_{It}, t = 1, 2, \ldots, T$

Where:

α (I = 1, ..., 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,

β_1 is the coefficient of the independent variable,

u_{It} is the error term

For the Random Effects model, an intercept-term is added to Eqn. 1. It is assumed that the fixed effect, α_I is independent of all regressors across all time periods. The Random Effect model used for State-Level Panel Regression uses State as the Entity and Year as the Time-Effect. $y_{It} = \beta_0 + \beta_1 x_{it1} + \beta_2 x_{it2} + \ldots + \beta_k x_{Itk} + \alpha_I + u_{It}, t = 1, 2, \ldots, T$

Table 4: Explanatory Variables (National Panel)

Term	Variable Description
$ln(SumOfLights)$	log of the sum of the radiance values of the pixels
$ln(SumElectricity)$	log of total Electricity Usage
$ln(SumOfLightsSq)$	log of the square of sum of lights
$ln(Population)$	log of Population
$ln(GDP)$	log of National GDP
$ln(GVA_{Sector})$	log of National Sectoral GVA

$$ln(GDP) = \beta_1 ln(SumOfLights)_t + \beta_2 ln(SumElectricity)_t + \beta_3 ln(SumOfLightsSq)_t + \beta_4 ln(Population)_t + \alpha_i + u_{it} \quad (3)$$

$$ln(GVA_{Sector}) = \beta_1 ln(SumOfLights)_t + \beta_2 ln(SumElectricity)_t + \beta_3 ln(SumOfLightsSq)_t + \beta_4 ln(Population)_t + \alpha_i + u_{it} \quad (4)$$

Table 5: Explanatory Variables (State-Level Panel)

Variable	Variable Description
$ln(GSVA)$	log of Gross State Value Added
$ln(SumOfLights)$	log of the sum of the radiance values of the pixels
$ln(Population)$	log of total state population
$ln(SumElectricity)$	log of total Electricity Usage
$ln(Area)$	log of total area (sq. km.)

$$ln(GSVA) = \beta_1 ln(SumOfLights)_t + \beta_2 ln(SumElectricity)_t + \beta_3 ln(Area)_t + \beta_4 ln(POpulation)_t + \alpha_i + u_{it}$$

Several Machine Learning - based algorithms were used with National GDP and State GSVA as the dependent variable. The set of models used were SVM with Radial Basis Function (RBF) Kernel, K-Nearest Neighbours, Random Forest extreme Gradient Boosting (Boost) which implements a Gradient Boosted Model and Lasso/Ridge Regression.

Table 6: Machine Learning Features (National/State Models)

Variable Scope	Variable	Variable Description
National	y	gdp (Gross Domestic Product)
State	y	gsva (Gross Domestic Product)
National, State	x_1	vnpyr_sol (sum of lights)
National, State	x_2	vnpareayr_sol (sum of pixels > 0)
National, State	x_3	sumelec (total electricity usage)
National, State	x_4	meanelec (mean electricity usage during the quarter/year)
National, State	x_4	pop (total population)
National	x_5	precip (total precipitation)

First, the features were scaled and centred to have mean 0 and a range between (0,1). Next the dataset was split into an 80/20 Training Set / Test Set sample. 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/targeted range of hyperparameters. The variable importance measure, which ranks the regressors in the order of most to least predictive was also estimated.

In the Pooled OLS model, the coefficient of Sum of Lights is 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 the Fixed-Effect models (3-6) with Quarterly FE, the coefficient of Sum of Lights is 0.189 when it is 384 the only independent variable.

The premium drops to 0.023 when population is added. Population is significant across all models 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 and the results are intuitive.

Table 10: Panel Regression (India Quarterly) with Robust SE

	Dependent variable:					
	log(Gross Domestic Product)					
	Pooled OLS	Pooled OLS	Qtr FE	Qtr FE	Qtr FE	Qtr FE
	(1)	(2)	(3)	(4)	(5)	(6)
log(Sum of Lights)	0.189***	0.113***	0.189***	0.023***	0.103*	0.018***
	(0.011)	(0.012)	(0.030)	(0.007)	(0.053)	(0.006)
log(Sum of Electricity)		0.313***			0.346	0.010
		(0.078)			(0.259)	(0.042)
log(Sum of Lights^2)				−0.001		
				(0.001)		
log(Population)				5.547***		5.569***
				(0.281)		(0.342)
Constant	11.297***	9.130***				
	(0.202)	(0.758)				
Observations	26	26	26	26	26	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

Diagnostics for State-Level Panel shown indicated presence of Cross-Sectional Dependency, Serial Auto-Correlation and Heteroskedasticity and accordingly, Robust SE values were estimated and reported in the output.

The variable, Sum of Lights, was significant at the 1% level across all models even after controlling for state, year, area and population effects. The elasticity of Sum of Lights, 0.735, in the Pooled OLS model where it is the only dependent variable overestimates the true effect and when the unobserved effect is removed, it drops to 0.263 but still remains statistically significant.

After controlling for State as the entity and Year as the time-effect, the coefficient drops further to 0.191. With population as an additional regressor, the coefficient becomes 0.167. When all variables are regressed (5), the variables, population and sum of lights remain significant at the 1% level.

At the sector level, the coefficients of all sectoral variables were significant at the 1% level with the effect of the Primary, i.e., agrarian sector being the highest at 1.24.

Table 11: Panel Regression (State Yearly) with Robust SE

	Dependent variable:					
	log(Gross State Value Added)					
	Pooled OLS	Pooled OLS	State/Year	St/Yr	St/Yr	St/Yr
	(1)	(2)	(3)	(4)	(5)	(6)
log(Sum of Lights)	0.735***	0.263***	0.191***	0.195***	0.167***	0.163***
	(0.056)	(0.056)	(0.060)	(0.005)	(0.006)	(0.010)
log(Sum of Electricity)		0.664		0.015		−0.005
log(Area)						−0.175
log(Population)					0.710	0.850***
						(0.139)
Constant	5.137***	6.634***		13.362	5.381***	5.716***
	(0.926)	(0.926)			(0.400)	(1.557)
Observations	221	218	221	218	220	217
R^2	0.738	0.895	0.883	0.777	0.889	0.900
Adjusted R^2	0.736	0.894	0.865	0.775	0.888	0.898
Note:						*p<0.1; **p<0.05; ***p<0.01

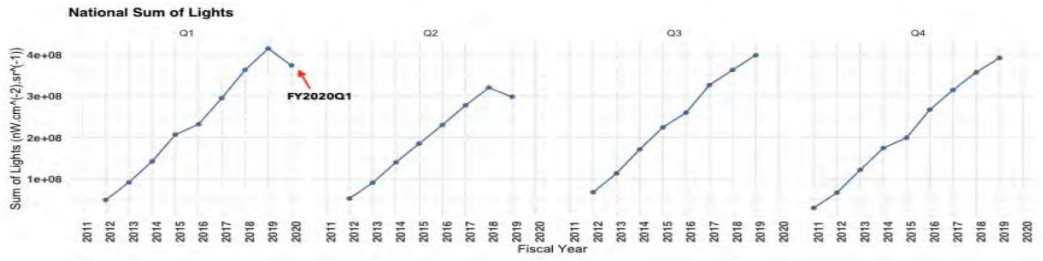


Figure 5: National Sum of Lights

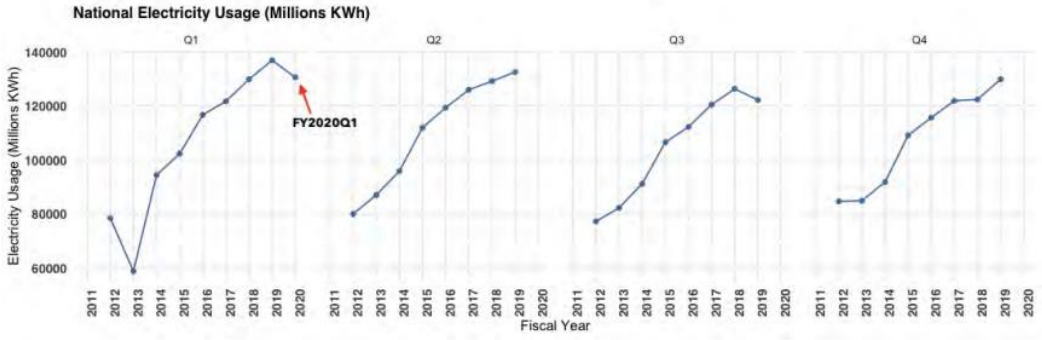


Figure 6: National Electricity Usage

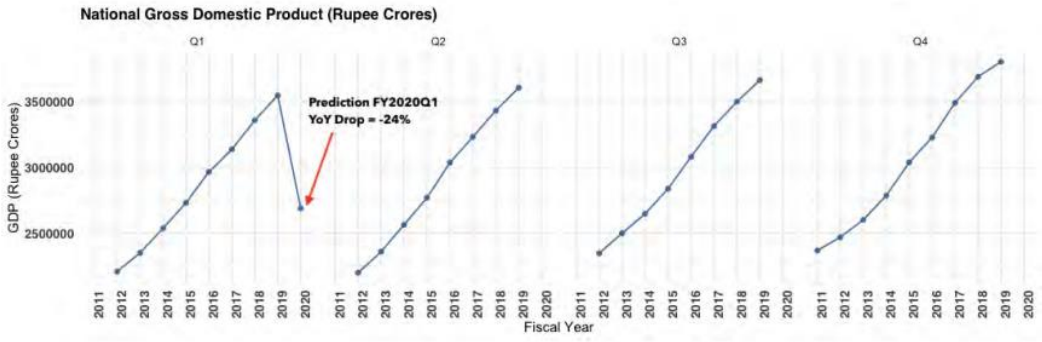


Figure 7: National GDP with FY2020Q1 Prediction

Analysis:

This study marks the first time the newly released VNP46A1 Black Marble Night Light Dataset has 406 been used for economic research.

In order to assess the quality of the extracted data, a comparison of the nightlight values between VNP46A1 and DNB Composite, the gold standard, were performed for 408 data between 2012-2019. The radiance values were found to be strongly correlated (0.85).

However, the results while similar were not perfect. There could have been several reasons for the difference.

First, VNP46A1 has a much higher nightlights detection limit and a much narrower confidence band of 0.5 nW.cm-2.sr-1 (± 0.10 nW.cm-2.sr-1) compared to 3.0 nW.cm-2.sr-1 (± 3.0 nW.cm-2.sr-1 411) in the erstwhile system.

This means VNP46A1 dataset is able to capture minute radiance values that the DNB Composite is unable to detect. Second, a wider temporal averaging window of 90-days could have also reduced extraneous sources of light that may remain on DNB Composites that use a 30-day averaging window.

It is too early to tell which estimate is closer to the actual radiance, but research by Román et al. 416 (2019) at NASA has indicated that VNP46-related dataset produces more accurate results.

Main Findings:

The radiance data extracted from the geocoding of raw nightlights across nearly 100,000 images (VNP46A1) were closely correlated with the official dataset called DNB Composite. This demonstrates that high-quality nightlight dataset can be developed at scale with parallel processing tools and publicly available high-performance servers.

Electricity usage had a high predictive value and since it is easier to obtain electricity usage data, it should be explored further, similar to studies that have been already conducted elsewhere.

Precipitation data was also found to improve model performance and linear models had better performance at the national level, confirming observations by other researchers.

Hence, nightlights and electricity usage could be invaluable 517 proxies and leading indicators of economic changes.

Table 12: 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

Table 13: Panel Regression (State-Level Sectoral GSVA): VNP46A1

	Dependent variable:			
	Log Sectoral Gross State Value Added (Primary, Secondary, Tertiary)			
	log(GSVA)	log(GSVA Pri)	log(GSVA Sec)	log(GSVA Tert)
	(1)	(2)	(3)	(4)
Log Sum of Lights	0.9907***	1.2445***	1.0447***	0.9668***
	(0.0363)	(0.0327)	(0.0476)	(0.0406)
Constant	2.9160***	−2.2617***	0.8477	2.5099***
	(0.4981)	(0.4493)	(0.6531)	(0.5571)
Observations	171	171	171	171
R^2	0.8152	0.8954	0.7405	0.7706
Adjusted R^2	0.8141	0.8948	0.7390	0.7693
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 15: State-Level Impact (₹ crores)

state	y2019q1	y2019q4	y2020q1	YoYPctChange	QoQPctChange	YoYPctChangeGSVA	Cases	CovCasePerM
Arunachal Pradesh	68,129.84	99,617.64	46,620.80	-46.14	-113.68	-8.81	187	135.14
Meghalaya	51,720.60	57,579.89	35,614.68	-45.22	-61.67	-8.64	47	15.84
Sikkim	7,700.26	12,828.56	5,479.83	-40.52	-134.11	-7.74	88	144.13
Jammu and Kashmir	162,609.54	368,190.35	126,728.21	-28.31	-190.54	-5.41	7,237	589.96
Uttar Pradesh	2,078,088.80	2,142,802.00	1,819,696.40	-14.20	-17.76	-2.71	22,828	114.25
Uttarakhand	164,452.50	217,405.83	147,815.29	-11.26	-47.08	-2.15	2,831	280.68
Himachal Pradesh	80,732.48	128,856.20	73,281.42	-10.17	-75.84	-1.94	942	137.23
Manipur	43,202.04	45,012.01	39,274.04	-10.00	-14.61	-1.91	1,227	477.36
Chhattisgarh	496,625.05	519,638.20	454,077.00	-9.37	-14.44	-1.79	2,761	108.08
Goa	33,836.77	32,541.02	31,013.05	-9.10	-4.93	-1.74	1,198	821.37
Madhya Pradesh	1,273,862.00	1,183,024.90	1,186,534.70	-7.36	0.30	-1.41	13,370	184.09
Odisha	545,926.60	510,866.10	512,998.90	-6.42	0.42	-1.23	6,859	163.41
Rajasthan	1,427,543.90	1,358,830.90	1,351,563.60	-5.62	-0.54	-1.07	17,660	257.63
Nagaland	30,544.35	34,680.18	29,074.39	-5.06	-19.28	-0.97	434	219.36
West Bengal	637,510.20	686,535.35	613,745.40	-3.87	-11.86	-0.74	17,907	196.18
Bihar	725,374.20	837,156.90	698,425.10	-3.86	-19.86	-0.74	9,640	92.60
Jharkhand	401,278.95	435,431.25	387,076.68	-3.67	-12.49	-0.70	2,426	73.54
Puducherry	9,582.87	8,290.61	9,266.92	-3.41	10.54	-0.65	619	496.01
Haryana	470,288.60	577,432.00	456,313.20	-3.06	-26.54	-0.58	14,210	560.52
Assam	270,025.73	275,095.15	262,512.55	-2.86	-4.79	-0.55	7,752	248.42
Mizoram	20,763.93	24,184.00	20,313.05	-2.22	-19.06	-0.42	148	134.89
Tripura	52,148.97	55,769.86	51,149.52	-1.95	-9.03	-0.37	1,380	375.62
Maharashtra	1,577,750.20	1,425,958.50	1,600,832.90	1.44	10.92	0.28	169,883	1,511.76
Andhra Pradesh	601,449.50	566,276.90	615,045.00	2.21	7.93	0.42	13,891	280.19
Chandigarh	14,046.83	16,507.29	14,376.50	2.29	-14.82	0.44	435	412.15
Gujarat	1,045,581.20	961,591.90	1,077,335.10	2.95	10.74	0.56	31,938	528.43
Karnataka	1,030,689.50	936,037.30	1,064,347.20	3.16	12.06	0.60	14,295	233.98
Punjab	434,427.95	598,657.75	456,904.30	4.92	-31.02	0.94	5,418	195.29
Tamil Nadu	774,875.80	710,981.95	821,821.55	5.71	13.49	1.09	86,224	1,195.12
Telangana	654,522.35	619,840.40	696,416.25	6.02	11.00	1.15	15,394	439.78
Kerala	151,349.03	159,738.25	171,775.85	11.89	7.01	2.27	4,189	125.40

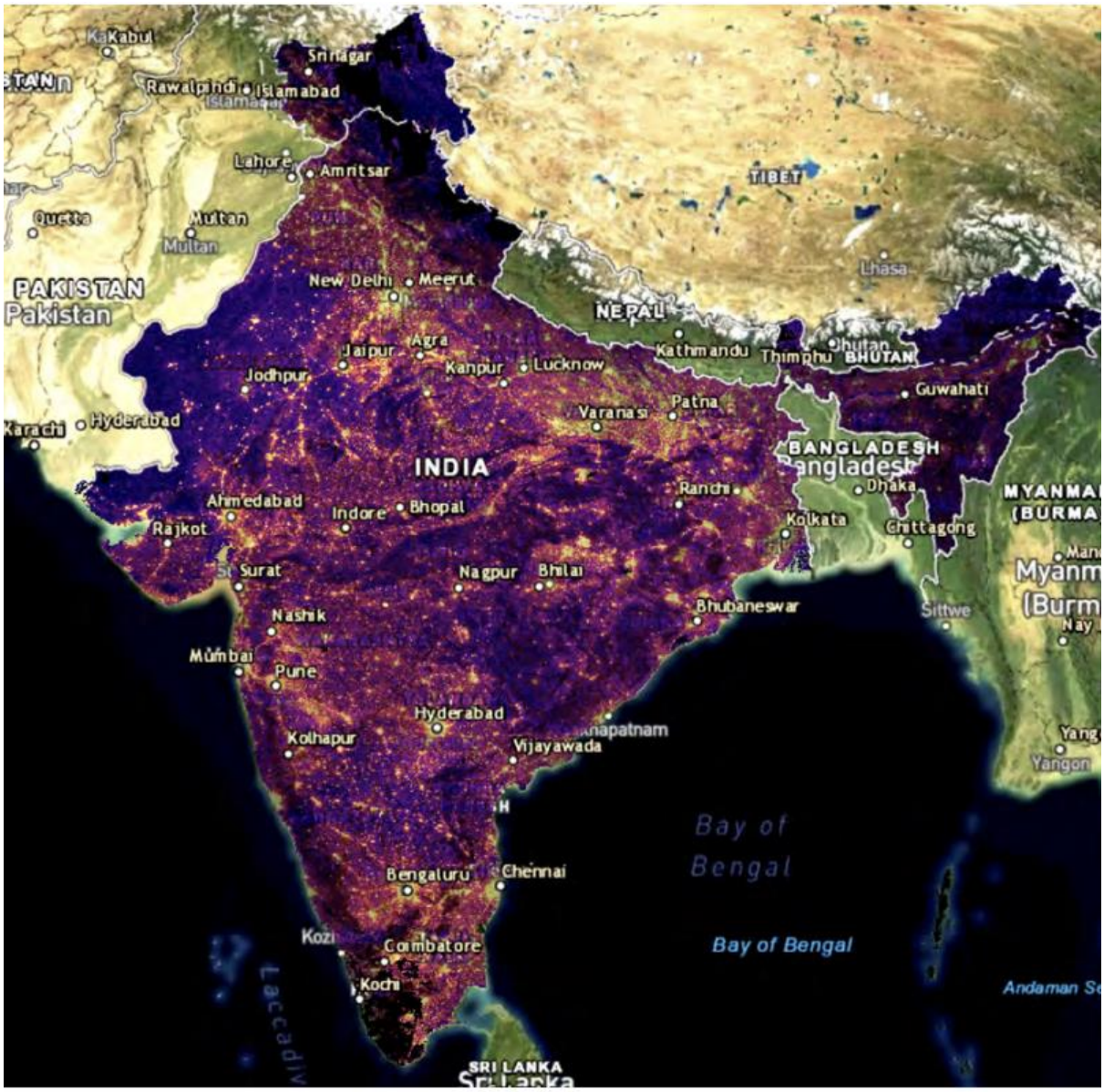


Figure 17: Visualisation of Nightlights in India using VNP46A1

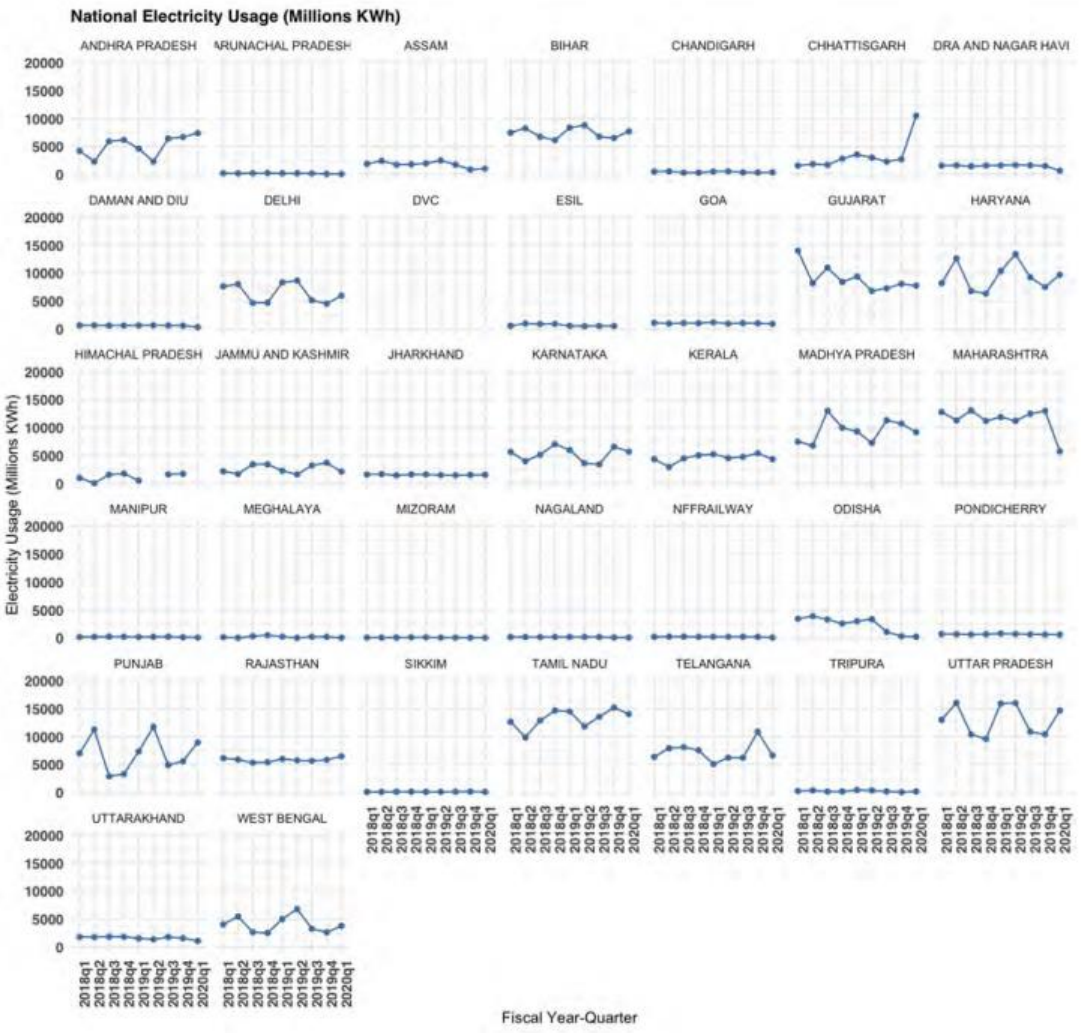


Figure 14: State Level: Electricity Usage

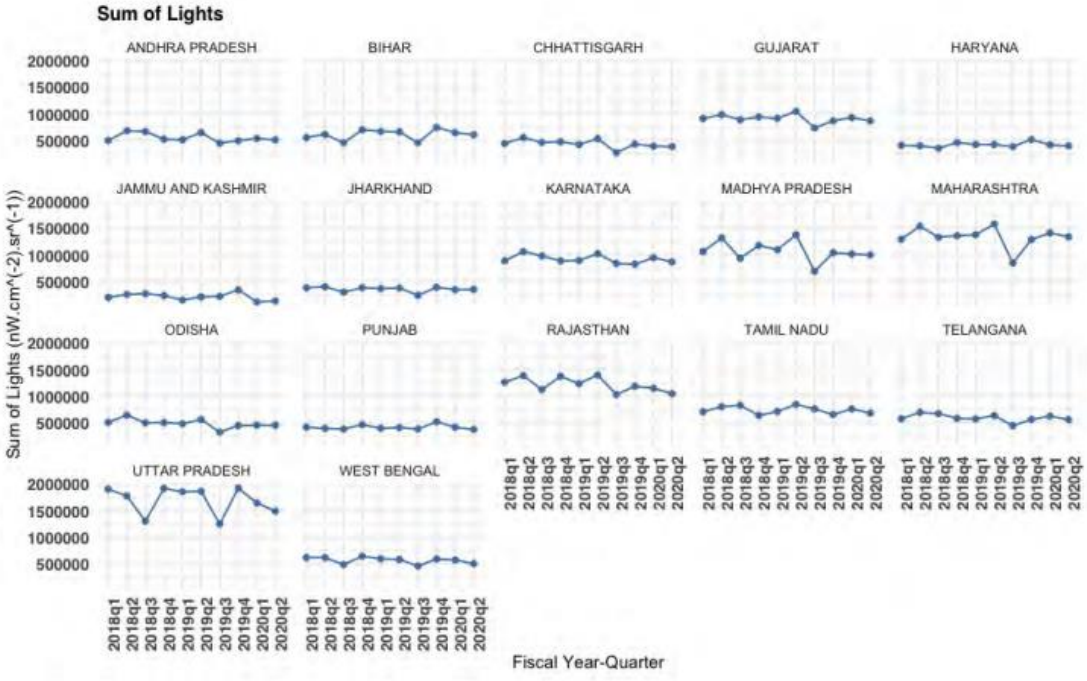


Figure 13: State Level: Sum of Lights (I of II)

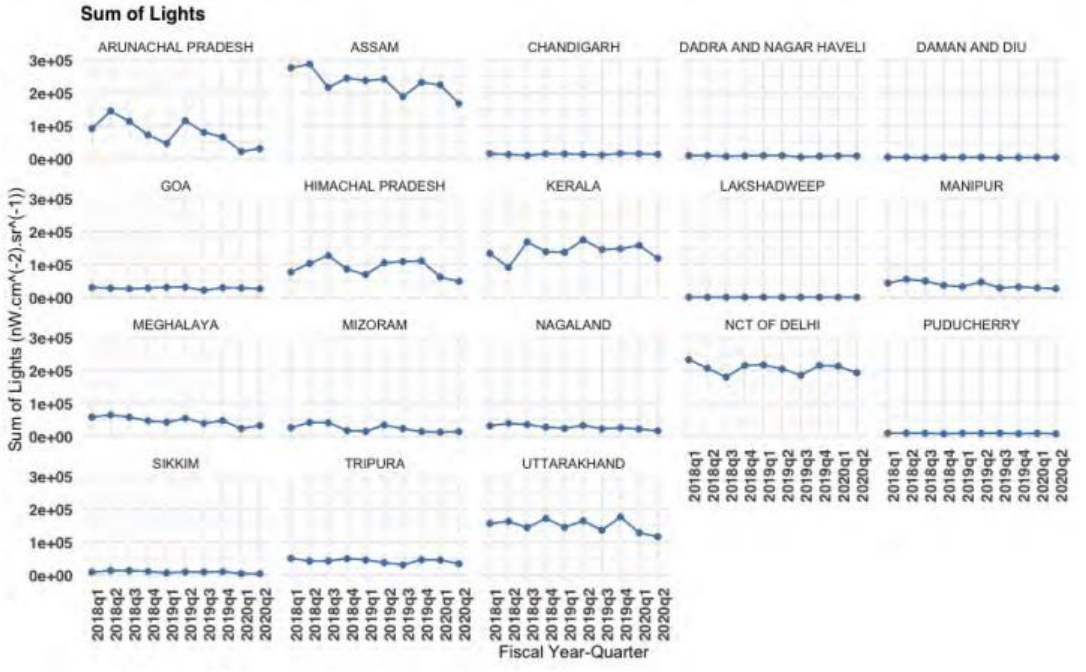


Figure 12: State Level: Sum of Lights (I of II)

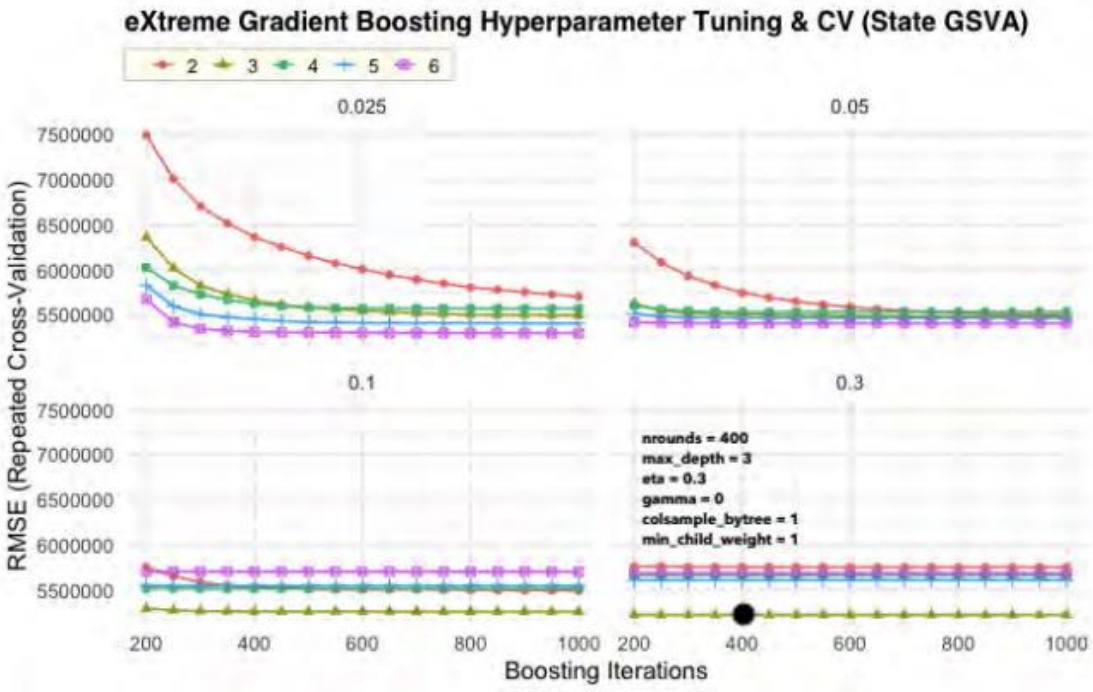


Figure 10: State eXtreme Gradient Boosting CV Results

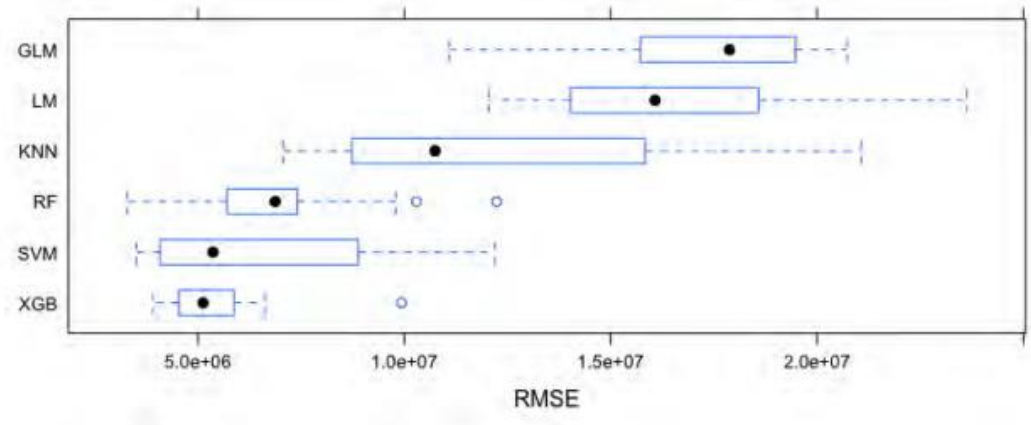


Figure 11: Comparative Model Performance: State RMSE

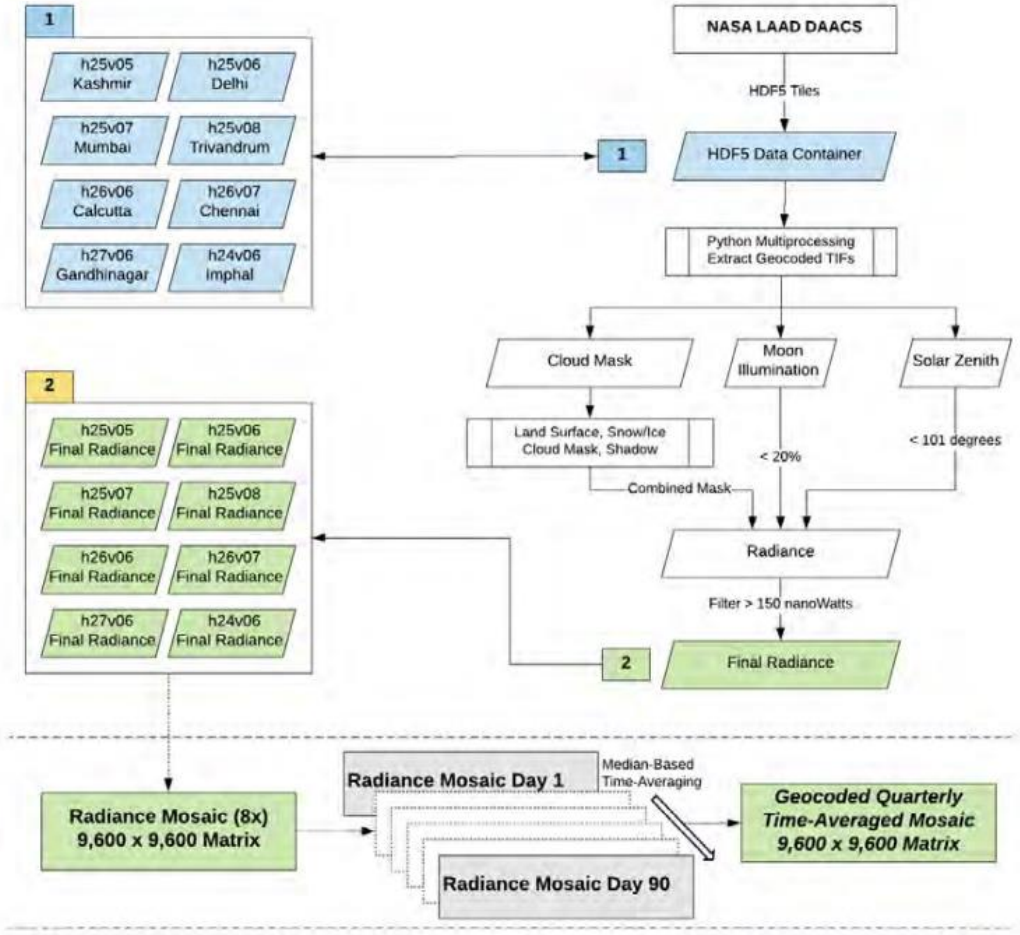


Figure 19: VNP46A1 Processing Workflow