Tracking Socio-Economic Development in Rural India over Two Decades Using Satellite Imagery

ANANT GULGULIA, AMAN GUPTA, AKSHAY P SARASHETTI, AADITYA SINHA, and AADITESHWAR SETH, Indian Institute of Technology Delhi, India

Longitudinal analysis of socio-economic development at sub-national scales can reveal valuable insights about which areas tend to develop faster than others and why. However, such analysis is difficult to conduct with traditional data sources such as censuses and surveys which are not repeated frequently and may require assumptions for imputation of values at non-surveyed locations. Indicators of socio-economic development based on satellite data have emerged as a proxy to track development at fine spatial and temporal scales. We build a model using daytime and nightlights satellite data to estimate an index of socio-economic development at the village level in India. We evaluate our model for temporal robustness and use it to produce estimates at three time points over a two-decade period. We then use these estimates to understand the effect on village-level development of factors such as the geographic distance of a village to hubs of economic activity and the inequality of development in the district. Our findings provide evidence of the possible impact that policy changes during this period have had on village development.

CCS Concepts: • **Applied computing** → **Economics**; • **Computing methodologies** → **Machine learning**; **Computer vision;**

Additional Key Words and Phrases: Poverty mapping, satellite data, nightlights, socio-economic development, inequality, census

ACM Reference format:
Anant Gulgulia, Aman Gupta, Akshay P Sarashetti, Aaditya Sinha, and Aaditeshwar Seth. 2023. Tracking Socio-Economic Development in Rural India over Two Decades Using Satellite Imagery. *ACM J. Comput. Sustain. Soc.* 1, 2, Article 12 (December 2023), 31 pages.
https://doi.org/10.1145/3615361

12

1 INTRODUCTION

Longitudinally tracking socio-economic development at sub-national scales can provide important insights into underlying development processes that arise from and lead to diversity in development levels within countries [1–3]. Regularly conducted surveys and censuses are standard tools for this purpose. However, surveys and censuses may sometimes not be comparable over successive rounds either due to changes in sampling strategies or the parameters that are collected, or may not be conducted regularly, or the data may not be made available in a timely manner. For example, India conducts a household census every 10 years, but the data takes several additional years to be released. The Indian census originally scheduled for 2021 was inordinately delayed due

Authors' address: A. Gulgulia, A. Gupta, A. P. Sarashetti, A. Sinha, and A. Seth, Indian Institute of Technology Delhi, IIT Campus, Hauz Khas, New Delhi, Delhi 110016, India; e-mails: anant9821@gmail.com, Aman.Gupta.cs119@cse.iitd.ac.in, aps1310@gmail.com, aadityasinha1019@gmail.com, aseth@cse.iitd.ac.in.

2834-5533/2023/12-ART12 $15.00
https://doi.org/10.1145/3615361

to COVID-19 disruptions. Partly to counter this lack of data availability, and to bring some standardization to development indicators that can be compared over several years, proxy variables are often used such as nightlights satellite data or machine learning models trained on daytime satellite imagery that produce estimates of wealth indexes and other variables [4–6]. A few gaps, however, remain in related research in this space, especially for studies in the Indian context.

First, models to estimate socio-economic indicators using daytime satellite imagery have not been adequately evaluated for temporal robustness, i.e., whether models trained on data from one year can be used to produce estimates for other years. In prior work, we attempted to study this at the district level in India [9]; in this article, we describe our studies at the village level. Second, whereas nightlights time series have been explored to track sub-national scale development over time [7], the low variability for rural areas does not render this method suitable by itself for village-level tracking. We build a method that uses both daytime satellite imagery and nightlights data to estimate a composite indicator of socio-economic development at the village level. Some unique aspects of our method to estimate socio-economic development are as follows. Similar to Jean et al. [6], we use pre-trained **Convolutional Neural Networks (CNNs)** on a variant of a ResNet architecture to learn a model that can produce first-level estimates of development variables using data from the 2011 Indian census as ground truth labels. We then improve the estimates by building a model to also take other features into account, namely the first-level estimates of development variables of neighbouring villages, nightlights-based features for the given village and neighbouring villages, and distance of a village to a nearest hub of economic activity (obtained also from nightlights data). We then do a feature selection specifically to ensure temporal robustness, by identifying those sets of features that produce the most accurate estimates for 2001—that is, we train the model on census data from 2011 and evaluate its accuracy on census data from 2001 on those indicators that are available for both of these census years. Given a satisfactory temporally robust model, we then produce standardized socio-economic development estimates for 2001 and 2019, and go on to test various hypotheses of village development dynamics.

We build a composite indicator that combines variables related to asset ownership, access to water, bathroom facilities, literacy, and so forth in a manner similar to how the Human Development Index is calculated by giving equal weight to economic, education, and health variables. We term this the **Aggregate Development Index (ADI)** and use the satellite data based models to produce village-level estimates for the years 2003, 2011, and 2019.[1] We then conduct an econometric analysis to explain village-level development changes over these (approximately) two decades in terms of covariates such as the distance of a village to hubs of economic activity, its development relative to other villages in the district, and the inequality of the district to which the village belongs. We find that villages farther away from economic hubs tend to have a slower pace of development. We also find that less developed villages in general tend to have a faster rate of development, indicating a catch-up phenomenon. At the district level though, our analysis reveals that villages in more unequal districts were developing faster during the 2003–2011 period, but this changed during the 2011–2019 period. We discuss that a likely reason may be changes in the Indian economic policies over these two decades.

To the best of our knowledge, our work is the first attempt at a methodology using satellite data to study village-level changes over such a long time span, and especially to study the changes in the

[1] As we explain later, the data source we used for daytime satellite imagery is available from 2003 onward. We consider this to be sufficiently close to the census year of 2001 and therefore use satellite data from 2003 to evaluate our estimates of socio-economic development against the census data for 2001. Both the census and satellite datasets coincide for 2011. We then choose 2019 as the next year for which we obtain satellite data to analyze socio-economic development over this time span of almost two decades.

context of India. We produce ADI estimates for the past two decades for 14 states, comprising more than 77.8% of the population of India, to serve as an important data product for other development economics research.

Our work is not without its limitations. More recent CNN-based architectures such as those using attention models may produce more accurate estimates than our models. An end-to-end neural network architecture instead of the two-stage setup that we developed may also prove to be better. Finally, instead of using models pre-trained on the ImageNet data such as ResNet, models pre-trained on labeled satellite data such as the DeepSat model [46] may perform better, and we invite other researchers to improve upon our work. All data and code from our effort has been made available as open source.[2]

2 RELATED WORK

With improvements in the availability of satellite data, and its use in machine learning based methods to obtain estimates of socio-economic variables, several novel applications have emerged in development economics. This includes inferences for population density [10], gridded estimates of poverty mapping [11], use in targeting of social welfare support [13], and so on. In this section, we describe some such efforts and the scope of using satellite data based estimation of socio-economic development variables to study some fundamental hypotheses in development research.

2.1 Use of Nightlights Data

Satellite imagery captured during night hours of the intensity of light emitted by human-made lighting sources on the earth's surface is called *nightlights data*. It is captured by the Suomi NPP satellite system at a spatial resolution of ~500m and a temporal resolution of 12 hours and is made openly available in the form of NPP-**Visible Infrared Imaging Radiometer Suite (VIIRS)** series (referred to as *VIIRS data* henceforth, available from 2013). An earlier **Defense Meteorological Satellite Program (DMSP)**-OLS series, referred to as *DMSP data* henceforth, is available from 1992 to 2013. These have been shown to correlate at the country level with a nation's GDP [33, 34], and have also been used in other domains such as to assess the impact of war and post-war recovery measures in Syria [35], study the relationship between electoral cycles in north India with selective electrification of certain constituencies to influence their voting patterns [37], and track the impact of international trade sanctions on economic activities in North Korea [7], among others. GDP estimation at sub-national scales of states in India has also been studied [5]. Nightlights data, however, suffers from over-saturation in urban centers, and unobservant light intensities in rural areas, which renders it inadequate by itself to study socio-economic factors at fine spatial scales of villages, and within cities [5, 8]. In our work, we use features from nightlights data in conjunction with daytime imagery to overcome such challenges.

2.2 Use of Daytime Satellite Data

Spectral data of surface reflection from the earth's surface is captured by satellite systems such as Landsat (at a 30m resolution, with re-visit times of <16 days) and Sentinel-2 (at a 10m resolution, with re-visit times of <10 days), and this openly available optical multi-spectral data has been shown to be effective for several tasks. A CNN model trained on such data to predict the wealth index was shown to outperform models using nightlights data alone [39]. A transfer learning approach to fine-tune CNN-based models on nightlights data, and then fit regression models to estimate the final output, improved such methods further [6]. Similar supervised learning

[2]Code: https://github.com/amangupt01/Village_Development_Model
Dataset: https://drive.google.com/drive/folders/1xtaTGiaPJxDLr2t4RRqHYyJBkS3NcXTm?usp=sharing

techniques are able to estimate population density and poverty at the village level, using higher-resolution data from Google Maps [40]. Another popular set of applications is Land Use and Land Cover maps, to use spectral data for each satellite pixel to classify land surface into built-up and non-built-up areas [42], and further into forests, croplands, water bodies, and so on [43]. Both pixel-based machine learning and CNN-based deep learning models for crop coverage and consumption in developing countries have also been studied through daytime satellite imagery [41]. In our work, we use similar methods to obtain village-level models for the estimation of census data indicators such as asset ownership, main source of lighting, and main source of water, with the expectation that land surface features predictive of built-up areas, water bodies, and so on will also help estimate related census-based indicators. Our work is different from that of Jean et al. [6], which uses CNN-based based models trained on geographic clusters where the DHS survey data was conducted, and then produces gridded estimates of the wealth index for non-surveyed locations. More recent work in this space has improved the prediction by also using building footprints to geo-locate the survey locations more precisely [21]. We instead use village-level census data and directly train CNN-based methods to predict the census variables. We also do a careful feature selection to ensure temporal robustness in our models.

2.3 Understanding Socio-Economic Development

Our goal is to use longitudinal estimates of socio-economic development at the village level to understand development patterns in terms of covariates such as the distance of a village to hubs of economic activity, its development relative to other villages in the same district, and the inequality in the district. We draw inspiration from the Kuznets hypothesis, which states that inequality (in terms of the Gini coefficient) within a country first rises and then drops with improvements in the development (in terms of GDP per capita) of the country [30]. Although the hypothesis has been heavily debated [18, 19], we are motivated to test it at the district level to understand factors that may cause districts to converge or diverge on measures of inequality. We also draw motivation from research that establishes the relationship between village development and road infrastructure [31], and studies in public health that explore the impact of the distance of a village to health facilities on the health indicators of the village [20]. Such studies find that far-flung villages tend to have worse development indicators than villages more closely connected to areas that are already well developed. Similarly, other studies have looked at the effect of the mix of economic activities in and around a location on the development of the location [1], and motivates us to study the relationship of such factors with the socio-economic development of villages.

Our work is unique in that most other research in this area has relied on census and survey data to study changes over time. We contribute towards building a machine learning model that uses satellite data to produce estimates of village-level socio-economic development over two decades, and to use these estimates to study development patterns at a large scale. Our findings largely agree with what other literature has also found but also lays down a foundation to conduct such large-scale studies easily and draw new insights.

3 DATASETS

We next describe various datasets we use to build our model and pre-processing steps applied on the data.

3.1 Census of India

The Government of India conducts a census every 10 years. A total of 15 rounds have been conducted since 1872, the last being in 2011. The next census was supposed to happen in 2021 but was inordinately delayed, and no plans have been announced as of January 2023. This increases the relevance of our study to produce development estimates in the absence of an official census. The 2011 census collected data at two levels: House listing and People enumeration:

- *House listing*: A list of 35 questions was used to capture household data such as the construction material used for the roof of the house (cemented, tin roof, thatched, etc.), material for the walls, the main type of fuel used by the household for cooking (**Liquified Petroleum Gas (LPG)**, kerosene, fuel wood, etc.), the availability of a bathroom facility, the main source of water for the household (tap water, handpump, rivers, etc.), and asset ownership in the household (possession of a television, radio, car, motorbike, bicycle, etc.).
- *People enumeration*: A list of 30 questions at an individual level was used to gather data and produce village-level estimates for the total population in a village, the literate population, gender distribution, and nature of employment.

This data is used to develop indicators at the village level on the percentage of population using *bathroom facilities* of different types, *fuel for cooking* of various types, the *main source of water* for households, *asset ownership* in households, and the *literacy rates*. These indicators available from the census data for 2011 serve to create labels for the supervised machine learning models that we describe later.

Most of the preceding variables were not available in the 2001 census, but a different set of common variables did exist, such as binary variables for the presence of *telephone services*, *bus services*, *primary schools*, and *middle schools* in the villages. We used these variables as surrogate outputs from our models to identify features that are most important for temporal robustness of these models trained on census data from 2011 and to predict them for 2001.

3.2 Village Shapefiles

We needed the shapefiles of villages to obtain satellite data for each village. The Government of India has not officially released the shapefiles that were used to delineate village boundaries for the 2011 census, but shapefiles from the 2001 census are available and hosted on the NASA-SEDAC website.[3] The SHRUG project has further mapped village ids from the 2001 census with village ids from the 2011 census [12]. This therefore allows us to obtain village shapefiles and map them with the 2011 census data. Since village boundaries may change based on local administrative decisions, the SHRUG researchers had to undertake a rigorous exercise to detect these changes to merge some villages or split a larger village into smaller units. They used district handbooks that contained details of this mapping in narrative format, built string matching algorithms, and used other sanity checks such as ensuring that the population estimates across the mapped pair of villages were of a similar order as expected based on standard population growth rates seen in India during this time. For this reason, not all the villages could be uniquely mapped by them. Several villages were also dropped as a result of pre-processing steps undertaken by us on the satellite data, such as villages for which cloud-free imagery was not available, or satellite data was missing due to faults in the Landsat 7 sensors. These statistics are shown in Table 1.

3.3 Landsat 7 Satellite Data

The Landsat 7 satellite system has been providing optical data since 2003 and was our ideal choice as an openly available data source that went back almost two decades. It provides data for nine spectral bands, including the visible RGB (red, green, and blue) bands, at a resolution of 30×30 m^2 per pixel [14]. We used Landsat 7 data for 2003 to closely correspond with the census year of 2001, for 2011 to correspond with the census year of 2011, and for 2019 to correspond to approximately

[3] https://sedac.ciesin.columbia.edu/data/set/india-india-village-level-geospatial-socio-econ-1991-2001

Table 1. State-Wise: The Count of Villages as Per the 2011 Census, Villages for Which Shapefiles Were Available from the SHRUG Data, and Villages We Used for the Model Training and Validation

State	No. of villages	No. of villages with shapefiles	No. of villages used by us	Villages used (%)
Andhra Pradesh	26,705	22,193	10,217	46.04
Bihar	39,468	39,161	27,259	69.60
Chattisgarh	19,625	18,235	13,722	75.26
Gujrat	18,169	15,956	13,382	83.87
Haryana	6,801	6,550	2,442	37.28
Jharkhand	29,682	29,424	15,590	52.98
Karnataka	27,760	26,535	9,346	35.22
Maharashtra	41,474	39,725	33,169	83.50
Madhya Pradesh	52,372	51,799	46,248	89.28
Orissa	47,919	43,972	23,378	53.17
Punjab	12,383	12,140	8,792	72.42
Rajasthan	43,503	34,239	27,763	81.09
Tamil Nadu	16,119	14,756	4,499	30.49
Uttarpradesh	98,946	97,154	69,650	71.69

a decade later so as to keep roughly a 10-year separation between the various measurements. The data was downloaded through the Google Earth Engine (GEE) and cut along village shapefiles that were obtained from the SHRUG mapping described earlier. Each pixel thus carries data for the nine spectral bands. We only used the RGB bands which had for each band a digital value in 8-bit units.

3.4 DMSP and VIIRS Nightlights Data

Nightlights data has been used extensively as a proxy for economic development at regional scales, generally at the state (province) level [4, 5, 7, 33, 34, 37], and is available from two satellite systems over mostly non-overlapping periods of time. From 1992 to 2013, the DMSP monitored meteorological, oceanographic, and solar-terrestrial physics for the U.S. Department of Defense and made available the processed time series of visible and near-infrared (VNIR) emission sources at night at a spatial resolution of 2.7km and temporal resolution of 12 hours. Since 2012, a new data source from the VIIRS instrument hosted on a different NASA mission has provided a similar time series at a ~500m spatial and 12 hours temporal resolution [15]. Further, efforts have been made to harmonize the two time series from DMSP and VIIRS so that a single metric can be built for a continuous time series of nightlights. For any year, each pixel thus carries the VIIRS radiance value at that location in units of nano watts per square centimeter per steradian. A global publicly available dataset uses other spectral indexes as covariates to calibrate the historic DMSP series and fit them to the more recent VIIRS series to produce an extended time series from 2000 to 2020 [32]. We use this harmonized product to obtain nightlights data for the various years of interest.

3.5 Pre-Processing: Discretization of Census Variables

Various house listing parameters collected in the census can be grouped into smaller categories. For example, the type of fuel used for cooking is described in terms of multiple parameters such as the percentage of households using firewood, those using cow dung, or kerosene, or LPG, **PNG (Piped Natural Gas)**, biogas, and so forth. In prior work at the district level [49], and following a similar methodology used by other researchers [16], we first clubbed these parameters into three broad

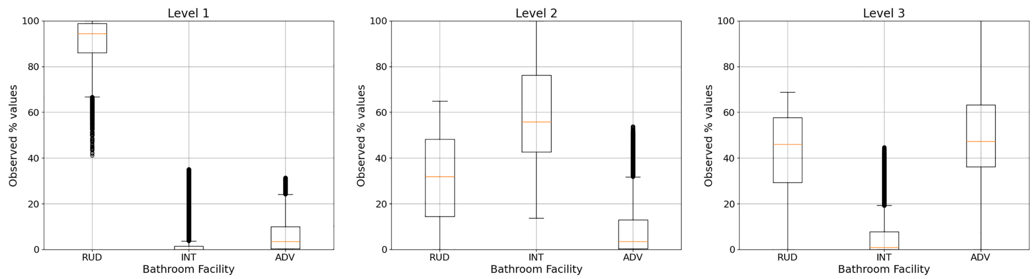


Fig. 1. Example of the clustering obtained to build labels for the type of fuel used for cooking. Cluster-1, Cluster-2, and Cluster-3 represent level 1 (low), level 2 (medium), and level 3 (high) developed villages in terms of fuel for cooking as an indicator for socio-economic development of the village.

types: **Rudimentary (RUD)** sources of fuel such as firewood and cow dung, **Intermediate (INT**) sources of fuel such as kerosene, and **Advanced (ADV)** sources of fuel such as LUG, PNG, and biogas. Next, we obtained clusters on a tuple of (the percentage of households using RUD sources of fuel, INT sources of fuel, and ADV sources of fuel). We found that using k-means clustering, with k=3, gave us three reasonably distinct clusters that could be ordinally arranged as level-1 villages that predominantly used RUD sources of fuel, level-2 villages that predominantly used INT sources of fuel, and level-3 villages that had the highest incidence of ADV sources of fuel. A box plot of this clustering is shown in Figure 1. A similar method was followed to build a categorization for level-1/2/3 villages for the bathroom facility in the house and for the main source of water. These are shown in the appendix. The grouping of parameters into RUD, INT, and ADV variables, and their distribution for the village levels obtained after clustering, is shown in Table 2. A slightly different method was required to build a similar level-1/2/3 categorization for asset ownership. Here, we performed a direct clustering on the percentage of families owning assets such as a television, a telephone, 2-wheeler vehicles, and 4-wheeler vehicles, because the ownership of these assets is not mutually exclusive in households. Similarly for literacy, we built clusters on the percentage of literate population in villages. These methods effectively allowed us to discretize multiple census parameters into a few categorical variables, lending it easier to formulate a supervised machine learning task with the village levels for various socio-economic indicators as labels.

3.6 Pre-Processing: Satellite Data

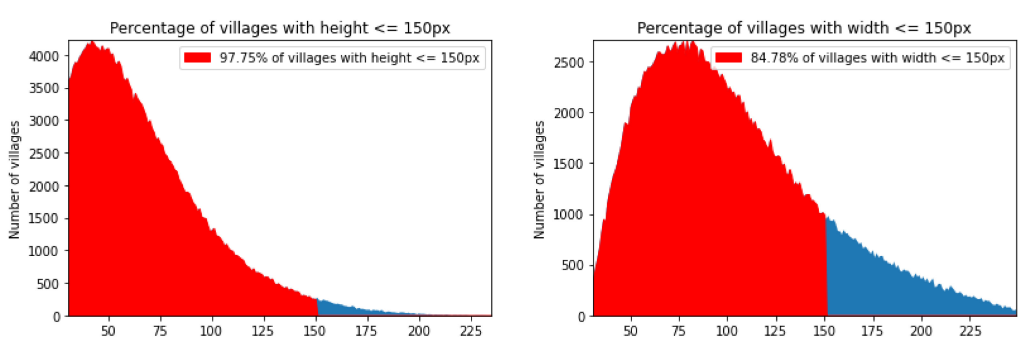
We download annual composites of the Landsat 7 satellite imagery data at the state level for the years 2003, 2011, and 2019 and then use village shapefiles to cut village images. We remove cases with high cloud cover or missing data due to sensor faults on Landsat 7 and finally crop the images to a standard size for the CNN-based models. To identify an appropriate image size to use, we checked several height and width combinations and finally decided to use 150x150-pixel images for which 85% of the villages could be accommodated in this size (Figure 2). Villages larger than 150x150 pixels were cropped around their centroid to fit into this image size.

4 PREDICTION MODELS

Our goal is to learn machine learning models that can take daytime and nightlights satellite data for a village as an input, and output the socio-economic development level for the village for various indicators. As discussed in the earlier sections, we label villages at levels 1/2/3 for five indicators: BF (bathroom facilities), FC (fuel for cooking), MSW (main source of water), LIT (literacy), and ASSET (asset ownership). We now need to train the models using labeled data for 2011, and produce

Table 2. Grouping of Census Variables to Categorize Villages in Terms of Levels for Different Indicators. Also shown is the Range of Percentages of RUD/INT/ADV Households for Villages Categorized at Different Levels

Indicator	Type	Resources	Level 1	Level 2	Level 3
Bathroom facility (BF)	Rudimentary	No Latrine Facility	86–98.9	14.4–48.2	29.3–57.6
	Intermediate	Pit Latrine	0–1.5	42.7–76.2	0–7.7
	Advanced	Pipe Sewer/Septic Tank	0.4–9.9	0.3–12.9	36.1–63.2
Fuel for cooking (FC)	Rudimentary	Firewood	89.7–100	0.4–19.1	39.8–61.2
	Intermediate	Cow Dung/Kerosene	0–3	72.5–97.1	11.2–44.3
	Advanced	LPG/PNG/Biogas	0–4.1	0–4.9	1.9–37.5
Main source of water (MSW)	Rudimentary	Well/Spring/River	57.9–97.7	0–10.4	0–6.9
	Intermediate	Tube Well/Hand Pump	0–36.1	79.2–99.1	0–25.7
	Advanced	Tap Water/Treated Water	0–2.4	0–3.1	64.2–97.4
Literacy (LIT)		Literate population percent	29.3–40.3	49.2–57.4	64.1–72.6
Asset ownership (ASSET)		TV	1–14.8	11.5–31.9	51–76.3
		Telephone	9.5–31.6	51.7–73.5	67.1–86.1
		Two wheeler (2w)	1.9–8.3	7.6–17.8	21.3–42.9
		Four wheeler (4w)	0–0.7	0.3–2.3	1.1–6.4



(a) Histogram for image height: *Y*-axis shows the number of villages with that height

(b) Histogram for image width: *Y*-axis shows the number of villages with that width

Fig. 2. Distribution of the number of villages with a given height and width.

outputs using the satellite data inputs for 2003 and 2019 (which we consider close enough to the decadal census years of 2001 and 2021). We next outline the architecture for various machine learning models that we built for this purpose and present their evaluation to identify the best model. Outputs from this model are used to estimate the census indicators over two decades and study the dynamics of development across the country.

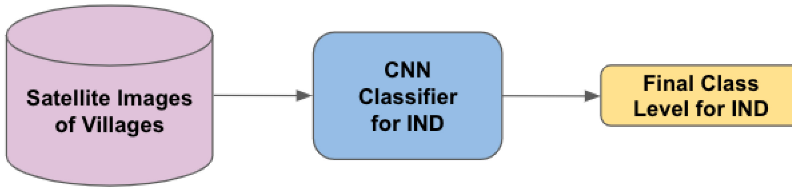


Fig. 3. Direct CNN-based classification where the model takes a village satellite image as an input and predicts the development class (level 1, level 2, or level 3) for the village. Five models were learned, one for each indicator, denoted by IND = {BF, FC, MSW, LIT, ASSET}.

4.1 Architecture-1: Direct CNN-Based Classification

We start with the vanilla ResNeXt-50 architecture and model pre-trained on the ImageNet dataset [24], and change the last layer to output the village levels for different indicators. We learn a different model for each of the five indicators Figure 3. None of the pre-trained layers were frozen. Data was split into train-test sets in an 80:20 ratio. Data augmentation was undertaken using standard methods such as image rotation and reflection. We attempted to address class imbalance by modifying the loss function with a penalizing weight attached to each class based inversely on the number of samples for that class. We found that the performance improved with computing the *effective number of samples* instead of using the actual number of samples, for which feature-space overlap is taken into account for the samples [25]. Finally, we also found that training the models by removing outliers (those villages for which the RUD/INT/ADV values lie beyond 1.5 times the inter-quartile range) from the k-means clustering to assign village labels performed better than models trained with outliers.

To check whether the trained CNN models were able to acquire meaningful features, we conducted occlusion and GradCAM studies [26, 27]. An occlusion study is done by blacking out portions of an image and then examining if this results in a drop in probability of the predicted class. GradCAM utilizes the gradients with respect to the CNN features to determine which parts of the image are important for classification. It is suggested to conduct occlusions during the training stage as well [28], for which we undertook a re-training step for approximately 10% of the dataset. Figure 4 shows the results where darker colours indicate greater relevance of the region for the prediction. Here, asset ownership and literacy predictions seem to focus on built-up regions within villages, and the main source of water prediction seems to be aligned with cultivable areas. This aligns sensibly with our anticipations and shows that the CNN models seem to be learning meaningful features.

4.2 Architecture-2: Two-Stage Model of CNN-Based Features Augmented with Nightlights Features

In this architecture, for each indicator, we take the softmax outputs from its corresponding Architecture-1 trained model as independent features, concatenate them with several more features, and use this expanded feature set to learn separate regression models for the constituent RUD, INT, and ADV parameters for the indicator. These parameters are then discretized by mapping them to the closest level-1/2/3 cluster centroids for that indicator to produce a final output label for the indicator level of the village. Figure 5 shows the overall setup.

The expanded feature set includes the following features:

- Softmax outputs from Architecture-1 for the target indicator and the target village
- Softmax outputs from Architecture-1 for other indicators as well

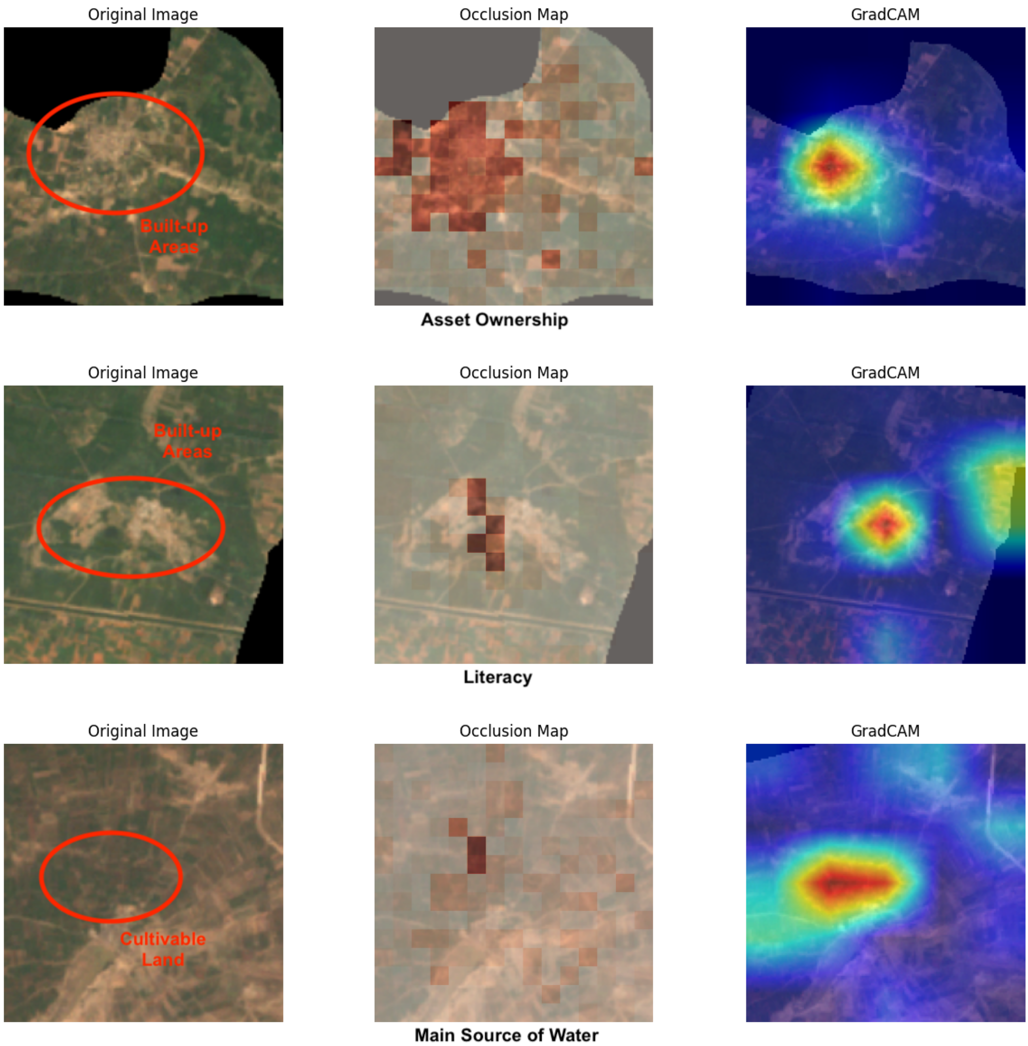


Fig. 4. The images in the first column are high-resolution maps of villages and indicate regions of relevance. The second column shows the occlusion heatmap overlayed on the village image. The third column shows the GradCAM output for the villages. The predictions made are for asset ownership, literacy, and the main source of water.

- Mean softmax outputs from Architecture-1 over the target village and its neighbours, for all indicators
- Nightlights-based features for the target village as described later, and derived features such as the logarithm and square root of these values
- Distance of the target village to economic hubs, where the economic hubs were delineated using a blob identification procedure on nightlights data
- Population features of the target village.

A feature selection is done to identify those features that appear most important for temporal robustness, tested against census data from 2001. The final set of features is then used to learn the

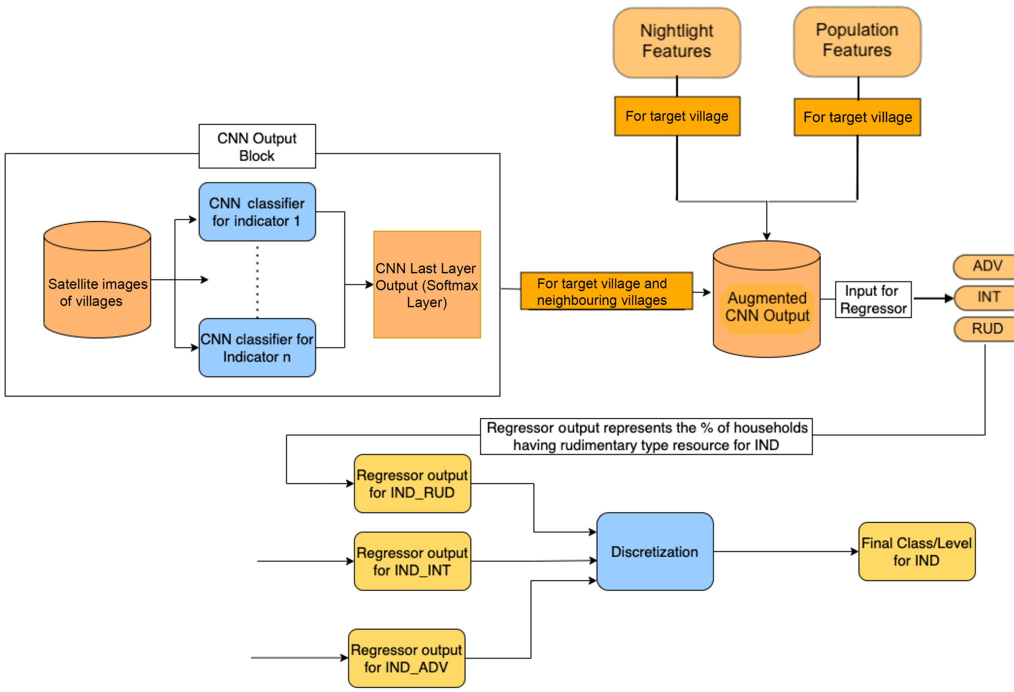


Fig. 5. Architecture-2 structure. This is a two-stage setup, with the first stage being the same as for Architecture-1, of a CNN-based model to predict the development class. The second stage takes the softmax outputs of the first stage as features; concatenates them with the same features obtained for neighboring villages, along with nightlights-based features; and uses this large set of features to learn regression models for each output. The output is then discretized to the development class (level 1, level 2, or level 3) for the village. Separate models are learned for each indicator, denoted by IND = {BF, FC, MSW, LIT, ASSET}.

regression models and go on to produce the final classification for the target village and indicator. We next describe the process in more detail.

4.2.1 Neighbour Identification. To identify the neighbours of a village, we first compute the centroid of the village and its radius as the maximum distance between the centroid and any boundary point, then we list as neighbours all villages whose centroid is at a distance less than twice the village radius. Figure 6 shows a histogram of the count of neighbours of a village. We find that 83.7% of the villages have five or more neighbours, and therefore we chose five of the nearest neighbours to include their features in the expanded feature set to predict variables for the target village.

4.2.2 Nightlights-Based Features. These features are generated by locating hubs of economic activity in the district where the target village is situated (Figure 7). To do this, we developed a blob identification procedure using Otsu thresholding on the nightlights values of pixels within a district [29]. Otsu thresholding automatically identifies a suitable value separating the data into binary classes that minimize the intra-class variance and maximize the inter-class variance for the classes. Patches of lit blobs (which we interpret as hubs of economic activity) are thus obtained for each district, with the thresholding done automatically and independently for each district. The boundary of the resulting patches on the binary raster mask of the district is traced to obtain polygon vectors of the hubs. Figure 8 shows these identified hubs. The mean number of hubs per district is 3.97 (with a standard deviation of 3.82), indicating that major economic hubs are indeed

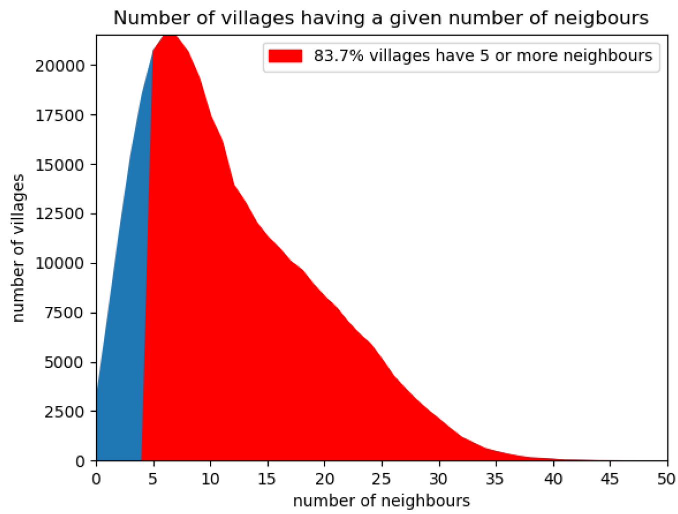


Fig. 6. Histogram of the number of neighbours of villages.

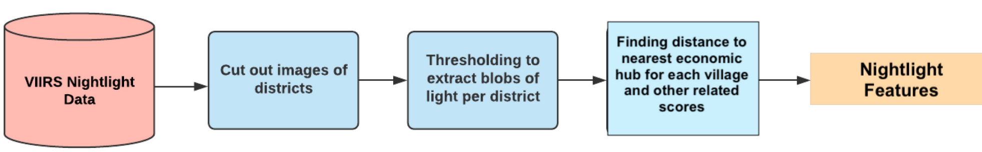


Fig. 7. Pipeline of obtaining the nightlights features.

identified through this procedure. The following values are then obtained: distances between villages and economic hubs (distance between the centroids of a village and hub pair), size of the economic hubs (number of pixels spanned by a hub), and intensity of the economic hubs (mean nightlights value of the hub). Combinations of these values are used to obtain features, such as the product of the inverse of the distance of a village to a hub and intensity of the hub, as shown in Table 3. Two variants of each of the preceding features are created after applying logarithm and square root wrapper functions. For any given feature, the largest value is used in the machine learning model.

4.2.3 Population Features. The total village population and the total number of households are also included as features. The values are already available from the census data for 2001 and 2011. For 2019, we estimated a crude value assuming the same population growth rate as between 2001 and 2011.

4.2.4 Feature Selection. A total of 132 features are computed for each target village, based on the CNN features of the village and its neighbours, nightlights-based features, and population-based features. A feature selection is then done to identify those features that appear most important for temporal robustness. As explained earlier, only a few census variables were available for both 2001 and 2011 to undertake this activity. We identified four variables: number of public primary schools, number of public middle schools, availability of landline service, and availability of public bus service. Other variables either had a large number of missing data (not available) entries or had a very high imbalance of zero values. Even for the variables we selected, we converted them to binary variables for presence/absence since very few villages had more than one school, for

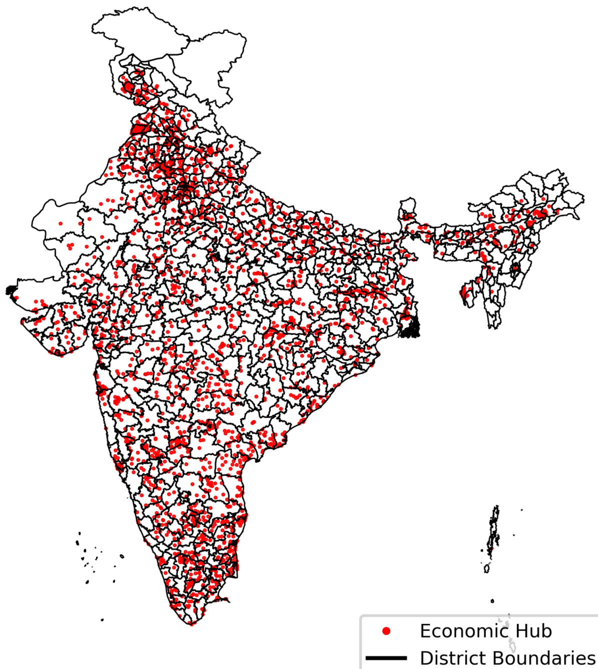


Fig. 8. Hubs of economic activities identified through nightlights.

Table 3. List of Nightlights-Based Features for Hubs of Economic Activity in the District of the Target Village

SI No.	Feature	Description
1	$\frac{1}{D}$	Nearest economic hub
2	$\frac{S}{D}$	Nearest and largest economic hub
3	$\frac{I}{D}$	Nearest and most intense economic hub
4	$\frac{S*I}{D}$	Nearest, largest, and most intense economic hub

Here, D is the distance between the village and the economic hub, S is the size of the hub, and I is the mean intensity of the hub.

example. Random forest binary classifiers were trained for each of these four variables on 2011 data, and the feature importance scores were obtained for all 132 features. The average score for each feature was computed from across the four models, and 40 of the top-ranked features were identified. The performance of the models re-trained on this smaller set of features is shown in Table 4, along with the accuracy of using the model trained on 2011 data to produce scores for 2003. The accuracy and F1 scores seem reasonable, and the models seem to be temporally robust. The most important features that were selected included those related to the hubs of economic activity, the village population, and CNN-based features from the village itself. This tallies with similar findings in related research that CNN-based features on daytime satellite imagery are better for poverty mapping at the village level than nightlights-based features [6, 11]. However, nightlights

Table 4. F1 Scores and Accuracy for Four Indicators Common between the 2001 and 2011 Censuses

Indicator	Year	Arch-2 Accuracy	Arch-2 Wtd Avg F1-score	Arch-3 Accuracy	Arch-3 Wtd Avg F1-score
Landline services	2011	0.80	0.80	0.78	0.78
	2001	0.72	0.72	0.70	0.70
Public bus services	2011	0.78	0.78	0.76	0.76
	2001	0.71	0.71	0.69	0.69
Govt. primary schools	2011	0.80	0.81	0.76	0.77
	2001	0.76	0.76	0.74	0.75
Govt. middle schools	2011	0.79	0.79	0.77	0.77
	2001	0.74	0.74	0.74	0.74

features are quite relevant to infer economic activity in the proximity of the village and hold good predictive potential for poverty mapping.

4.2.5 Regression Models. The features selected in the previous step are used to build linear regression models to estimate the constituent RUD, INT, and ADV parameters for the five socio-economic development indicators of interest. For example, for the MSW indicator, three models are built, namely to estimate the percentage of households in a village having RUD sources of water, INT sources of water, and ADV sources of water. These estimates are then discretized by mapping them to the closest level-1/2/3 cluster centroids for that indicator to produce a final output label for the village for that indicator. Section 4.4 presents the overall results from this setup.

4.3 Architecture-3: CNN-Based Regression Output Instead of a Categorical Output

This architecture is only slightly different from the previous architecture. The CNN at the first stage is used to produce a continuous valued output for the RUD/INT/ADV percentage of households following the methodology of deep regression [23] instead of the categorical output for level 1/2/3 for the village. The rest of the setup remains the same. Instead of using the outputs directly, the last-layer softmax features from the CNN model are combined with features from neighouring villages, nightlights-based features, and population features, followed by the same feature selection process as before, and a final discretization step to obtain the village level for the given indicator. We also experimented with directly using the CNN outputs to obtain the village development level but found that the additional features were useful and provided an improved performance.

4.4 Evaluation

We next present results in Table 5 comparing the three architectures with one another. In general, apart from ASSET, Architecture-2 and Architecture-3 perform better than Architecture-1, indicating the relevance of additional features beyond the CNN-based features alone. Overall, Architecture-2 performs better than Architecture-3. We also compared the regression R2 scores before the final discretization step and found that Architecture-2 gave better scores than Architecture-3.

However, we can also see that the accuracy for some classes can be quite low. This motivated us to develop an aggregate index on the lines of the Human Development Index by aggregating multiple indicators with the hope that errors for the indicators might not compound upon aggregation. We term this the *ADI* and calculate it as a simple sum of the levels of a village for the five different

Table 5. Class-Wise F1 Scores (First 3 Rows in the Table), Weighted Avg F1 Score (Second Last Row in the Table), and Accuracy (Last Row in the Table)

	ASSET			BF			MSW			FC			LIT		
	A1	A2	A3	A1	**A2**	A3	A1	**A2**	A3	A1	**A2**	A3	A1	**A2**	A3
Level 1	**0.79**	0.73	0.71	0.81	**0.89**	0.89	0.60	**0.63**	0.56	0.84	**0.85**	0.84	0.45	**0.51**	0.46
Level 2	**0.77**	0.72	0.72	0.30	**0.37**	0.36	0.77	**0.81**	0.81	0.61	**0.65**	0.59	0.58	**0.70**	0.69
Level 3	**0.78**	0.73	0.73	0.41	**0.49**	0.48	0.66	**0.69**	0.65	0.42	**0.41**	0.41	0.62	**0.69**	0.66
Weighted Avg F1 Score	**0.78**	0.73	0.72	0.70	**0.77**	0.77	0.71	**0.75**	0.73	0.71	**0.72**	0.71	0.57	**0.66**	0.65
Accuracy	**0.78**	0.73	0.72	0.69	**0.80**	0.80	0.71	**0.74**	0.73	0.72	**0.73**	0.72	0.57	**0.67**	0.64

The best performance for each category is shown in bold font.

Table 6. RMSE and Normalized RMSE for the Three Architectures

	A1	**A2**	**A3**
RMSE	1.551	1.527	1.585
Normalized RMSE	0.172	0.170	0.176

indicators. The ADI thus ranges from 5 (for villages at level 1 for all indicators) to 15 (for villages at level 3 for all indicators). Table 6 shows the RMSE and NRMSE scores for the ADI computation, and these seem reasonable within two units of the ADI, on an ADI range between 5 and 15. We delve further to analyze for which kind of villages the ADI is more error prone. Figure 9 shows the frequency of villages that encountered an error greater than two units of the ADI, for villages at different levels of the ADI. We can see that the error tends to be higher for villages already at high levels of the ADI. Villages at low levels of the ADI have proportionately lower error, which may be expected because of their lower base. This indicates that the NRMSEs are likely to be uniformly distributed across villages at different levels of development. This is further validated when we analyze the errors across different states. Table 7 shows the NRMSE and MAE for various states, and we can see that the errors are uniformly spread across the states. This broadly convinces us that the ADI may be a satisfactory measure as an indicator of socio-economic development, and we next proceed to analyze the development dynamics of Indian villages over two decades by using ADI outputs for 2003 and 2019 produced by the Architecture-2 models, along with the actual ADI for 2011 computed from the census data.

5 ANALYSIS

We begin with a brief characterization study. Figure 10 shows the histogram for the ADI of villages in 2003, 2011, and 2019. We can see that the development has consistently improved over the years. During 2003–2011, 42.14% villages showed a positive change in the ADI and 29.30% showed no change, whereas during 2011–2019, these statistics were 36.17% and 30.06%, respectively. State-wise development statistics are shown in Table 8 to reveal the state of socio-economic development in 2003 and 2011, and the change during the 2003–2011 and 2011–2019 periods. The wide

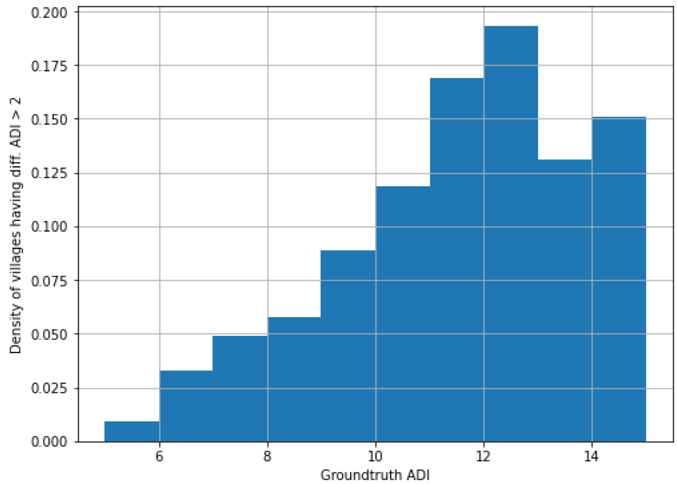


Fig. 9. Histogram of the frequency of error for villages at different levels of development.

Table 7. State-Wise Normalized RMSE for Architecture-2

State	Normalized RMSE	MAE	State	Normalized RMSE	MAE
Andhra Pradesh	0.183	1.173(1.257)	Maharashtra	0.187	1.318(1.279)
Bihar	0.155	0.934(0.984)	Madhya Pradesh	0.183	0.967(1.055)
Chattisgarh	0.157	0.771(0.915)	Orissa	0.168	0.830(1.005)
Gujrat	0.194	1.405(1.408)	Punjab	0.114	1.027(1.073)
Haryana	0.145	1.217(1.311)	Rajasthan	0.171	1.001(1.088)
Jharkhand	0.189	0.887(1.003)	Tamil Nadu	0.128	0.998(1.111)
Karnataka	0.153	1.033(1.045)	Uttar Pradesh	0.166	1.124(1.107)

differences across states is evident from these results, with states like Tamil Nadu and Punjab showing strong improvement even after starting from a comparatively high base of socio-economic development, and some of the poorest states states like Jharkhand, Madhya Pradesh, and Odisha showing improvement but still unable to catch up with other states. Two states of Bihar and Chattisgarh also show a negative growth.

Upon further examination of villages showing negative ADI change, we found this to be happening mostly in more developed villages. We attribute this either to a higher chance of error in this region, as explained in the previous section, or perhaps an actual degradation in the development of these villages. We did attempt to analyze covariates such as the change in nightlights for these villages which showed a negative ADI change but were unable to conclude anything since nightlights increased on average by approximately 53% during 2003–2011 and 78% during 2011–2019 for not only villages with a negative change in the ADI but also those which saw a positive or no change in the ADI. We therefore proceed with these ADI outputs for subsequent analysis.

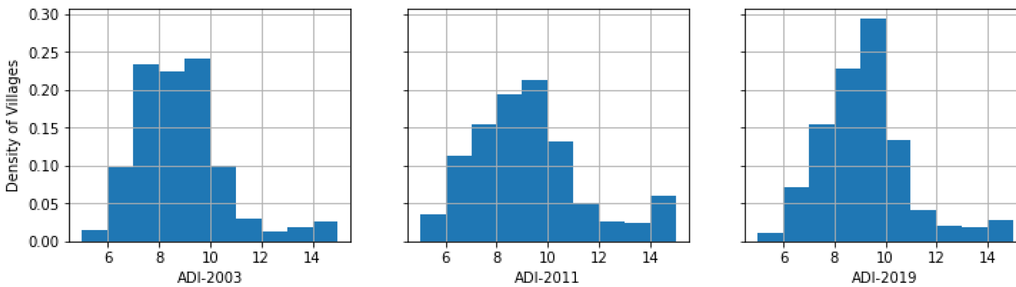


Fig. 10. Distribution of ADI for all villages in India for the years 2003, 2011, and 2019 (from left to right).

Table 8. State-Wise Statistics of Village-Level Socio-Economic Development

State	**Villages in the low and medium ADI range in 2003 (%)**	**Low and medium range villages that had a positive development in the 2003–2011 period (%)**	**Villages in the low and medium ADI range in 2011 (%)**	**Low and medium range villages that had a positive development in the 2011–2019 period (%)**
Andhra Pradesh	75.43	78.32	62.64	53.17
Bihar	78.22	79.18	88.26	54.96
Chhattisgarh	85.99	64.78	98.13	64.95
Gujarat	89.98	87.05	46.48	57.75
Haryana	14.21	92.8	9.99	38.52
Jharkhand	99.62	75.57	98.59	53.76
Karnataka	71.89	95.43	35.56	82.06
Maharashtra	85.79	87.67	47.86	66.76
Madhya Pradesh	95.06	61.23	92.55	44.32
Odisha	93.64	77.07	79.82	54.77
Punjab	17.27	99.93	0.03	33.33
Rajasthan	82.88	74.1	80.2	49.93
Tamil Nadu	44.4	99.51	1.95	64.04
Uttar Pradesh	75.52	84	69.16	50.6

We further show visualizations of two sample districts and how they have changed over the years. The colour coding is done by categorizing low ADI villages as those with an ADI value below 9, medium ADI villages with a value of 9, and high ADI villages for those with an ADI value greater than 9. This is also justified from a CDF plot of the ADI as of 2011 across all the villages in our dataset, where each of the low/medium/high ADI categories as determined previously end up with a roughly equal number of villages in each category. The two districts of Bharatpur in Rajasthan and Jamui in Bihar show a steady improvement, and we also notice that villages closer to the district center tend to develop faster (Figures 11 and 12).

We also study the distribution of the distances of villages to their nearest hubs of economic activity. Figure 13 shows the CDF for 2003, 2011, and 2019. The distances decreased somewhat from 2003 to 2011 but rapidly after that. The median distance in 2003 was 19.4 km, which reduced to 18.4 km in 2011 and 11.2 km in 2019. This can likely be attributed to a spread of economic

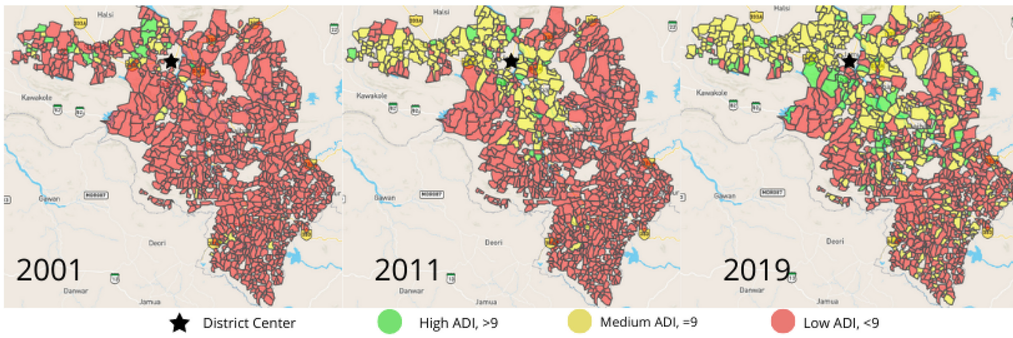


Fig. 11. ADI categorization for villages in the Jamui District of Bihar.

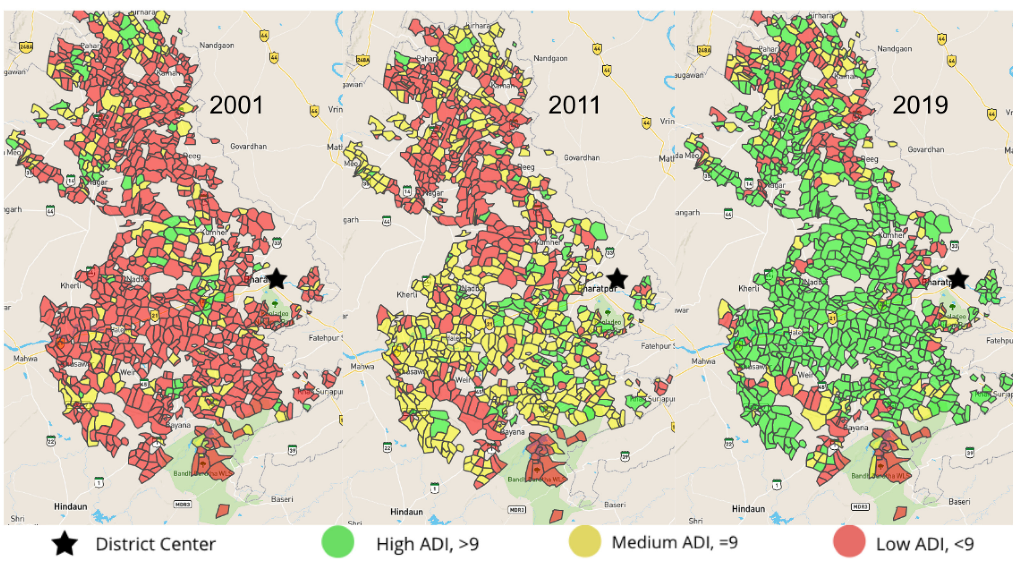


Fig. 12. ADI categorization for villages in the Bharatpur District of Rajasthan.

activity in more areas, bringing it closer to the villages, and also due to an increase in electrification. According to World Bank statistics, electrification rates in India went up from 64% in 2003 to 67.6% in 2011 and 97% in 2019 [36], although these statistics are disputed [38].

This motivates us to conduct two sets of analyses. First, we study how the ADI as of 2003, 2011, and 2019 is related to the distance of villages to hubs of economic activity. Similarly, we study how the decadal change in the ADI across these two periods is also related to the distance of villages to economic hubs, along with other variables such as the inequality in the district to which the village belongs, and the relative difference between the initial ADI of the village and the average ADI of other villages in the district. Second, we study whether the inequality within districts follows the Kuznets hypothesis—that is, when ADI is low, then inequality increases with an improvement in ADI, but when ADI has improved and continues to improve further, then inequality begins to fall.

5.1 Variables Linked with the ADI and Change in the ADI

5.1.1 Villages That Are More Developed Tend to Be Closer to an Economic Hub. We build OLS models separately for the years 2003, 2011, and 2019, with the ADI as the dependent variable and

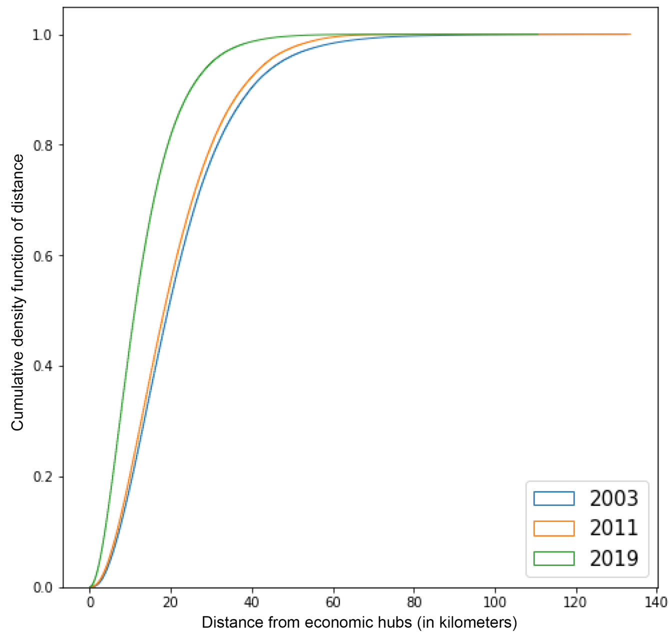


Fig. 13. Distribution of distance from economic hubs of villages for the years 2003, 2011, and 2019. The Haversine formula was used to approximate the ground distance between two geospatial points.

distance to the closest economic hub as the independent variable. We build models both without and with fixed effects for states. The fixed effects models are made with reference to the state of Uttar Pradesh by introducing dummy variables for the other 13 states, to avoid problems with multi-collinearity.

We can see from Table 9 that both with and without fixed effects, there is a statistically significant negative relationship between village ADI and the distance of the village from its closest economic hub. Villages farther away tend to be less developed, and this holds for all 3 years of 2003, 2011, and 2019.

5.1.2 Explaining the Rate of Change in the Development of Villages. We next create dependent variables for change in the ADI between 2003 and 2011, and between 2011 and 2019, and try to explain this change based on three variables. The first independent variable we use, as before, is the distance of the village to its nearest economic hub in the base year. Two additional variables are constructed as well.

Mean District ADI. We calculate a population weighted ADI for each district based on the ADIs of its constituent villages. We then introduce an independent variable for the OLS models as the difference between the village ADI and the mean district ADI. This is meant to examine whether villages that are lagging further behind in development than their cohort (other villages in the same district) are catching up with them.

Gini Coefficient for the District. The Gini coefficient is a measure of inequality and has been used extensively to quantify the deviation in the distribution of income in a population from a perfectly equal distribution of income among everybody. It has a value between 0 and 1, with 0 denoting perfect equality. Using the same idea of a population-level Gini coefficient, we build a district-level Gini coefficient based on distribution of the ADI among the villages of the district. We weigh each village by its population. For the OLS models, we introduce this district Gini coefficient as an

Table 9. OLS Models for ADI in 2003, 2011, and 2019 with and without Fixed Effects for States

Without Fixed Effects			
ADI in	**2003**	**2011**	**2019**
Distance from economic hub	−0.0297*** (0.000)	−0.0485*** (0.000)	−0.0509*** (0.000)
Intercept	9.0247*** (0.006)	10.0656*** (0.008)	9.3554*** (0.006)
R-Squared	0.049	0.067	0.067
No. of Villages	295,800	295,800	295,800
With Fixed Effects			
ADI in	**2003**	**2011**	**2019**
Distance from economic hub	−0.0174*** (0.000)	−0.0314*** (0.000)	−0.0295*** (0.000)
Andra Pradesh	0.1397 (0.016)	0.0151 (0.020)	−0.4221*** (0.015)
Bihar	−0.2221*** (0.011)	−0.6310*** (0.014)	−0.0549*** (0.010)
Chhattisgarh	−0.9674*** (0.014)	−1.7995*** (0.018)	−1.6922*** (0.013)
Gujarat	−0.9471*** (0.014)	0.8456*** (0.018)	−0.0851*** (0.013)
Haryana	3.7198*** (0.034)	3.2064*** (0.044)	1.9575*** (0.033)
Jharkhand	−2.3049*** (0.014)	−2.3133*** (0.018)	−1.6920*** (0.014)
Karnataka	0.1965*** (0.017)	0.6535*** (0.022)	0.2379*** (0.017)
Maharastra	−0.7089*** (0.010)	0.6690*** (0.013)	−0.7062*** (0.010)
Madhya Pradesh	−1.2058*** (0.009)	−1.3919*** (0.012)	−1.0681*** (0.009)
Orissa	−1.7255*** (0.011)	−1.5297*** (0.014)	−1.3991*** (0.011)
Punjab	3.4295*** (0.017)	3.3976*** (0.022)	3.5813*** (0.017)
Rajasthan	−0.5308*** (0.011)	−0.7404*** (0.014)	−0.4264*** (0.010)
Tamilnadu	1.1397*** (0.023)	1.9554*** (0.030)	0.7412*** (0.022)
Intercept (For Uttar Pradesh)	9.2580*** (0.007)	10.0687*** (0.009)	9.5051*** (0.006)
R-Squared	0.351	0.315	0.339

independent variable to find out how the change in development of a village varies with differences in the inequality of the district in which it is located.

We first separately examine the relationship between each of the independent variables and the dependent variable of change in the ADI. We find that the distance to the closest economic hub is statistically significantly and negatively related to change in the ADI—that is, villages that are farther away from economic hubs are not only less developed (as shown in the previous section) but also are slower to develop than closer villages. With respect to the second variable of the difference between the village ADI and the mean district ADI, we find that villages that are lagging further behind do have a faster rate of change in development. This is encouraging and indicates a process of catch-up for villages that lag behind others in their cohort. For the third variable of the district Gini coefficient, we find a statistically significant positive relationship—that is, villages in more unequal districts tend to develop faster. This too seems encouraging and may imply that the gap will narrow down over time.

We next build a comprehensive OLS model with all three variables together and also build an equivalent model with fixed effects for states. The results are shown in Table 10. Most of the relationships we had identified through separate models for each variable still hold and persist when state fixed effects are introduced. The relationship with the Gini coefficient, however, changes in an interesting manner. During 2003–2011, villages in more unequal districts seemed to have developed faster, but this reversed during 2011–2019—that is, villages in unequal districts did not develop as fast as villages in more equal districts. This interesting observationmight have to do with policy changes during the two time periods. Indian politics underwent a significant change in 2014 when the incumbent UPA government was replaced by the NDA government in the Center, and led to a difference in how welfare and industrial policies were designed. For example, welfare funds expenditure was increased for direct cash transfer schemes at the cost of other schemes that were built around in-kind transfers and asset creation in rural areas [17]. A new Goods and Services Tax regime was also brought in to increase formalization of the Indian economy [45]. Finally, the demonetization event of November 2016, which has remained under judicial scrutiny until recently, squeezed liquidity out of the informal economy and is claimed by many economists to have negatively impacted socio-economic development [44]. It is indeed possible that such changes in the governance approach could have led to the observations we make through our models, wherein unequal districts in which villages were developing faster may have slowed down because industrialization-oriented economic activity that tends to be concentrated in unequal districts may in general have slowed down. Less developed villages still seem to be developing faster to catch up, but the advantage that unequal districts earlier had during the 2003–2011 period for their villages to develop faster seems to have dissipated during the 2011–2019 period.

5.2 Testing the Kuznets Hypothesis

Kuznets [30] hypothesized that as countries develop economically, their inequality first rises and then falls. This was assumed to happen when an economy moves from agricultural to non-agricultural employment: an influx of cheap rural labour first leads to diminished wages and rising inequality, but as the economy industrializes further to absorb surplus labour, and aided also by social welfare mechanisms, the inequality eventually decreases. Kuznets, however, did not evaluate this hypothesis on longitudinal data. Rather, his samples were based on cross-sectional estimates of different countries which seemed to lie on an inverted U-shaped curve. Subsequent studies that examined longitudinal data found mixed results [18, 19], indicating that the interplay between inequality and economic development depends on a variety of factors, especially policy-related aspects such as welfare spending, taxation, and broad-based or concentrated industrialization. We examine whether the Kuznets hypothesis holds at the district level in India, in two parts: during

Table 10. OLS Models for Change in ADI between 2003–2011 and 2011–2019 with and without Fixed Effects for States

Without Fixed Effects		
Change in ADI between	**2003–2011**	**2011–2019**
Initial Gini	4.9413^{***} (0.136)	-1.7343^{***} (0.116)
Distance from development hub	-0.0177^{***} (0.000)	-0.0086^{***} (0.000)
(Village_ADI-Avg_district_ADI)	-0.7531^{***} (0.003)	-0.7821^{***} (0.002)
Intercept	0.5192^{***} (0.012)	-0.2758^{***} (0.012)
R–Squared	0.225	0.383
No. of Villages	295,800	295,800
With Fixed Effects		
Change in ADI in	**2003–2011**	**2011–2019**
Initial Gini	3.3718^{***} (0.1158)	-5.9750^{***} (0.137)
Distance from development hub	-0.0170^{***} (0.000)	-0.0128^{***} (0.000)
(Village_ADI-Avg_district_ADI)	-0.7519^{***} (0.003)	-0.8127^{***} (0.002)
Andra Pradesh	-0.4405^{***} (0.019)	-1.1643^{***} (0.016)
Bihar	-0.6186^{***} (0.013)	0.6206^{***} (0.011)
Chhattisgarh	-0.7663^{***} (0.017)	0.0050 (0.014)
Gujarat	1.5210^{***} (0.017)	-1.3277^{***} (0.015)
Haryana	-0.6542^{***} (0.041)	-1.3399^{***} (0.034)
Jharkhand	-0.0451^{**} (0.017)	0.3165^{***} (0.015)
Karnataka	0.4255^{***} (0.021)	-0.6761^{***} (0.017)
Maharastra	1.1521^{***} (0.013)	-1.8222^{***} (0.011)
Madhya Pradesh	-0.3632^{***} (0.011)	0.2464^{***} (0.010)
Orissa	0.1878^{***} (0.013)	0.0502^{***} (0.011)
Punjab	-0.2248^{***} (0.021)	0.2684^{***} (0.018)
Rajasthan	-0.2578^{***} (0.013)	0.0196 (0.011)
Tamilnadu	0.5399^{***} (0.028)	-1.4687^{***} (0.023)
Intercept(For UttarPradesh)	1.0967^{***} (0.014)	-0.7088^{***} (0.013)
R - Squared	0.297	0.500

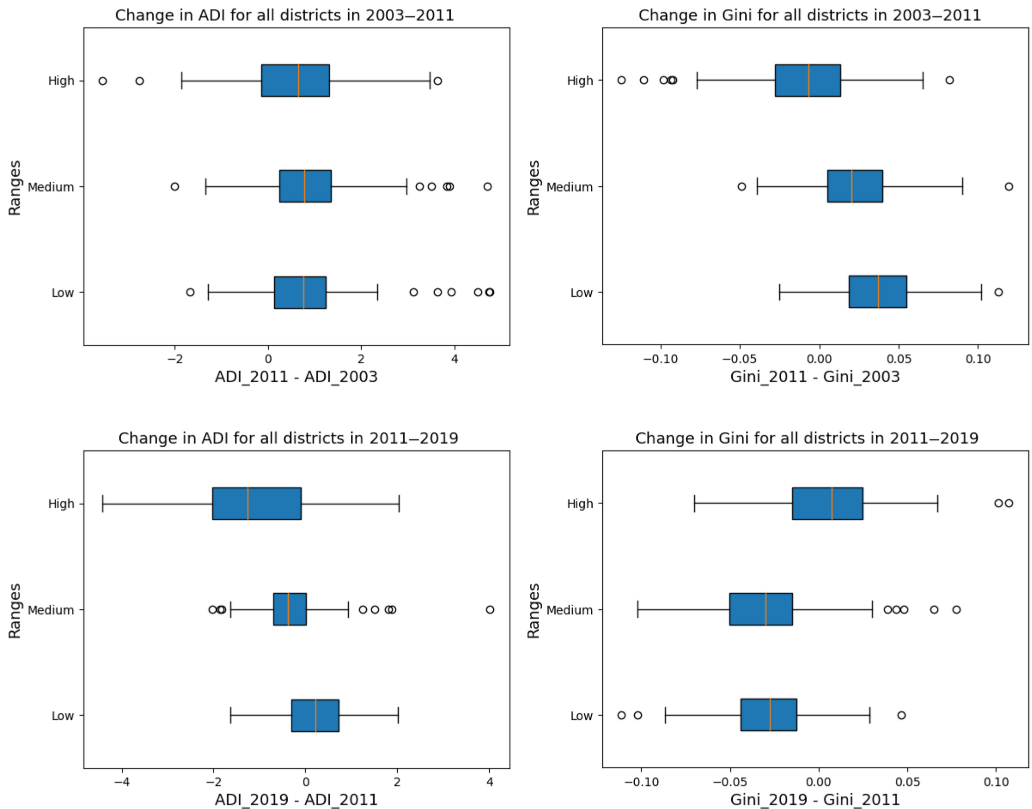


Fig. 14. Box plots of change in the ADI and change in the Gini for all districts across the periods of 2003–2011 and 2011–2019.

the expected rising curve of the hypothesis when inequality increases with economic development, and during the falling part of the curve when inequality reduces with economic development. Although there is little reason to consider districts as equivalent to countries, most significantly due to easy movement of people across districts as compared to across countries, we are primarily motivated to study the interplay between inequality and development and hence conduct this analysis.

We begin with an analysis shown in Figure 14 of change in the ADI and the Gini coefficient at the district level, between the 2003–2011 and 2011–2019 periods. Although changes in the ADI during the 2003–2011 period do not seem to have depended on the ADI in 2003, clear differences do seem to exist during the 2011–2019 period when less developed districts seem to have developed faster than more developed districts. This appears to be a positive sign indicating that less developed districts may be catching up faster. However, observing changes in the Gini coefficient shows that inequality reduced in more developed districts during 2003–2011, but this trend reversed during 2011–2019 when inequality increased in more developed districts. The Kuznets hypothesis, which seems to have held true during 2003–2011, reversed during the 2011–2019 period. We evaluate this more closely next.

5.2.1 Starting Phase of the Kuznets Curve. We use the mean district ADI as a proxy for economic development and calculate the slope between the Gini coefficient (Y axis) and the ADI (X axis)

Table 11. Mean Value of Slope and z-Scores for Districts Having Positive ADI Change and Positive Gini Change

	Mean slope for low ADI districts	Mean slope for medium ADI districts	Mean slope for high ADI districts	Low and medium slope diff: z-test	Low and high slope diff: z-test
2003–2011	0.474	0.074	0.005	1.294	1.522
2011–2019	0.040	0.085	0.351	−0.902	−1.577

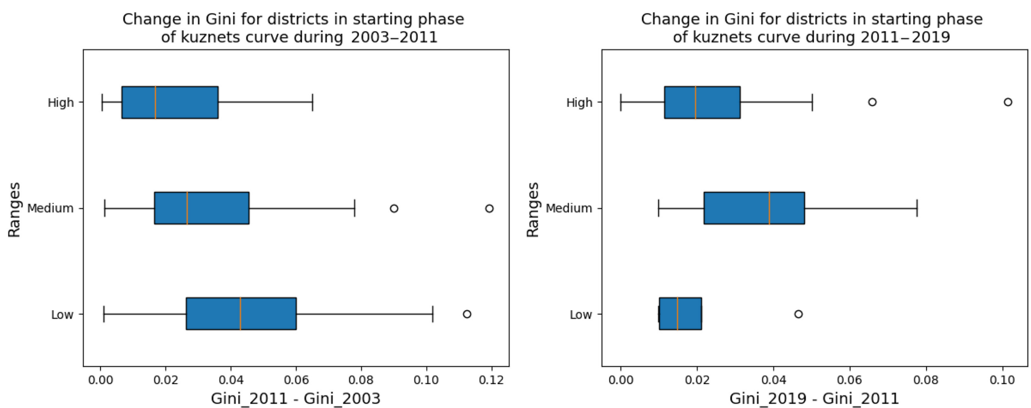


Fig. 15. Box plots of change in the Gini for districts in the starting phase of the Kuznets curve during 2003–2011 and 2011–2019.

during the 2003–2011 period and during the 2011–2019 period. We then hypothesize that for the Kuznets hypothesis to be valid, the slope should be higher in magnitude (and positive) for low ADI districts than for medium and high ADI districts, for districts which see a positive Gini change (increase in inequality) and positive ADI change (increase in economic development). We discretize the mean district ADI values into low/medium/high levels by using the 33 and 67 percentiles as thresholds, similar to how we had earlier discretized village ADIs for their visualizations. The null and alternative hypotheses can then be framed as follows. We test them through a two-sample z-test on the slope values for the low/medium/high district ADI levels. We take care to exclude districts which showed a negative or no change in the ADI, possibly due to errors in ADI estimation as explained earlier, and conduct the hypothesis test on 56.7% of districts during 2003–2011 and 7.88% of districts during 2011–2019 which showed a positive change in both the ADI and Gini.

Null Hypothesis

- Mean((Gini Change/ADI Change) for Low ADI districts) = Mean((Gini Change/ADI Change) for High ADI districts)
- Mean((Gini Change/ADI Change) for Low ADI districts) = Mean((Gini Change/ADI Change) for Medium ADI districts)

Alternate Hypothesis

- Mean((Gini Change/ADI Change) for Low ADI districts) > Mean((Gini Change/ADI Change) for High ADI districts)
- Mean((Gini Change/ADI Change) for Low ADI districts) > Mean((Gini Change/ADI Change) for Medium ADI districts)

Table 12. Mean Value of Slope and z-Scores for Districts Having Positive ADI Change and Negative Gini Change

	Mean slope for low ADI districts	**Mean slope for medium ADI districts**	**Mean slope for high ADI districts**	**High and low slope diff: z-test**	**High and medium slope diff: z-test**
2003–2011	−0.018	−0.028	−0.066	−2.756	−1.899
2011–2019	−0.183	−1.414	−0.061	1.857	1.606

In Table 11, the absolute value of the z-scores being less than 1.645 (for a 5% significance) for both periods indicates that the Kuznets hypothesis cannot be accepted with confidence, although the relative mean values of the slope during the 2003–2011 period agree with the Kuznets hypothesis: more developed districts have a lower slope. However, the negative sign during the 2011–2019 period is curious and indicates that inequality during this time increased at a faster rate with the ADI in medium and high ADI districts than in low ADI districts. This, when coupled with the analysis in the previous section, seems to suggest that policy changes may have slowed down the development of less developed villages in more developed districts (likely to be those which are more industrialized) and thus not reduced inequality in these districts.

Figure 15 shows box plots for the change in the Gini coefficient for districts in their starting phase of the Kuznets curve. The same trends are visible: inequality increased more in less developed districts than in more developed districts during 2003–2011, as hypothesized by Kuznets, but this changed during 2011–2019.

5.2.2 Ending Phase of the Kuznets Curve. We similarly construct hypotheses for the ending phase of the Kuznets curve, where slope of the Gini (Y axis) and ADI (X axis) is expected to be higher in magnitude (and negative) for high ADI districts than for medium or low ADI districts, for districts which see a negative Gini change (decrease in inequality) and positive ADI change (increase in economic development). As before, we exclude districts that showed a negative or no change in the ADI, and conduct the test on 19.7% of the districts during 2003–2011 and 28.07% during 2011–2019 for which the ADI change is positive and the Gini change is negative.

Null Hypothesis

- Mean((Gini Change/ADI Change) for High ADI districts) = Mean((Gini Change/ADI Change) for Low ADI districts)
- Mean((Gini Change/ADI Change) for High ADI districts) = Mean((Gini Change/ADI Change) for Medium ADI districts)

Alternate Hypothesis

- Mean((Gini Change/ADI Change) for High ADI districts) < Mean((Gini Change/ADI Change) for Low ADI districts)
- Mean((Gini Change/ADI Change) for High ADI districts) < Mean((Gini Change/ADI Change) for Medium ADI districts)

From Table 12, during the 2003–2011 period, there seems to be sufficient evidence to not reject the Kuznets hypothesis: high ADI districts seem to have a greater reduction in inequality than low and medium ADI districts. However, this seems to reverse (reasonably strong statistical significance) during the 2011–2019 period, again hinting towards a possible effect of policy changes resulting in slower development of villages in high ADI districts (more industrialized) than in low and medium ADI districts.

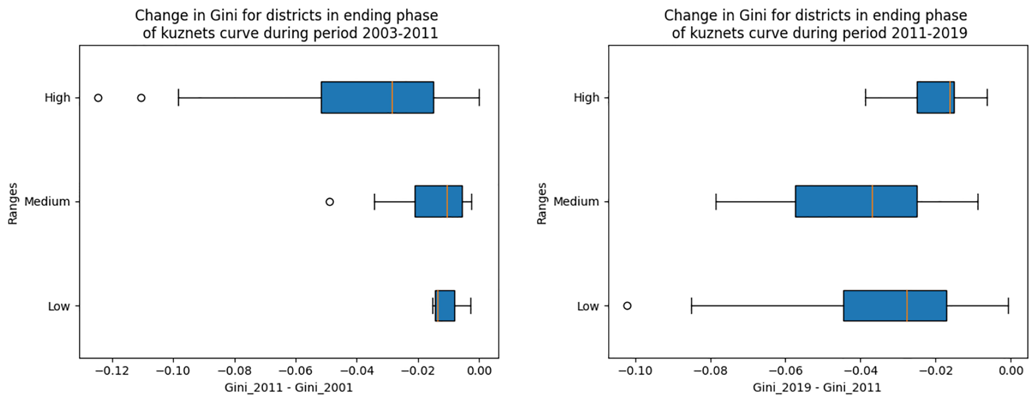


Fig. 16. Box plots of change in Gini for districts in the ending phase of the Kuznets curve during 2003–2011 and 2011–2019.

Figure 16 shows box plots for the change in the Gini coefficient for districts in their ending phase of the Kuznets curve. Inequality reduced more in more developed districts than in less developed districts during 2003–2011, as hypothesized by Kuznets, but this changed during 2011–2019.

6 DISCUSSION AND CONCLUSION

We demonstrated the feasibility of using satellite data to estimate socio-economic development at the village scale in a temporally robust manner. Our method combines data from both daytime satellite imagery as well as nightlights. We then showed that this estimate can be used to study development patterns and validated several such observations made in other studies. We showed that villages farther away from hubs of economic activities tend to less developed, and develop slower, than villages closer to economic activities. We also showed that a catch-up pattern exists wherein villages that are less developed tend to develop faster to catch up with other villages in the district. At the same time, a positive relationship is seen during the 2003–2011 period between village development and the initial inequality in a district: villages in more unequal districts tend to develop faster. This relationship, however, reversed during the 2011–2019 period. We hypothesize that this is likely due to policy changes in India during this time wherein economic activities such as industrialization that would have been more prevalent in unequal districts may have slowed down. This is consistent with our interpretation of testing the Kuznets hypothesis during these two periods. According to Kuznets, inequality is likely to first increase with industrialization-led economic growth and then decrease as wages and social welfare improve. Something similar seems to have happened in India during the 2003–2011 period, where industrialization-dominated economic activity followed by more equitable development would have led to districts following the Kuznets hypothesis. However, policy changes that disrupted the informal economy during 2011–2019 seem to have led to a pattern where unequal districts saw a slowdown in economic activity and thereby slower development of their villages, and hence also a reversal of the Kuznets curve. With 80% of the non-agricultural employment in India mostly in the informal sector [47], the recent policy push towards formalization using sudden treatments like demonetization and GST, coupled with a change in the social welfare strategy towards cash transfers from asset creation in rural areas, seems to have concentrated growth in already developed areas and thereby increased inequality. Gradual formalization coupled with improvements in rural asset creation would perhaps be a more appropriate policy to balance growth and equality [48].

Through this experience, we believe that the ADI outputs we have produced, and our methodology that can be used to produce similar outputs for other years, can be used to study development patterns at the village level and higher aggregate geographical scales. Changes in indicators related to health, vaccination coverage, education, welfare expenditure, and so on can be studied with the ADI as a covariate or an outcome variable to understand co-movement and causal patterns. In future work, we plan to study the impact of welfare expenditure on socio-economic development to understand how much and what kind of welfare expenditure is more impactful for socio-economic development.

APPENDIX

A DISCRETIZATION OF CENSUS SOCIO-ECONOMIC INDICATORS

A box plot for clustering census indicators FC, BF, LIT, MSW, and ASSET is shown in Figures 17, 18, 19, 20, and 21, respectively.

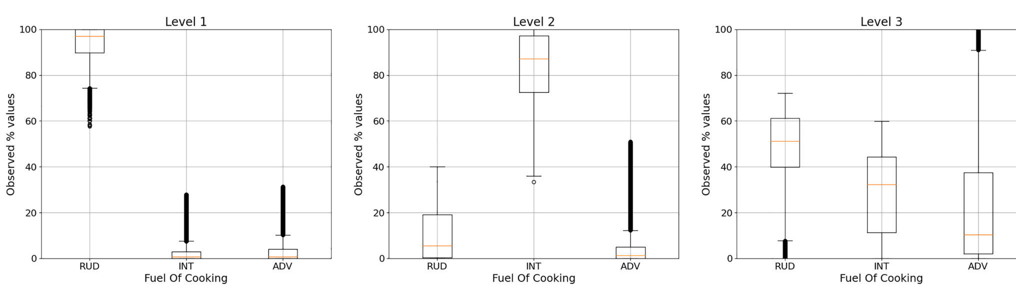


Fig. 17. Clustering of the FC indicator into level-1/2/3 villages.

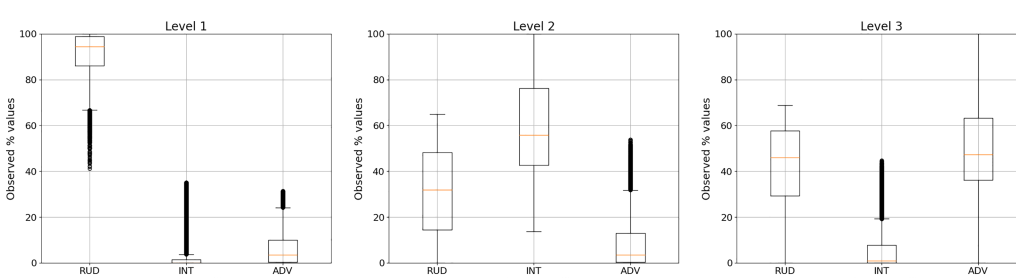


Fig. 18. Clustering of the BF indicator into level-1/2/3 villages.

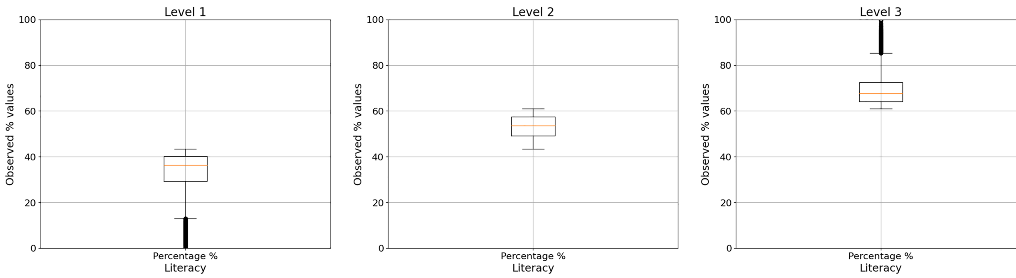


Fig. 19. Clustering of the LIT indicator into level-1/2/3 villages.

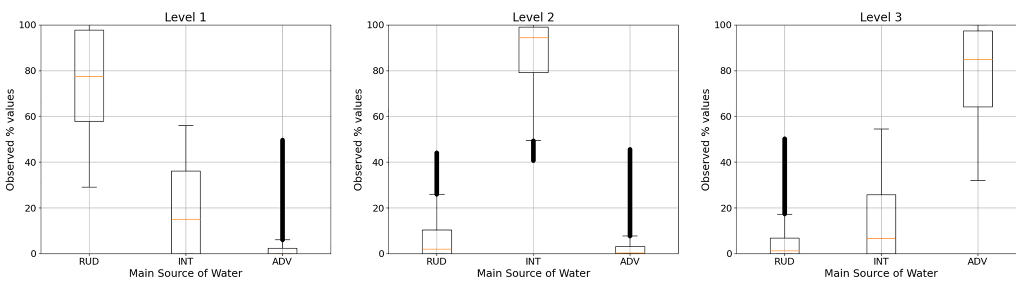


Fig. 20. Clustering of the MSW indicator into level-1/2/3 villages.

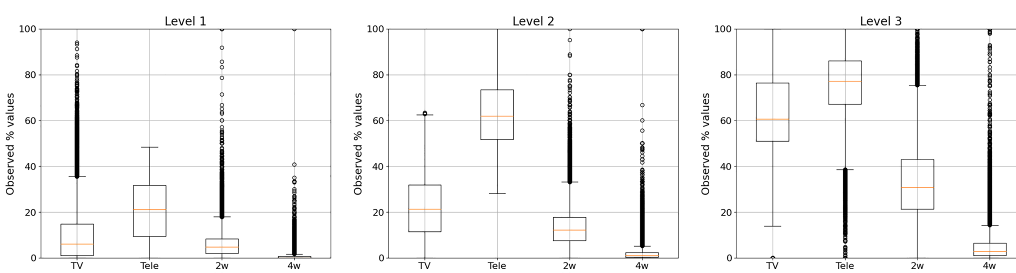


Fig. 21. Clustering of the ASSET indicator into level-1/2/3 villages.

ACKNOWLEDGMENTS

We are grateful to the High Performance Computing (HPC) infrastructure team of IIT Delhi for their support. We also want to express our gratitude to Chahat Bansal and Badrinath Padmanabhan for support with a few technical aspects, and anonymous reviewers for their feedback in strengthening the article.

REFERENCES

[1] D. J. Peters. 2012. Income inequality across micro and meso geographic scales in the midwestern United States, 1979–20091. *Rural Sociology* 77, 2 (2012), 171–202. https://doi.org/10.1111/j.1549-0831.2012.00077.x

[2] A. Kalaiyarasan and M. Vijayabaskar. 2021. *The Dravidian Model: Interpreting the Political Economy of Tamil Nadu.* Cambridge University Press. https://doi.org/10.1017/9781108933506

[3] A. Kohli. 2012. State and redistributive development in India. In *Growth, Inequality and Social Development in India: Is Inclusive Growth Possible?*, R. Nagaraj (Ed.). Palgrave Macmillan, London, UK. https://doi.org/10.1057/9781137000767_7

[4] C. Hall. 2022. Nighttime lights. *Earthdata*. Retrieved September 5, 2023 from https://www.earthdata.nasa.gov/learn/backgrounders/nighttime-lights
[5] A. Prakash, A. K. Shukla, C. Bhowmick, R. Carl, and M. Beyer. 2019. Night-time luminosity: Does it brighten understanding of economic activity in India? *Reserve Bank of India Occasional Papers* 40, 1. https://www.researchgate.net/publication/334811462_Nighttime_Luminosity_Does_it_Brighten_Understanding_of_Economic_Activity_in_India
[6] N. Jean, M. Burke, M. E. Xie, W. M. Davis, D. B. Lobell, and S. Ermon. 2016. Combining satellite imagery and machine learning to predict poverty. *Science* 353, 6301 (2016), 790–794. https://doi.org/10.1126/science.aaf7894
[7] Y. S. Lee. 2018. International isolation and regional inequality: Evidence from sanctions on North Korea. *Journal of Urban Economics* 103 (2018), 34–51. https://doi.org/10.1016/j.jue.2017.11.002
[8] F. Bickenbach, E. Bode, P. Nunnenkamp, and Mareike Soder. 2016. Night lights and regional GDP. *Review of World Economics* 152 (2018), 425–447. https://doi.org/10.1007/s10290-016-0246-0
[9] C. Bansal, A. Jain, P. Barwaria, A. Choudhary, A. Singh, A. Gupta, and A. Seth. 2020. Temporal prediction of socio-economic indicators using satellite imagery. In *Proceedings of the 7th ACM IKDD CoDS and 25th COMAD*. 73–81. https://doi.org/10.1145/3371158.3371167
[10] W. Hu, J. Patel, Z. Robert, P. Novosad, S. Asher, Z. Tang, M. Burke, D. Lobell, and S. Ermon. 2019. Mapping missing population in rural India: A deep learning approach with satellite imagery. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. https://doi.org/10.1145/3306618.3314263
[11] K. Ayush, B. Uzkent, M. Burke, D. B. Lobell, and S. Ermon. 2020. Generating interpretable poverty maps using object detection in satellite images. In *Proceedings of the 29th International Joint Conference on Artificial Intelligence Special Track on AI for CompSust and Human Well-Being*. 4410–4416. https://doi.org/10.24963/ijcai.2020/608
[12] S. Asher, T. Lunt, R. Matsuura, and L. Novosad. 2021. Development research at high geographic resolution: An analysis of night-lights, firms, and poverty in India using the SHRUG Open Data platform. *World Bank Economic Review* 35, 4 (2021), 845–871. https://doi.org/10.1093/wber/lhab003
[13] E. Aiken, S. Bellue, D. Karlan, C. Udry, and J. E. Blumenstock. 2022. Machine learning and phone data can improve targeting of humanitarian aid. *Nature* 603, 7903 (2022), 864–870. https://doi.org/10.1038/s41586-022-04484-9
[14] Landsat NASA. 2022. Landsat 7 | Landsat Science. Landsat Science | A Joint NASA/USGS Earth Observation Program. Retrieved September 5, 2023 from https://landsat.gsfc.nasa.gov/satellites/landsat-7/
[15] M. Wong. 2023. Visible Infrared Imaging Radiometer Suite (VIIRS). *Earthdata*. Retrieved September 4, 2023 from https://earthdata.nasa.gov/earth-observation-data/near-real-time/download-nrt-data/viirs-nrt
[16] D. H. Fabini, D. P. De Leon Barido, A. Omu, and J. Taneja. 2014. Mapping induced residential demand for electricity in Kenya. In *Proceedings of the 5th ACM Symposium on Computing for Development (ACM DEV-5'14)*. ACM, New York, NY, 43–52. https://doi.org/10.1145/2674377.2674390
[17] Deepanshu Mohan, Soumya Marri, Bilquis Calcuttawala, Malhaar Kasodekar, Aniruddh Bhaskaran, and Hemang Sharma. 2023. Modi govt's fiscal policy on welfare: Trends so far and what to expect. *The Wire*. Retrieved September 4, 2023 from https://thewire.in/economy/modi-govts-fiscal-policy-on-welfare-trends-so-far-and-what-to-expect
[18] G. S. Fields. 2001. *Distribution and Development: A New Look at the Developing World*. MIT Press, Cambridge, MA. https://doi.org/10.7551/mitpress/2465.001.0001
[19] R. W. Fogel. 1987. *Some Notes on the Scientific Methods of Simon Kuznets*. NBER Working Paper No. w2461. National Bureau of Economic Research. https://doi.org/10.3386/w2461
[20] C. E. Utazi, J. Thorley, V. A. Alegana, M. J. Ferrari, S. Takahashi, B. Roche, J. Lessler, and A. J. Tatem. 2018. High resolution age-structured mapping of childhood vaccination coverage in low and middle income countries. *Vaccine* 36, 12 (2018), 1583–1591. https://doi.org/10.1016/j.vaccine.2018.02.020
[21] A. Elmustafa, E. Rozi, Y. He, G. Mai, S. Ermon, M. Burke, and D. B. Lobell. 2022. Understanding economic development in rural Africa using satellite imagery, building footprints and deep models. In *Proceedings of the 30th International Conference on Advances in Geographic Information Systems*. https://doi.org/10.1145/3557915.3561025
[22] A. Krizhevsky, I. Sutskever, and G. E. Hinton. 2012. ImageNet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems (NIPS'12)*. 1097–1105. http://books.nips.cc/papers/files/nips25/NIPS2012_0534.pdf
[23] S. Lathuilière, P. Mesejo, X. Alameda-Pineda, and R. Horaud. 2020. A comprehensive analysis of deep regression. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42, 9 (2020), 2065–2081. https://doi.org/10.1109/tpami.2019.2910523
[24] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He. 2017. Aggregated residual transformations for deep neural networks. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'17)*. 5987–5995. https://doi.org/10.1109/cvpr.2017.634
[25] Y. Cui, M. Jia, T. Lin, Y. Song, and S. Belongie. 2019. Class-balanced loss based on effective number of samples. In *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'19)*. 9260–9269. https://doi.org/10.1109/cvpr.2019.00949

[26] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. 2017. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV'17)*. 618–626. https://doi.org/10.1109/iccv.2017.74

[27] M. D. Zeiler and R. Fergus. 2014. Visualizing and understanding convolutional networks. In *Computer Vision—ECCV 2014*. Lecture Notes in Computer Science, Vol. 8689. Springer, 818–833. https://doi.org/10.1007/978-3-319-10590-1_53

[28] N. Jethani, M. Sudarshan, Y. Aphinyanaphongs, and R. Ranganath. 2021. Have we learned to explain?: How interpretability methods can learn to encode predictions in their interpretations. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*. 1459–1467. https://doi.org/10.48550/arXiv.2103.01890

[29] N. Otsu. 1979. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics* 9, 1 (1979), 62–66. https://doi.org/10.1109/tsmc.1979.4310076

[30] S. Kuznets. 1955. Economic growth and income inequality. *American Economic Review* 45, 1 (1955), 1–28. http://www.jstor.org/stable/1811581

[31] S. Asher, K. Nagpal, and P. Novosad. 2018. *The Cost of Distance: Geography and Governance in Rural India*. World Bank Working Paper. World Bank. https://thedocs.worldbank.org/en/doc/499571527990962709-0010022018/original/B1administrativeremotenesspaper20180223.pdf

[32] Z. Chen, B. Yu, C. Yang, Y. Zhou, S. Yao, X. Qian, C. Wang, B. Wu, and J. Wu. 2021. An extended time series (2000–2018) of global NPP-VIIRS-like nighttime light data from a cross-sensor calibration. *Earth System Science Data* 13, 3 (2021), 889–906. https://doi.org/10.5194/essd-13-889-2021

[33] X. Chen and W. D. Nordhaus. 2010. *The Value of Luminosity Data as a Proxy for Economic Statistics*. Cowles Foundation Discussion Paper No. 1766. Cowles Foundation. https://doi.org/10.2139/ssrn.1666164

[34] C. D. Elvidge, K. E. Baugh, S. B. Anderson, P. J. Sutton, and T. K. Ghosh. 2012. The night light development index (NLDI): A spatially explicit measure of human development from satellite data. *Social Geography* 7, 1 (2012), 23–35. https://doi.org/10.5194/sg-7-23-2012

[35] Giorgia Giovannetti and Elena Perra. 2019. *Syria in the Dark: Estimating the Economic Consequences of the Civil War through Satellite-Derived Night Time Lights*. Working Papers—Economics wp2019. Universita degli Studi di Firenze. https://ideas.repec.org/p/frz/wpaper/wp2019_05.rdf.html

[36] World Bank. 2023. World Bank Open Data. Retrieved September 5, 2023 from https://data.worldbank.org/indicator/EG.ELC.ACCS.ZS

[37] T. Baskaran, B. Min, and Y. Uppal. 2015. Election cycles and electricity provision: Evidence from a quasi-experiment with Indian special elections. *Journal of Public Economics* 126 (2015), 64–73. https://doi.org/10.1016/j.jpubeco.2015.03.011

[38] A. Sharma. 2023. Electrification still a challenge in rural India. *Frontline*. Retrieved September 5, 2023 from https://frontline.thehindu.com/the-nation/electrification-still-remains-a-challenge-in-rural-india/article66493576.ece

[39] A. Perez, C. Yeh, G. Azzari, M. Burke, D. Lobell, and S. Ermon. 2017. Poverty prediction with public Landsat 7 satellite imagery and machine learning. In *Proceedings of the NIPS 2017 Workshop on Machine Learning for the Developing World*. http://arxiv.org/abs/1711.03654

[40] S. Pandey, T. Agarwal, and N. C. Krishnan. 2018. Multi-task deep learning for predicting poverty from satellite images. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence, the 30th Innovative Applications of Artificial Intelligence Conference, and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (AAAI'18/IAAI'18/EAAI'18)*. Article 957, 6 pages. https://doi.org/10.1609/aaai.v32i1.11416

[41] R. Rustowicz, R. Y. Cheong, L. Wang, S. Ermon, M. Burke, and D. B. Lobell. 2019. Semantic segmentation of crop type in Africa: A novel dataset and analysis of deep learning methods. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*. 75–82. https://openaccess.thecvf.com/content_CVPRW_2019/papers/cv4gc/Rustowicz_Semantic_Segmentation_of_Crop_\Type_in_Africa_A_Novel_Dataset_CVPRW_2019_paper.pdf

[42] R. Goldblatt, A. R. Ballesteros, and J. Burney. 2017. High spatial resolution visual band imagery outperforms medium resolution spectral imagery for ecosystem assessment in the Semi-Arid Brazilian Sertão. *Remote Sensing* 9, 12 (2017), 1336. https://doi.org/10.3390/rs9121336

[43] P. Helber, B. Bischke, A. Dengel, and D. Borth. 2019. EuroSAT: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 12, 7 (2019), 2217–2226. https://doi.org/10.1109/jstars.2019.2918242

[44] W. Wu, Z. Lin, P. Oghazi, and P. C. Patel. 2022. The impact of demonetization on microfinance institutions. *Journal of Business Research* 153 (2022), 1–18. https://doi.org/10.1016/j.jbusres.2022.08.009

[45] A. Sen, D. Ghatak, K. Kumar, G. Khanuja, D. Bansal, M. Gupta, K. Rekha, S. Bhogale, P. Trivedi, and A. Seth. 2019. Studying the discourse on economic policies in India using mass media, social media, and the parliamentary question hour data. In *Proceedings of the 2nd ACM SIGCAS Conference on Computing and Sustainable Societies (COMPASS'19)*. ACM, New York, NY, 234–247. https://doi.org/10.1145/3314344.3332489

[46] S. K. Basu, S. Ganguly, S. Mukhopadhyay, R. DiBiano, M. Karki, and R. R. Nemani. 2015. DeepSat: A learning framework for satellite imagery. In *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems (SIGSPATIAL'15)*. ACM, New York, NY, Article 37, 10 pages. https://doi.org/10.1145/2820783.2820816

[47] Internal Labour Organization. n.d. Informal Economy in South Asia. Retrieved September 5, 2023 from https://www.ilo.org/newdelhi/areasofwork/informal-economy/lang--en/index.htm

[48] J. Drèze and A. Sen. 2013. *An Uncertain Glory: India and Its Contradictions*. Princeton University Press, North Oxford, UK. https://press.princeton.edu/books/hardcover/9780691160795/an-uncertain-glory

[49] D. Goswami, S. B. Tripathi, S. Jain, S. Pathak, and A. Seth. 2019. Towards building a district development model for India using census data. In *Proceedings of the ACM SIGCAS Conference on Computing and Sustainable Societies (COMPASS'19)*. https://doi.org/10.1145/3314344.3332491

Received 14 February 2023; revised 15 June 2023; accepted 6 July 2023