Using Satellite Images to Track Relative Socioeconomic Development in India

Aatif Nisar Dar*
Badrinath Padmanabhan*
aatif.dar11@gmail.com
badrinath95@gmail.com
Indian Institute of Technology, Delhi
Delhi, India

Mrunal Atul Kadane
Mrunal.Atul.Kadhane.cs121@cse.iitd.ac.in
Indian Institute of Technology, Delhi
Delhi, India

Aaditeshwar Seth
aseth@cse.iitd.ac.in
Indian Institute of Technology, Delhi
Delhi, India

ABSTRACT

Studying how regions develop over time can offer valuable insights, but it's challenging with traditional data like censuses and surveys, which aren't frequent and face a lot of delays. Using satellite imagery for socio-economic indicators has become a useful alternative to track development at fine spatial and temporal scales. In this paper, we train a model using satellite imagery to estimate socio-economic development at the village level in India. We test its consistency over time and use it to analyze development trends over a two-decade period. Our study looks at how factors like the geographic distance of a village to economic hubs and the inequality of development in the district affect village development. Our results provide evidence of the possible impact that policy changes during this period have had on village development.

KEYWORDS

Satellite Images, Socioeconomic status, Poverty, Nightlights, Inequality.

ACM Reference Format:
Aatif Nisar Dar, Badrinath Padmanabhan, Mrunal Atul Kadane, and Aaditeshwar Seth. 2018. Using Satellite Images to Track Relative Socioeconomic Development in India. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX).* ACM, New York, NY, USA, 4 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Studying how villages develop over time is crucial at sub-national scales as it provides important insights into underlying development processes. Census data can potentially provide valuable insights into socioeconomic development but it's not without its limitations. Census data faces delays in conducting and publishing, resulting in outdated information. For example, in India, the most recent census data was conducted in 2011. Usually, this information is collected every ten years, but the census scheduled for 2021 was postponed due to COVID-19 disruptions. So, to fill in the gaps, and standardize development indicators, researchers often resort to proxy variables like nightlights satellite data or machine learning models trained on satellite imagery. But there are still some issues, especially in India. Firstly, the temporal robustness of models utilizing daytime satellite imagery to estimate socio-economic indicators hasn't been sufficiently evaluated. It is not sure if the models using satellite images can predict development changes well over time, especially at the village level. Secondly, while nightlight time series data has been explored for tracking sub-national development, their limited variability in rural villages isn't enough to understand how villages are changing. To address these challenges, we propose a method, combining both daytime satellite images with night-time light data to get a better idea of a composite indicator of socio-economic development at the village level.

Our approach leverages pre-trained Convolutional Neural Networks (CNNs) on a variant of a ResNet architecture to generate initial estimates of development variables based on 2011 Indian census data as ground truth labels. We then refine these estimates by building a model that incorporates additional features, namely; the first-level estimates of development variables of neighboring villages, nightlight-based features for the given village and neighboring villages, and the distance of a village to the nearest hub of economic activity (Obtained also from the nightlight 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. This generated standardized socio-economic development estimates for 2001 and 2019 enables us to explore various hypotheses regarding village development dynamics.

For evaluations, Gulgulia et. al. [5] have used the Aggregate Development Index (ADI), 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. In our work, we use the Relative Wealth Index (RWI) as it is a standard metric to evaluate socioeconomic development. The Relative Wealth Index (RWI) serves as a pivotal tool in assessing socioeconomic status, constructed from a compilation of vital asset ownership variables. It operates as a proxy measure, offering valuable insights into the relative prosperity of households within a given population or geographical area. By analyzing the distribution of assets, the RWI enables nuanced comparisons, facilitating

*Both authors contributed equally to this research.

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

targeted interventions and policy decisions to address disparities and foster inclusive development initiatives.

2 RELATED WORK

Traditionally, researchers have relied on household surveys or censuses to track socioeconomic development, but these methods come with limitations [11, 7, 13]. Recent advancements in machine learning have led to alternative approaches, such as using mobile phone metadata or satellite imagery. Blumenstock et al. [2] demonstrated the effectiveness of mobile phone data in predicting poverty in Rwanda, while nighttime satellite imagery has been utilized to estimate metrics like GDP and electricity access [4]. However, challenges such as noise in nighttime data have prompted the exploration of daytime satellite imagery and combinations of both [8, 12, 6]. Meanwhile, studies analyzing visual data at a granular level, such as Google Street View images, have shown promise in predicting neighborhood income levels and poverty [9, 10, 1]. Fatehkia et al. introduced a unique approach using Facebook's Advertising Data to map socioeconomic development, showing comparable results to satellite data in some contexts [3]. Gulgulia et. al [5] proposed an altogether different work where they utilized both daytime and nighttime to track the socio-economic development in rural India. They build a composite indicator known as the Aggregate Development Index (ADI) that combines variables related to asset ownership, access to water, bathroom facilities, literacy, and so forth. In our work, we will also utilize both daytime and nighttime satellite imagery but use Relative Wealth Index (RWI) as indicators.

3 DATASET

Various datasets are utilized in this study, including the Census of India, village shapefiles [1], Landsat 7 satellite data, and DMSP and VIIRS nightlights data[2]. The Census of India provides valuable socio-economic indicators such as housing materials, fuel sources, water access, asset ownership, literacy rates, and employment patterns at the village level, serving as labels for machine learning models. Village shapefiles are obtained from the 2001 census and mapped with 2011 census data using SHRUG [3], allowing for spatial analysis. Landsat 7 satellite data spanning 2001, 2011, and 2019 are used to track changes in land use and cover over time, aligning with census years for better comparison. Alongside satellite imagery, nightlight data is used to understand economic development trends by examining nighttime illumination patterns. Satellite data is subjected to some preprocessing, which involves breaking down census data into categories and then using them to categorize the satellite data. 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, kerosene, or LPG (Liquified Petroleum Gas), PNG (Piped Natural Gas), biogas, etc. These parameters are clubbed into three broad types: rudimentary (RDV), intermediate (INT), and advanced (ADV). This thorough preparation aims to reveal detailed socio-economic trends and enhance our understanding of rural development in India.

[1]https://sedac.ciesin.columbia.edu/data/set/india-india-village-level-geospatial-socio-econ-1991-2001/data-download
[2]https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/YGIVCD
[3]https://www.devdatalab.org/shrug$_{d}ownload/$

4 METHODOLOGY AND RESULTS

Our goal is to train machine learning models that can take daytime and nightlight satellite data for a village as an input, and then output the socio-economic development level for the village for various indicators. As discussed in the Dataset section, 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). A ResNeXt-50 architecture, trained on ImageNet data, is tailored to output village levels for various socio-economic indicators. We split the data into training and test sets in an 80:20 ratio and employed data augmentation techniques such as image rotation and reflection to enhance model generalization. Weighted Cross entropy loss function is used to address the data imbalance issue, which penalizes weights to each class based on the effective number of samples. Table 1 contains the results of the CNN model trained on the five indicator villages.

Table 1: CNN Model Training Results

Indicator	Train Accuracy (%)	Validation Accuracy (%)
MSW	75.34	74.10
ASSET	77.06	78.93
FC	72.09	79.47
BF	77.09	79.47
LIT	66.09	79.47

Next, for each indicator, softmax outputs from its CNN-trained model are taken as independent features and are concatenated with several more features to form an expanded feature set. The features include softmax outputs from CNN-trained models for other indicators, mean softmax outputs over the target village and its neighbors, nightlights-based features for the target village, derived features such as logarithm and square root transformations of these values, distances to economic hubs, and population features.

To identify neighbors, we compute the centroid and radius of each village and consider villages within twice the radius as neighbors. Nightlight-based features are generated by identifying economic hubs within the district where the target village is situated using a blob identification procedure. This procedure automatically thresholds nightlight values to delineate hubs of economic activity, providing features such as distances to hubs, hub size, and hub intensity. Population features, including total village population and number of households, are also included.

Feature selection is performed to identify the most important features for temporal robustness, considering census data from 2001. Random forest binary classifiers are trained on variables available for both 2001 and 2011, with feature importance scores used to select the top-ranked features. Linear regression models are then built using these selected features to estimate each indicator's RUD, INT, and ADV parameters. These estimates are discretized by mapping them to the closest cluster centroids for the indicator, producing final output labels for the village. The selected features primarily include those related to economic activity hubs, village population, and CNN-based features from the village itself. Regression models are trained to estimate parameters for each indicator, which are then mapped to produce final output labels. This comprehensive approach ensures the models' temporal robustness and accuracy in predicting village socio-economic development indicators over time. Table 2 contains the regression results after training the model.

Table 2: RMSE and Normalized RMSE Results

RMSE	Normalized RMSE
1.527	0.170

The DHS conducts data collection encompassing various household assets and characteristics. These variables encapsulate asset ownership, housing conditions, access to basic services, and possession of household items. We denote these variables as $X_1, X_2, ..., X_n$ for n households and M variables. Columns from the Census 2011 are matched with the ones provided in the DHS to identify the set of features used to understand wealth in a household. These features include material of walls, floor material, whether the resident owns a house, etc. We select 15 indicator features and specify levels as (RUD, INT, ADV) as mentioned in the study by [5], and proceed to create an augmented dataset containing only the indicator columns from the census data. Our augmented data contained the percentage of households with RUD, INT, and ADV respectively in each indicator feature.

Following the standard procedure in constructing the Relative Wealth Index (RWI), we normalize the data and compute the covariance matrix. We then perform Principal Component Analysis (PCA) on the data and take the first principal component (PC1). The weight along each dimension represents the factor loading for that particular column. To obtain the RWI value for the villages, we use the factor loadings of the first principal component and multiply them with the rows individually (here rows represent the data of a given village). The final Wealth Index of a village is obtained by summing up these products over all the dimensions.

We use the RWI values obtained and further categorize them into quintiles representing the percentile position of a village based on its wealth index. The first quintile represents the top 20 percentile (80-100), while the second quintile (60-80), the third quintile (40-60), the fourth quintile (20-40), and the fifth quintile (0-20) represent different percentile categories. To ensure that our calculation of RWI made sense, we compared the values obtained with the Asset Deprivation Index (ADI) obtained in the study by [5]. We found the Pearson's Coefficient to be **0.8395 (High)**. We made a scatterplot of ADI vs RWI and found that the values are meaningful. In Figure 1, we can see that the RWI values increase with an increase in ADI, which is further numerically emphasized by its high correlation coefficient.

The RWI values are ordered in quintiles (percentiles from 0 to 20 (quintile 5), 20 to 40 (quintile 4), 40 to 60 (quintile 3), 60 to 80 (quintile 2), 80 to 100 (quintile 1)). From Figure 2, we can see that the majority of villages with ADI less than 9 (low ADI) are in quintiles 5 and 4. For villages with medium ADI, the villages lie in the second and third quintiles. Similarly, for villages with high ADI, their RWI lies mostly in quintiles 1 and 2.

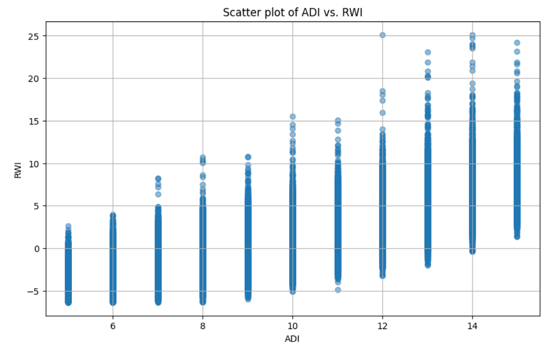

Figure 1: The scatterplot of ADI vs RWI.

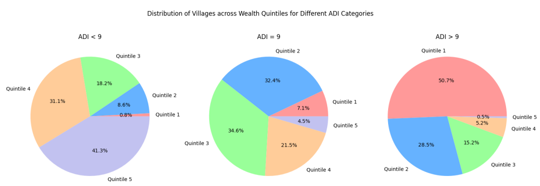

Figure 2: Distribution of villages across wealth quintiles for different ADI categories.

5 CONCLUSION

We utilized satellite data to create a socio-economic indicator metric at the village level using the census data of 2011. We aim to find the Relative Wealth Index for other years like 2001 and 2019 to analyze the progress of villages over time. This research highlights the potential of satellite data to track development trends and informs future studies aimed at understanding the impact of welfare expenditure on socio-economic progress. In the future, we aim to replace the CNN-based architecture with a Transformer-based architecture.

REFERENCES

[1] Ambika Acharya, Helen Fang, and Shubha Raghvendra. 2017. Neighborhood watch: using cnns to predict income brackets from google street view images. *URL: http://cs231n. stanford. edu/reports/2017/pdfs/556. pdf*.

[2] Joshua Blumenstock, Gabriel Cadamuro, and Robert On. 2015. Predicting poverty and wealth from mobile phone metadata. *Science*, 350, 6264, 1073–1076.

[3] Masoomali Fatehkia, Benjamin Coles, Ferda Ofli, and Ingmar Weber. 2020. The relative value of facebook advertising data for poverty mapping. In *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 14, 934–938.

[4] Tilottama Ghosh, Sharolyn J Anderson, Christopher D Elvidge, and Paul C Sutton. 2013. Using nighttime satellite imagery as a proxy measure of human well-being. *sustainability*, 5, 12, 4988–5019.

[5] 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 Journal on Computing and Sustainable Societies*, 1, 2, 1–31.

[6] J Vernon Henderson, Adam Storeygard, and David N Weil. 2012. Measuring economic growth from outer space. *American economic review*, 102, 2, 994–1028.

[7] STUDYING VILLAGE ECONOMIES IN INDIA. [n. d.] A note on household income surveys in india.

[8] Neal Jean, Marshall Burke, Michael Xie, W Matthew Davis, David B Lobell, and Stefano Ermon. 2016. Combining satellite imagery and machine learning to predict poverty. *Science*, 353, 6301, 790–794.

[9] Nikhil Naik, Jade Philipoom, Ramesh Raskar, and César Hidalgo. 2014. Streetscore-predicting the perceived safety of one million streetscapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 779–785.
[10] Nikhil Naik, Ramesh Raskar, and César A Hidalgo. 2016. Cities are physical too: using computer vision to measure the quality and impact of urban appearance. *American Economic Review*, 106, 5, 128–132.
[11] Luisa Natali and Chris De Neubourg. 2014. A tale of two short consumption modules.
[12] Ye Ni, Xutao Li, Yunming Ye, Yan Li, Chunshan Li, and Dianhui Chu. 2021. An investigation on deep learning approaches to combining nighttime and daytime satellite imagery for poverty prediction. *IEEE Geoscience and Remote Sensing Letters*, 18, 9, 1545–1549. DOI: 10.1109/LGRS.2020.3006019.
[13] Hans Pettersson, O Ajayi, G Kalton, K Krotki, J Lepkowski, H Pettersson, and A Turner. 2005. Survey design and sample design in household budget surveys. *Household sample surveys in developing and transition countries*, 2005, 71e94.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009