Using Satellite Data to predict Socio-Economic Development in Indian Villages

Manoov R, Adhya Dagar

Vellore Institute of Technology,Vellore

March 3, 2020

Overview

Conducting census is an enormous and expensive task:

- Growth of a country can be measured through socio-economic parameters which cover qualitative and economic well being
- Census is held every 10 years(2001 and 2010 in India), includes extensive survey of household and takes several years for successful compilation
- Difficulties in attributing socio-economic activities to a finer geo-spatial unit(villages). How do we know a particular district is outperforming others? We study its constituents.
- Various satellite based imagery prediction methods have not yet been used at village level

Introduction

- Various indicators such as GDP (Gross Domestic Product), literacy, employment, electricity access, drinking water access, bathroom facilities, asset ownership, condition of household, forest cover etc are used to measure the socio- economic development of a region.
- In India, the national census covers a wide range of these development indicators at different spatial granularity (country/ state/ district/ village).
- Recent advances in machine learning systems have facilitated this analysis of processing high-resolution satellite data for the prediction of socio-economic indicators.
- Temporal transferability of satellite-data based models has not been evaluated for socio-economic indicators so far for village level.

Prediction from Nightlights,Daytime Satellite Imagery and combination of both

- Nightlights are satellites that are responsible for light intensity measurement during night hours and measurements show strong correlation with GDP
- GDP at country level is not a good measure of well being due to lack of accounting for different indicators and we need to find results at the sub national scale
- GDP tells nothing about distribution of development within the country
- Estimations at sub-national scales - villages have also been made but inaccurate due to blooming effect where diffusion of nightlight over long distances in certain topographies and almost unobservant intensities in rural areas make it unusable.

Prediction from Nightlights,Daytime Satellite Imagery and combination of both-continued

- Poverty based mapping techniques using daylight satellite imagery have given better results than nightlight based predictions
- At village level in India a few supervised learning techniques have been used to predict population density and poverty
- Large training datasets have been used to build CNN based regression models for other socio-economic indicators like education, literacy, and health
- However, none of these works have been tested for prediction over time, a likely reason being the unavailability of ground-truth data at different points in time for which the satellite data is also available
- Other transfer-learning approaches which use labeled datasets like ImageNet , and DeepSat for pre-training deep learning models would also worked poorly with Indian district level dataset as it is too small to fine-tune the prediction models however this approach is yet to be tested on village level models.

Methods

We would use the following approach for prediction of socio-economic indicators:

- Satellite Imagery is an acceptable proxy for tracking development through census based data;
- We start with the cross-sectional prediction of eight socio-economic indicators for each village by using a model where the input is given as the satellite imagery of that particular village.
- We have to do this for the census data of 2001 as well as 2011, and demonstrate a reasonable accuracy of our model;

Methods continued..

- For each of the eight socio-economic indicators (Assets, BF, CHH, FC, MSW, MSL,LIT), every district is assigned a label of level-1/2/3, which indicates their level of development for that indicator.
- Hence our prediction task can be formulated as a **multi-class classification problem** for each indicator; and
- We use Landsat7 satellite system for daytime imagery since it is available since 1999, which matches the years of 2001 and 2011 for which we have the ground-truth census data. We would use the freely available spectral data via the Google Earth Engine (GEE) platform, at a 100m resolution, capturing the tier-1 top-of-atmosphere reflectance.

Classification Probelm

We would be looking at this machine learning problem as a classification problem. Each image of a village will be classified into 1/2/3 label depending on its level of development. 1 being the most rudimentary (RUD) or least developed; 2 being intermediate (INT) or moderately developed and 3 being advanced (ADV) or the most developed

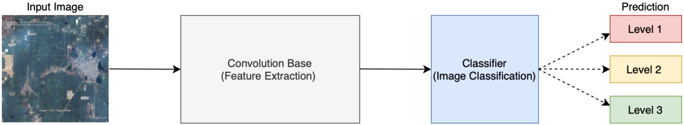


Figure: Model Architecture

Transfer Learning

- When we train the network on a large dataset(for example: ImageNet) , we train all the parameters of the neural network and therefore the model is learned. It may take hours on your GPU.
- We can give the new dataset to fine tune the pre-trained CNN. Consider that the new dataset is almost similar to the orginal dataset used for pre-training. Since the new dataset is similar, the same weights can be used for extracting the features from the new dataset.

Pretraining Models

Pretraining ResNet using ImageNet Dataset

ImageNet Dataset: ImageNet, is a dataset of over 15 millions labeled high-resolution images with around 22,000 categories.

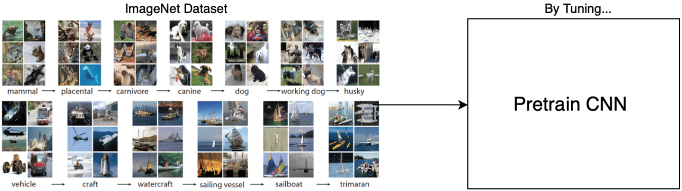


Figure: Pretraining using ImageNet

Pretraining Models

Pretraining ResNet using DeepSat Dataset

DeepSat: SAT-6 consists of a total of 405,000 image patches each of size 28x28 and covering 6 landcover classes - barren land, trees, grassland, roads, buildings and water bodies.

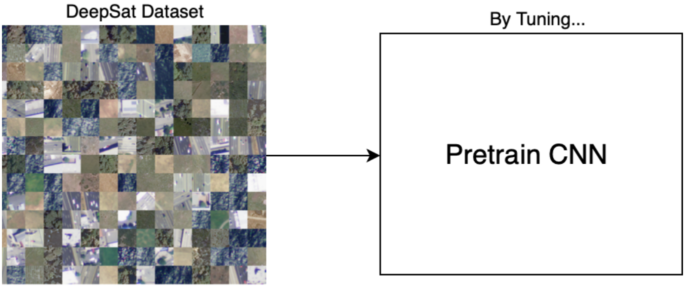


Figure: Pretraining using DeepSat

Training LandSat7 images of Indian villages on Fine-tuned model

We will be using the Village level Shapefiles in SHRUG dataset to cut out villages from the Landsat7 data which is downloaded from Google Earth Engine.

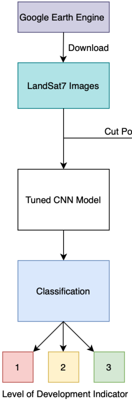
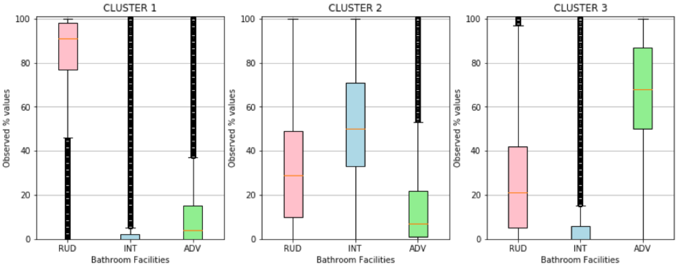


Figure: Training using LandSat on fine-tuned model

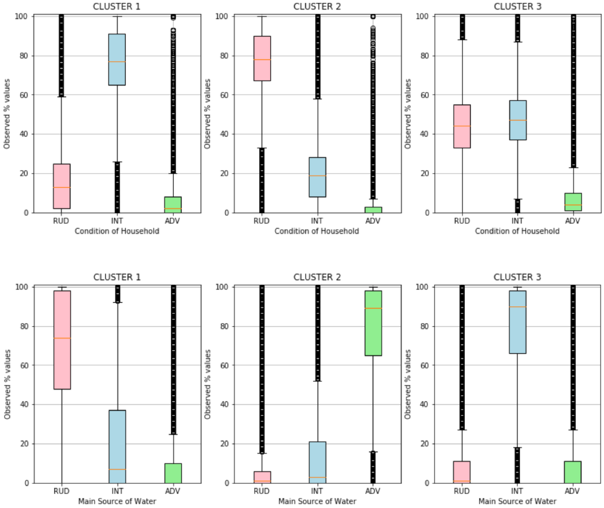
Ground Truth for village level indicators

After performing Village level labelling for all indicators we get the final results- We can see each cluster represents distinct level of development. Eg. In clusters of Bathroom Facility(BF)-In cluster 1 RUD type of villages are dominant, in cluster 2 INT types are dominant and in cluster 3 ADV types are dominant.



Ground Truth for village level indicators

Similarly clustering is performed for all the 7 indicators. Some of the results are-



Pre-Processing of Satellite Images

We download Landsat7 images (cloud free) for different states for the year 2011. The images have been downloaded twice, once as medium composite and the other as min. The state images are then cut with the help of village shapefiles, which gives us the images of each village. After this, we crop all the village images. 150 x 150 pixels from the centre.

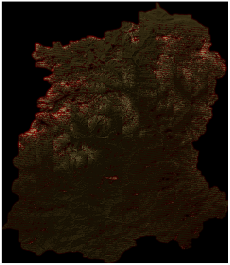


Figure: LandSat7 image of Sikkim

Hyperparameters and Evaluation

With each iteration the weights are updated and saved in .hd5 files.We will be using F2 score and R2 scores as evaluation metrics. The evaluation metric for each model is average F2 score, which is defined for one sample as: The F2 score prefers recall to precision, as can be seen in Figure 4. This means we punish false negatives more severely than false positives. This makes sense for a problem where we are trying to detect some rare phenomena: we would prefer to identify all occurrences even if we end up making some false positive mistakes.

$$F_2 = (1 + 2^2) \cdot \frac{\text{precision} \cdot \text{recall}}{(2^2 \cdot \text{precision}) + \text{recall}}$$

where Precision = $\frac{TP}{TP+FP}$ and Recall = $\frac{TP}{TP+FN}$.

Figure: Predicting F2 Score

Hyperparameters and Evaluation

In a previous study on deep learning models trained on ImageNet, VGG-16 was not able to break 0.90 F2 in the first 40 epochs while both Inception-v3 and ResNet-50 were. ResNet-50 was the best, achieving an F2 of 0.907 after 50- 60 epochs. Hence, following proven results we choose to implement ResNet50 model.

Model	Train F2	Val F2
Baseline	0.875	0.836
VGG-16	0.903	0.897
Inception-v3	0.912	0.901
ResNet-50	0.921	0.907
ResNet-50 (data aug/ensemble)	**0.922**	**0.908**

Figure: Comparison of various models on ImageNet

Hardware and Software Requirement

- Hardware Used- Super Computer Cluster at IIT Delhi to train Deep Learning Models and data preparation
- Software- Google Earth Engine: Used to download Landsat data
- Software- Google Colaboratory: Used to run python scripts for data pre-processing
- Software- Scikit-learn library: Scikit-learn is a free software machine learning library for python programming
- Software- TensorFlow: TensorFlow is an end-to-end open source platform for machine learning

We have referred to the following papers-

- 2019. Census of India Website: Office of the Registrar General and Census Commissioner, India. (2019). http://censusindia.gov.in/
- Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B., Ermon, S. (2016). Combining satellite imagery and machine learning to predict poverty. Science, 353(6301), 790–794. https://doi.org/10.1126/science.aaf7894
- Xie, M., Jean, N., Burke, M., Lobell, D., Ermon, S. (2016). Transfer learning from deep features for remote sensing and poverty mapping. 30th AAAI Conference on Artificial Intelligence, AAAI 2016, 3929–3935.
- (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), 1336. https://doi.org/10.3390/rs9121336
- Head, A., Manguin, M., Tran, N., Blumenstock, J. E. (2017). Can Human Development be Measured with Satellite Imagery? 1–11. https://doi.org/10.1145/3136560.3136576
- Seth, A. (n.d.). Temporal Prediction of Socio-economic Indicators

We have referred to the following papers-

- Sen, A., Ghatak, D., Kumar, K., Khanuja, G., Bansal, D., Gupta, M., . . . Seth, A. (2019). Studying the discourse on economic policies in India using mass media, social media, and the parliamentary question hour data. COMPASS 2019 - Proceedings of the 2019 Conference on Computing and Sustainable Societies, 234–247. https://doi.org/10.1145/3314344.3332489
- Dibyajyoti Goswami, Shyam Bihari Tripathi, Sansiddh Jain, Shivam Pathak, and Aaditeshwar Seth. 2019. Towards Building a District Development Model for India Using Census Data. (2019).
- Donaldson, D., Storeygard, A. (2016). The view from above: Applications of satellite data in economics. Journal of Economic Perspectives, 30(4), 171–198. https://doi.org/10.1257/jep.30.4.171
- Suraj, P. K., Gupta, A., Sharma, M., Paul, S. B., Banerjee, S. (2017). On monitoring development indicators using high resolution satellite images. 1–36. Retrieved from http://arxiv.org/abs/1712.02282
- Dugoua, Eugenie Kennedy, Ryan Urpelainen, Johannes. (2018). Satellite data for the social sciences: Measuring rural electrification

We have referred to the following papers-

- Elvidge, C. D., Baugh, K. E., Anderson, S. J., Sutton, P. C., Ghosh, T. (2012). The Night Light Development Index (NLDI): A spatially explicit measure of human development from satellite data. Social Geography, 7(1), 23–35. https://doi.org/10.5194/sg-7-23-2012
- Henderson, J. V., Storeygard, A., Weil, D. N. (2009). Nber Working Paper Series Measuring Economic Growth From Outer Space. Retrieved from http://www.nber.org/papers/w15199
- Xi Chen and William D Nordhaus. 2010. The value ofluminosity data as a proxy for economic statistics. Technical Report. National Bureau of Economic Research
- Giovannetti, G., Perra, E. (2019). Syria in the Dark: Estimating the Economic Consequences of the Civil War through Satellite-Derived Night-Time Lights Syria in the Dark:
- Zhaoxin Dai, Yunfeng Hu, and Guanhua Zhao. 2017. The suitability of different nighttime light data for GDP estimation at different spatial scales and regional levels. Sustainability 9, 2 (2017), 305.

We have referred to the following papers-

- Frank Bickenbach, Eckhardt Bode, Peter Nunnenkamp, and Mareike Söder. 2016. Night lights and regional GDP. ReviewofWorld Economics 152, 2 (2016), 425–447.
- 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 (2016), 790–794.
- Wenjie Hu, Jay Harshadbhai Patel, Zoe-Alanah Robert, Paul Novosad, Samuel Asher, Zhongyi Tang, Marshall Burke, David Lobell, and Stefano Ermon. 2019. Mapping Missing Population in Rural India: A Deep Learning Approach with Satellite Imagery. arXiv preprint arXiv:1905.02196 (2019)
- Shailesh MPandey, Tushar Agarwal, and Narayanan C Krishnan. 2018. Multi-task deep learning for predicting poverty from satellite images. In Thirty-SecondAAAI Conference on Artificial Intelligence.
- Potnuru Kishen Suraj, Ankesh Gupta, Makkunda Sharma, Sourabh Bikash Paul, and Subhashis Banerjee. 2017. On monitoring development using high resolution satellite images. arXiv preprint

We have referred to the following papers-

- Gary R Watmough, Peter M Atkinson, Arupjyoti Saikia, and Craig W Hutton. 2016. Understanding the evidence base for poverty–environment relationships using remotely sensed satellite data: an example from Assam, India. World Development 78 (2016), 188–203.
- Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. 2017. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. arXiv preprint arXiv:1709.00029 (2017).
- Saikat Basu, Sangram Ganguly, Supratik Mukhopadhyay, Robert DiBiano, Manohar Karki, and Ramakrishna Nemani. 2015. Deepsat: a learning frame- work for satellite imagery. In Proceedings ofthe 23rd SIGSPATIAL international conference on advances in geographic information systems. ACM, 37.
- Gyanesh Chander, Brian L Markham, and Dennis L Helder. 2009. Summary of current radiometric calibration coefficients for Landsat MSS, TM, ETM+, and EO-1 ALI sensors. Remote sensing ofenvironment 113, 5 (2009), 893–903.

The End