Temporal Prediction of Socio-economic Indicators Using Satellite Imagery

Chahat Bansal
Indian Institute of Technology
Delhi, India
chahat.bansal@cse.iitd.ac.in

Arpit Jain
Indian Institute of Technology
Delhi, India
arpit2602@gmail.com

Phaneesh Barwaria
Indian Institute of Technology
Delhi, India
phaneesh.barwaria.mcs18@cse.iitd.ac.in

Anuj Choudhary
Indian Institute of Technology
Delhi, India
anuj.choudhary27@gmail.com

Anupam Singh
Indian Institute of Technology
Delhi, India
anupam191196@gmail.com

Ayush Gupta
Indian Institute of Technology
Delhi, India
ayushabg@gmail.com

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

ABSTRACT

Machine learning models based on satellite data have been actively researched to serve as a proxy for the prediction of socio-economic development indicators. Such models have however rarely been tested for transferability over time, i.e. whether models learned on data for a certain year are able to make accurate predictions on data for another year. Using a dataset from the Indian census at two time points, for the years 2001 and 2011, we evaluate the temporal transferability of a simple machine learning model at sub-national scales of districts and propose a generic method to improve its performance. This method can be especially relevant when training datasets are small to train a robust prediction model. Then, we go further to build an aggregate development index at the district-level, on the lines of the Human Development Index (HDI) and demonstrate high accuracy in predicting the index based on satellite data for different years. This can be used to build applications to guide data-driven policy making at fine spatial and temporal scales, without the need to conduct frequent expensive censuses and surveys on the ground.

CCS CONCEPTS

• **Computing methodologies** → **Machine Learning**; • **Applied computing** → *Computing in government*.

KEYWORDS

poverty mapping, socio-economic development, satellite imagery, Landsat, census, temporal prediction

CoDS COMAD 2020, January 5–7, 2020, Hyderabad, India

ACM ISBN 978-1-4503-7738-6/20/01…$15.00
https://doi.org/10.1145/3371158.3371167

ACM Reference Format:
Chahat Bansal, Arpit Jain, Phaneesh Barwaria, Anuj Choudhary, Anupam Singh, Ayush Gupta, and Aaditeshwar Seth. 2020. Temporal Prediction of Socio-economic Indicators Using Satellite Imagery. In *7th ACM IKDD CoDS and 25th COMAD (CoDS COMAD 2020), January 5–7, 2020, Hyderabad, India.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3371158.3371167

1 INTRODUCTION

Socio-economic development of a society is the improvement in its qualitative well-being and economic status. Different indicators like GDP (Gross Domestic Product), literacy, employment, electricity access, drinking water access, etc are used to measure the socio-economic development of a region. In India, the national census [1] covers a wide range of these development indicators at different spatial granularity (country/ state/ district/ village). However, the census is an expensive exercise involving an enumeration of every household and is repeated only at a gap of every ten years. Subsequent releases of the compiled data take several more years. Hence, there is a need to develop methods for more frequent data collection to assess development at fine spatial and temporal scales, and then use the insights for data-driven policy making. Satellite imagery has been suggested as a viable data-source that can serve as a proxy for census data to predict different socio-economic indicators at national and sub-national scales [17, 20, 28, 29, 39, 44]. Recent advances in machine learning systems have facilitated this analysis of processing high-resolution satellite data for the prediction of socio-economic indicators.

Most machine learning models learn to classify the satellite imagery data into socio-economic indicators based on the ground-truth available for a particular year. It is unclear though, whether these models would transfer well over time, i.e. if models learned on a particular year can make accurate predictions on the satellite data from another year. Transferability of a well-known deep-learning model [25] across countries was shown to be sensitive to hyper-parameter tuning [20]. Factors such as extensive hyper-parameter tuning are likely to surface in the transferability of models over time also. To the best of our knowledge, temporal transferability

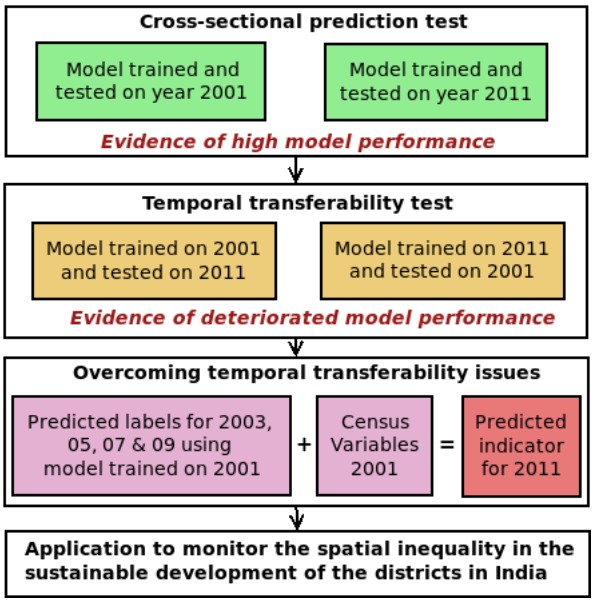

Figure 1: Overall paper summary

of satellite-data based models has not been evaluated for socio-economic indicators so far, especially with models trained on small datasets. A likely reason may be the unavailability of village or district level ground-truth data for different years, for which the satellite data from the same source is available.

In our study, we specifically focus on the *issue of evaluating the temporal transferability of satellite data based machine learning models for the prediction of socio-economic indicators*. We curate district-level ground-truth data from the Indian censuses of 2001 and 2011 to show that models learned on either year are not able to accurately predict indicators on the other year. We then develop a method to mitigate this transferability problem and make robust predictions for socio-economic indicators across different years. The predictions can be further used to monitor aggregate development across a broad range of different social and economic indicators, and we are able to visually show changes in the spatial inequality of development in India between 2001, 2011, and the current year of 2019.

Figure 1 summarizes the flow of the paper. Our work is divided into four parts:

(1) We begin with the cross-sectional prediction of six socio-economic indicators for each district by using a machine learning model which takes the satellite imagery of that district as an input. We do this for both the census years 2001 and 2011, and demonstrate a reasonable accuracy of our model.
(2) We then test the performance of our model across years: Using the model trained on data from 2001, we try to predict the development indicators for 2011, and vice versa. Results show that the model does not transfer well over time.
(3) We then present a novel method which mitigates this transferability problem to predict socio-economic indicators for each district in 2011, using the model trained on data from 2001. This is achieved by predicting the indicators at every alternate year between 2001 and 2011, and then aggregating these results as an error correction technique. A further improvement is achieved by using some variables of the census data from the base year of 2001 as an additional input to the model.
(4) In the final segment, we apply this model to show how spatial inequality in the development of India has changed between the years 2001 and 2019.

We next present related work in section 2, followed by a description of our dataset in section 3. In section 4 we present our model for cross-sectional prediction, and in section 5 we test the transferability of this model over different years. Our method to achieve temporal generalizability is described in section 6, and its application to track aggregate development over the years is shown in section 7. We finally conclude our study with discussions about future work in section 8. Our work is relevant and timely in using big-data techniques to aid policy makers in making data-driven decisions for socio-economic development.

2 RELATED WORK

Satellite data has been shown to have considerable potential in serving as a proxy to assess socio-economic development at fine spatio-temporal scales [11]. We describe related studies in this area, and call attention to a pressing need for a comprehensive evaluation of the transferability of various methods over time, as also highlighted by other researchers [12, 39].

2.1 Predictions from nightlights

Light intensity measurements done by satellites during night hours, called nightlights, have been shown to have a strong correlation with GDP at the country level [9, 13, 22, 34]. This correlation has led to important applications, such as attempts to assess the impact of war and post-war recovery efforts in Syria [16], where any ground-level censuses or surveys have not been conducted in recent years. Concerns have, however, also been raised about an over-estimation of GDP-fall, highlighting the need for stronger evaluations over time. Efforts have also been made for estimation of GDP at sub-national scales [10, 27, 30], but issues like the blooming effect where nightlights diffuse over long distances in certain topographies [23], and almost unobservant intensities in rural areas [4, 34], have raised concerns about the use of nightlights at fine spatial granularity. State-level GDP predictions have been attempted in India [30, 38], but district-level GDP data is largely unavailable to make stronger assessments.

2.2 Predictions from daytime satellite imagery

Given the problems with nightlights such as the blooming effect and lack of useful observations for rural areas, the use of multi-spectral daytime satellite imagery has seen considerable research interest. It has been anticipated that indicators for drinking water could be related to spectral features of surface-level water bodies, indicators for asset ownership could be related to the density of residential construction observable in the visible bands of daytime satellite imagery, etc. Some national-level studies for poverty mapping using daytime imagery have outperformed nightlights

based models [33, 40]. At sub-national levels, supervised learning techniques have been used to predict population density [24] and poverty [31] at the village-level in India. CNN-based regression models have also been built for other socio-economic indicators like education, literacy, and health [39] using a large dataset of 2,18,000 images for training. Semi-supervised learning techniques using generative adversarial networks [32] have been explored to predict poverty in the absence of sufficient labeled training data, but these models are hard to train. Some other applications have studied the relationship between poverty and environmental data [43], the ability to classify built-up and non-built-up areas [17], and finer categories of land-use classification into industrial areas, residential areas, cropland, and forests [21]. None of these works have, however, 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. Robinson et al. [36] use a deep learning approach on a training dataset of 8 million pixel values to predict the population in the US at a county-level. Even though their model trained on one year is shown to work across ten years, we cannot expect a similar result with any dataset, especially with small datasets.

2.3 Combination of nightlights and daytime satellite imagery

A well-known transfer learning approach has trained deep-learning models to use daytime imagery to predict nightlight intensities, and then use the mined features to predict poverty indicators [25]. This study reported reasonably good accuracy at a cross-sectional level and indicated the need to evaluate the model across time as well. Subsequent research [20] showed that the models did not transfer well across different countries without explicit hyper-parameter tuning. These models again have, however, not been tested for reliability over time. Other transfer-learning approaches which use labeled datasets like ImageNet [25], and DeepSat [3] for pre-training deep learning models would also work poorly with our dataset as it is too small to fine-tune these prediction models.

In our work, we build a simple model using daytime multi-spectral satellite data to predict six different socio-economic indicators at the district level. Our key contribution is then to show that this model, which performs well for a given year, does not transfer well to predict indicators for a different year. We come up with a method to address the transferability of our model; this method is generic and can be applied to different prediction models. As part of future work, we plan to evaluate the same method for other models as well.

3 DATASET

3.1 Satellite data

We use the 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 downloaded the freely available spectral data via the Google Earth Engine (GEE) platform, at a 100m resolution, capturing the tier-1 top-of-atmosphere reflectance [5, 42]. This data contains nine primary bands, as shown in Table 1. These primary bands can also be used to derive several other useful bands. Cloud cover in the images was removed through a standard process in GEE, which filters out the images having high cloud-cover values and then takes the median of the band values at a pixel-level for the remaining images in a year [18].

Band	Type	Resolution
B1	Blue	30m
B2	Green	30m
B3	Red	30m
B4	Near Infrared	30m
B5	Shortwave Infrared 1	30m
B6_VCID_1	Low-gain Thermal Infrared	30m
B6_VCID_2	High-gain Thermal Infrared	30m
B7	Shortwave Infrared 2	30m
B8	Panchromatic	15m
(B4-B3)/(B4+B3)	Normalized Difference Vegetation Index (derived)	30m
(B2-B5)/(B2+B5)	Modified Normalized Difference Water Index (derived)	30m
(B5-B4)/(B5+B4)	Normalized Difference Built Index (derived)	30m

Table 1: Landsat 7 bands

3.2 Census of India: 2001 and 2011

The Government of India conducts a population census every ten years. We use data from the 2001 and 2011 censuses, available from the official census website [6]. The census reports the number of households in each spatial unit (village, district, state), belonging to 90 different categories such as the type of construction of the house, the cooking fuel used, assets owned by the household, type of employment, sector of employment, and many others. The 2001 data was available only at the district level; hence we do our analysis at the district level only. Between 2001 and 2011, 47 districts were split into smaller ones due to administrative changes, and for our study we grouped them back into the original 593 districts that existed as of 2001. We use the district-level shape-files for all 593 districts to demarcate the corresponding satellite images.

3.2.1 Discretization of socio-economic variables. For an indicator variable like the type of fuel used for cooking, the census reports multiple constituent parameters such as the number of households that use firewood, kerosene, LPG (Liquefied Petroleum Gas), PNG (Piped Natural Gas), bio-gas, etc. To avoid building separate prediction models for each individual parameter, we need to compress these multiple parameters into a single value for each indicator. This is done in the following way: First, the parameters are grouped into three broad types - Rudimentary, intermediate, and advanced types. For example, firewood is considered as a rudimentary type of fuel for cooking, kerosene and cow-dung as intermediate types, and PNG and LPG as advanced types.

Next, a k-means clustering is performed for each indicator based on the percentage of households in a district that is of type rudimentary, intermediate, and advanced. As an example, Figure 2 shows a box-plot for the distribution of districts across three levels (k = 3) in terms of their use of different types of fuel for cooking. This clustering allows us to label each district as a level-1/2/3 district, where level-1 districts predominantly use rudimentary types of fuel

Variable	Using/Access to	Level-1 (in %)	Level- 2 (in %)	Level -3 (in %)
Asset Ownership	TV	15-30	30-50	60-85
	Telephone	35-55	40-60	50-60
	2 Wheeler	5-12	5-18	20-40
	4 Wheeler	0-2	0-5	2-12
Bathroom Facility (BF)	No Latrine facility	65-82	20-40	18-40
	Pit Latrine	0-5	30-45	0-10
	Piped Sewer/Septic Tank	15-28	25-40	50-70
Condition of Household (CHH)	Dilapidated house	5-10	0-5	0-5
	Livable house	55-65	40-50	25-35
	Good house	30-40	45-55	65-75
Fuel for Cooking (FC)	Firewood	60-80	0-12	10-25
	Cow Dung/Kerosene	30-50	40-60	5-20
	LPG/PNG/Bio gas	15-40	5-20	45-65
Main Source of Light (MSL)	No source of light	0-5	0-5	0-5
	Kerosene oil/Other oil	70-80	30-50	5-15
	Electricity/Solar Light	20-30	50-70	85-95
Main Source of Water (MSW)	Well/Spring/River	40-70	2-20	5-15
	Hand Pump/Tube Well	2-25	55-80	10-28
	Tap Water/Treated water	20-40	10-28	60-85

Table 2: Census variables: Range of % of households across all districts using or with access to different amenities [19]

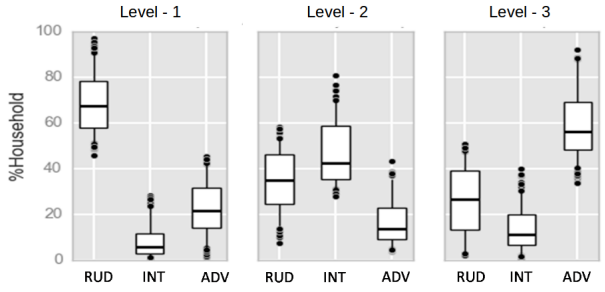

Figure 2: Fuel for cooking: District Box-plots for 3 levels [19]

for cooking, level-2 districts predominantly use intermediate types, and level-3 districts largely use advanced types of fuel for cooking. In this manner, we are able to map each district to levels 1/2/3 for every indicator. Table 2 summarizes this grouping along with the range of percentage value of households for different indicators and districts at different levels. A detailed explanation of the discretization method and robustness for different values of k can be found in [19].

We are thus able to frame our machine learning problem as a classification task where we use the spectral data of a district to predict it's level for an indicator. A separate classifier is learned for each indicator. Other than avoiding having to learn different models for each parameter of an indicator, this method is also useful for several additional reasons. First, as shown in the book *Factfulness* by Hans Rosling [37], a similar 4-level coarse mapping reflecting the different stages of development of a region, is easy for people to interpret and helps them compare different regions with one another. Second, it aggregates the constituent parameters to a single category without assigning arbitrary weights to combine together the various parameters for a variable. Third, it simplifies the training of classification models. Finally, the census data can have errors, and in such situations as explained by Ganguli et al. [14], a classification problem can help in eliminating noise which could otherwise get amplified if we were to build a regression model for each variable or its parameters.

4 CROSS-SECTIONAL CLASSIFICATION

As discussed in the previous section, for each of the six socio-economic indicators (Assets, BF, CHH, FC, MSW, and MSL), 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. In this section, we first present the feature extraction technique used to represent the spectral values for each district and then we proceed towards a classification model which uses these feature vectors to predict the district labels for the indicators. This cross-sectional classification is conducted for the census years 2001 and 2011 independently to test the model's performance for both the years.

4.1 Feature extraction

We build a simple model at this stage, by using histogram-based features for all the 12 (primary and derived) bands. We do this as follows. First, a *quantile binning* method is used to determine the bin-intervals for each band by taking the band-values for all pixels across all the districts. We determine these bin-intervals for a different number of bins (experimenting with values of 5, 10, 15, 20, 25 and 30 bins). Next, for each district, we find the frequency distribution of the band-values according to the bin-intervals computed for the band. This frequency distribution is then normalized with the count of the total number of pixels in the district. Thus, for each district, we are able to obtain 12 vectors, each of size equal to the number of bins used for that band. We experimented with a different number of bins for each band and finally chose the value of 10 which gave a high f1_score across all indicators when tested with different machine learning models. It is also intuitive not to set the bin-count too high, as it leads to a higher dimensional representation which increases the model complexity [26].

We chose this feature extraction method because it is sensitive to the tonal distribution of an image and is invariant to transformations like rotation and scaling. To demonstrate that it is able to capture relevant differences between districts, we show an example in Figure 3 of four districts, Moga and Balaghat which have a high vegetation cover due to large forests, and Jaipur and Nagpur which are highly urbanized districts with low vegetation cover. The histogram vector of these districts for the Normalized Difference Vegetation Index (NDVI) derived band clearly shows a difference between the high-vegetation and low-vegetation districts. However, this method does not capture any spatial features, and we leave it to future work to build more sophisticated models.

4.2 Classification model

Having created the feature vectors for each district, we next evaluate various classifiers for the multi-class classification problem. We experimented with different families of ML models including kernel-based (support vector machines), tree-based (decision trees),

Indicator	Census year 2001						Census year 2011					
	[25] Baseline	Majority Baseline	SVC	RF	XGBoost model		[25] Baseline	Majority Baseline	SVC	RF	XGBoost model	
	Weighted F1	Weighted F1	Weighted F1	Weighted F1	Weighted F1	Accuracy	Weighted F1	Weighted F1	Weighted F1	Weighted F1	Weighted F1	Accuracy
ASSETS	0.41	0.69	0.73	0.73	**0.79**	0.79	0.36	0.20	0.59	0.64	**0.67**	0.67
BF	0.42	0.56	0.71	0.76	**0.79**	0.77	0.39	0.41	0.56	0.67	**0.69**	0.69
CHH	0.38	0.31	0.52	0.58	**0.62**	0.61	0.36	0.25	0.55	0.59	**0.64**	0.63
FC	0.38	0.44	0.65	0.70	**0.74**	0.74	0.42	0.41	0.61	0.69	**0.76**	0.76
MSL	0.35	0.17	0.59	0.60	**0.64**	0.63	0.37	0.38	0.63	0.64	**0.72**	0.70
MSW	0.38	0.21	0.62	0.63	**0.66**	0.65	0.37	0.28	0.69	0.71	**0.76**	0.76

Table 3: District-level weighted F1 scores and accuracy for the year 2001 and 2011. The weighted F1 scores of the XGBoost model are compared against the scores of RF (Random Forest), SVC (Support Vector Classifier), Majority Baseline, and Jean et al. [25] baseline which refers to the model created using training data with shuffled labels

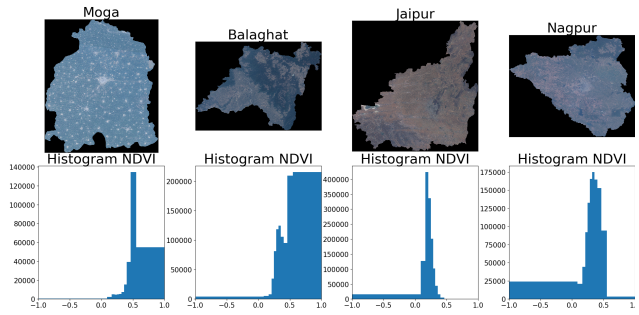

Figure 3: Histogram comparison for the NDVI band for four district examples

neural network (multi-layer perceptron) and ensemble models (random forest, XGBoost, AdaBoost). Among these models, XGBoost [8] gave the best results for all the socio-economic indicators, and we use this model for further analysis. Our results are in line with the observations made by [15] regarding the higher robustness of ensemble techniques to noise and small training datasets (593 districts in our case). Although CNN-based methods are known to give better results with images, we do not use them here because of our small dataset at the district level. In the future, we plan to improve the results by using more sophisticated machine-learning models over a bigger training dataset of village-level images.

While using XGBoost, we use the SMOTE (Synthetic Minority Oversampling Technique) method [7] to address class imbalance issues. SMOTE creates new minority class instances (synthetic) between existing (real) minority instances. All the hyper-parameters of the XGBoost model are set using the RandomSearch method followed by GridSearch in python's scikit-learn module. We use the 5-fold cross-validation technique for evaluating the performance of the model, and the resulting accuracies and the weighted F1-scores are summarized in Table 3 for the years 2001 and 2011.

4.3 Results and analysis

As can be seen from Table 3, even at a coarse resolution of spectral values at 100m, our simple model is able to attain a fairly reasonable performance in classifying several socio-economic indicators for both the years 2001 and 2011. The performance is different for different indicators and this trend persists even when using other classifiers like SVM, decision tree, random forest, and AdaBoost.

To demonstrate the statistical significance of our results, we use an approach to generate a comparison baseline similar to the one used by Jean et al. [25]. We arbitrarily shuffle the labels of the training data and then perform the classification. This classification task using the shuffled training dataset is done for both the years 2001 and 2011. The results are summarized in Table 3. We observe that the F1-scores for each indicator on the shuffled training data are much lesser, showing that the classification accuracy of our model is not coincidental. Further, we compare our results against a majority baseline where, for each indicator, every district is assigned the most frequently occurring label as its predicted label for that indicator. As shown in Table 3, the performance of XGBoost model surpasses the baseline results for each indicator.

5 TEMPORAL TRANSFERABILITY ANALYSIS

Unlike with most related work where ground-truth data was not available at multiple points in time, we are in a position to evaluate whether models learned on some year are able to perform well on data from another year. This is an essential requirement if satellite data models are to serve as an effective proxy for census data. We therefore use the XGBoost models described in the previous section learned on the data from 2001 to predict the labels in 2011, and then vice versa to learn the models on data from 2011 to predict the labels in 2001.

5.1 Results and analysis

Figure 4 shows a comparison between the performance of models trained on data from 2001 to classify the districts in 2011. Weighted F1-scores are used as the performance metric. Similarly Figure 5 indicates the performance of the prediction model trained on spectral data of 2011, to predict for 2001.

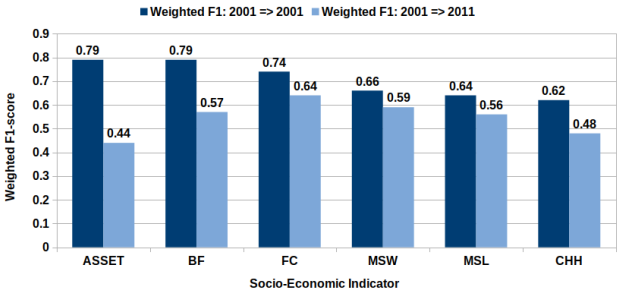

Figure 4: Weighted F1-scores of the models trained on 2001 to predict labels for 2011

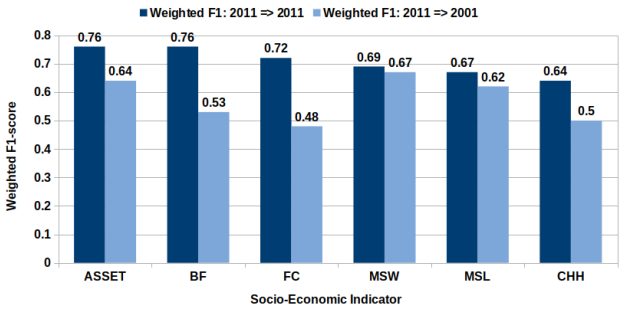

Figure 5: Weighted F1-scores of the models trained on 2011 to predict labels for 2001

We observe that the scores consistently dip for all the indicators, indicating that our model does not translate in a straightforward manner across ten years. Some of the possible explanations can be:

- 593 district samples are too small a dataset to train a robust model that is scalable across time. A small collection of images can capture only a limited set of tonal distributions and hence the model potentially under-performs when tested over a gap of ten years, during which districts could have dramatically changed in their profile.
- Data quality of a satellite can degrade over the years, and in fact a failure of the Scan Line Corrector in Landsat7 since the year 2003 has been documented to cause some data gaps. The median-based filtering is believed to address this issue to a significant extent [41].
- Over-sensitivity of the models to values of various hyper-parameters can also affect its transferability.

In light of the poor transferability evidence, the next section proposes a method to mitigate some of the above mentioned shortcomings to make our model perform better over different years.

6 IMPROVING TEMPORAL TRANSFERABILITY

Figure 9 shows the districts in darker shades of red on which the two-step classification works incorrectly when going from the base year of 2001 to the target year of 2011. We observe that only **1.34%** of the districts have three or more indicators predicted incorrectly, **29.67%** have two indicators predicted incorrectly, and **63.4%** of the d A small training dataset and potential errors in the satellite data can act as barriers to the temporal transferability of prediction models. To overcome some of these issues, we propose an error correction/smoothing method to improve the model transferability over different years. This method applies the model to predict labels for several intermediate years between the base year (whose ground truth is available) and the target year (whose labels are finally to be predicted), and uses these as features for a second classifier to make the final prediction. We call this second classifier the *forward classifier*. This improved two-step classification is further complemented by including some census variables of the base year as features in the forward classifier, which are likely to affect the socio-economic

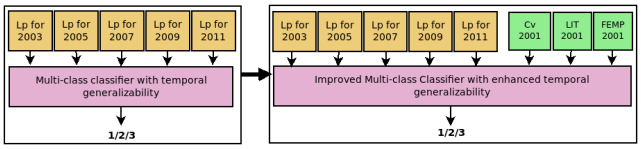

Figure 6: Forward classifier: Basic (left) and improved (right) feature extraction. Here, L_p represents the predicted label of an indicator and C_v stands for the true label of the variable from the census

change that would be taking place in a district. We next describe this methodology in more detail.

6.1 Feature extraction

Since Landsat7 data is annually available since 1999, we use the 2001 classification model to predict the indicator labels for the alternate intermediate years between 2001 and 2011. Even if not highly accurate for a specific intermediate year, our hope is that using all these predicted labels together can act as an error correction or smoothing mechanism to predict the eventual label for the target year. Such a method may be able to handle sporadic noise due to cloud-cover or other artifacts of satellite-based errors that might hamper the prediction for a specific year but could get neutralized when given values for several years. Labels are predicted for every two years between 2001 and 2011 (2003, 2005, 2007, 2009, 2011), and used as an input feature vector to train the forward classifier, as explained in the left sub-part of Figure 6.

Goswami et al. [19] discovered several relationships between various census variables. In particular they found that variables related to discretionary spending by households, like assets and bathroom facilities, were related with the level of literacy (LIT) and formal employment (FEMP) in a district. Districts with higher literacy and higher formal employment saw a more rapid change in most of the discretionary variables over the years. They also found that districts at intermediate levels of development improved faster than districts at lower levels of development. We therefore use this domain knowledge to include additional variables as features for change classification, and evaluate its performance with variables from 2001 for the literacy rate, formal employment, and the current status of an indicator. The right sub-part of Figure 6 shows this enhanced feature vector.

6.2 Classification model

We use an XGBoost classifier as before to evaluate the two models for forward classification, using only the predicted labels for the intermediate years, and also including features for the current status of the indicator, the literacy rate and the formal employment in the base year. The hyper-parameters of the model are set using the RandomSearch method followed by GridSearch in python's scikit-learn module, and the class imbalance of the training data is handled using SMOTE. The performance is evaluated using 5-fold cross-validation.

6.3 Results and analysis

Weighted F1-scores of the improved two-step classification method to predict the district levels for 2011 are shown in Figure 7. We

witness a significant increase in the performance of the improved two-step method over the original direct application of the 2001 model on 2011 data. There in an average increase of **22.8%** in the performance between these methods. We also test the backward compatibility of the improved two-step classification method by keeping the year 2011 as the base year and the year 2001 as the target year to build a backward classifier. These results are shown in Figure 8, and a similar pattern of improved performance is observed in this case as well with a **21.5%** rise in the weighted F1-scores. The increase in performance by including census variables highlights the importance of domain knowledge in machine learning tasks.

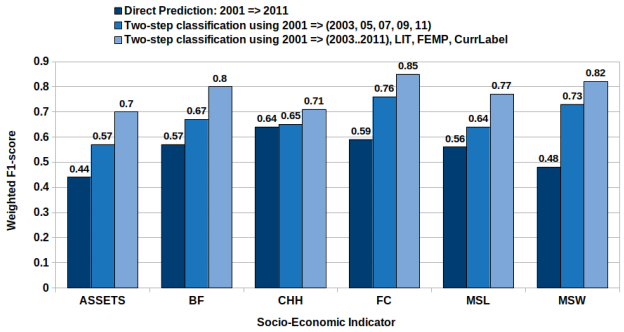

Figure 7: Forward classification: Performance of the improved model trained on the year 2001 to predict for 2011. Weighted F1-scores are used as the performance metric. Features (2003..2011) indicate the predicted labels for intermediate years, LIT and EMP denotes literacy and formal employment respectively, and CurrLabel denotes the status of an indicator in 2001

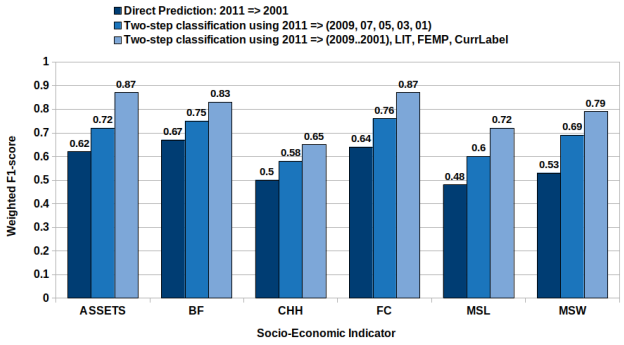

Figure 8: Backward classification: Performance of the improved model trained on the year 2011 to predict for 2001. Weighted F1-scores are used as the performance metric. Features (2009..2001) indicate the predicted labels for intermediate years, LIT and FEMP denotes literacy and formal employment respectively, and CurrLabel denotes the status of an indicator in 2001

Figure 9 shows the districts in darker shades of red on which the two-step classification works incorrectly when going from the base year of 2001 to the target year of 2011. We observe that only **1.34%** of the districts have three or more indicators predicted incorrectly, **29.67%** have two indicators predicted incorrectly, and **63.4%** of the districts are correctly classified for each and every indicator for 2011. This encourages us to further build an aggregate assessment of the development of a district as a single index which is simply the sum of that district labels over all the indicators.

7 MONITORING AGGREGATE DISTRICT DEVELOPMENT OVER TIME

The HDI (Human Development Index) is a method to build an aggregate index for development by giving equal weightage to indicators for economic development (per capita GDP), education (literacy rate), and health (life expectancy) [35]. We similarly build an **aggregate development index (ADI)** as the sum of the levels of all the indicators for a district. The value of this index ranges from 6 (all six indicators at the lowest level 1) to 18 (all six indicators at the highest level 3) for every district. Having observed a good performance of our improved two-step classification, we now try to predict the ADI for every district. On comparing the values of ADI for 2011 predicted by the forward classifier with the actual values computed from the census data, we get a normalized RMSE (root mean square error) value of 0.0413 across the districts. Similarly, a normalized RMSE value of 0.0352 is achieved using the predictions from the backward classifier for 2001.

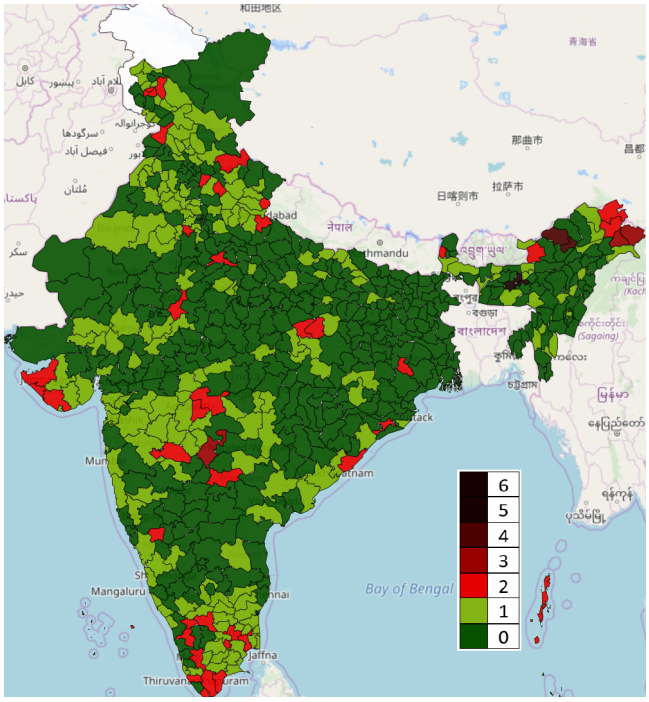

Figure 9: Count of mis-classified indicators in 2011: Districts with fewer indicators predicted correctly are shown in darker shades of red.

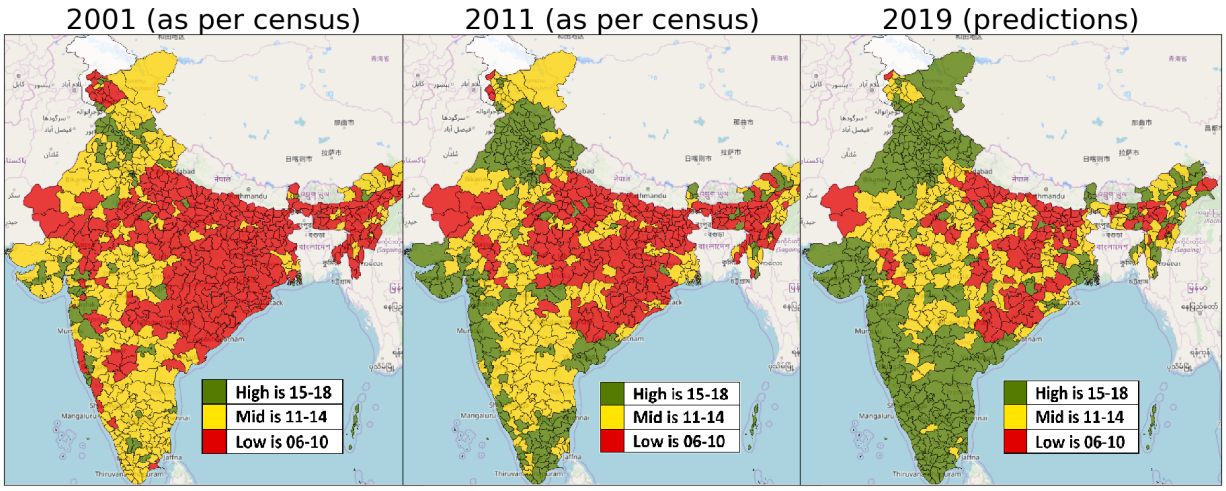

Figure 10: Aggregate district development in 2001 (as per census), 2011 (as per census), and 2019 (as per our predictions made from satellite data)

7.1 Visualizing district development in India

Given the low RMSE values for the aggregate development index, we apply the same method to predict the index for the current year of 2019. We learn a new model based on the ground-truth data for 2011, and we then use it to predict labels for the intermediate years of 2013..2019. These labels are fed into the forward change classifier (trained on the predicted labels between 2001..2011) to make the final predictions of various indicators for 2019, which are then aggregated to build the index. Note that we use 2011 as the year to learn the cross-sectional model, but the change classifier needs an input over ten years from 2009 to 2019. We therefore need to make an assumption that the forward change classifier is robust to the choice of year used for the cross-sectional model. We believe that this is reasonable given that most of the indicators we are studying are typically slow-moving indicators that may not have changed substantially between two years.

Figure 10 visualizes the aggregate development of districts over the period of almost two decades, for the years 2001, 2011, and 2019. Districts with an aggregate development index between 6 to 10 are coloured red, between 11 to 14 are coloured yellow, and more than 14 are coloured green. Between 2001 and 2011, we observe that states from the eastern part of India (such as Orissa, Bihar, Jharkhand, and West Bengal), large parts of north and central India (Uttar Pradesh and Madhya Pradesh), and the north-eastern districts, show little change in development. These states indeed have been the poorest states of the country. On the other hand, states like Gujarat, Maharashtra, Tamil Nadu, and Andhra Pradesh, saw many districts improve substantially during this time. These observations tally with potential explanatory factors such as the degree of industrialization in these states: Industrialized districts are known to see more rapid growth as compared to non-industrialized and predominantly agricultural districts [19].

Between 2011 to 2019, based on the predicted values for 2019, there is an indication of more widespread growth in some of the poorest states like West Bengal and Madhya Pradesh. However, states like Jharkhand, Bihar, and Orissa, and large parts of Uttar Pradesh, have not progressed substantially even now. These findings seem to tally with some of the latest data from the Niti Aayog based on state level surveys [2]. These observations illustrate the kind of applications that can be developed based on the use of satellite images to predict socio-economic indicators at the district-level.

8 DISCUSSION AND CONCLUSIONS

We presented an analysis of the potential to use satellite data for the prediction of socio-economic indicators over time, at the spatial scales of districts. We found that while our simple classification model performed robustly in a cross-sectional analysis, the model was unable to satisfactorily predict indicators for a different year than what was used for its training. This problem in transferability could arise either because of the small size of the training dataset that could capture limited variations, or could possibly be due to some year-specific effects such as clouds, rainfall, or satellite-based artifacts like changes in sensor calibrations in specific years. It is probably for this reason that the use of data points for multiple consecutive years is able to perform better in making predictions over time. This method is generic and can be applied to improve the temporal transferability of other kinds of prediction models as well. We are also able to achieve a good accuracy in predicting over ten years an aggregate development index calculated as the sum of values of multiple socio-economic indicators. This application can be useful to identify anomalous districts that should be investigated further, such as outliers that progressed rapidly or did not progress at all, over many years. As part of future work, we plan to analyze mass media and other datasets about these outlier districts in an attempt to explain their behavior. We also plan to build more sophisticated machine learning models and try to make predictions at village-level over time.

REFERENCES

[1] 2019. Census of India Website: Office of the Registrar General and Census Commissioner, India. http://censusindia.gov.in/

[2] NITI Aayog. 2018. SDG India Index Baseline Report.

[3] Saikat Basu, Sangram Ganguly, Supratik Mukhopadhyay, Robert DiBiano, Manohar Karki, and Ramakrishna Nemani. 2015. Deepsat: a learning framework for satellite imagery. In *Proceedings of the 23rd SIGSPATIAL international conference on advances in geographic information systems*. ACM, 37.

[4] Frank Bickenbach, Eckhardt Bode, Peter Nunnenkamp, and Mareike Söder. 2016. Night lights and regional GDP. *Review of World Economics* 152, 2 (2016), 425–447.

[5] 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 of environment* 113, 5 (2009), 893–903.

[6] C Chandramouli and Registrar General. 2011. Census of India 2011. *Provisional Population Totals. New Delhi: Government of India* (2011).

[7] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research* 16 (2002), 321–357.

[8] Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. ACM, 785–794.

[9] Xi Chen and William D Nordhaus. 2010. *The value of luminosity data as a proxy for economic statistics*. Technical Report. National Bureau of Economic Research.

[10] 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.

[11] Dave Donaldson and Adam Storeygard. 2016. The view from above: Applications of satellite data in economics. *Journal of Economic Perspectives* 30, 4 (2016), 171–98.

[12] Eugenie Dugoua, Ryan Kennedy, and Johannes Urpelainen. 2018. Satellite data for the social sciences: measuring rural electrification with night-time lights. *International journal of remote sensing* 39, 9 (2018), 2690–2701.

[13] Christopher D Elvidge, Kimberly E Baugh, Sharolyn J Anderson, Paul C Sutton, and Tilottama 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.

[14] Swetava Ganguli, Jared Dunnmon, and Darren Hau. 2016. Predicting food security outcomes using cnns for satellite tasking.

[15] Bardan Ghimire, John Rogan, Víctor Rodríguez Galiano, Prajjwal Panday, and Neeti Neeti. 2012. An evaluation of bagging, boosting, and random forests for land-cover classification in Cape Cod, Massachusetts, USA. *GIScience & Remote Sensing* 49, 5 (2012), 623–643.

[16] Giorgia Giovannetti, Elena Perra, et al. 2019. *Syria in the Dark: Estimating the Economic Consequences of the Civil War through Satellite-Derived Night Time Lights*. Technical Report. Universita'degli Studi di Firenze, Dipartimento di Scienze per l'Economia e âĂę.

[17] Ran Goldblatt, Alexis Rivera Ballesteros, and Jennifer 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.

[18] Google. 2019. Landsat Algorithms. https://developers.google.com/earth-engine/landsat

[19] 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).

[20] Andrew Head, Mélanie Manguin, Nhat Tran, and Joshua E Blumenstock. 2017. Can Human Development be Measured with Satellite Imagery?. In *ICTD*. 8–1.

[21] 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).

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

[23] Tengyun Hu, Jun Yang, Xuecao Li, and Peng Gong. 2016. Mapping urban land use by using landsat images and open social data. *Remote Sensing* 8, 2 (2016), 151.

[24] 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).

[25] 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.

[26] David A Landgrebe. 2005. *Signal theory methods in multispectral remote sensing*. Vol. 29. John Wiley & Sons.

[27] Charlotta Mellander, José Lobo, Kevin Stolarick, and Zara Matheson. 2015. Night-time light data: A good proxy measure for economic activity? *PloS one* 10, 10 (2015), e0139779.

[28] Brian Min and Kwawu Gaba. 2014. Tracking electrification in Vietnam using nighttime lights. *Remote Sensing* 6, 10 (2014), 9511–9529.

[29] Brian Min, Kwawu Mensan Gaba, Ousmane Fall Sarr, and Alassane Agalassou. 2013. Detection of rural electrification in Africa using DMSP-OLS night lights imagery. *International journal of remote sensing* 34, 22 (2013), 8118–8141.

[30] KN Nischal, Radhika Radhakrishnan, Sanket Mehta, and Sumit Chandani. 2015. Correlating night-time satellite images with poverty and other census data of India and estimating future trends. In *Proceedings of the Second ACM IKDD Conference on Data Sciences*. ACM, 75–79.

[31] Shailesh M Pandey, Tushar Agarwal, and Narayanan C Krishnan. 2018. Multi-task deep learning for predicting poverty from satellite images. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[32] Anthony Perez, Swetava Ganguli, Stefano Ermon, George Azzari, Marshall Burke, and David Lobell. 2019. Semi-supervised multitask learning on multispectral satellite images using wasserstein generative adversarial networks (gans) for predicting poverty. *arXiv preprint arXiv:1902.11110* (2019).

[33] Anthony Perez, Christopher Yeh, George Azzari, Marshall Burke, David Lobell, and Stefano Ermon. 2017. Poverty prediction with public landsat 7 satellite imagery and machine learning. *arXiv preprint arXiv:1711.03654* (2017).

[34] Anupam Prakash, Avdhesh Kumar Shukla, Chaitali Bhowmick, and Robert Carl Michael Beyer. 2019. Night-time Luminosity: Does it Brighten Understanding of Economic Activity in India? (2019).

[35] United Nations Development Programme. 2019. Human Development Reports. http://hdr.undp.org/en/content/human-development-index-hdi

[36] Caleb Robinson, Fred Hohman, and Bistra Dilkina. 2017. A deep learning approach for population estimation from satellite imagery. In *Proceedings of the 1st ACM SIGSPATIAL Workshop on Geospatial Humanities*. ACM, 47–54.

[37] Hans Rosling. 2019. *Factfulness*. Flammarion.

[38] SP Subash, Rajeev Ranjan Kumar, and KS Aditya. 2018. Satellite data and machine learning tools for predicting poverty in rural India. *Agricultural Economics Research Review* 31, 347-2019-571 (2018), 231–240.

[39] Potnuru Kishen Suraj, Ankesh Gupta, Makkunda Sharma, Sourabh Bikash Paul, and Subhashis Banerjee. 2017. On monitoring development using high resolution satellite images. *arXiv preprint arXiv:1712.02282* (2017).

[40] Binh Tang, Ying Sun, Yanyan Liu, and David S Matteson. 2018. Dynamic Poverty Prediction with Vegetation Index. (2018).

[41] USGS. 2019. Landsat Missions. https://www.usgs.gov/land-resources/nli/landsat/landsat-7?qt-science_support_page_related_con=0#qt-science_support_page_related_con

[42] USGS/Google. 2019. USGS Landsat 7 Collection 1 Tier 1 and Real-Time data TOA Reflectance. https://developers.google.com/earth-engine/datasets/catalog/LANDSAT_LE07_C01_T1_RT_TOA

[43] 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.

[44] Michael Xie, Neal Jean, Marshall Burke, David Lobell, and Stefano Ermon. 2016. Transfer learning from deep features for remote sensing and poverty mapping. In *Thirtieth AAAI Conference on Artificial Intelligence*.