Temporal Prediction of Socio-economic Indicators Using Satellite Imagery

Chahat Bansal, Arpit Jain, Phaneesh Barwaria, Anuj Choudhary, Anupam Singh, Ayush Gupta, and Aaditeshwar Seth

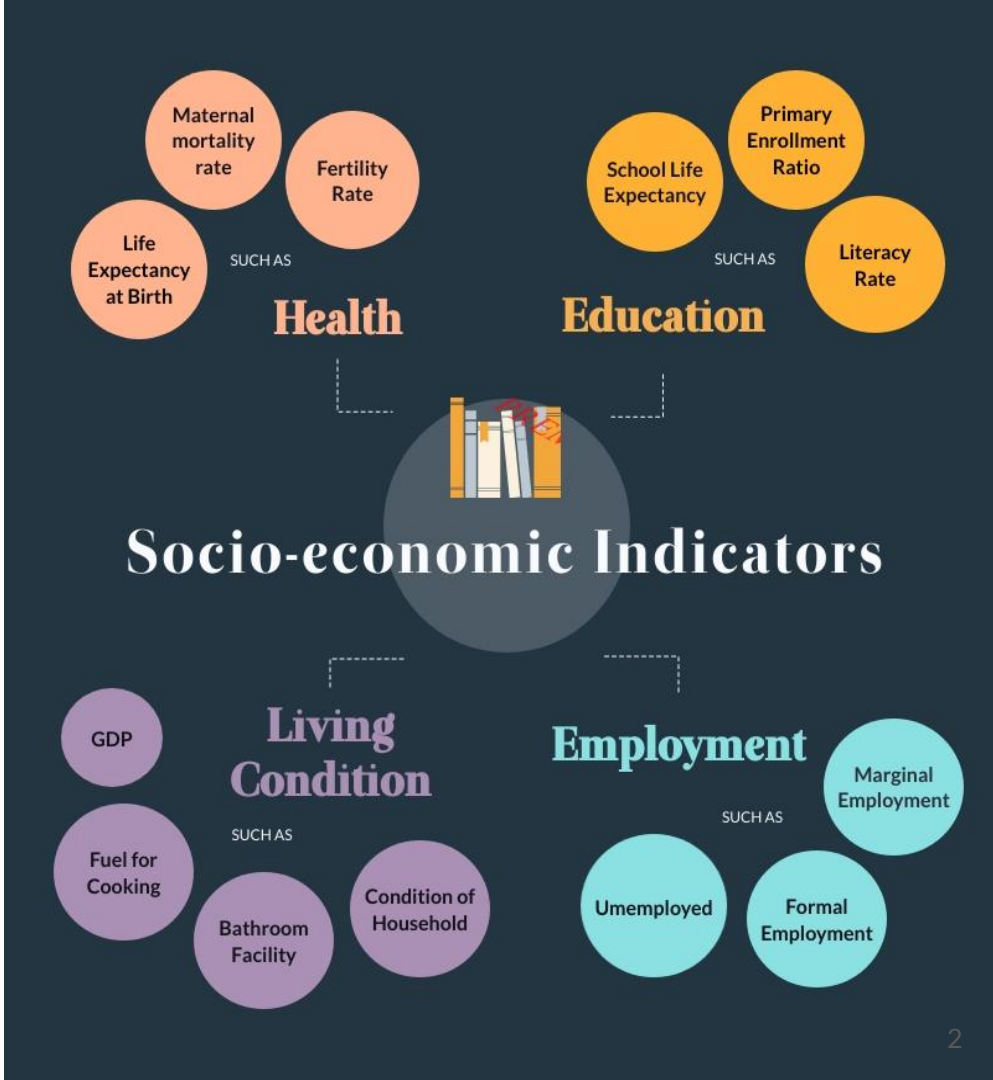
School of Information Technology, Indian Institute of Technology- Delhi

JANUARY 6, 2020

CoDS COMAD 2020

What are Socio-economic Indicators ?

Socio-economic indicators gauge the economic conditions and socio-cultural environment in which people live



The Indian Census

The Indian Census is among the most credible sources of information on the socio-cultural and demographic data since 1872

The Indian Census

The Indian Census is among the most credible sources of information on the socio-cultural and demographic data since 1872

This is one of the source of primary data at the village, town, and ward level collected once every 10 years

The Indian Census

The Indian Census is among the most credible sources of information on the socio-cultural and demographic data since 1872

This is one of the source of primary data at the village, town, and ward level collected once every 10 years

Census is the basis for reviewing the country's progress in the past decade, monitoring the ongoing schemes of the Government and most importantly, use the data to plan for the future

Limitations of Census

Limitations of Census

EXPENSIVE

Limitations of Census

EXPENSIVE

TIME CONSUMING

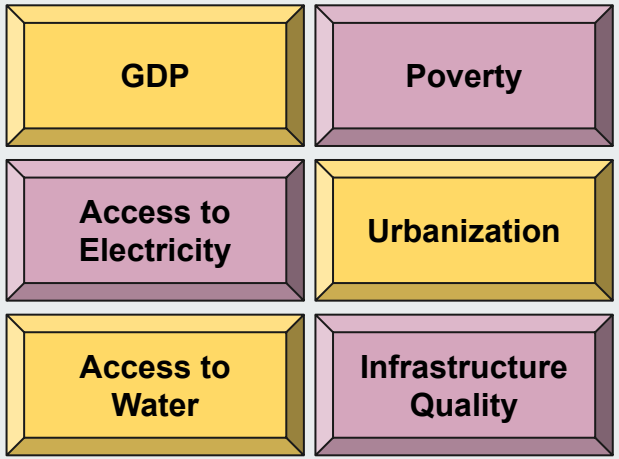
Limitations of Census

EXPENSIVE → TIME CONSUMING → INFREQUENT

SATELLITE DATA has been identified as a proxy to predict different socio-economic indicators

© BCCL 2019. ALL RIGHTS RESERVED.

Related Studies



- Some indicators like poverty, GDP, and access to electricity can be predicted using nightlights [1]
- Urbanization, access to water, and infrastructure quality can be predicted using daytime spectral imagery [2]
- Combination of nightlights and daytime imagery has been used through a transfer learning approach like by Jean et al. [3] to predict poverty

[1] J Vernon Henderson, Adam Storeygard, and David N Weil. 2012. Measuring economic growth from outer space. American economic review 102, 2 (2012), 994–1028
[2] 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
[3] 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

Related Studies

Concerns have been raised over transferability of models to different countries [4]

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

Related Studies

Concerns have been raised over transferability of models to different countries [4],

But

Temporal transferability has not been studied possibly because ground truth data has not been available for different years

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

Research Question

Do these prediction models transfer well over time?

i.e. do models learned on a particular year make accurate predictions on the satellite data from another year

Input

District Images from **Landsat-7** satellite available on a monthly basis since 1999

Input

District Images from **Landsat-7** satellite available on a monthly basis since 1999

Ground Truth

Values of socio-economic indicators from the **Indian Census 2001 and 2011**

- Asset Ownership
- Bathroom Facility (BF)
- Condition of Household (CHH)
- Fuel for Cooking (FC)
- Main Source of Light (MSL)
- Main Source of Water (MSW)

Input

District Images from **Landsat-7** satellite available on a monthly basis since 1999

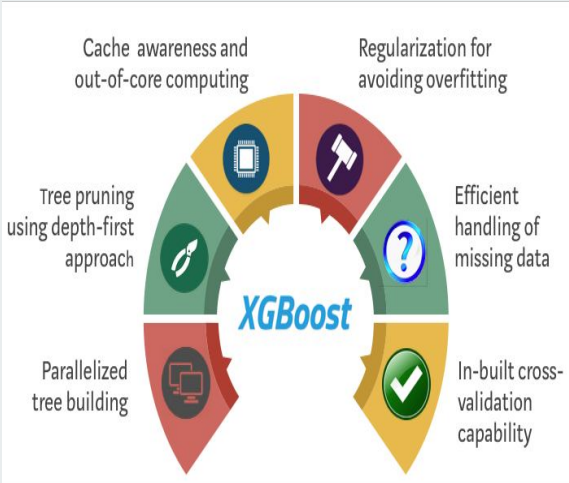
Ground Truth

Values of socio-economic indicators from the **Indian Census 2001 and 2011**

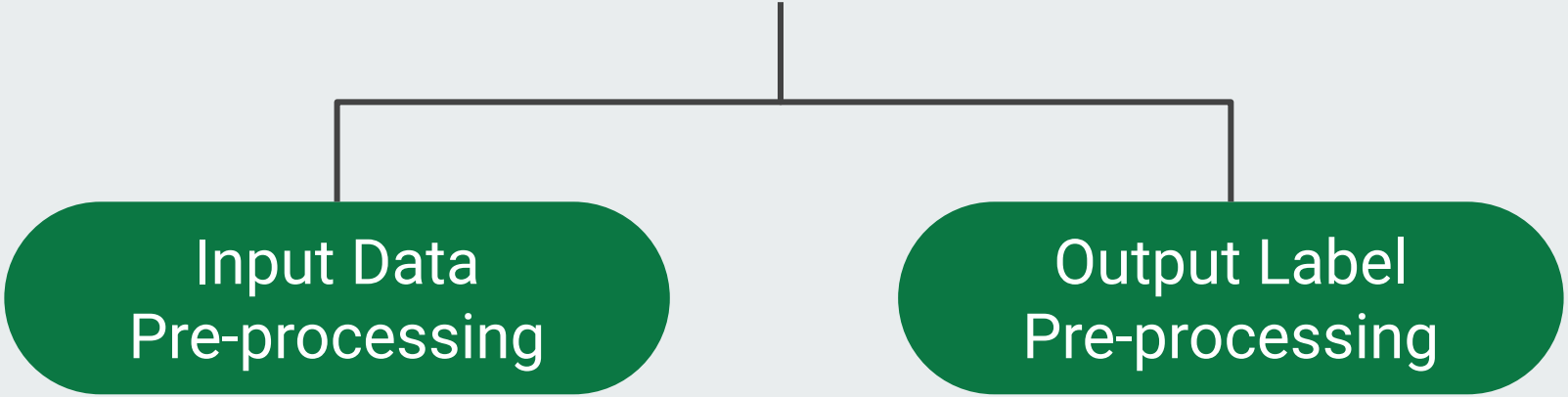
- Asset Ownership
- Bathroom Facility (BF)
- Condition of Household (CHH)
- Fuel for Cooking (FC)
- Main Source of Light (MSL)
- Main Source of Water (MSW)

ML Model

XGBoost



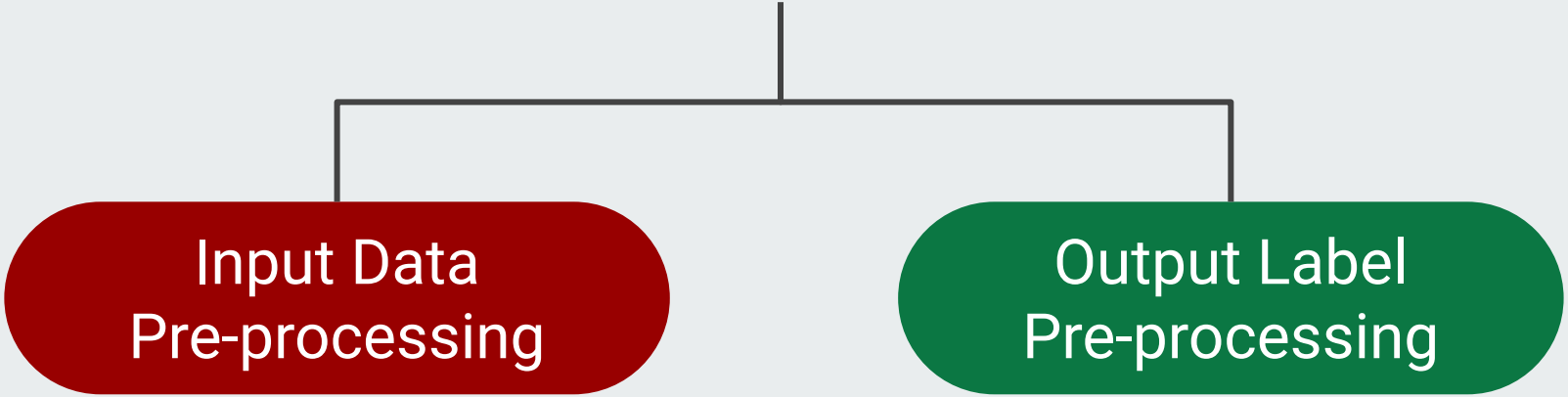
Data Pre-processing



Feature Extraction from Satellite Data

Discretization of Socio-Economic Indicators

Data Pre-processing



Feature Extraction from Satellite Data

Discretization of Socio-Economic Indicators

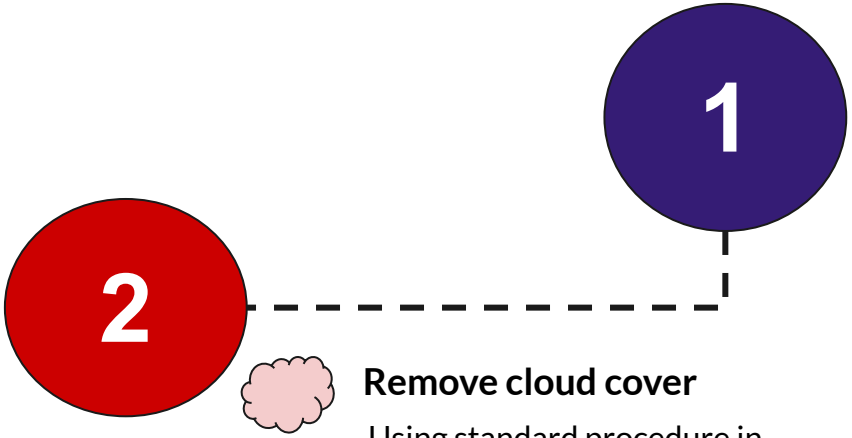
Feature Extraction

1

Download district images

Spectral data of 12 bands is downloaded from Google Earth Engine (GEE)

Feature Extraction



Download district images

Spectral data of 12 bands is downloaded from Google Earth Engine (GEE)

Remove cloud cover

Using standard procedure in GEE

Feature Extraction

1

Download district images

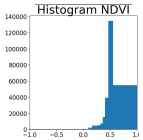
Spectral data of 12 bands is downloaded from Google Earth Engine (GEE)

2

Remove cloud cover

Using standard procedure in GEE

3



Quantile Binning

Use quantile binning to create bins for each of 12 bands. Find the frequency distribution of the band values according to the bin intervals computed for the band

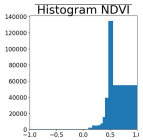
Feature Extraction

Download district images

Spectral data of 12 bands is downloaded from Google Earth Engine (GEE)

Remove cloud cover

Using standard procedure in GEE



Quantile Binning

Use quantile binning to create bins for each of 12 bands. Find the frequency distribution of the band values according to the bin intervals computed for the band

Feature Extraction

Normalize the frequency distribution and create a feature vector for each district

Agricultural Districts

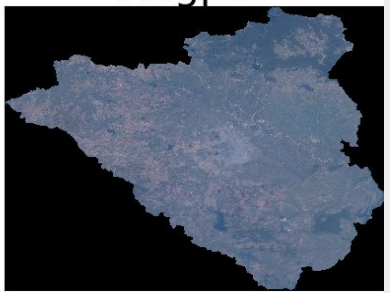
Non-agricultural Districts

Moga

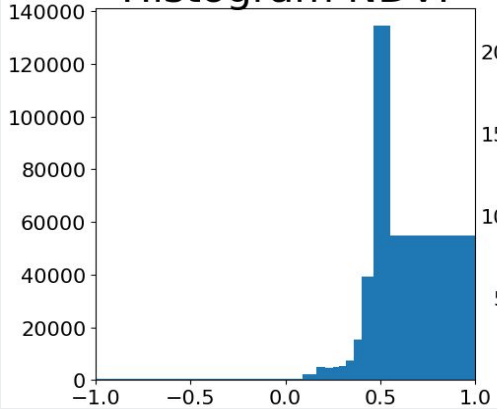
Balaghat

Jaipur

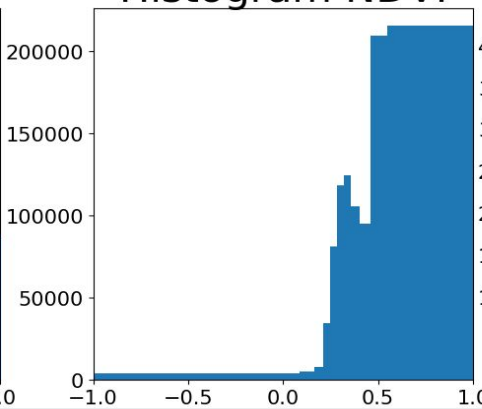
Nagpur



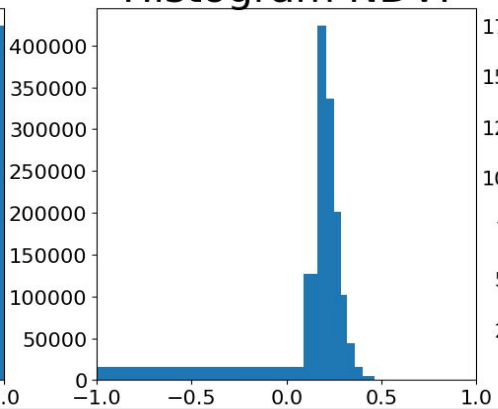
Histogram NDVI



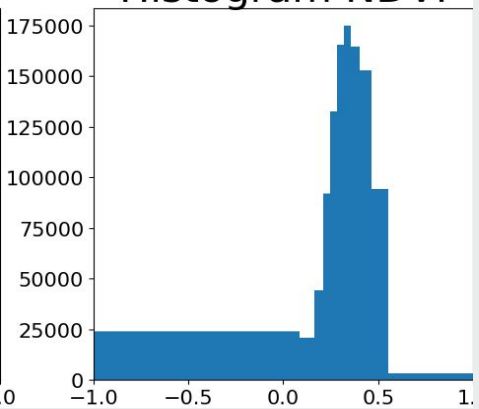
Histogram NDVI



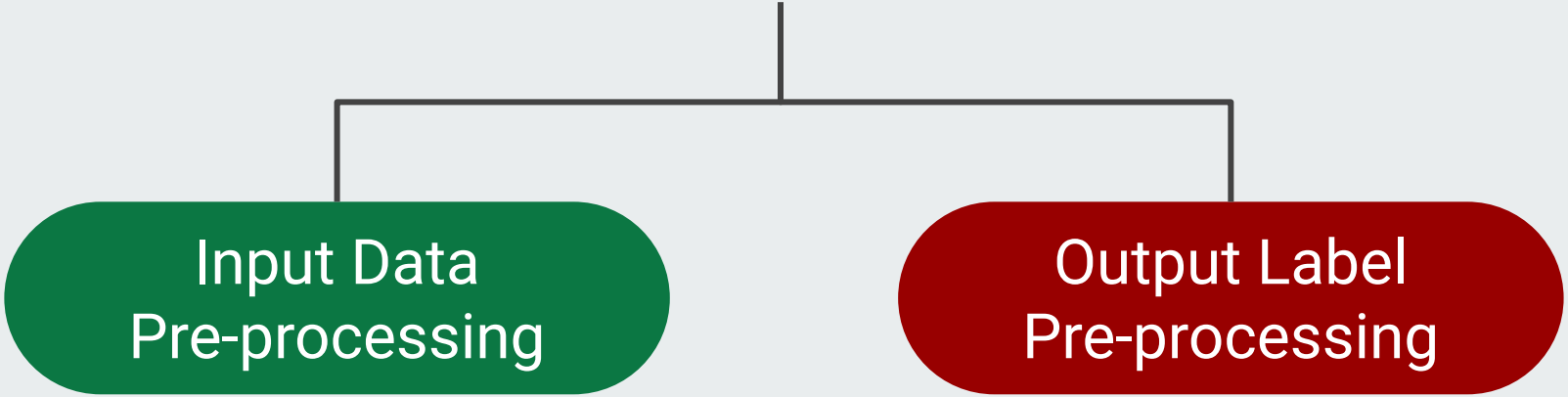
Histogram NDVI



Histogram NDVI



Data Pre-processing

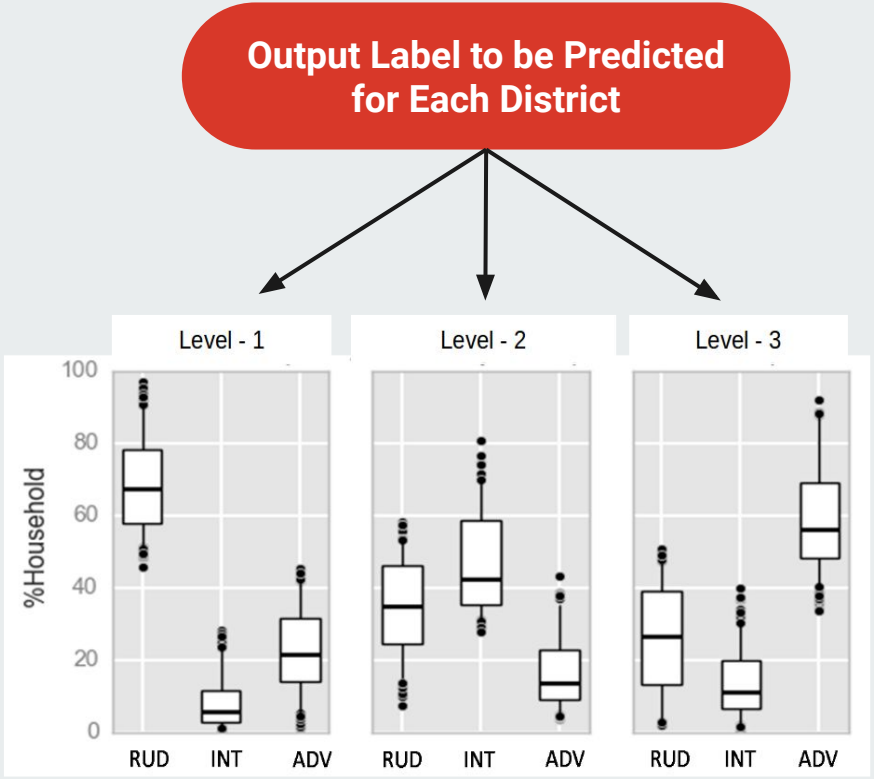


Feature Extraction from Satellite Data

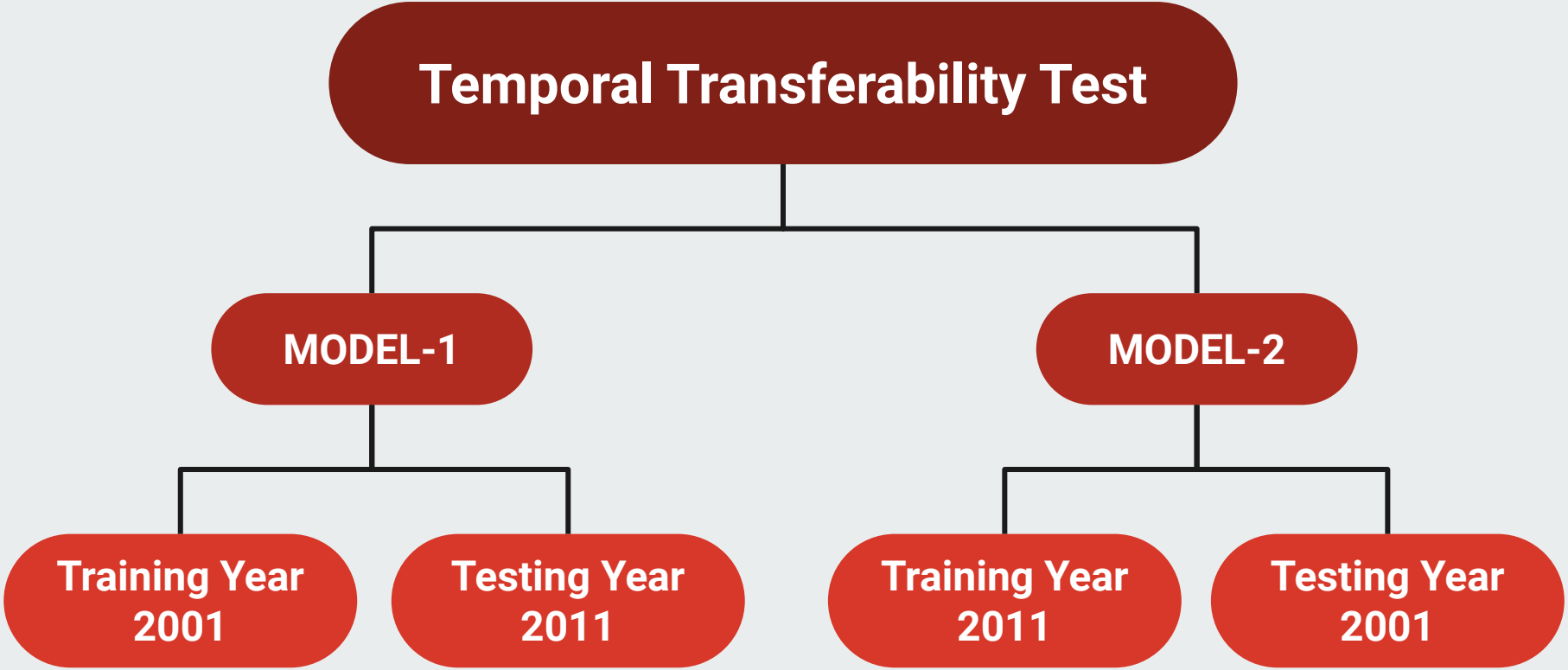
Discretization of Socio-Economic Indicators

Discretization of socio-economic indicators

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

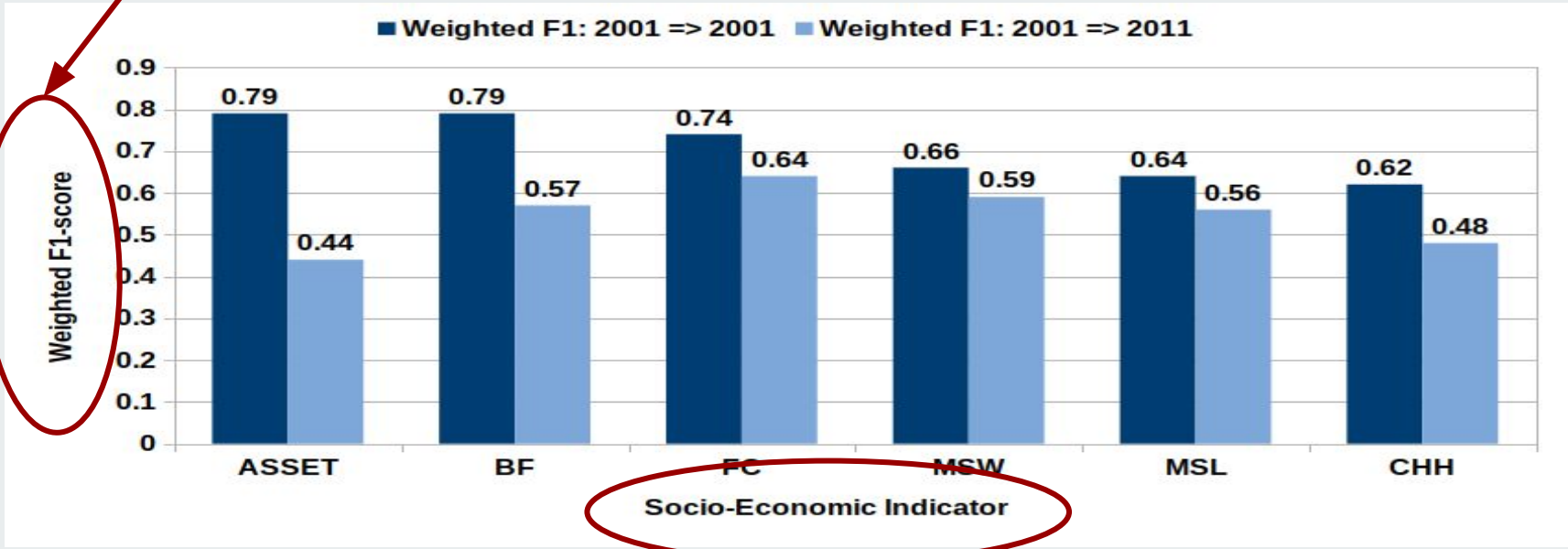


Fuel for cooking

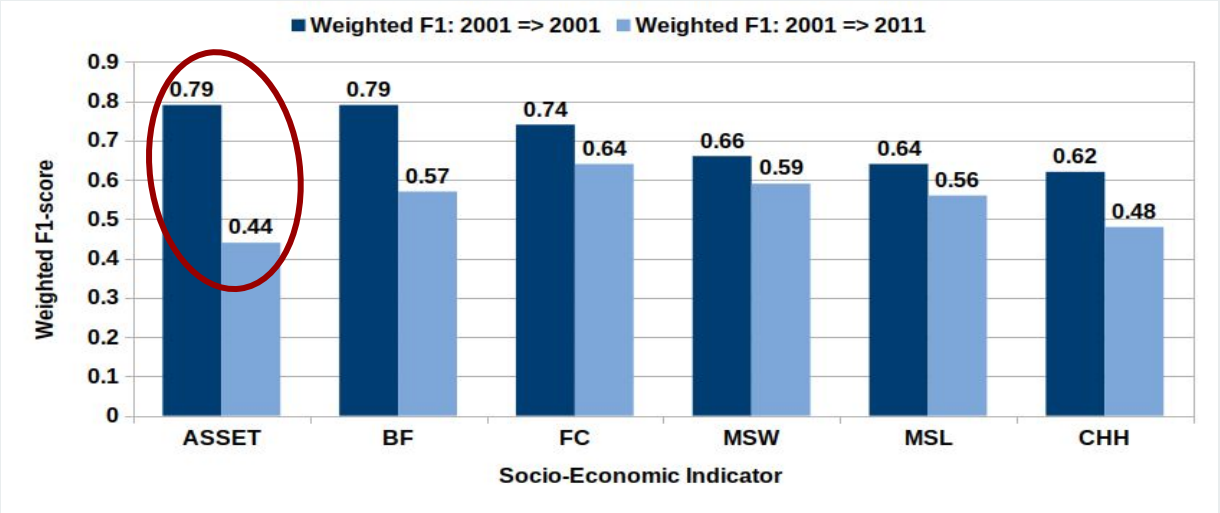

Temporal Transferability Test
MODEL-1
MODEL-2
Training Year 2001
Testing Year 2011
Training Year 2011
Testing Year 2001

Visualizing Test Results

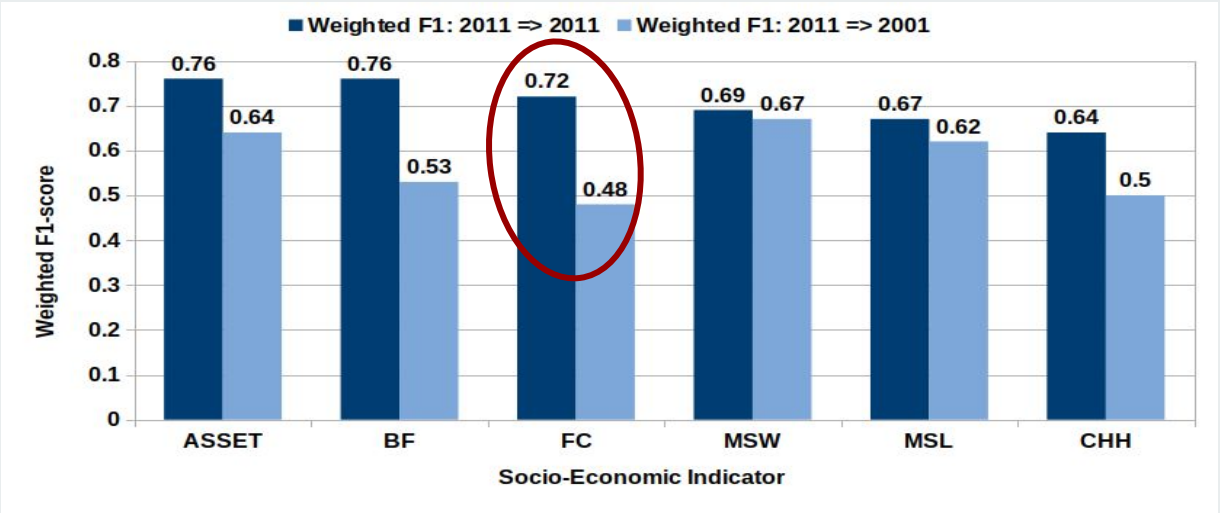
Calculated using 5-fold cross validation method



MODEL-1: Trained on year 2001



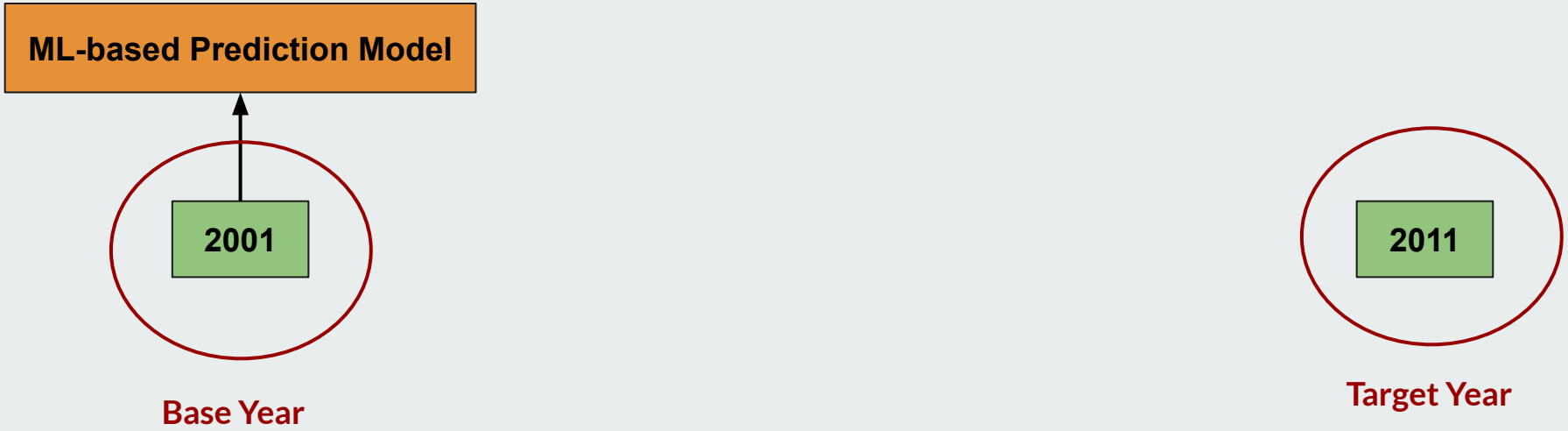
MODEL-2: Trained on year 2011



Possible Explanations of Deteriorated Performance

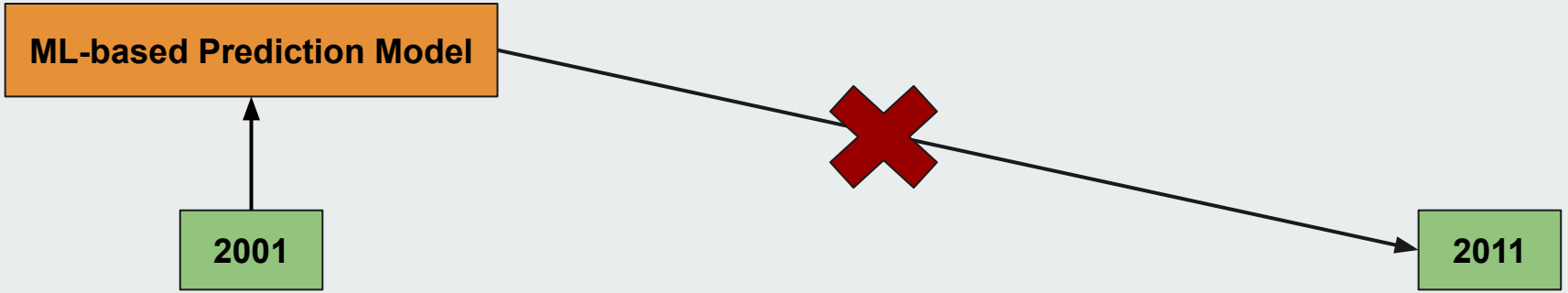
- Small training dataset of 593 districts: Low robustness of ML-model
- Degradation of satellite data over the years
- Sensitivity of model to hyper-parameters

Forward Classifier

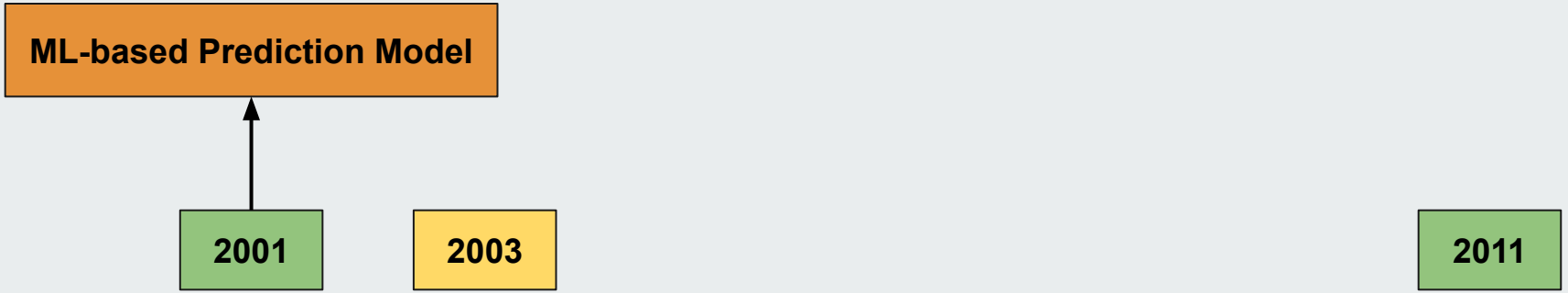


Forward Classifier

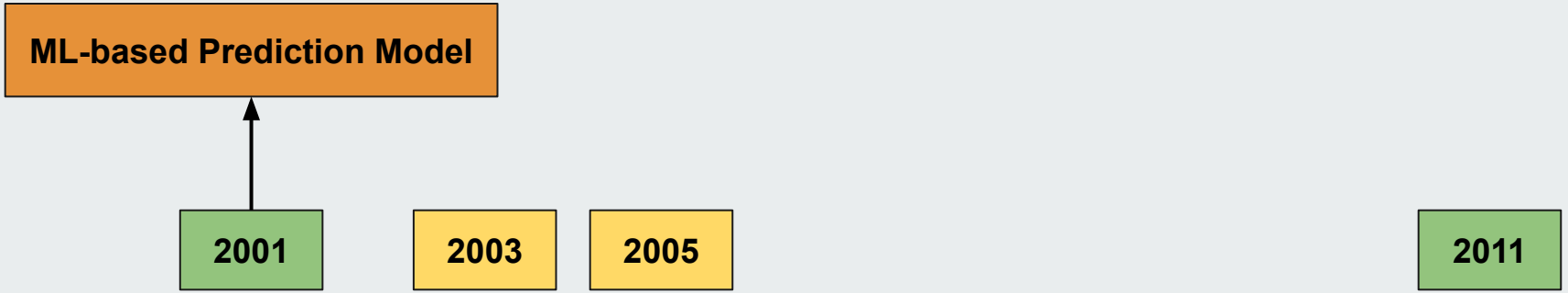
Forward Classifier



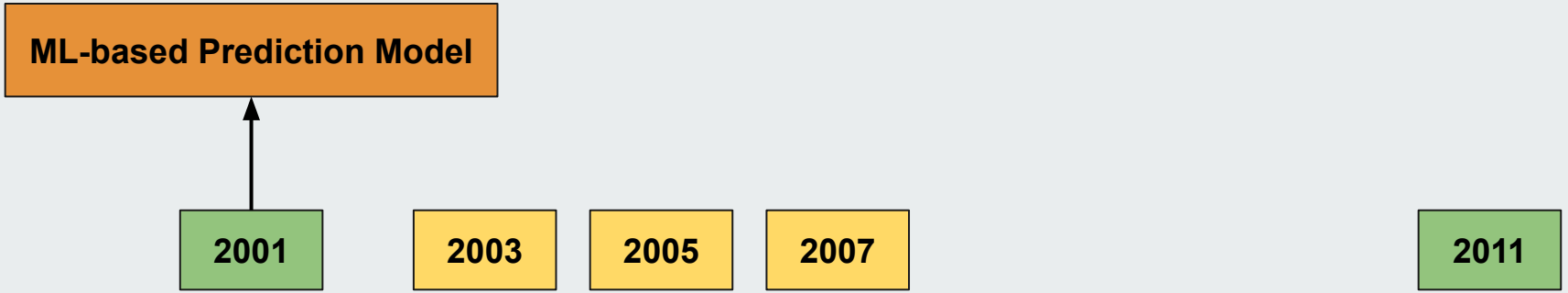
Forward Classifier



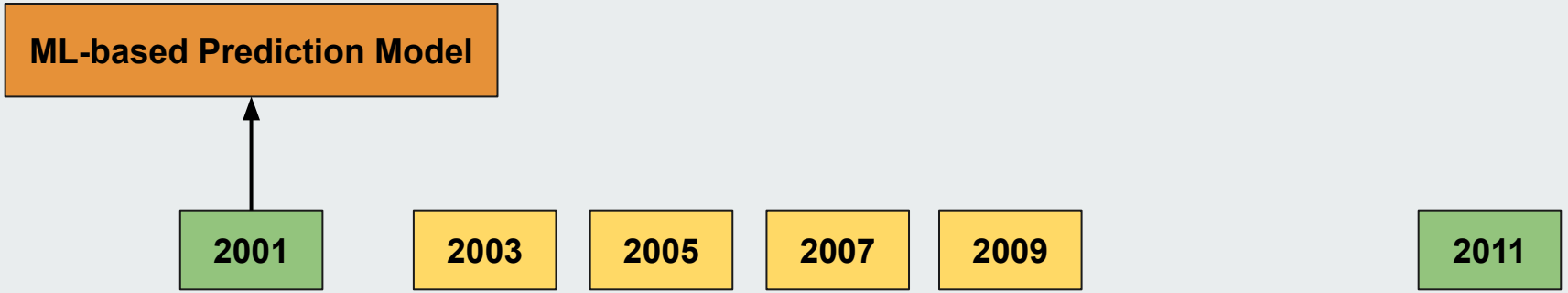
Forward Classifier



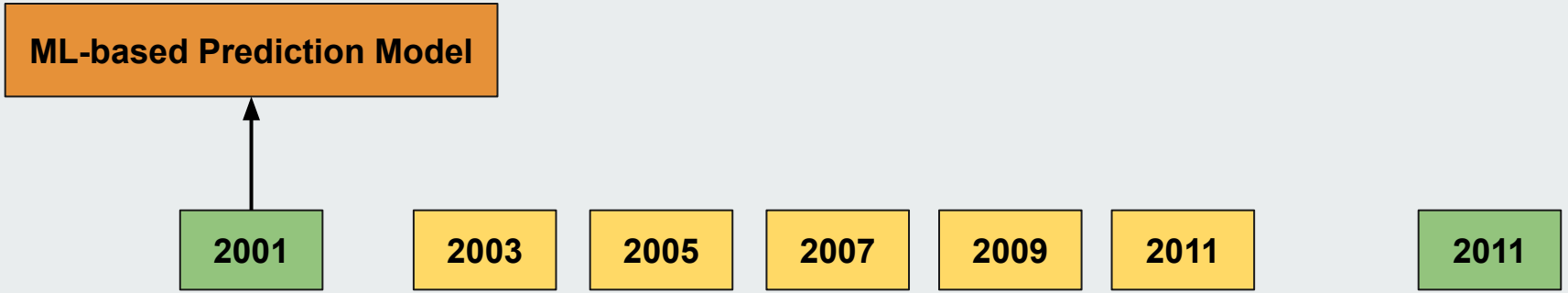
Forward Classifier



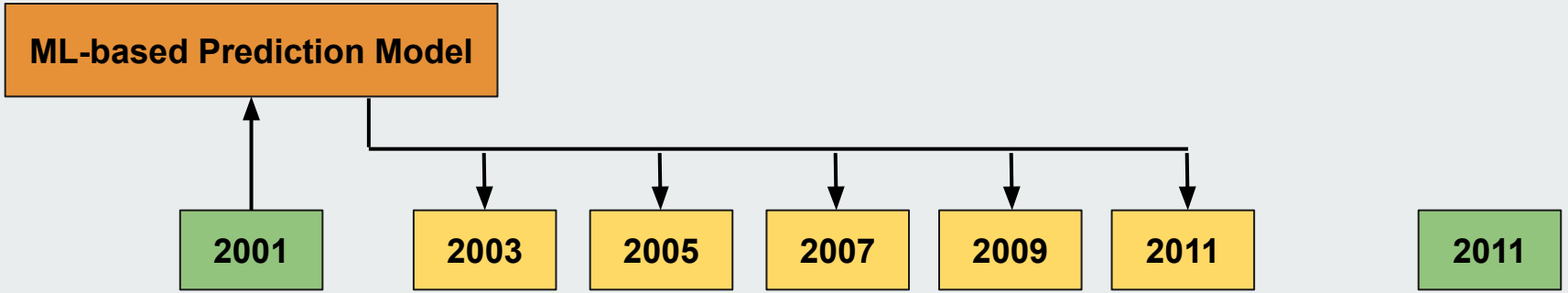
Forward Classifier



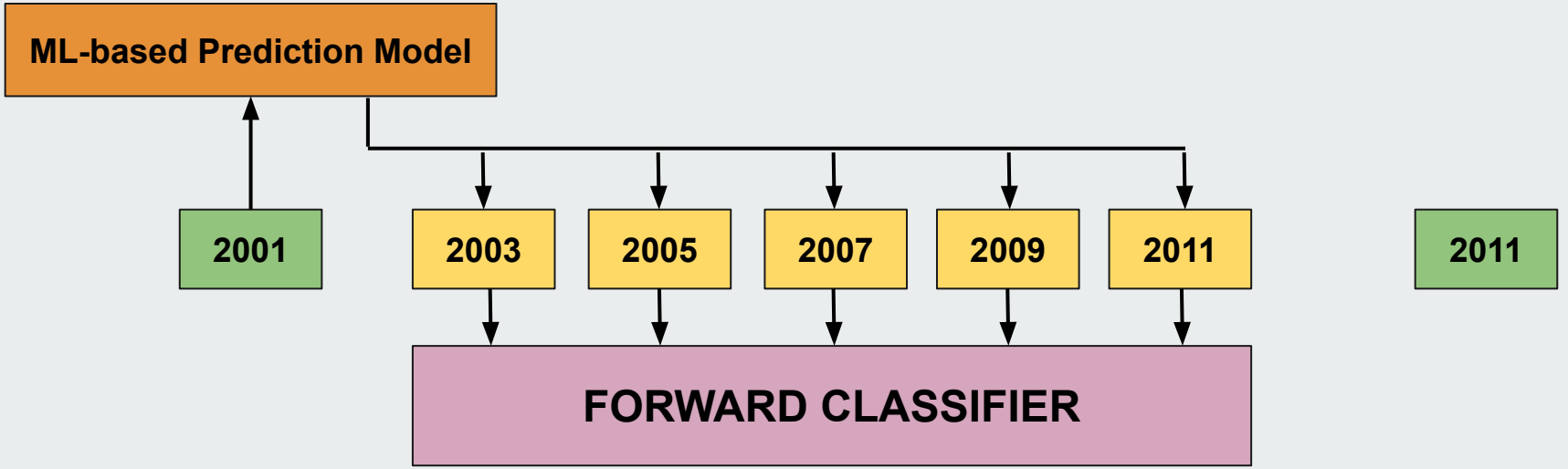
Forward Classifier



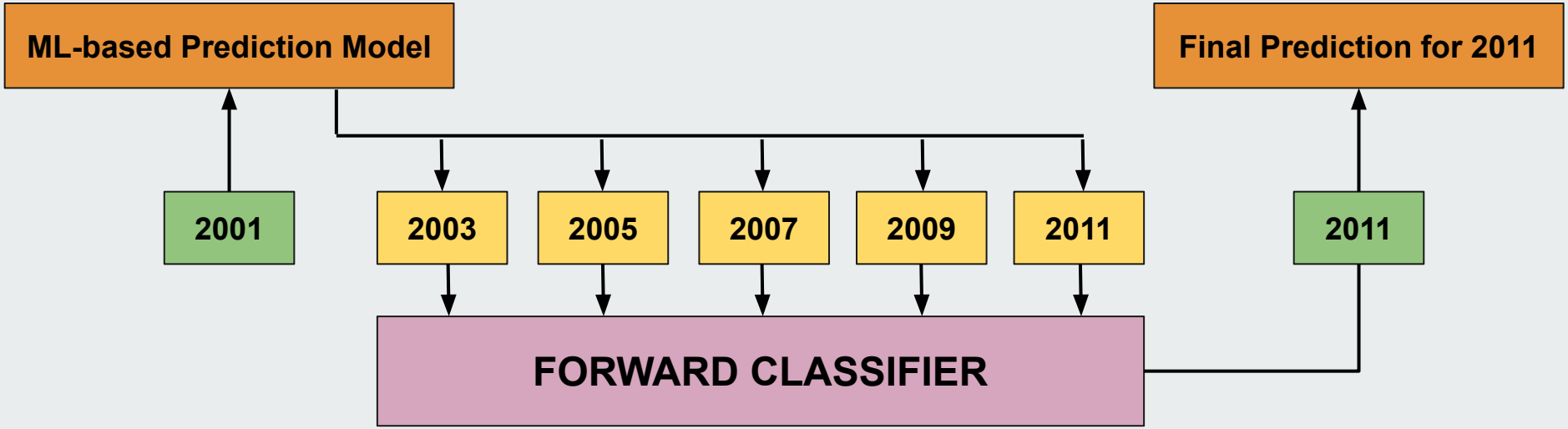
Forward Classifier



Forward Classifier



Forward Classifier



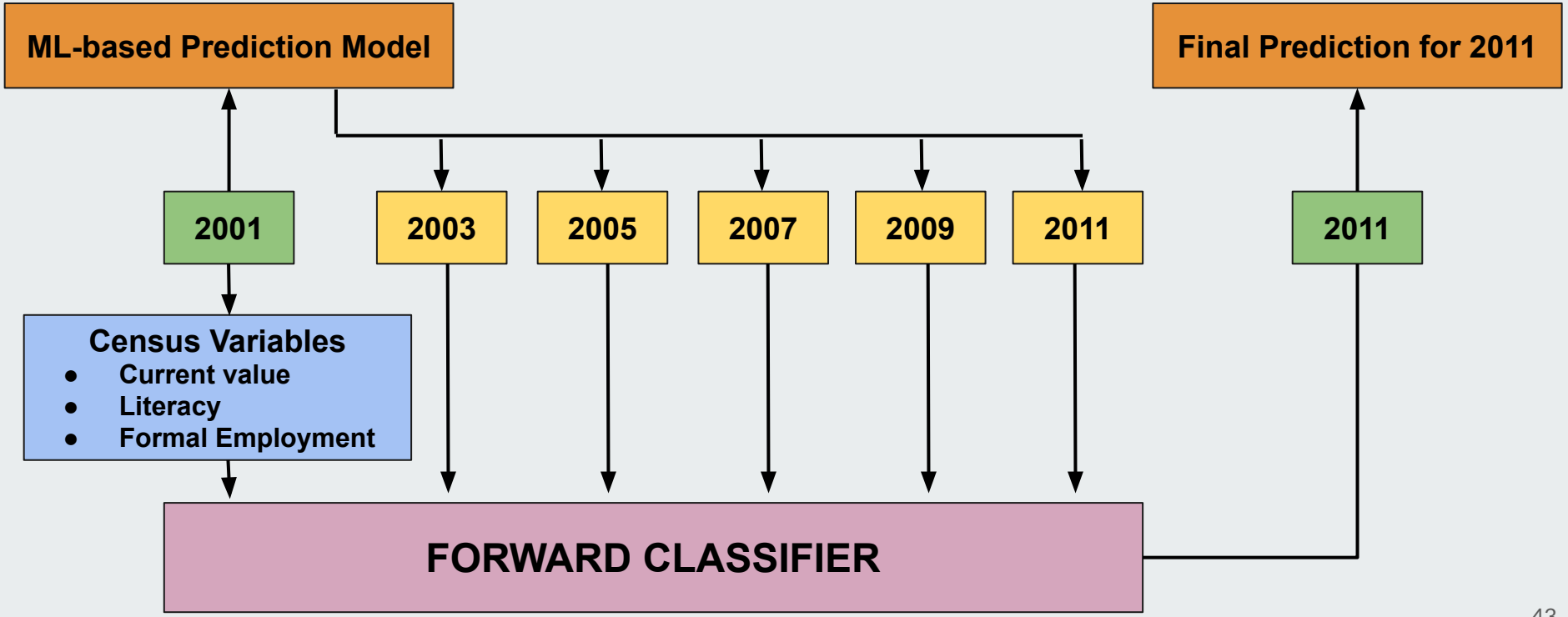
Improved Forward Classifier

It captures several relationships [5] between various census variable and uses this domain knowledge to improve the prediction of forward classifier

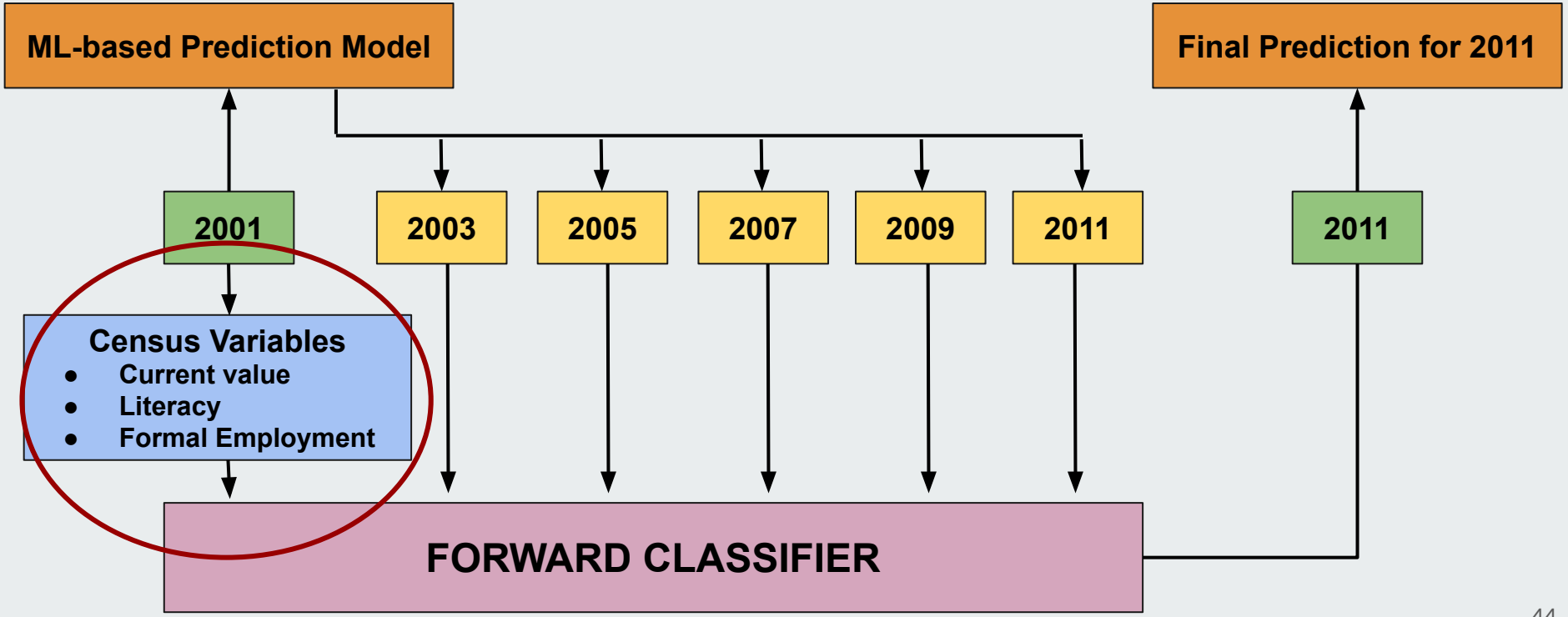
- Districts with higher literacy (LIT) and formal employment (FEMP) see a more rapid change in most of the discretionary variables over time
- Districts at intermediate levels of development improve faster than districts at lower levels of development

[5] 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)

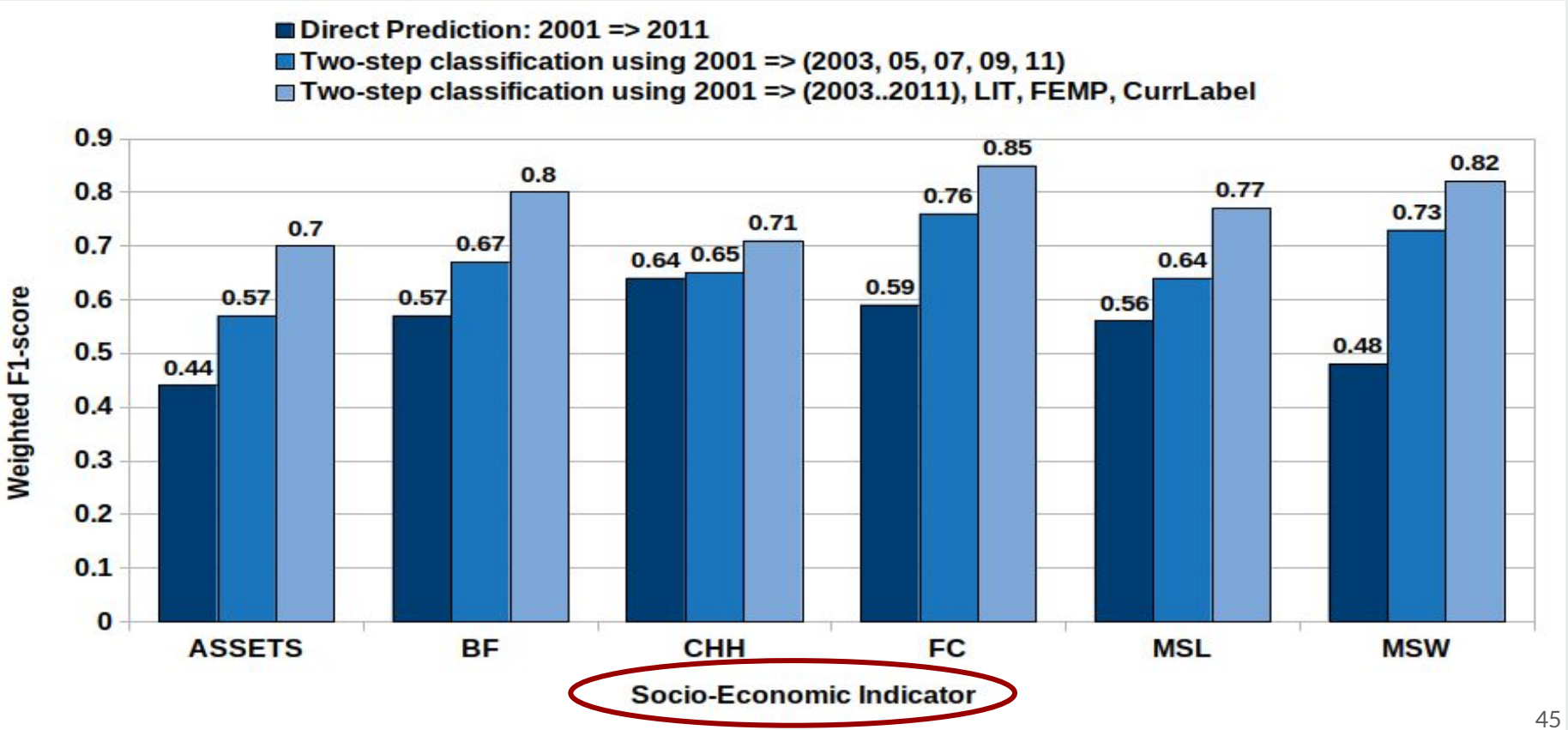
Improved Forward Classifier



Improved Forward Classifier

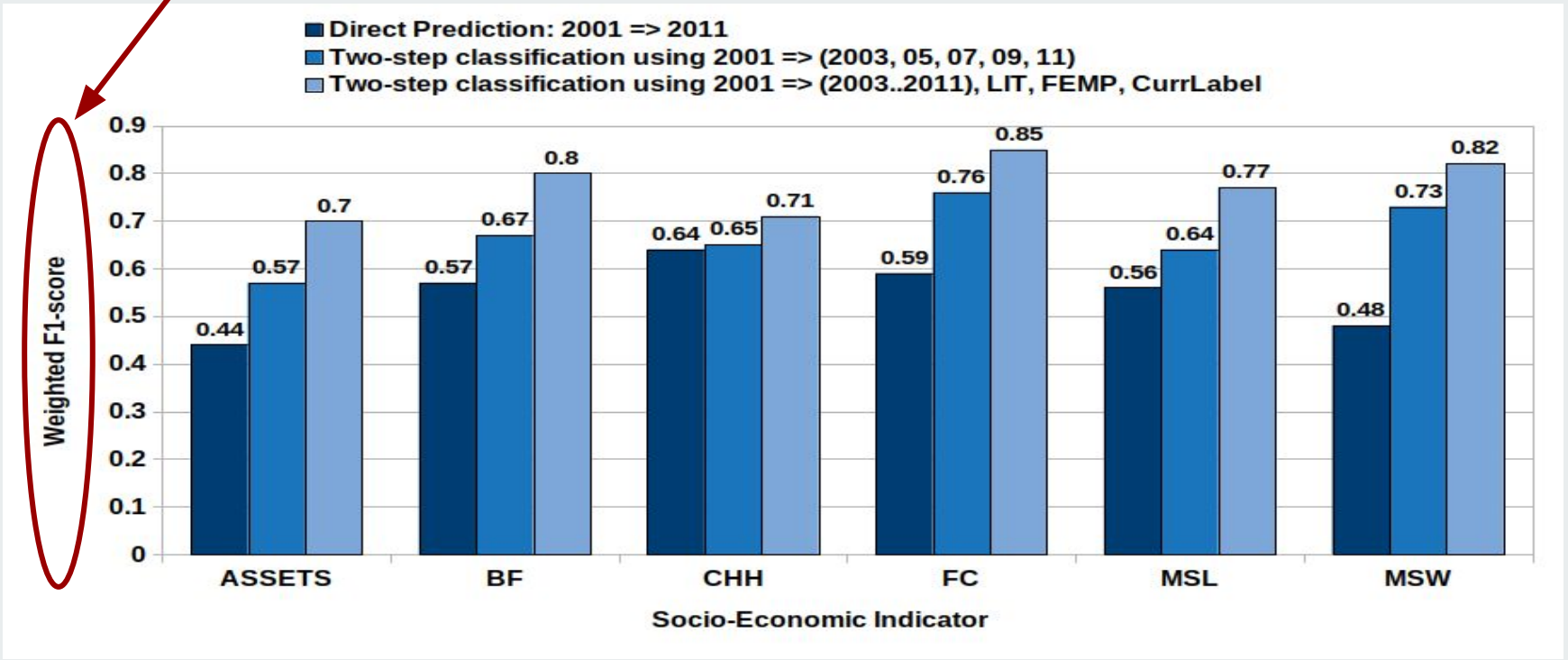


Sample Results: 22% Improvement in Accuracy

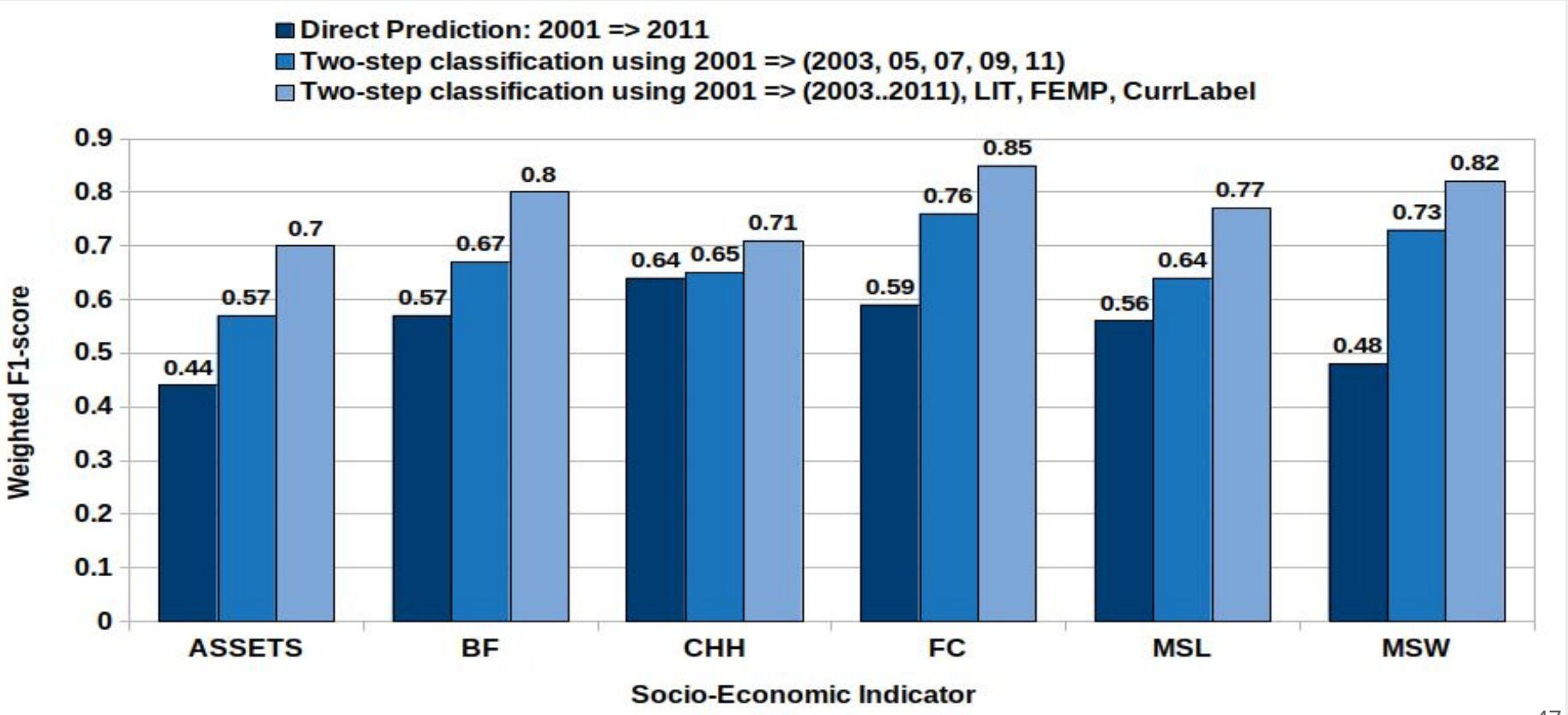


Sample Results: 22% Improvement in Accuracy

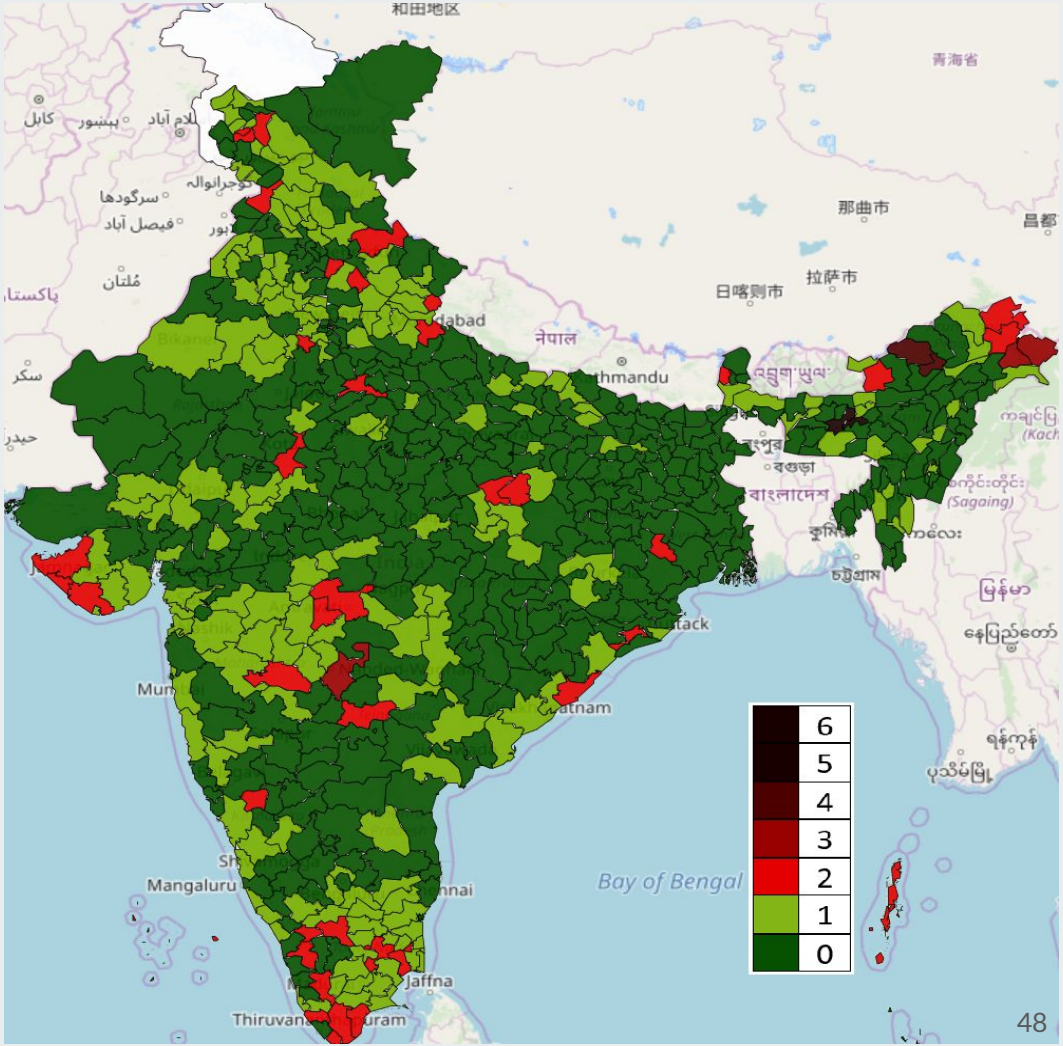
Calculated using 5-fold cross validation method



Sample Results: 22% Improvement in Accuracy



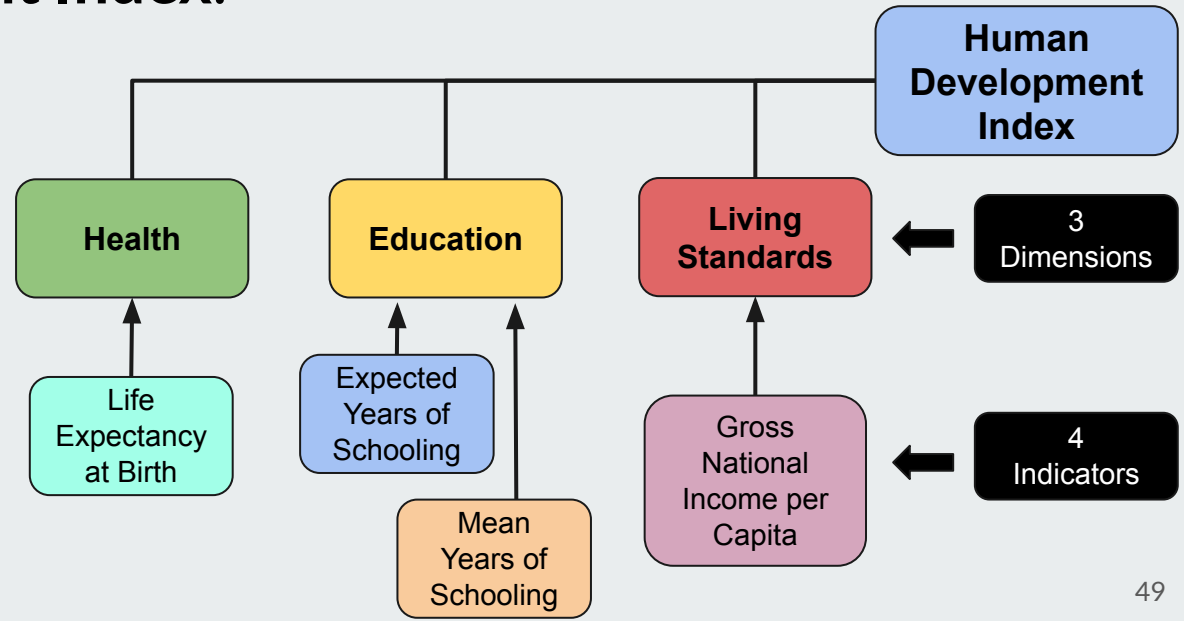
Number of Misclassified Indicators in 2011



Aggregate Development Index (ADI)

To understand an **aggregate behaviour of development** in terms of the different socio-economic indicators, we introduce an **Aggregate Development Index**.

It is similar to the concept of **Human Development Index (HDI)**



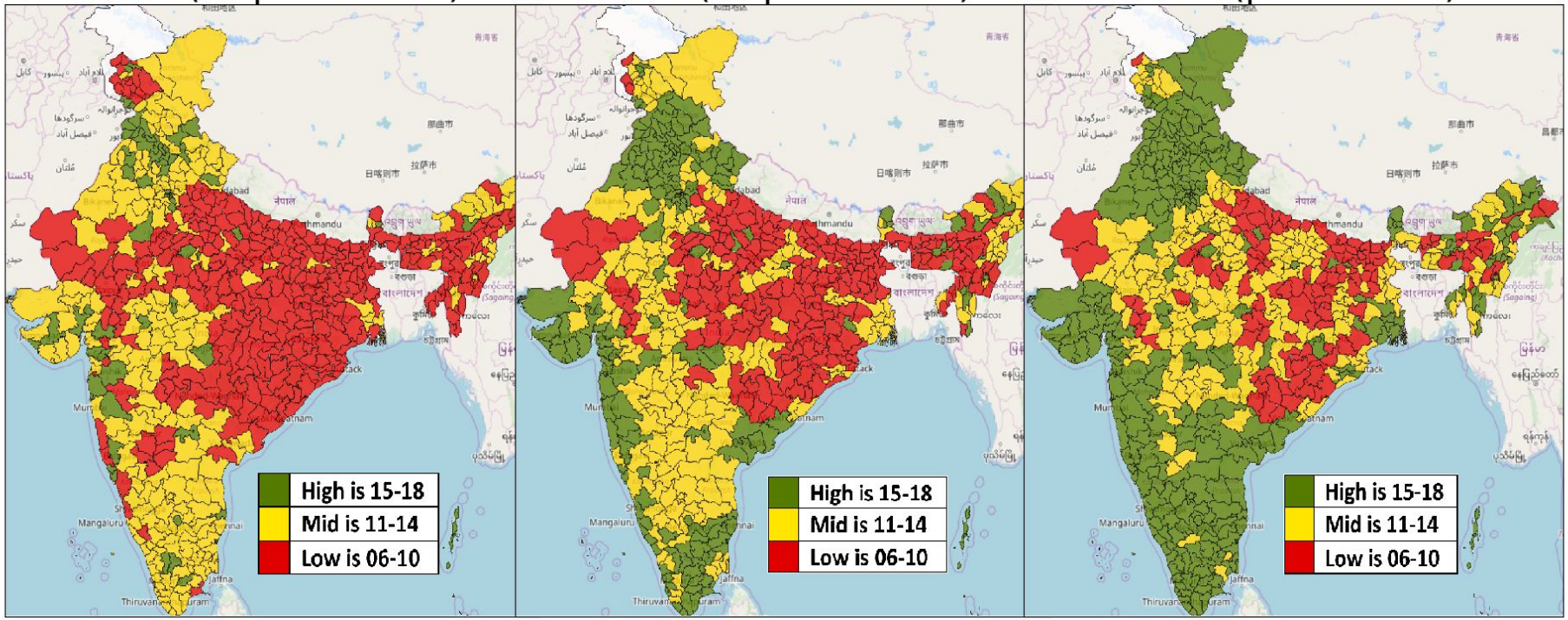
Aggregate Development Index (ADI)

ADI = sum of the levels of all the indicators for a district

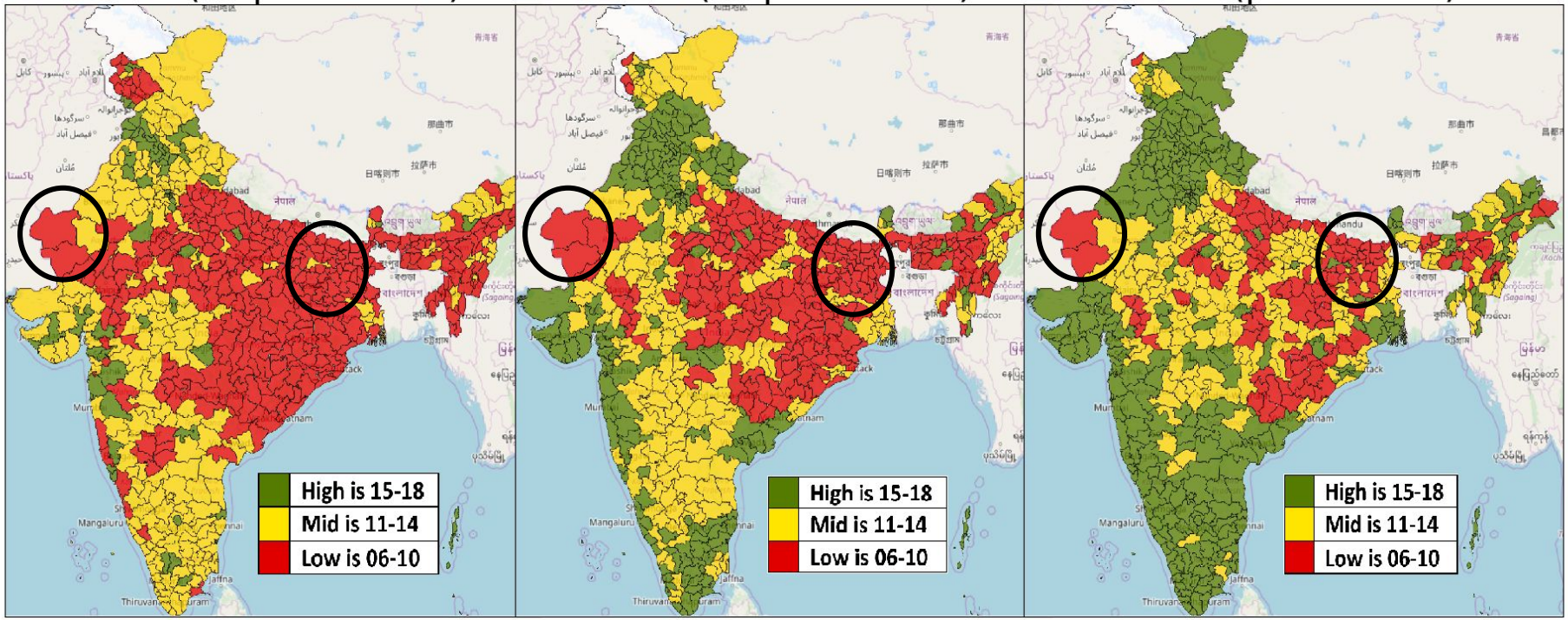
The value of ADI ranges from **6** to **18** for every district

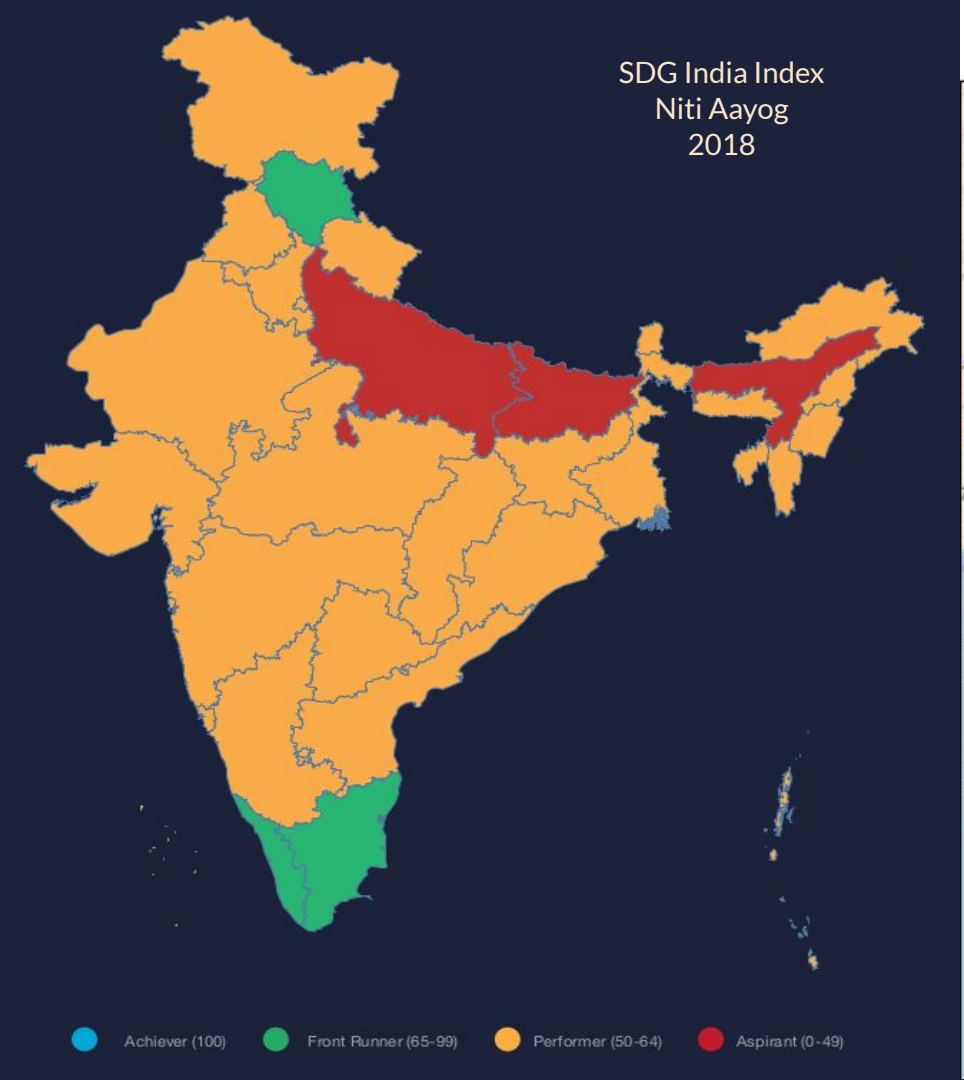
A Normalized Root Mean Square Error (RMSE) of 0.0413 is achieved when the ADI value of districts is predicted using satellite data for 2011

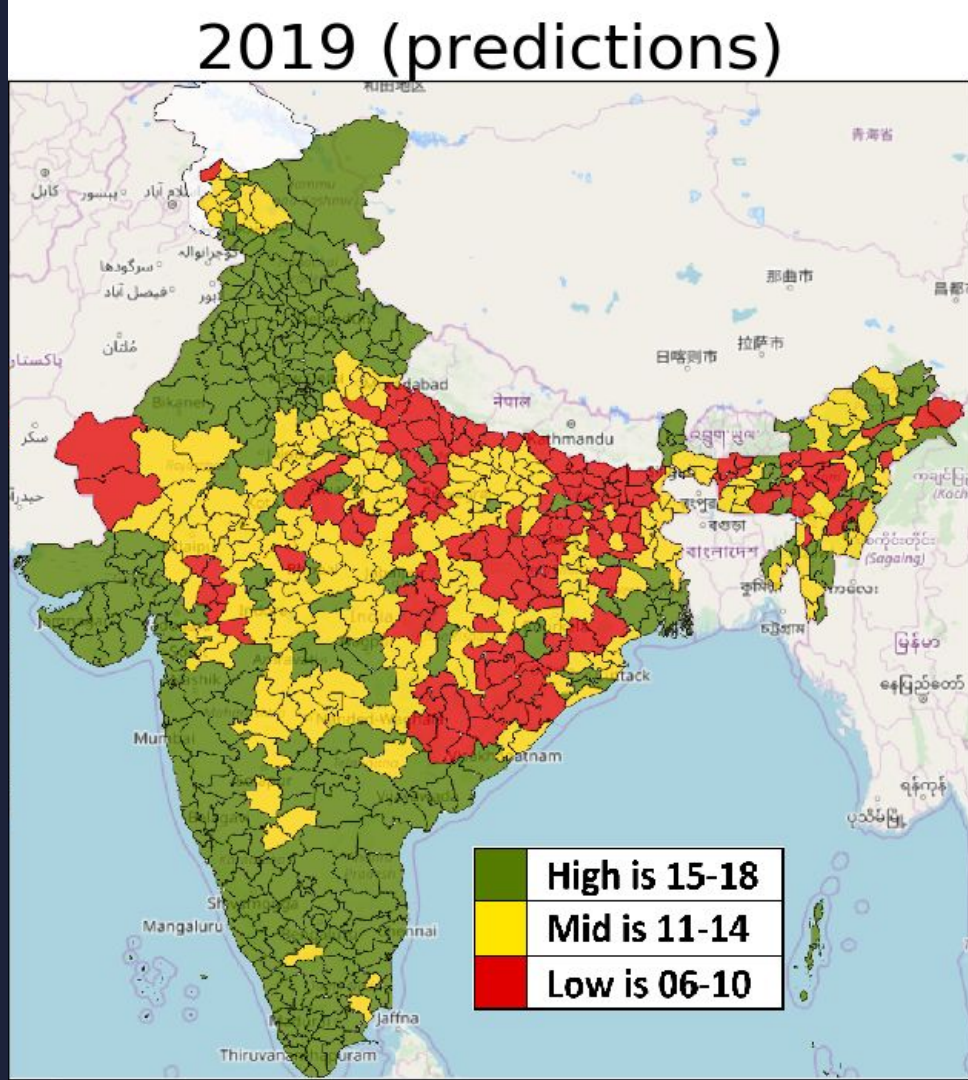
Predicting ADI for 2019

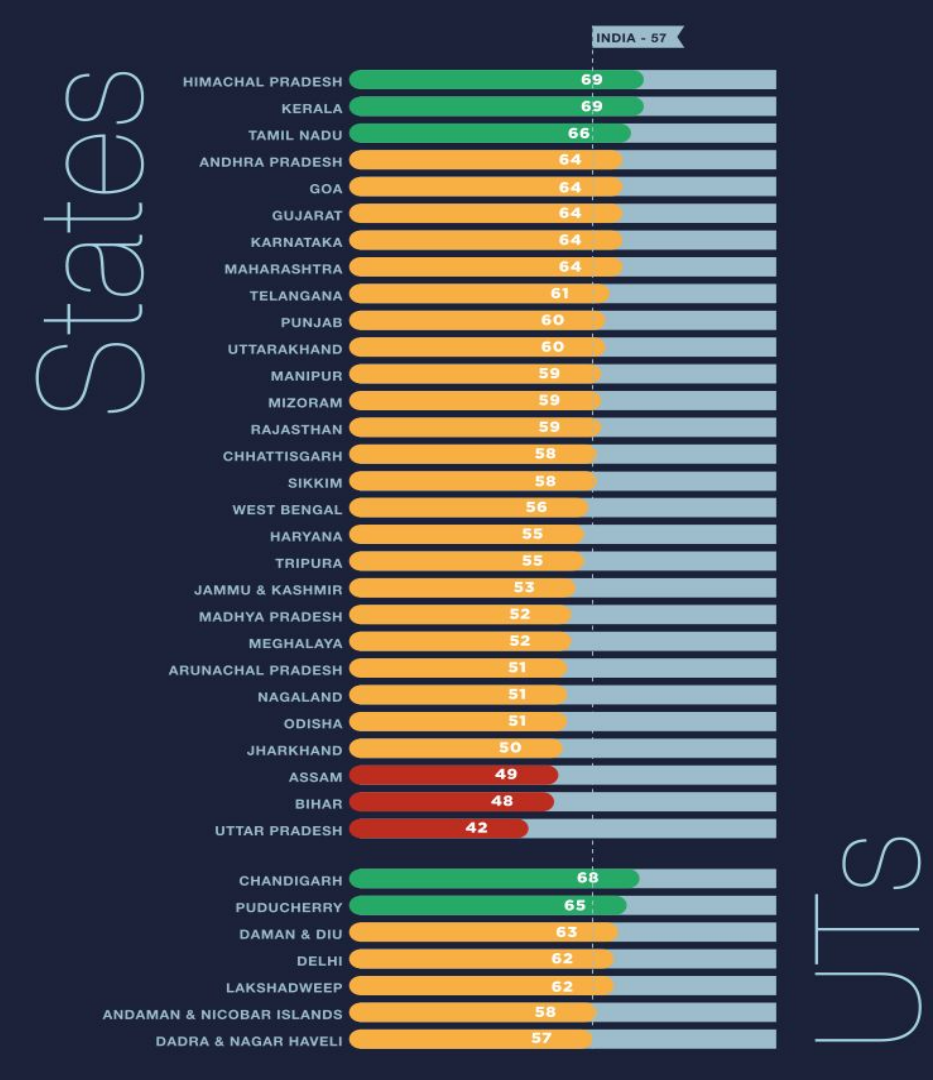


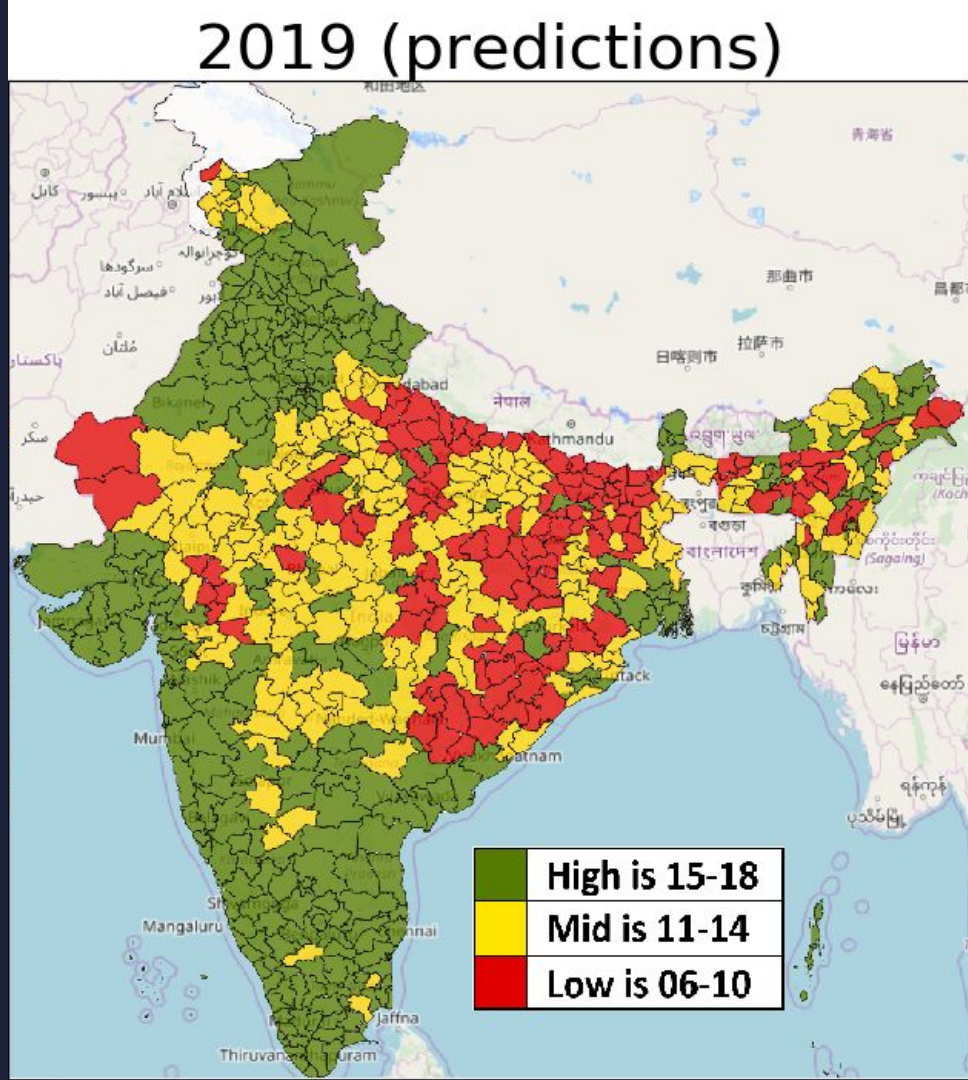
Predicting ADI for 2019




SDG India Index
Niti Aayog
2018
Achiever (100)
Front Runner (65-99)
Performer (50-64)
Aspirant (0-49)


2019 (predictions)
High is 15-18
Mid is 11-14
Low is 06-10


States
INDIA - 57
HIMACHAL PRADESH 69
KERALA 69
TAMIL NADU 66
ANDHRA PRADESH 64
GOA 64
GUJARAT 64
KARNATAKA 64
MAHARASHTRA 64
TELANGANA 61
PUNJAB 60
UTTARAKHAND 60
MANIPUR 59
MIZORAM 59
RAJASTHAN 59
CHHATTISGARH 58
SIKKIM 58
WEST BENGAL 56
HARYANA 55
TRIPURA 55
JAMMU & KASHMIR 53
MADHYA PRADESH 52
MEGHALAYA 52
ARUNACHAL PRADESH 51
NAGALAND 51
ODISHA 51
JHARKHAND 50
ASSAM 49
BIHAR 48
UTTAR PRADESH 42
UTs
CHANDIGARH 68
PUDUCHERRY 65
DAMAN & DIU 63
DELHI 62
LAKSHADWEEP 62
ANDAMAN & NICOBAR ISLANDS 58
DADRA & NAGAR HAVELI 57


2019 (predictions)
High is 15-18
Mid is 11-14
Low is 06-10

Conclusions

- We presented an analysis of the potential **use of satellite data for the prediction of socio-economic indicators** over time, at the spatial scales of districts.
- The prediction models tend to **lack generalizability** over time.
- The **proposed method is generic** and can be applied to improve the temporal transferability of other kinds of prediction models as well.
- We predict over ten years an **aggregate development index**. This application can be useful to identify outlier districts that should be investigated further.

Thank You