Estimation of Urbanisation in Indian Districts

Aditi Singla & Prachi Singh

Introduction

- Urbanisation
  - Rapid increase in population, especially in developing nations like India
  - Urbanisation has led to expansion of existing cities & settlement areas
  - Improved living conditions, but health, ecological balance compromised

- Need of Urbanisation Index
  - Better health planning and infrastructure expansion
  - No well-established or justified way to estimate landcover or landuse

Background

Remotely sensed imagery extensively used to identify land cover changes:

- Use of spectral bands like NIR, SWIR, etc. to decide land cover categories
- Use of combinations of these to give indices like NDVI, NDWI and NDBI
- Use of classification models using these indices as input parameters
- Goldblatt et al (2016):
  - Constructed a comprehensive dataset of ~20k points (30mx30m) for India
  - 4682 marked as builtup (BU) and 16,348 marked as non-builtup (NBU)
  - Used various classifiers and used cross-validation
- Goldblatt et al (2018):
  - Used GDLA & manually labelled data, for classification in Vietnam

Goldblatt, Ran, et al. "Detecting the boundaries of urban areas in india: A dataset for pixel-based image classification in google earth engine." Remote Sensing 8.8 (2016): 634.

Background

Issues with the classifiers by Goldblatt:

- Overall Accuracy reported : 87%
- Dependent on the input satellite images, sensitive to cloud cover, months selected
- Observation:
  - Groundtruth data chosen from highly populated areas (above 40 persons per hectare) and its peripherals till an enclosing polygon of double the size
  - Errors in classification in remote regions, like rocky terrains with no habitation, hamlets
- Preferable to perform a temporal analysis to record for the changes, instead of a cross-sectional analysis, on the classified images

Goldblatt, Ran, et al. "Detecting the boundaries of urban areas in india: A dataset for pixel-based image classification in google earth engine." Remote Sensing 8.8 (2016): 634.

Methodology

- Obtaining BU/NBU classified images
  - Run a CART classifier on satellite imagery on Google Earth Engine (GEE)
  - Observe noisy and unsatisfactory results on the raw images
  - Use of corrective methods to clean the satellite images

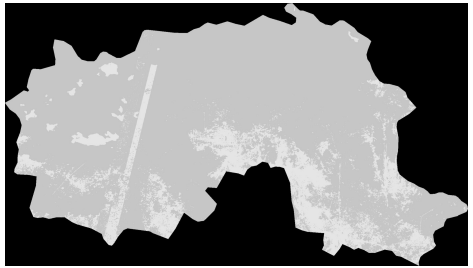
- Generating urbanisation indices using the classified images
  - Urbanisation Index:
    - Generate parameters to represent extent and patterns of change in builtup areas
  - Extent of change using pixel-wise Temporal Analysis
  - Patterns in change using Blob Detection

Obtaining BU/NBU classified images

- **Classifier on Google Earth Engine**
  - Use of GEE CART classifier trained on Goldblatt's data, on LANDSAT 8 satellite images
  - We observe noisy & unsatisfactory results

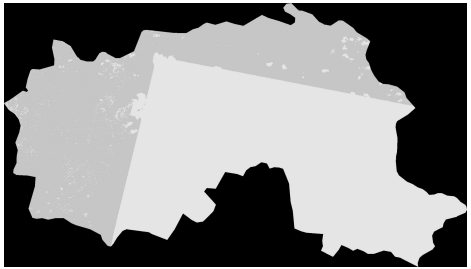
2014

2015



2016

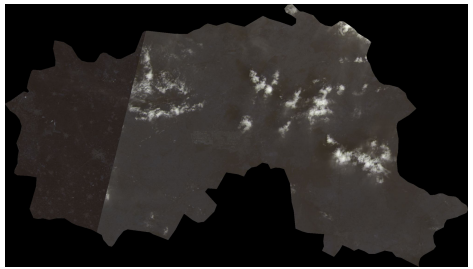
2017



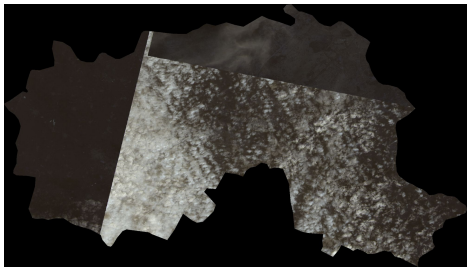
2018

Obtaining BU/NBU classified images

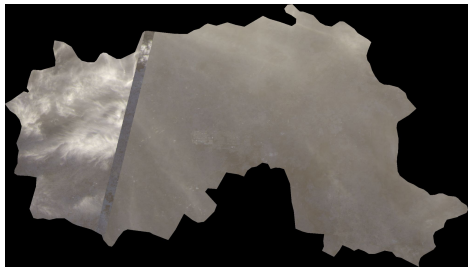
- **Classifier on Google Earth Engine**
  - Noise can be accounted by the variable cloud cover in the satellite images



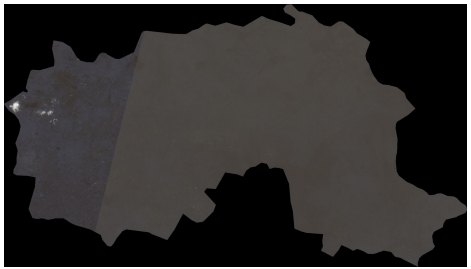
2014



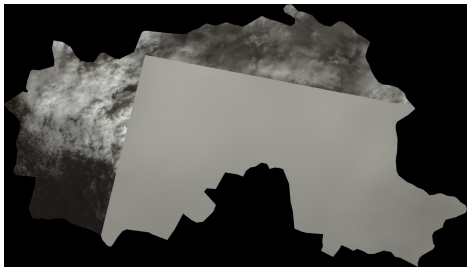
2015



2016



2017



2018

Obtaining BU/NBU classified images

- **Accounting for Cloud Cover**
  - Clouds masked away using Rodrigo Principe's cloud masking module
  - Median of all available pixels for the given duration
  - State-level analysis:
    - Districts with extremely high absolute values of cloud cover were removed
      - Threshold $t_h = 3$
      - Mostly hilly areas affected [Assam: 15/22, Bihar: 1/37, Uttarakhand: 3/13]
    - Using the technique of variance plots:
      - 2016 images have relatively high cloud cover
      - Skipped for our analysis

Obtaining BU/NBU classified images

- Final classification results after corrective measures

2014

2015

2017

2018

Generating urbanisation indices

Urbanisation: Change in the spread of built-up regions, measured as the extent and patterns in change of urban areas

- Extent of change using Temporal Analysis
  - Spatial Smoothing and Temporal Analysis
  - Thresholding and Classification
  - Validation of classification results
- Patterns in change using Blob Detection
  - Identification of Blobs for landuse concentration
  - Generating parameters using Blob Detection

Extent of change using Temporal Analysis

Spatial Smoothing and Temporal Analysis

- **Weighted Neighbourhood Score**

$$p_{ij} = \Sigma_{((a,b) \in N)} x_{ab} * (1/w) + x_{ij}$$

Example of mask for 1 Level Neighbours:

1/8	1/8	1/8
1/8	1	1/8
1/8	1/8	1/8

- **Simple Box Blur**

$$p_{ij} = \Sigma_{((a,b) \in N)} x_{ab} + x_{ij}$$

Example of mask for 1 Level Neighbours:

1	1	1
1	1	1
1	1	1

Three levels of spatial smoothing:

- 1 level Neighbourhood
- 2 level Neighbourhood
- 3 level Neighbourhood

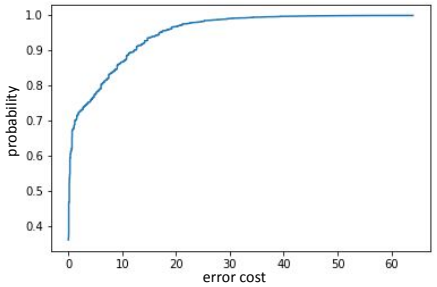
Spatial Smoothing and Temporal Analysis

- **Temporal Analysis:**
  - Performing linear regression on the value of each pixel across given years
  - Finding mse of the given pixel upon prediction on the obtained line
  - Plotting CDF curves for standard error obtained for each pixel

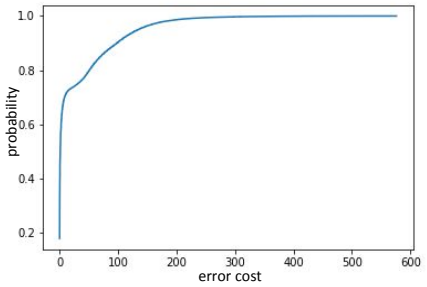
Spatial Smoothing and Temporal Analysis

- **Temporal Analysis post Weighted Neighbourhood Score:**

1 Level Neighbours

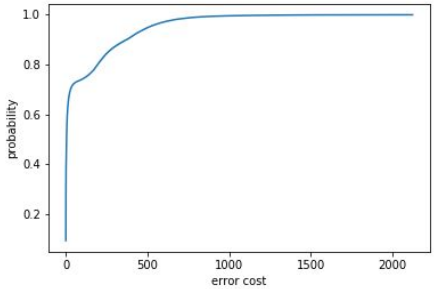


2 Level Neighbours

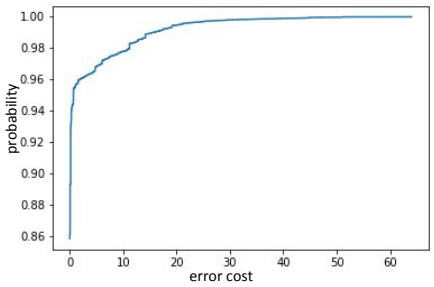


Gurgaon

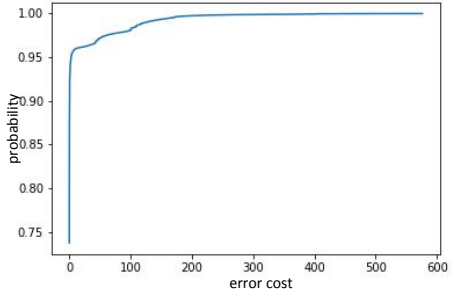
3 Level Neighbours



1 Level Neighbours

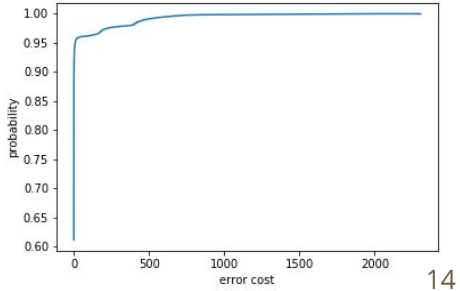


2 Level Neighbours



Jaipur

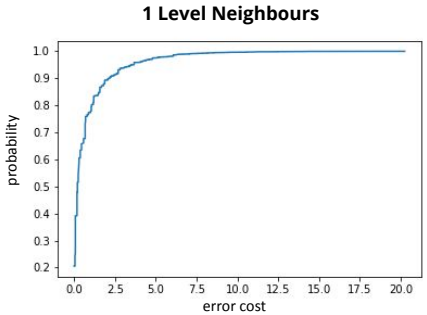
3 Level Neighbours

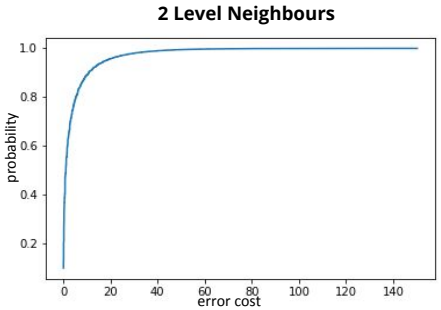


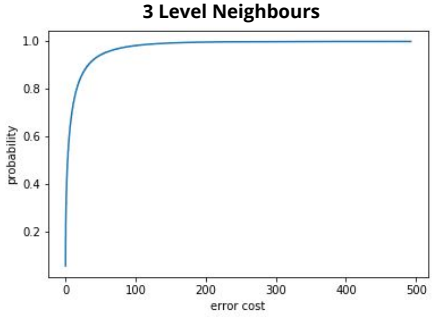
Spatial Smoothing and Temporal Analysis

- **Temporal Analysis post Simple Box Blur:**

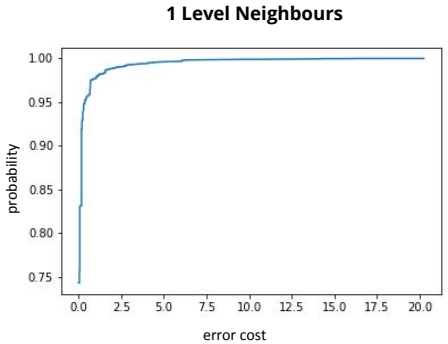
Gurgaon

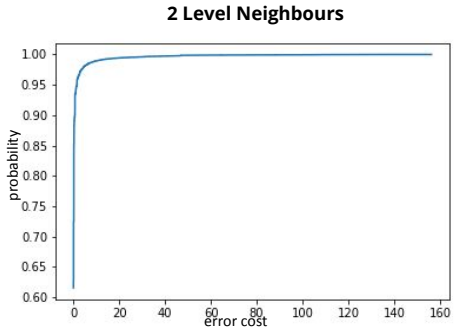


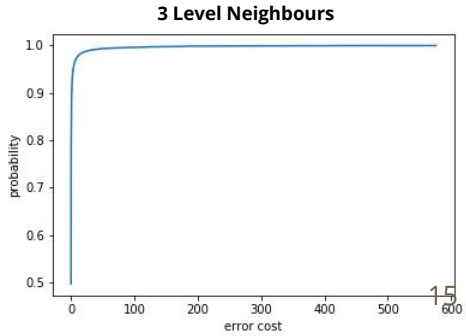




Jaipur

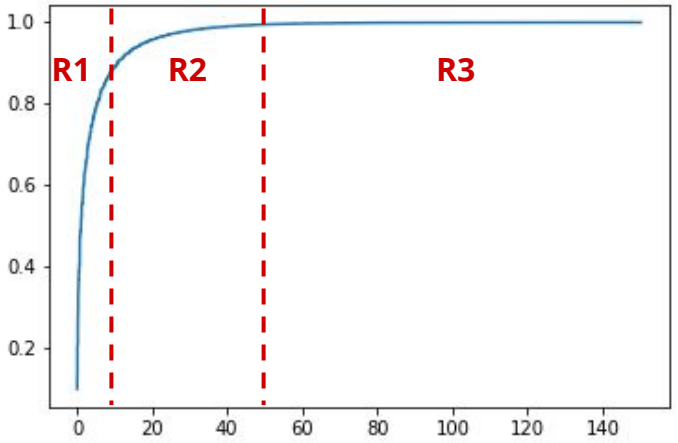






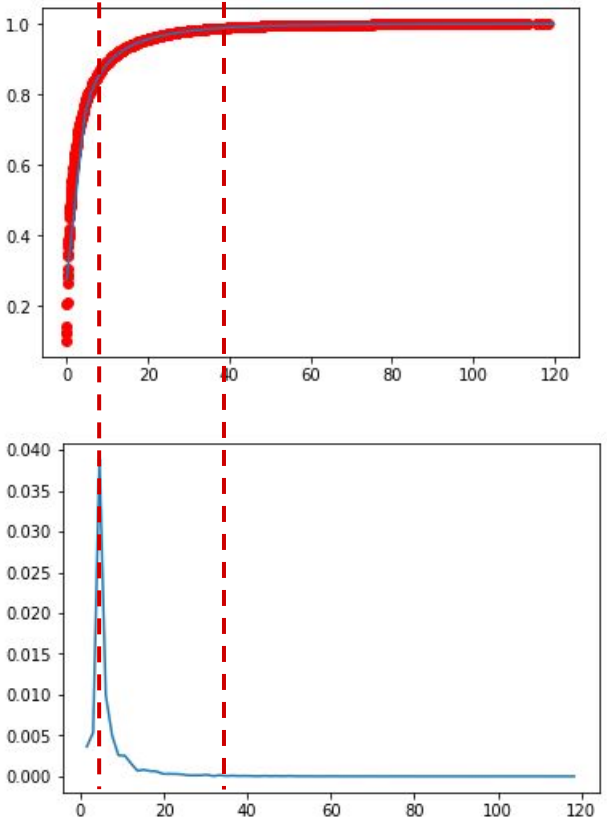
Thresholding and Classification

- Three types of regions in the cdf plot:
  - **Region 1:** low mean squared error of prediction, mostly because these pixels have zero or small slope i.e. constantly built-up or constantly non built-up throughout the years in consideration
  - **Region 2:** medium mean squared error of prediction, mostly because these pixels have some slope >=0 i.e. these points have changed throughout the years
  - **Region 3:** high mean squared error of prediction, mostly because these pixels follow erratic pattern in their BU and NBU values



Thresholding and Classification

- Automatic thresholding using double derivative
- Three major classes:
  - **Constantly NBU:** Error < threshold1 and NBU
  - **Constantly BU:** Error < threshold1 and BU
  - **Changing:** threshold1 < Error< threshold2
- Points in region 3 unreliable and not taken into consideration for index calculation

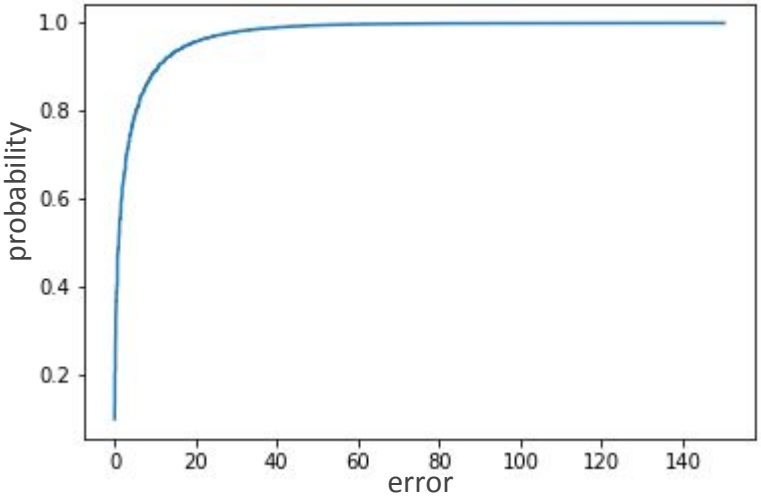


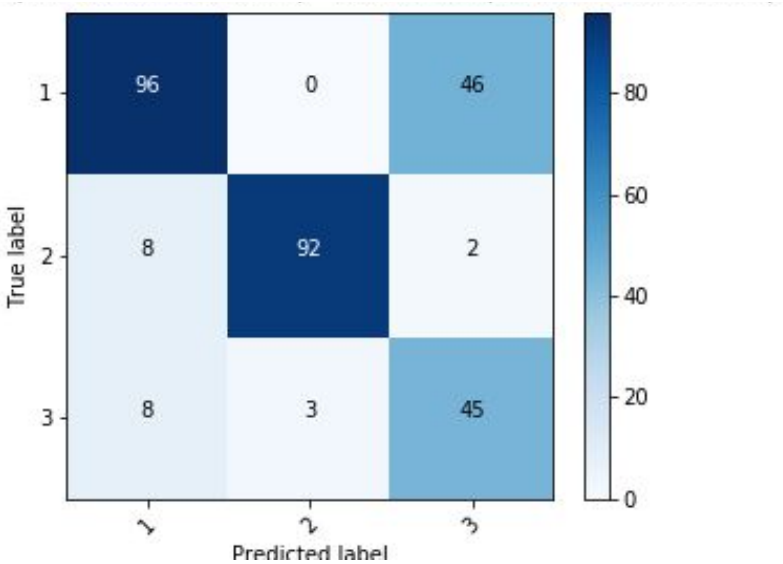
Validation of Classification Results

- Created ground truth of 300 points each for Gurgaon and Jaipur, marking points in classed:
  - Constantly NBU (1)
  - Constantly BU (2)
  - Changing (3)
- Distribution of points across districts as:
  - 100 from rural farmlands
  - 100 from urban clusters
  - 60 from urban peripheral
  - 40 from rocky terrains
- Choosing optimal cutoffs for thresholds using accuracies over this dataset

Validation of Classification Results

Threshold 1: 8.033820061818183, Threshold 2: 54.60872727272728



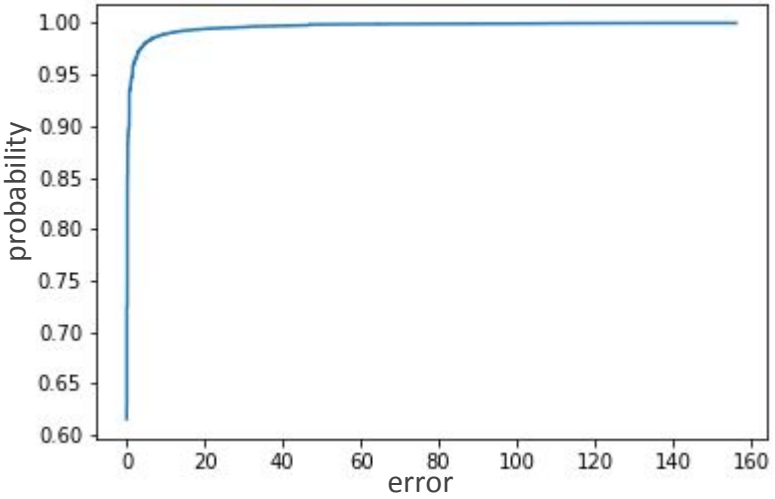


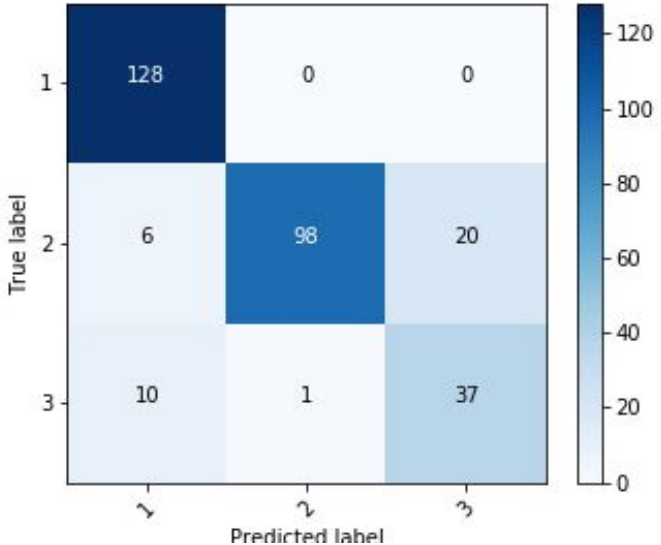
Gurgaon

Prediction accuracy: 77.67%

Validation of Classification Results

Threshold 1: 6.578284708282828, Threshold 2: 26.83080808080808



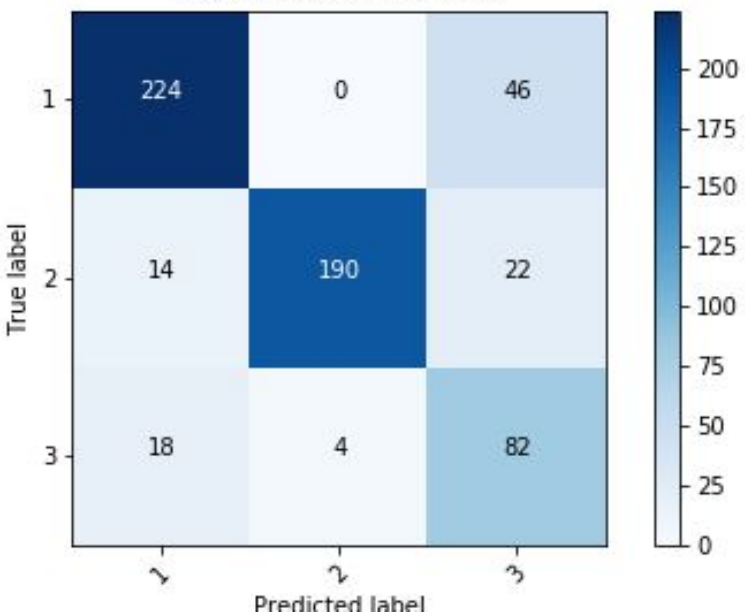


Jaipur

Prediction accuracy: 87.67%

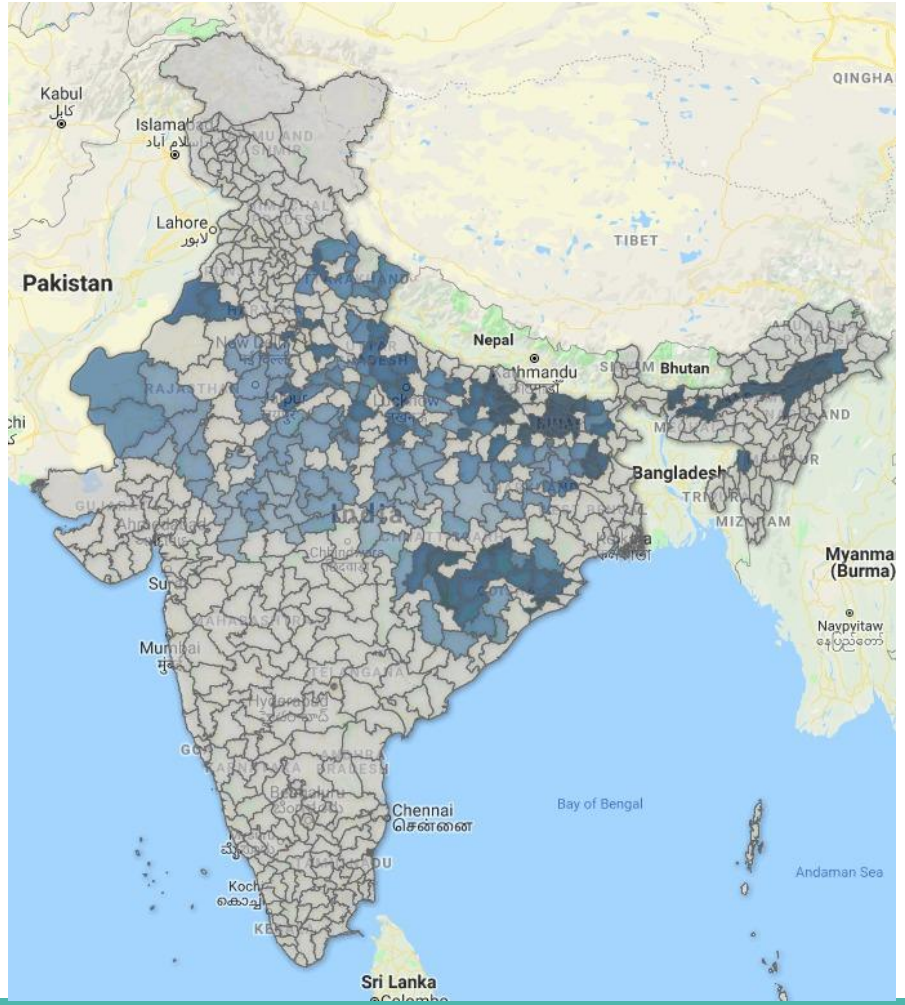
Validation of Classification Results

Gurgaon and Jaipur combined dataset



Prediction accuracy: 82.67%

Extent of change



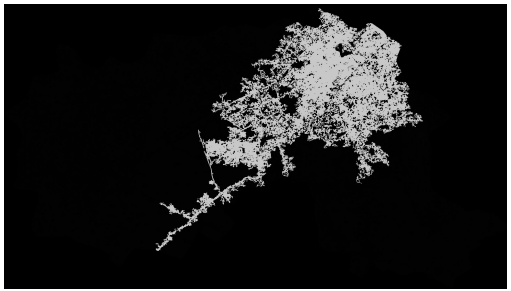
Patterns in change using Blob Detection

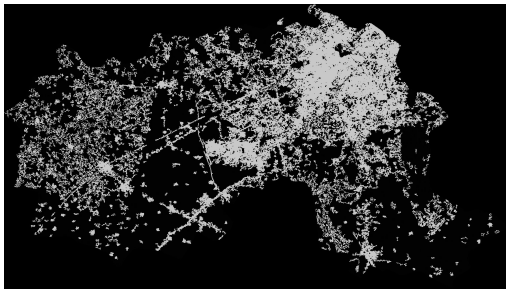
Identification of Blobs for land use concentration

- Urbanisation not just percentage change, but also patterns of change
  - Expansion around existing urban clusters
  - Emergence of new settlement areas altogether
- Overall distribution of population (for eg. monocentric vs polycentric)
- Blob detection:
  - Breadth first search approach to identify blobs in the district
  - Use of *Herfindahl-Hirschman Index (HHI)* to measure the concentration
  - Change in concentration using blob detection on:
    - Regions which are constant throughout the 5 years
    - Regions which remained constant or changed over the 5 years

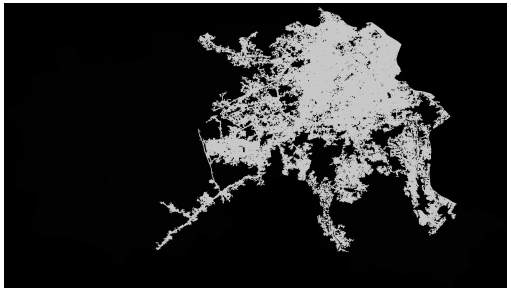
Identification of Blobs for land use concentration

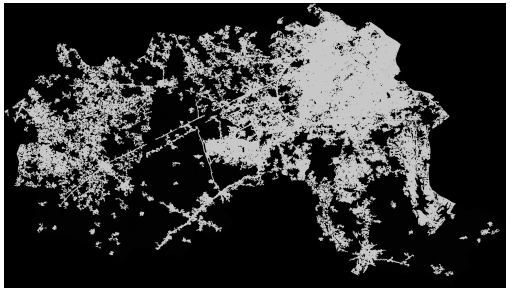
- Regions which are constant throughout the 5 years (Largest & All blobs)





- Regions which remained constant or changed over the 5 years (Largest & All blobs)

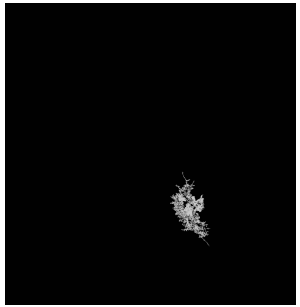
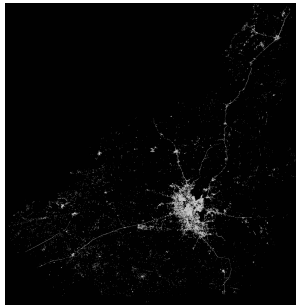




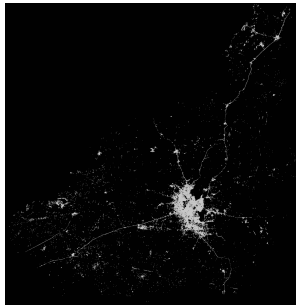
Gurgaon

Identification of Blobs for land use concentration

- Regions which are constant throughout the 5 years (Largest & All blobs)

- Regions which remained constant or changed over the 5 years (Largest & All blobs)

Jaipur

Generating parameters using Blob Detection

Parameters generated from blob detection:

1. **Num of blobs** for constant regions
2. **%BU in largest blob** for constant regions
3. **HHI Index** for constant regions
4. **Num of blobs** for constant + changing region
5. **%BU in largest blob** for constant + changing region
6. **HHI Index** for constant + changing region

Urbanisation Indices

Parameters	Gurgaon	Jaipur
% of Constantly NBU	78.1858	97.36003
% of Constantly BU	8.2863	1.5895
% of Changing	13.5279	1.0505
Num of blobs for constant regions	522	5182
%BU in largest blob for constant regions	56.2209	43.5371
HHI Index for constant regions	3240.73	1905.69
Num of blobs for complete region	127	3891
%BU in largest blob for complete region	66.9556	46.3907
HHI Index for complete region	4839.49	2169.8

Health and Urbanization

Bipul

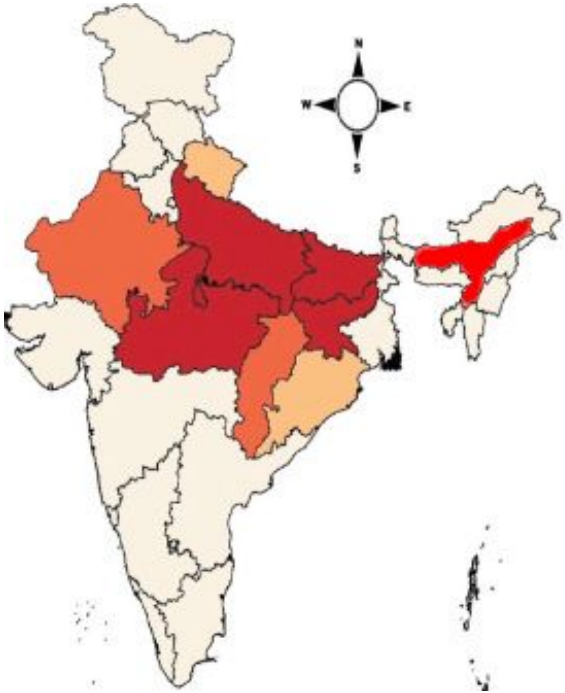
Motivation

- Policy makers need feedback, they have limited resources and wants to bring maximum impact by creating policies.

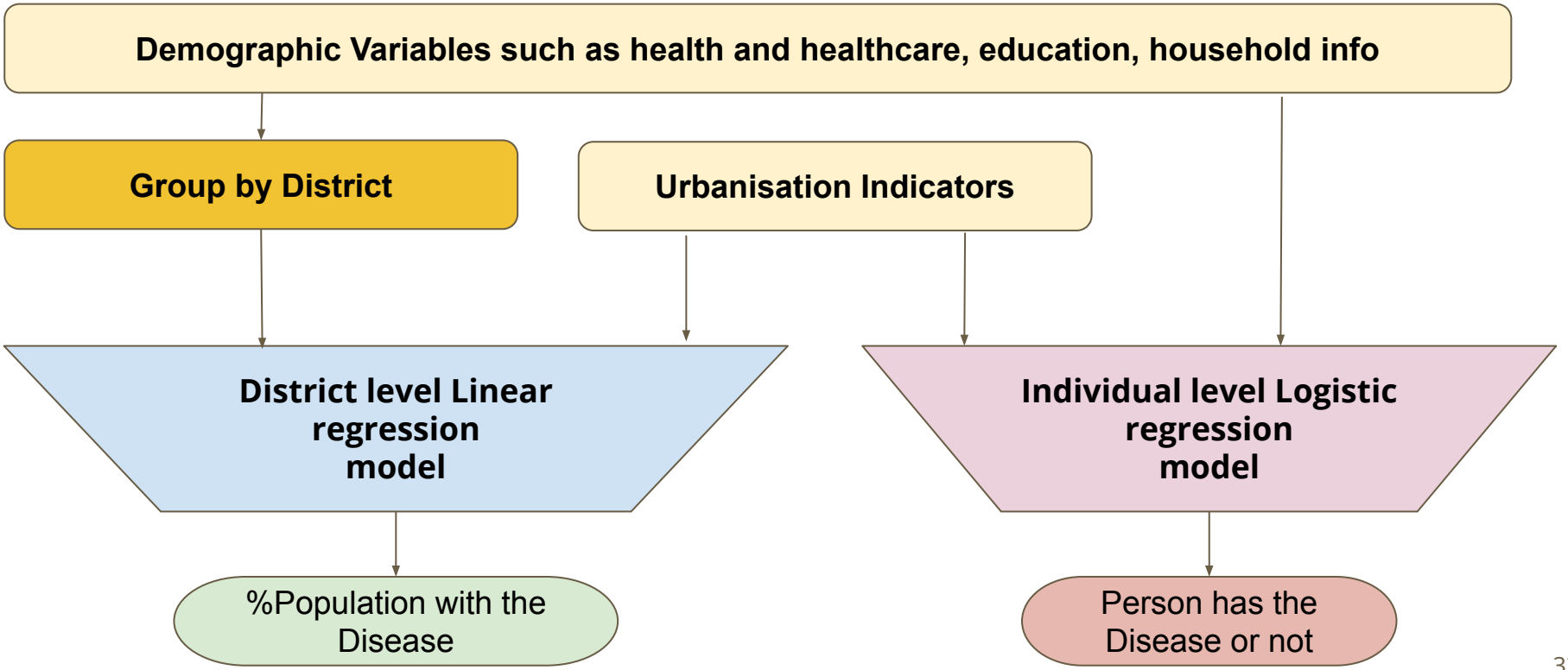
- So our main research objectives are:
  - Create individual and district level regression models using demographic variables and diseases as target variable
  - Use the urbanization indicators created by Aditi-Prachi team and note their role in our model

About data

- Annual Health Survey, a GOI initiative
- 11M women interviewed
- Questions related to **health** and **healthcare**, **pregnancy**, and other demographic variables such as **education**, **marriage**, etc.



Methodology



Data preprocessing

- Data cleaning
- Feature creation:
  - Some variables such as Age have 30 categories, so we need to re-categorize by creating newer boundaries
  - Variables such as Age, Highest Education may have ordinal number. E.g Highest Education value of 2 has less education than that of 3.
  - While variables such as Religion or Social Code has no such ordering.
  - One hot encoding:

	religion_1	religion_2
Person	1	0

Equal frequency binning of Highest education

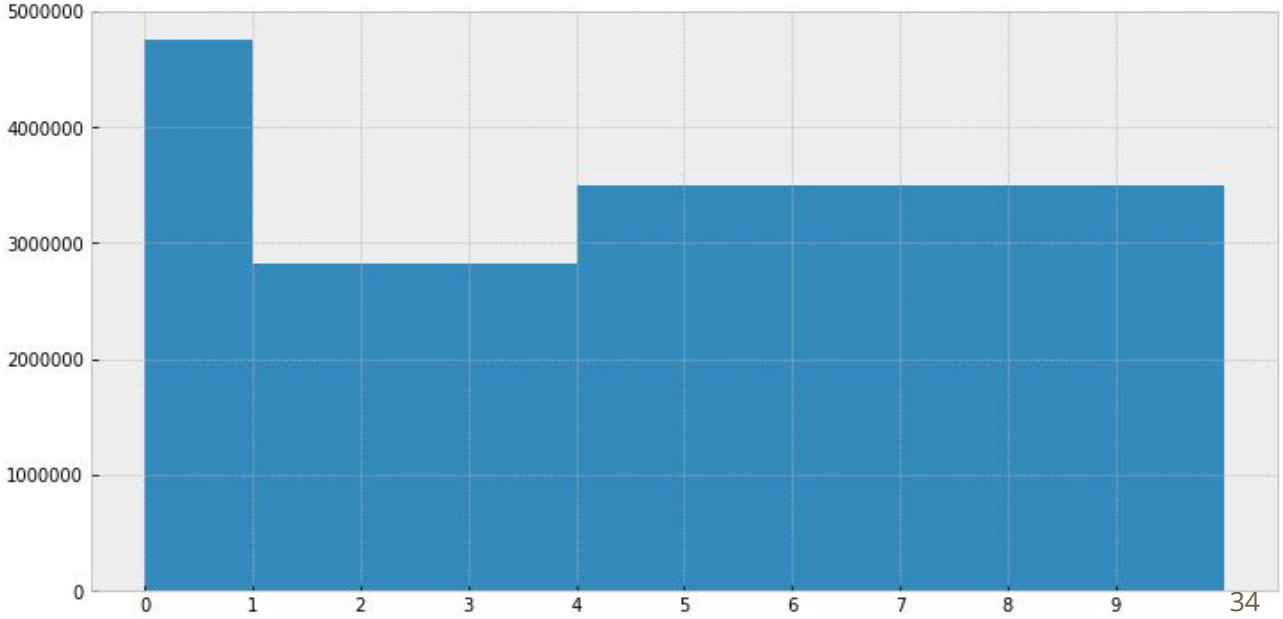
0: illiterate

1 & 2: education upto primary standard either at home or school

3 and above: middle school, and higher education going upto phd.

Y-axis : Frequency of population

X-axis: Education level



District level regression model

- We chose Linear regression:
  - Gives out coefficient corresponding to each feature
  - p-value may be used for feature selection
- **target variables:**- Diarrhoea/Dysentery
- We predicted district-wise %population having diarrhoea or dysentery
- We have 22 features and 174 sample districts

District level regression model

Feature selection

- Used the below two methods back and forth and removed one variable at a time
  a. Backward Feature selection based on p-values
     - Null hypothesis (x) : x has no effect on target variable
     - Rejected null hypothesis of variable with the highest p-values
     - Repeat until all p-values are less than alpha = 0.05

  b. Remove multicollinearity among features based on Variance inflation factor
     - Calculated VIF for each variable and removed the one with highest VIF
     - And ran the Backward Feature selection again

Result of District level regression model without Urbanization indicators

Feature	p-value	coeff
Relation to head	0.000	-0.0248
Treatment place	0.000	-0.0181
Smoking	0.0010	0.0706

District level regression model

After feature selection,the relevant features are:

1. **Relation to head**: Positive correlation with diarrhoea/dysentery
    - This variable may take values as:
        1: Family head or wife of family head
        2: Daughter or Daughter-in-law of family head
        3: Others
    - Power dynamic plays a huge role. So the closer the one is to family head, the better health amenities she will get

District level regression model

After feature selection,the relevant features are:

2. **% of population getting treated in hospital during last year:**
   Negatively correlates with the spread of diseases
   - Explanation: If someone have access to hospital, the likelihood of access to better sanitation is much more higher

District level regression model

After feature selection,the relevant features are:

3. **Smoke**: Positive correlation between % of population who smoke, regularly or offhanded, and % of population having target diseases.
    - Explanation: if one smokes, he is being ignorant to his health, which sips in other areas of healthcare as well

Other variables were removed in maximizing Adj. R-squared

- R-squared tells us the fraction of the variance in data explained by feature vector

Result of District level regression model with Urbanization indicators

High p-value of smoking

Feature	p-value	coeff
Relation to head	0.000	-0.0211
Treatment place	0.000	-0.0203
Smoking	0.094	0.0349
Urbanization Rate	0.0001	0.0135

Result of District level regression model with Urbanization indicators

Feature	p-value	coeff
Relation to head	0.000	-0.021
Treatment place	0.000	-0.0196
Urbanization Rate	0.0001	0.0137

District level regression model

When urbanization indicators **are added**:

- **Rate of Urbanization:** It positively correlates to % of population having diarrhoea/dysentery
  - Explanation: Increasing urbanization doesn't result in a simultaneous decrease in diseases because, in the new areas which are undergoing urbanization, the sanitation & healthcare systems have not been developed yet. But the increase in the population density has lead to the easy spread of contaminated diseases such as diarrhea, dysentery.

Result comparison:

Feature	p-value	coeff
Relation to head	0.000	-0.0248
Treatment place	0.000	-0.0181
Smoking	0.0010	0.0706

Feature	p-value	coeff
Relation to head	0.000	-0.021
Treatment place	0.000	-0.0196
Urbanization Rate	0.0001	0.0137

Features	Adj R-squared
without Urbanization	0.147
with urbanization and smoking	0.233
with Urbanization and without smoking	0.225

Individual level regression model

- 5M data points
- The diseases we want to model is diarrhoea/dysentery
- Skewed data: Only 0.7% people had diarrhoea/dysentery during last year

- Class imbalance: Using SMOTE, over-sample the minority class of the training dataset
- 19 + 9 (urbanization) features
- Feature selection: Chi-squared test

Feature	p-value
Injury treatment at home	8.7e-267
Alcohol	1.2e-197
Highest Education	5.6e-128
Injury treatment at hospital	3.9e-71
Rural	1.2e-53

Individual level regression model

- We chose Logistic regression:
  - target variable has only two values instead of real numbers
  - The values of 0 and 1 are arbitrary. What's important is the probabilities of success or failure.

- **target variables:**- Diarrhoea/Dysentery

- We predicted whether an individual will have diarrhoea or dysentery within one year

- We have 31 features and 174 sample districts

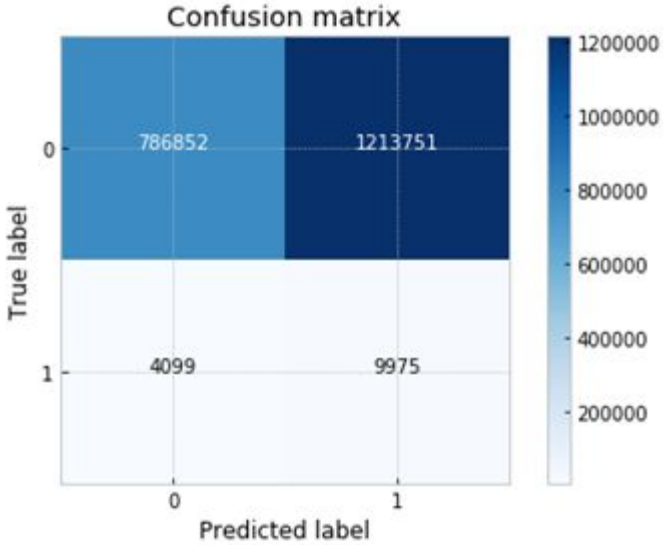
Individual level regression model

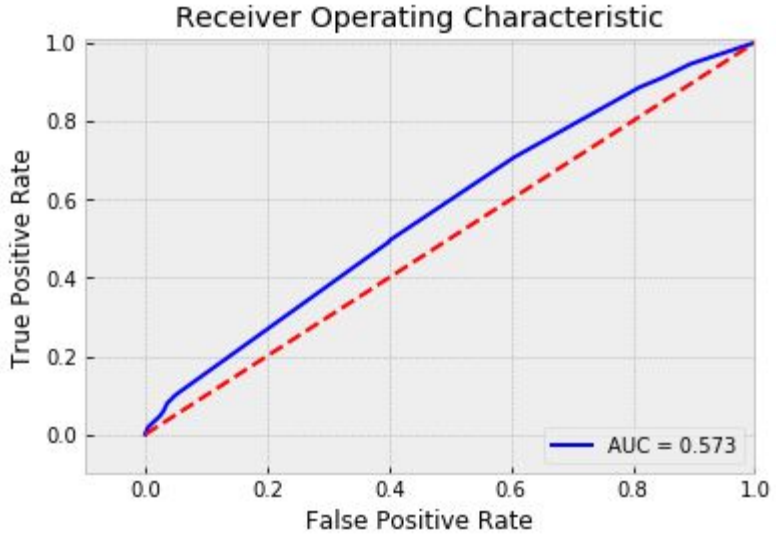
- **Regularisation parameter (1/C)**:
  - Larger the C, smaller the regularization will be and the model will be complex.

- Grid search with 5 fold cross validation to find best parameters for Logistic Regression.

Result of Individual level regression model

- **Modelling diseases without urbanization variables:**
- Logistic regression results:
  - Area Under Curve : 0.573
  - F1 score : 0.016



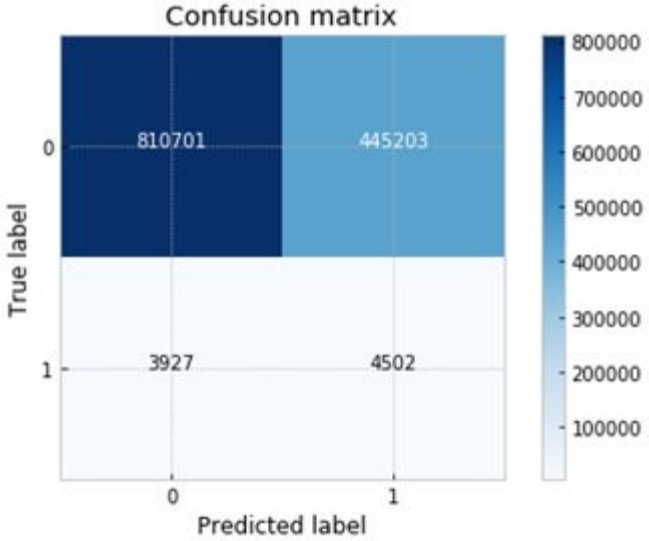


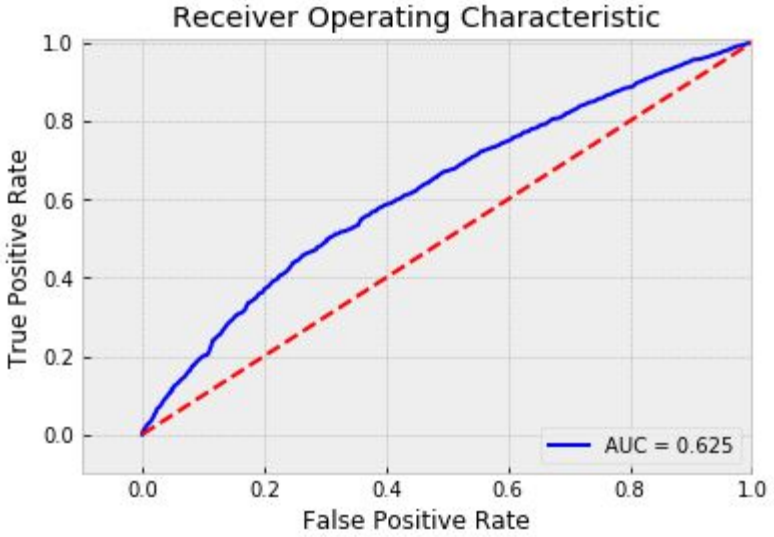
Individual level regression model

- **Modelling diseases with urbanization variables:**
- Since urbanization variable is available only district-wise, we copied the values for every individual of that district
- Feature selection:
  - Chi-squared test: Using this method we ranked features based on their p-values and selected the k features having least p-values
    - K = 14 had to be taken as optimal number of features because the this time all 9 urbanization indicators were having p-value the lowest among all
- Also performed SMOTE, in which the minority class variable were increased.

Result of Individual level regression model

- **Modelling diseases with urbanization variables:**
- Logistic regression results:
  - Area Under Curve : 0.625
  - F1 score : 0.020





Result comparison:

Without Urbanization		
	Predicted 0	**Predicted 1**
Real 0	39.3%	60.7%
Real 1	29.6%	70.4%

With Urbanization		
	Predicted 0	**Predicted 1**
Real 0	63.9%	36.1%
Real 1	46.5%	53.5%

	Area Under Curve	F1 score
Without Urbanization	**0.573**	**0.016**
With Urbanization	**0.625**	**0.02**

Conclusions

- At district level, rate of urbanization is the only indicator which have significant impact on our model performance and it positively correlates with the above diseases

- At individual level, although all urbanization indicators are selected in our model, but the model still remains very weak.

Future Work

- Linear & logistic regressions are entry level models for regression and classification respectively. Powerful ensemble methods such as Random Forest may be better suited because it is equipped to deal with lots of attributes
- The data is extremely skewed so I believe that the domain knowledge could help us identify really important features early on.

Thank you