# ESTIMATION OF URBANISATION IN INDIAN DISTRICTS

*Thesis submitted by*

**Aditi Singla**
2014CS50277

**Prachi Singh**
2014CS50289

*under the guidance of*

**Prof. Aaditeshwar Seth**

*in partial fulfilment of the requirements*
*for the award of the degree of*

**Bachelor and Master of Technology**

**Department Of Computer Science and Engineering**
**INDIAN INSTITUTE OF TECHNOLOGY DELHI**

**June 2019**

# THESIS CERTIFICATE

This is to certify that the thesis titled **ESTIMATION OF URBANISATION IN INDIAN DISTRICTS**, submitted by **Aditi Singla** and **Prachi Singh**, to the Indian Institute of Technology, Delhi, for the award of the degree of **Bachelor and Master of Technology**, is a bonafide record of the research work done by them under our supervision. The contents of this thesis, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

**Prof. Aaditeshwar Seth**
Professor
Dept. of Computer Science & Engineering
IIT-Delhi, 600 036

Place: New Delhi
Date: 27th June 2019

# ACKNOWLEDGEMENTS

We would like to wholeheartedly thank our advisor Prof. Aaditeshwar Seth for guiding us at each and every step of our project work. He helped us in learning on how to approach a problem by looking from different perspectives and taking a step back to see the bigger picture, which we would carry for the rest of our careers.

We would also like to thank Bipul, Ayush, Dibyajyoti, Arpit and Shyam with whom we had long and valuable discussions on various issues that we faced and the possible approaches we could take. We are also grateful to the Google Earth community for answering our numerous queries on their discussion forums and helping us in every way possible. We would also like to extend our thanks to the staff of CSE and SIT departments.

# ABSTRACT

KEYWORDS: Urbanisation; LANDSAT8; Settlement Delineation; Spatial Analysis; Extent of Urbanisation; Patterns of Urbanisation; India.

Recently, there has been a rapid increase in the population, especially in the developing nations like India. The existing cities have expanded and new settlement areas have been established. Although this has lifted the people in rural regions out of poverty, many factors like health, resource availability, etc. have been compromised. We here attempt to understand urbanisation and give a mechanism to estimate it, which can then help the government in better health planning and resource allocation.

This problem has got a lot of attention in the community and various attempts have been made to use census data, satellite imagery data and social media data collection, to obtain the above mentioned attributes and give an urbanisation estimate. We use the technique of spatial analysis using remotely sensed satellite data to give and urbanisation estimate. We define urbanisation as the extent and the distribution of change in built-up regions. To achieve this, we perform settlement delineation by applying pixel-wise classification models on the satellite imagery to generate an estimate of built-up regions at district level. Once this is done, we perform a temporal analysis on the spatial distribution of the built-up regions to generate these parameters to represent the urbanisation index.

# Contents

# Chapter 1

# INTRODUCTION

In the past few decades, there has been a rapid increase in the world's population, especially in the developing nations like India, which is accompanied by a drastic increase in the landarea covered by cities and towns. The existing cities have expanded and new settlement areas have been established. This rapid urbanisation comes with its own boons and banes. While it has helped in reducing poverty by improving the living conditions of thousands, many other factors like health, ecological balance have been compromised due to the disturbed balance. The living conditions of the lower sections of the society have worsened due to the inability of the concerned authorities to address these rapid changes.

We understand that a region-wise estimate of the extent of urbanisation can actually help the government in better health planning and infrastructure expansion. But there is no well-established or justified way to actually estimate the extent and change in landcover or landuse. The census data collected by the Registrar General and Census Commissioner of India, gives a fair overview on the population distribution, but is collected once every ten years and thus can't be used for urbanisation estimation on a more granular basis.

This problem has got a lot of attention from research groups at the intersection of data science and public administration. Urbanisation has been defined in different ways, based on the type of landcover (water bodies, vegetation, settlement areas), population distribution patterns, different social attributes (like health, diseased, etc.) or a combination of all of these. Various attempts have been made to use census data, satellite imagery data and social media data collection, to obtain the above mentioned attributes and give an urbanisation estimate. We have defined urbanisation as the extent and the distribution of change in built-up regions. To achieve this, we perform settlement delineation by applying pixel-wise classification models on the satellite imagery to generate an estimate of built-up regions at district level. Once this is done, we perform a temporal analysis on the spatial distribution of the built-up regions to generate the said urbanisation index.

In the next section, we give a brief literature review on the attempts that have been made to give an appropriate estimate for urbanisation and the motivation for our approach. Section 3 gives an outline for our proposed methodology to address the problem, which has been then described in details in the following Section 4. Section 5 details the experimental validation and results. Finally, the concluding remarks and scope of future work are presented in Section 6.

# Chapter 2

# BACKGROUND

Remotely sensed imagery has been extensively used to identify landcover changes occurring in the urban areas and its peripheral rural areas. A lot of efforts have been put in performing spatial analysis on satellite imagery like LANDSAT and Sentinel. There has been quite a development in the methods used to estimate the landcover areas and hence urbanisation. New indices have been devised to indicate different aspects of a human settlement, like Normalized Difference Vegetation Index (NDVI), Normalized Difference Built Index (NDBI), Normalized Difference Water Index (NDWI), etc., which can typically be calculated using only four bands, namely red, green, near infrared (NIR) and short-wave infrared (SWIR). Most of the initial work involving satellite data used these parameters to give estimates on the settelment areas, and hence determine urbanisation. Macarof et al.[MS17] have used NDBI and NDVI as indicators for Surface Urban Heat Island Effect. Zha et al.[YJS03] have used NDBI in automatically mapping urban areas from TM imagery.

Further, these estimates have been used to classify the land into vegetation land, barren land and a builtup region. Szabo et al.[SGBB16] worked on specific features of NDVI, NDWI and NDBI as reflected in land cover categories. These spectral indices were used to determine land cover types: water body (W); plough land (PL); forest (F); vineyard (V); grassland (GL) and built-up areas (BU) using Landsat-7 ETM+ data. They identified that the categories of BU, GL and F could be calculated using NDVI values, but the other land cover types differed significantly. Further, they investigated the ranges of these spectral indices from the aspect of land cover types. Similarly, Faridatul et al.[FW18] worked on estimating urban landcovers based on Novel Spectral Indices which were defined as MNDBI, TCWVI, ShDI. These indices were calculated using the basic bands and the initial indices of NDVI, NDBI and NDWI. They then devised an approach to classify four major urban land types: impervious, bare land, vegetation, and water.

Further, there have been approaches to automate the process of classfication by building models over the existing bands and indices as the input parameters. And it has been one of the most important applications developed using Earth observation satellites. Phiri et al.[PM17], in their paper, review the developments in landcover classification methods for Landsat images in the last four decades. This review suggests that initial approached involved visual analyses, followed by unsupervised and supervised pixel-based classification methods using maximum likelihood, K-means, etc. In the year 2015, Erzhu Li et al.[LDS$^+$15] gave an automatic approach for urban land cover classification from Landsat 8 OLI data. They used object oriented methods, instead of pixel based classification and introduced new parameters over the existing ones like NDVI, NDWI, the modified normalized difference water index (MNDWI), the soil adjusted vegetation index (SAVI) etc., which then used a non-linear SVM to train over these data points.

Goldblatt et al.[GYHK16] gave some revolutionary work on remote sensing to identify the boundaries of urban areas using pixel wise classification. They constructed and validated a large-scale and comprehensive dataset of 21,030 points (4682 marked as builtup(BU) and 16,348 marked as non-builtup) designed for mapping urban areas in India, and then demonstrated its applicability for mapping urban areas in India. They also give a classifier to perform the task of classification in Google Earth Engine (GEE) and evaluate its spatial generalizability, using a spatial k-fold cross-validation procedure. They utilize the full spectral imagery available in Landsat, and NDVI and NDBI indices(Normalized difference Built-up Index) indices, and incorporate the agro-climatic zones found in the large majority of developing countries. These agro-climatic zonesare geographical regions characterized by relatively homogenous environmental-physical characteristics, such as soil type, rainfall, temperature, and water resources. While this approach has proven to give promising results, upon observation we identify that the classification is deeply affected by the inherent noise present in the input satellite data, discussed in section 4.1.2. Also, we further observe that since the classification approach has its own limitations, it mightbe preferable to use the results for a temporal analysis to record for the changes, instead of a cross-sectional analysis.

Based on similar lines, Goldblatt et al.[GDH18] did some work using satellite data for urban classification in Vietnam. They designed a tool to map built-up landcover (LC)/ landuse (LU) regions in Ho Chi Minh City, Vietnam, using the publicly available satellite data and a cloud-based computational platform. They then mapped the temporal changes in the extent of built-up land cover in the province over the period 2000 to 2015. They used GDLA administrative cadastral data (polygons) and 15,945 hand-labeled examples (points) as reference data for supervised pixel-based image classification into “not-built-up” land cover, “residential” and “non-residential” land use. This analysis also gives a reasonable estimate on the distribution of the urban regions, and not urbanisation. Also, detailed GDLA data being used here is available only for Vietnam, and hence cannot be extended to India.

We therefore attempt to give an approach to estimate urbanisation, by providing an end to end pipeline, each component of which addresses one or the other shortcomings of the work that has been done so far and closely aligns with the aim and the methodology of this work.

# Chapter 3

# PROPOSED METHODOLOGY: OVERVIEW

The aim of our work is to establish methods and options for supervised training and classification of built up (BU) and non-built up (NBU) pixels in various districts of India, for years 2014 to 2018. We then establish methods to draw conclusions on temporal changes in the land-use pattern for different districts. These changes account for the urbanisation that has happened in the concerned region in the given duration of time. And hence as discussed before, to get a better estimate of the urbanisation, it is important to account for the extent of change as well as the distribution pattern of the change over the region. Then we generate a groundtruth dataset to validate our temporal analysis and generate a set of 9 parameters to represent the urbanisation index. Throughout our analysis, we have used Gurgaon and Jaipur to show the results, unless mentioned otherwise. The outline of the pipeline is as follows:

1. **Obtaining BU/NBU classified images:** We train a classifier on an available groundtruth dataset and run it on the Landsat-8 images of years 2014 to 2018. We identify that performing classification on the raw images leads to noisy and unsatisfactory results, which can be owed to the variable cloud cover index spatially and temporally. We then apply some correction methods to improve the results.

2. **Generating urbanisation indices using the classified images:** We generate a total of 9 parameters to account for urbanisation, 3 of which represent the extent (amount) of change, while the rest 6 parameters represent the patterns in change during the given period.

   - **Extent of change using Temporal Analysis:** Since the Landsat images available have their own limitations and that the classifier trained mightnot be the most accurate, we understand that the absolute pixel wise classification results maynot align with the ground truth. To address this, we apply some spatial smoothing techniques on the classification results to address the noisy pixels, and then do a temporal analysis over the span of 5 years, to obtain the said 3 parameters that can represent the extent of change for urbanisation. Since at various steps, we need to decide on threshold values, it is important to justify them by validating them against the ground truth. We therefore generate this dataset and perform the validation.
   - **Patterns in change using Blob Detection:** Urbanisation doesn't just account for the percentage change across a period of time, but also the patterns of change in the distribution. Urbanisation might be expansion around the existing urban clusters, or emergence of new settlement areas altogether. The overall distribution of the population in a district (for example, monocentric vs polycentric) also represent urbanisation. We therefore generate a few parameters using the method of blob detection.

# Chapter 4

# PROPOSED METHODOLOGY: COMPONENT-WISE

## 4.1 Obtaining BU/NBU classified images

### 4.1.1 Classifier on Google Earth Engine

We use a pixel-based classification approach (into builtup and non-builtup categories) for the districts of India, utilizing the full spectral imagery available in Landsat, as well as the NDVI and NDBI indices. For this, we use the GEE built-in Classification and Regression Trees (CART) classifier [BFSO84] based on the remotely sensed imagery from Landsat 8 (LANDSAT/LC08/C01/T1_TOA i.e. USGS Landsat 8 Collection 1 Tier 1 TOA Reflectance [LS8]) as input to predict built up and non-built up areas. The training data fed to the classifier is a FeatureCollection with a property storing the class label as 1 for BU and 2 for NBU and the bands ['B1','B2', 'B3', 'B4', 'B5', 'B6', 'B7', 'B8', 'B10', 'B11', 'NDBI', 'NDVI'] as predictor variables. This FeatureCollection is imported from a groundtruth dataset available in the form of a Fusion Table [DAT], generated by Goldblatt et al.[GYHK16] As discussed previously, they have created and validated a new dataset for India, consisting of 21,030 polygons (30 m x 30 m in size), randomly distributed throughout India and manually labeled as BU or as NBU. Among these 21,030 polygons, 4682 polygons are labeled as BU and 16,348 polygons are labeled as NBU. This data incorporates different agro-climatic zones found in majority of developing countries like India, zones being characterized by their relatively homogenous environmental-physical characteristics, such as soil type, rainfall, temperature, etc.

We run this classifier on the districts of Gurgaon and Jaipur for the years from 2014 to 2018, for which the Landsat 8 imagery is available. The original google earth images of these districts can be seen in Fig. 4.1 and Fig. 4.2. Further, Fig. 4.3 and Fig. 4.4 show the pixel-wise classified images obtained from the CART classifier. Here, the dark gray pixels represent BU, while the white ones are NBU.

From the classified images obtained, we observe that the classification isn't visibly accurate with large patches of white region throughout the districts. On deeper investigation, we identify that the classification is highly influenced by the cloud cover index in the Landsat images and can account for the noisy results obtained. The dense and non-uniformly distributed cloud cover for Gurgaon and Jaipur can be clearly observed in Fig. 4.5 and Fig. 4.6.

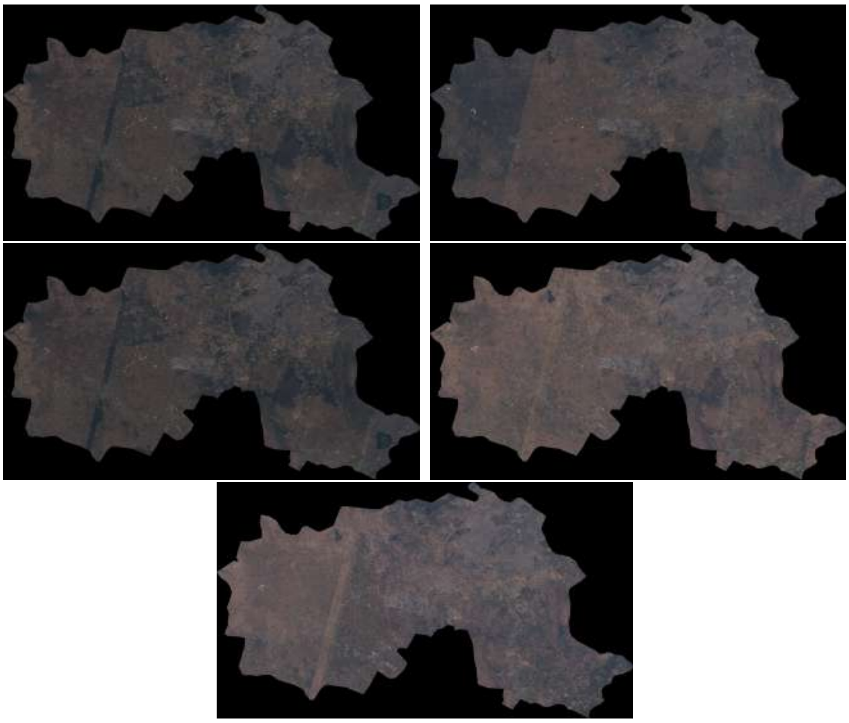


Figure 4.1: Gurgaon: Original Google Earth images, years 2014, 2015, 2016, 2017 & 2018

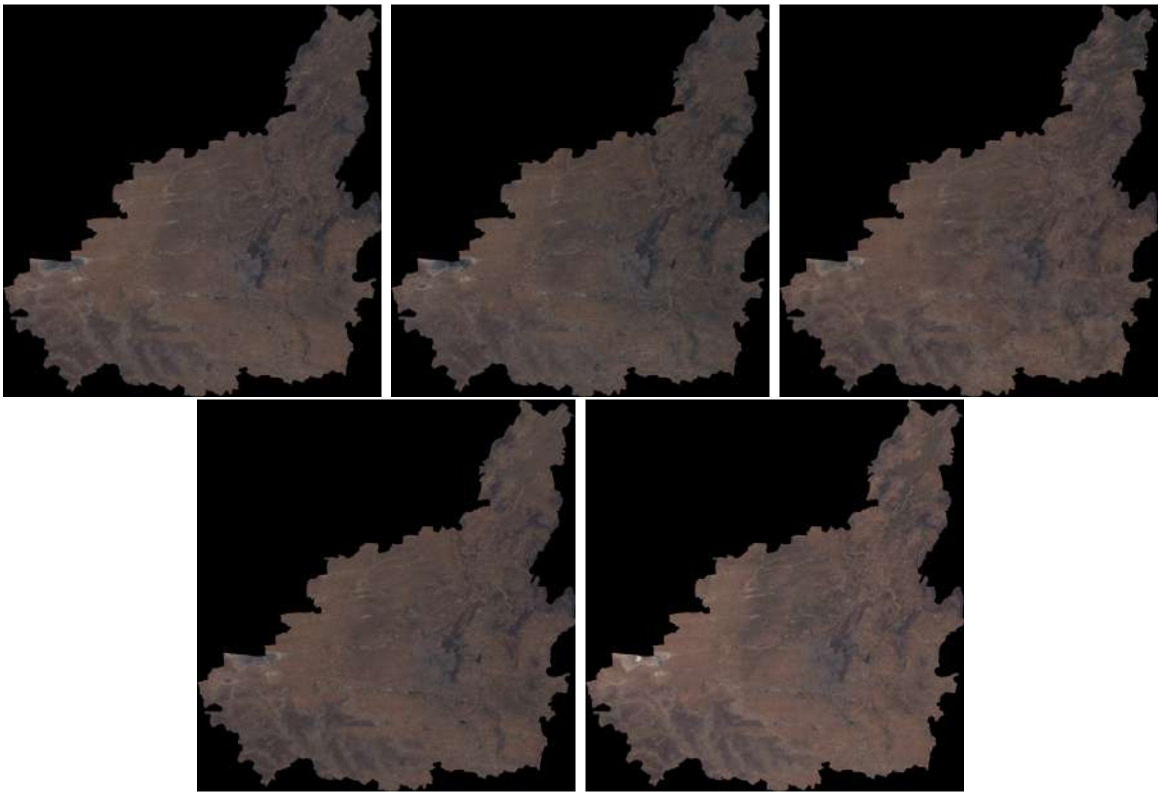


Figure 4.2: Jaipur: Original Google Earth images, years 2014, 2015, 2016, 2017 & 2018

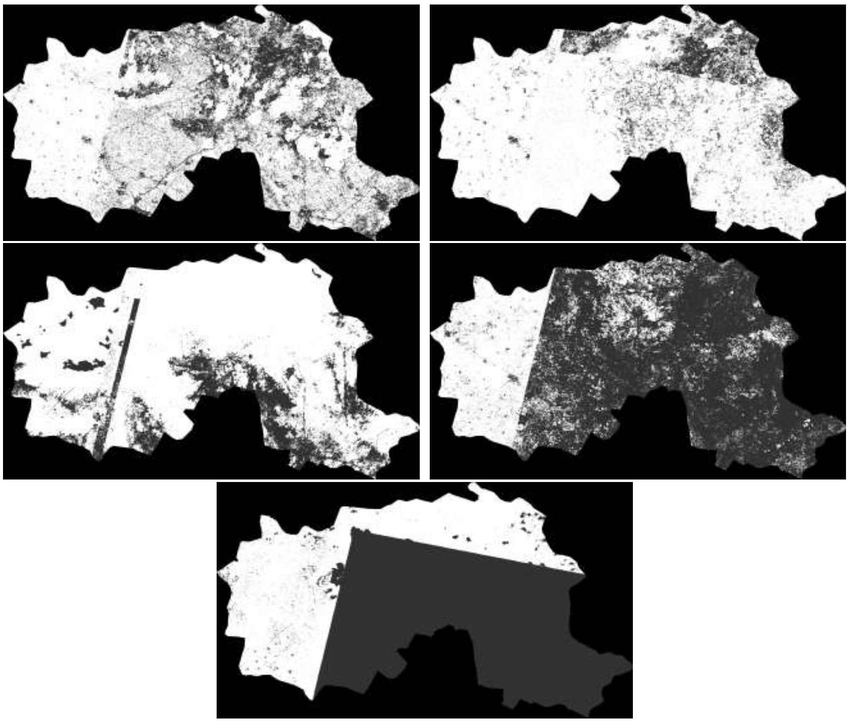


Figure 4.3: Gurgaon: Classified images, years 2014, 2015, 2016, 2017 & 2018

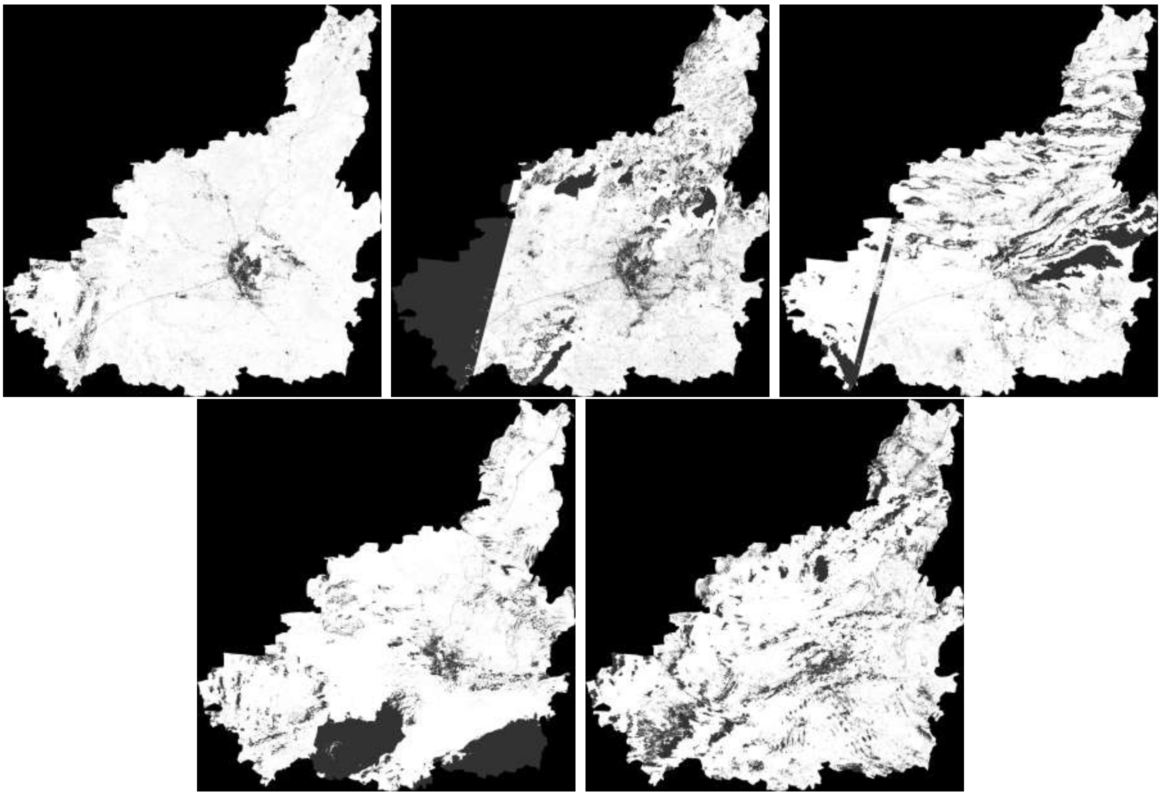


Figure 4.4: Jaipur: Classified images, years 2014, 2015, 2016, 2017 & 2018

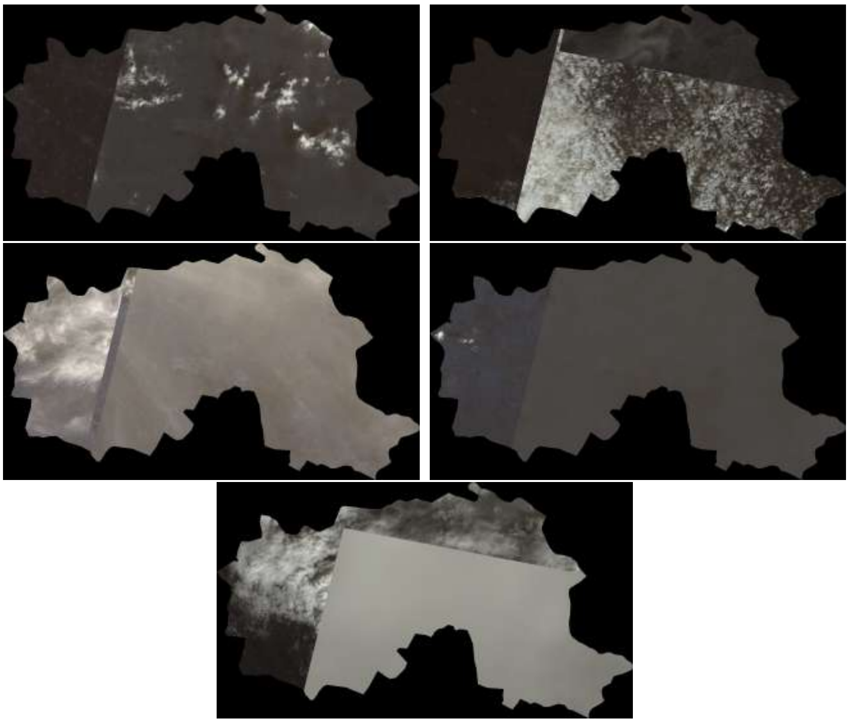


Figure 4.5: Gurgaon: Cloud Cover in Earth images, years 2014, 2015, 2016, 2017 & 2018

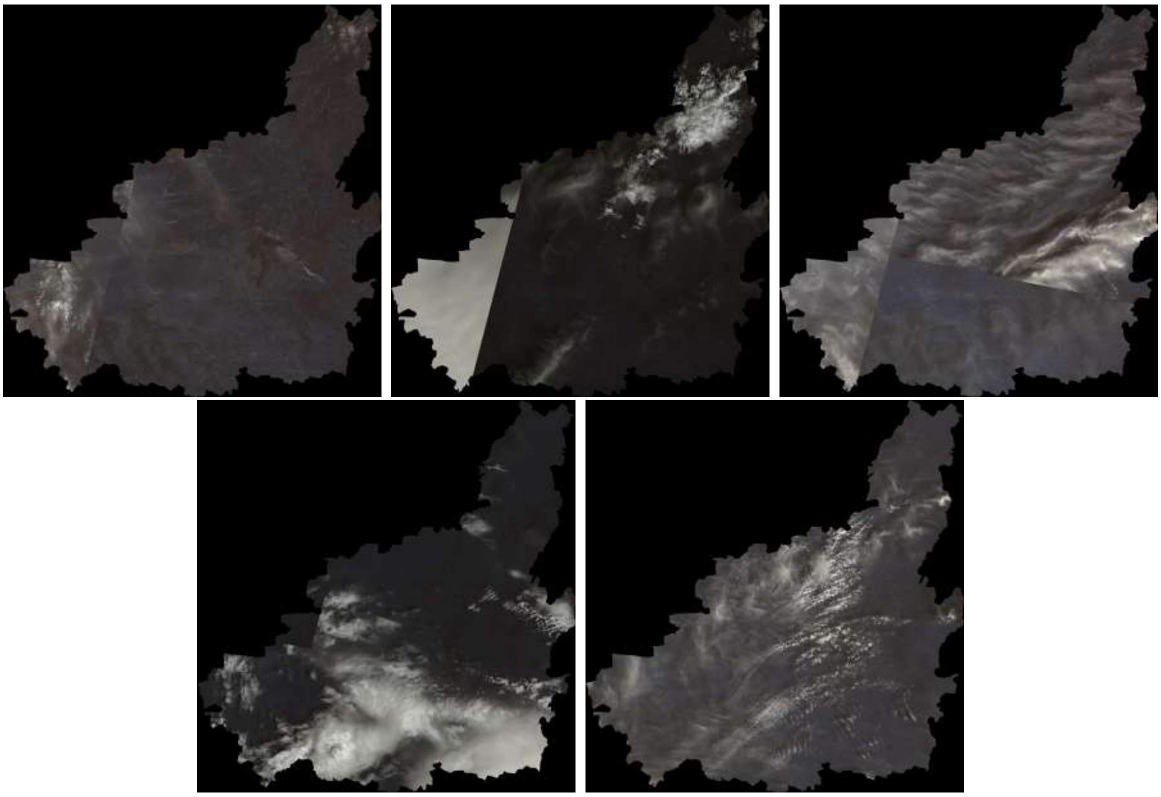


Figure 4.6: Jaipur: Cloud Cover in Earth images, years 2014, 2015, 2016, 2017 & 2018

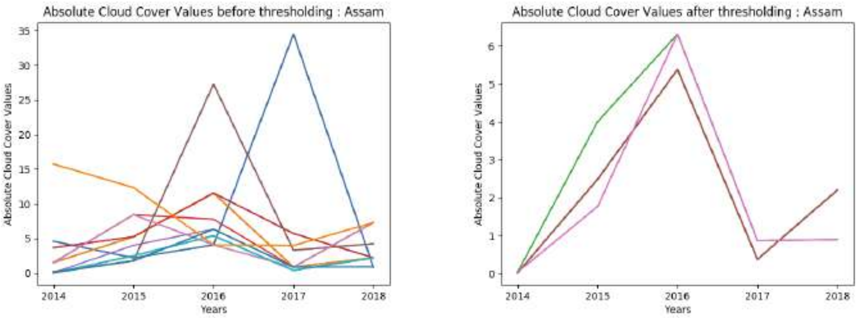
### 4.1.2 Accounting for Cloud Cover

1. As we observe the high cloud cover content randomly spread in the area, we use cloud masking module by Rodrigo Principe [RPM] to mask away the majority of clouds and apply maximum correction on the input Landsat images.

2. We try out two approaches to address the issue of high cloud cover indices. The first technique involves creating a composite by selecting the pixels from the least cloudy scene, by taking a median of all available values (multiple images are available for the given duration) for each pixel. While in the second technique, we calculate the cloud score of each pixel and use that as the quality band for a quality mosaic. On further experimentation, we find that taking median gives us better results and thus, we stick to the technique of median in further experimentation and analysis.

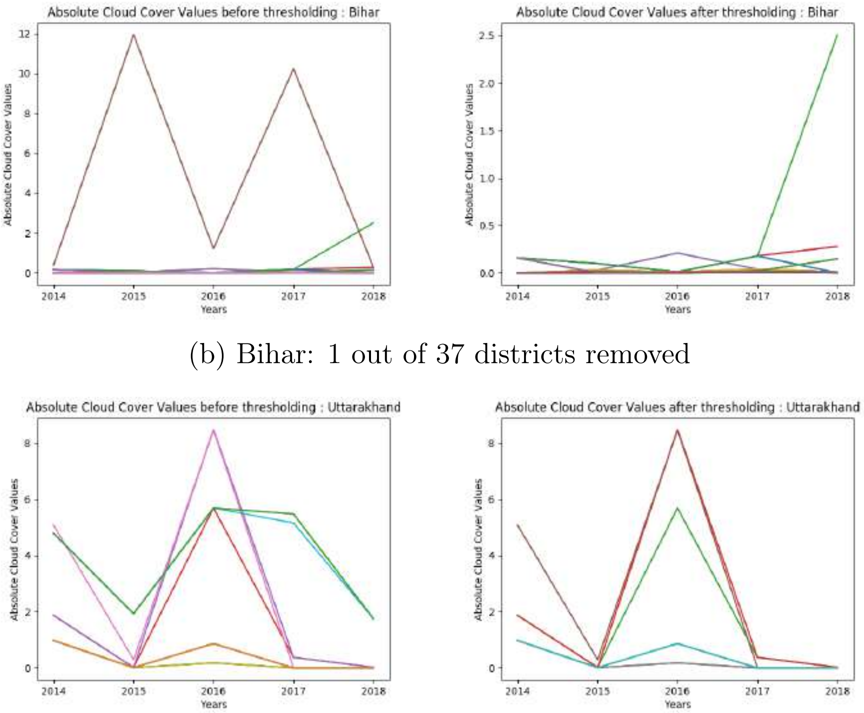
3. Since these masking techniques are quite generic without hampering other attributes available in the imagery, we observe that there are still many districts with high and noisy values of cloud cover. In most districts, we observe that the values for some years are particularly higher than those for other years. So we perform a state-level analysis to remove the districts that have extremely high absolute values of cloud cover and establish rules for selecting years to take for further analysis:

   - **Absolute Cloud Cover Values:** We put a threshold on the absolute value of cloud cover over the years, to eliminate the extremely high cloud cover districts, which can otherwise act as noise in our further analysis. The threshold value is then set to $t_h = 3$. This leads to the removal of any district with the average cloud cover value (over the 5 years) more than $t_h$, from further analysis. Among the 9 states being monitored, 6 states go undisturbed, while in the rest 3, a few districts are dropped [Assam: 15/22, Bihar: 1/37, Uttarakhand: 3/13 are dropped]. The absolute value plots before and after thresholding for these three districts can be found in Fig. 4.7.



(a) Assam: 15 out of 22 districts removed

Figure 4.7: Absolute Cloud Cover plots before and after thresholding

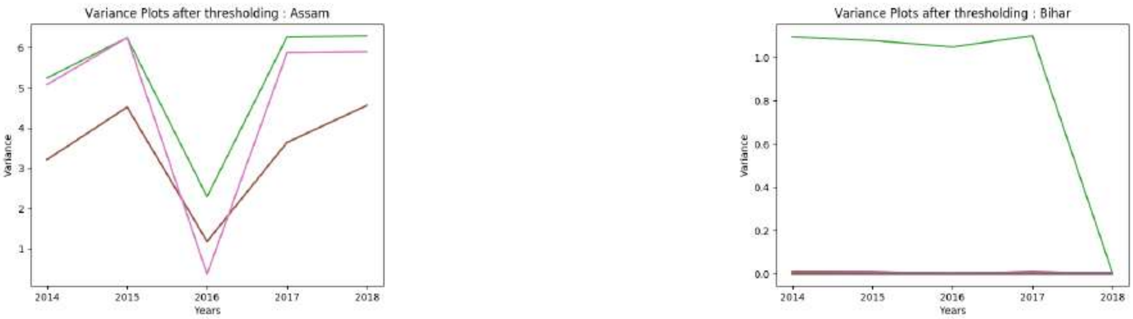


(b) Bihar: 1 out of 37 districts removed

(c) Uttarakhand: 3 out of 13 districts removed

Figure 4.7: Absolute Cloud Cover plots before and after thresholding (contd.)

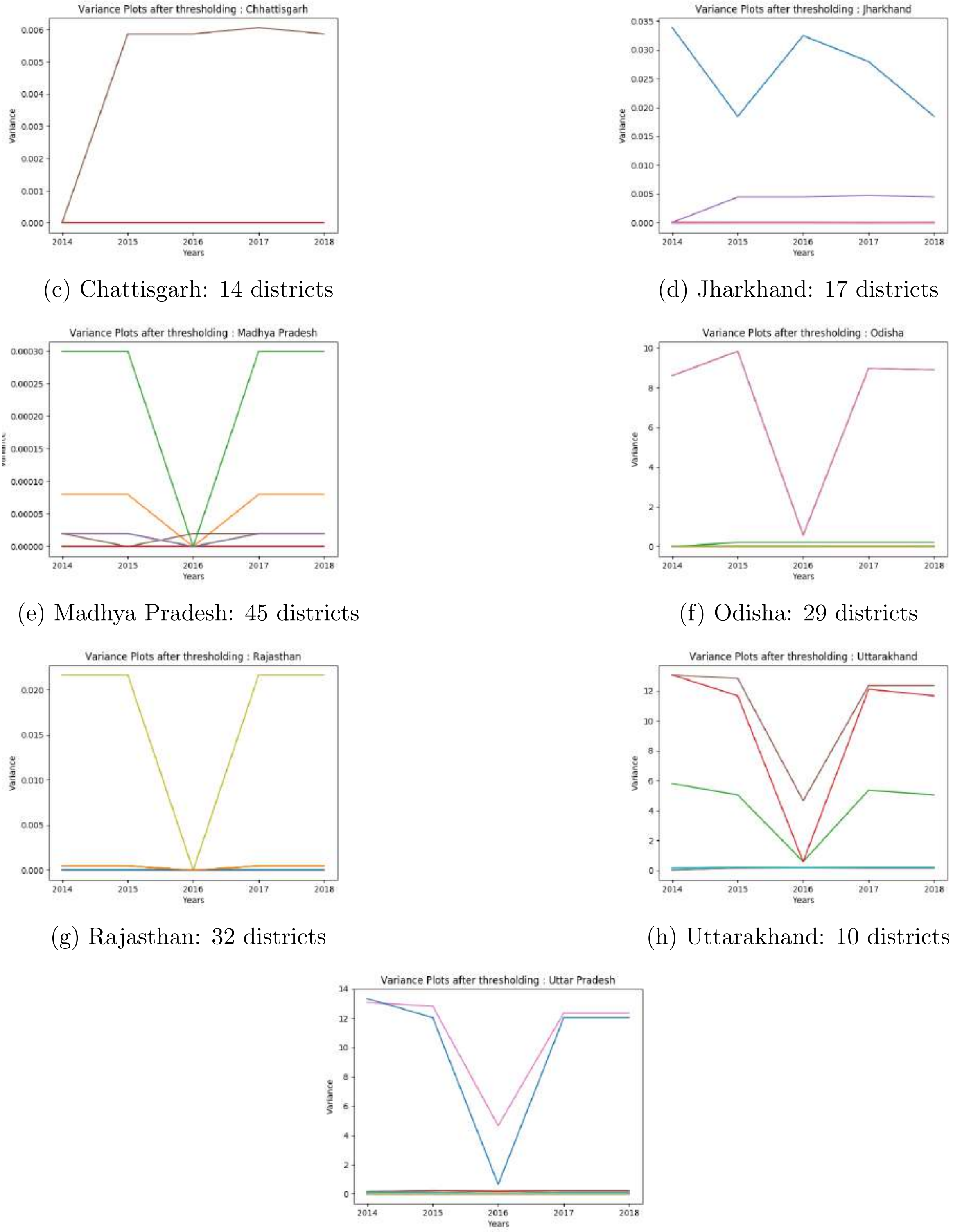
- **Variance in Cloud Cover Values:** After removing the highly cloudy districts, we further run a deeper analysis on each of the states to check for any outlier values. For each state, we plot the variance of the cloud cover values for every district, calculated by removing a year one by one. We understand that if for some year, the variance is lower than the other values, then that year is an outlier and likely to serve as a noise. From this analysis, we observe a dip at year 2016 for most of the states (Fig. 4.8), which is confirmed by visually looking at the Earth images and hence we skip this year from our further analysis. Further, to have minimum cloud cover, we do our analysis on the summer months of March and April.



(a) Assam: 7 districts

(b) Bihar: 36 districts

Figure 4.8: Year wise variance plots after thresholding



(c) Chattisgarh: 14 districts

(d) Jharkhand: 17 districts

(e) Madhya Pradesh: 45 districts

(f) Odisha: 29 districts

(g) Rajasthan: 32 districts

(h) Uttarakhand: 10 districts

(i) Uttar Pradesh: 69 districts

Figure 4.8: Year wise variance plots after thresholding (cont.)

The final classified images of these districts for the years 2014, 2015, 2017 and 2018, can be seen in Fig. 4.9 and Fig. 4.10. Here, the black pixels represent BU, while the white ones are NBU.

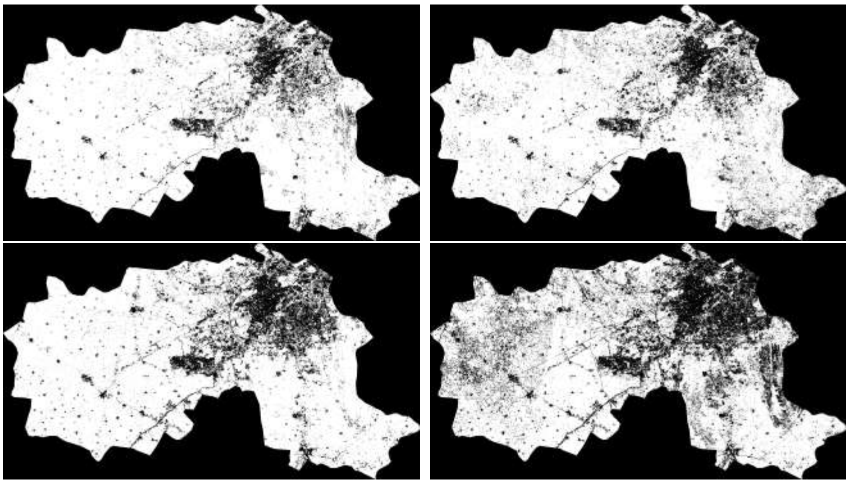


Figure 4.9: Gurgaon: Final classified images, years 2014, 2015, 2017 & 2018

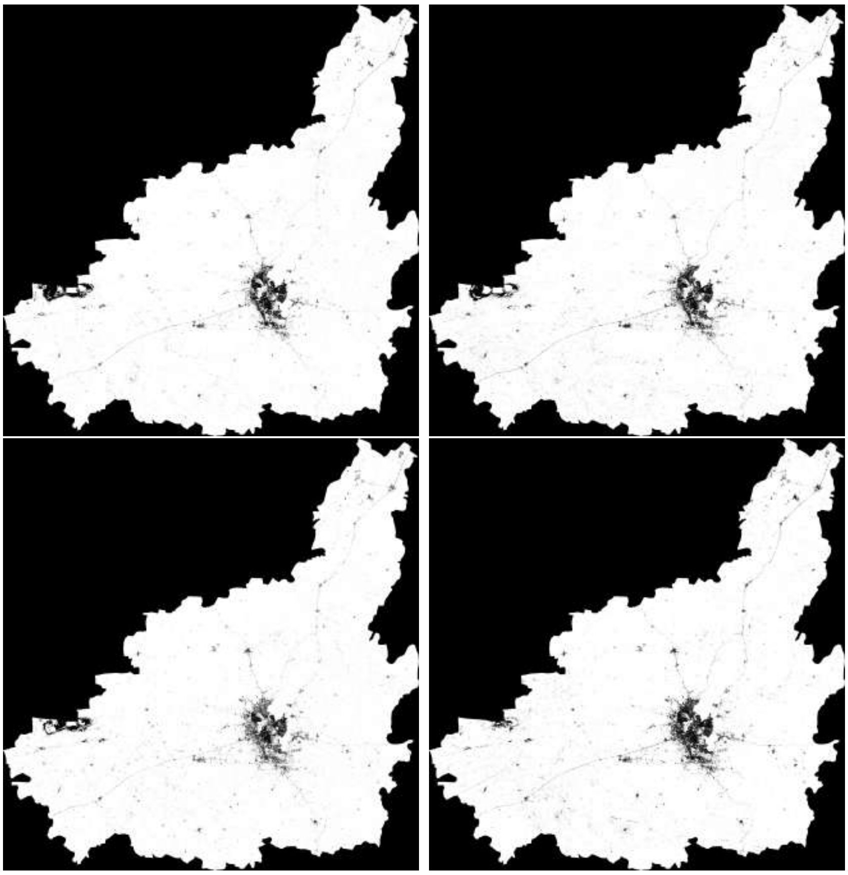


Figure 4.10: Jaipur: Final classified images, years 2014, 2015, 2017 & 2018

## 4.2 Generating urbanisation indices using classified images

### 4.2.1 Extent of change using Temporal Analysis

Now we proceed with analyzing the behaviour of our results from classification over the years 2014, 2015, 2017 and 2018. For this we run a pixel-wise analysis on classified images. To measure extent of change, we follow the following three steps:

1. Spatial Smoothing of classified images
2. Temporal Analysis
3. Thresholding and Classification
4. Validation of classification results

As a measure of extent of urbanisation, we get 3 parameters for each district as the percent of Constantly Non-Built up pixels, Constantly Built up pixels and Changing pixels.

#### 4.2.1.1 Spatial Smoothing of classified images

For each district, each pixel has a value of 0 and 1 for built-up and non built-up consecutively. Since the classification has been pixel-wise, there is no accounting for neighbours. In order to remove discrepancies where a pixel class differs from it's surrounding classes, we perform spatial smoothing on the classified images. After smoothing, each pixel value becomes a number within a range, depending upon the type of smoothing we apply. This is no more a binary classification and we call this as the measure of built-up value. We perform 2 kinds of post-classification smoothing on Gurgaon and Jaipur based on each pixel till levels of neighbours i.e. 1 level neighbours, 2 level neighbours and 3 level neighbours.

**Weighted Neighbourhood Score**

In this method, we replace each pixel by its neighbourhood count calculated as:

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

where:
$w = (2n+1)^2 - 1$
N is the set of all n level adjacent neighbours of $x_{ij}$

We apply this scoring for 1,2 and 3 levels of adjacent neighbours.

- **1 Level Adjacent Neighbours**

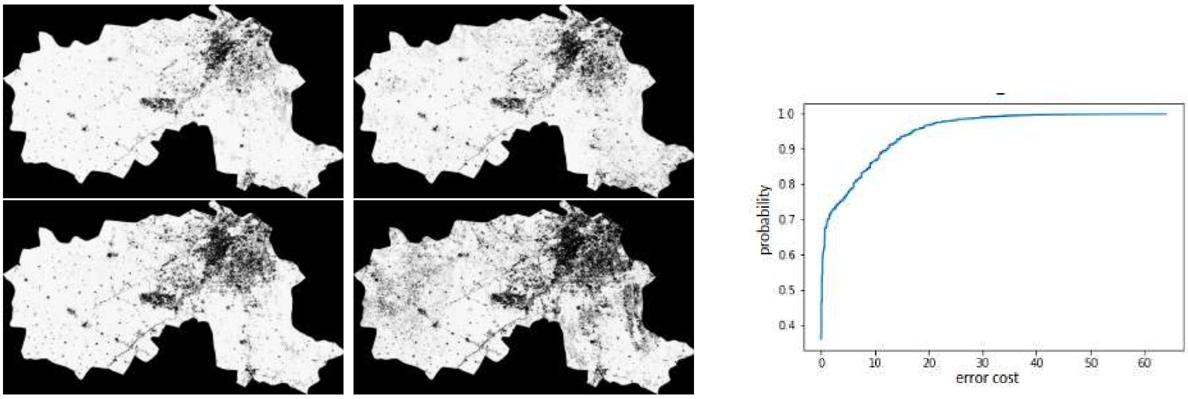


Figure 4.11: Gurgaon: After spatial smoothing using 1-level Weighted Neighbourhood Score (a) Classified results (b) CDF plot

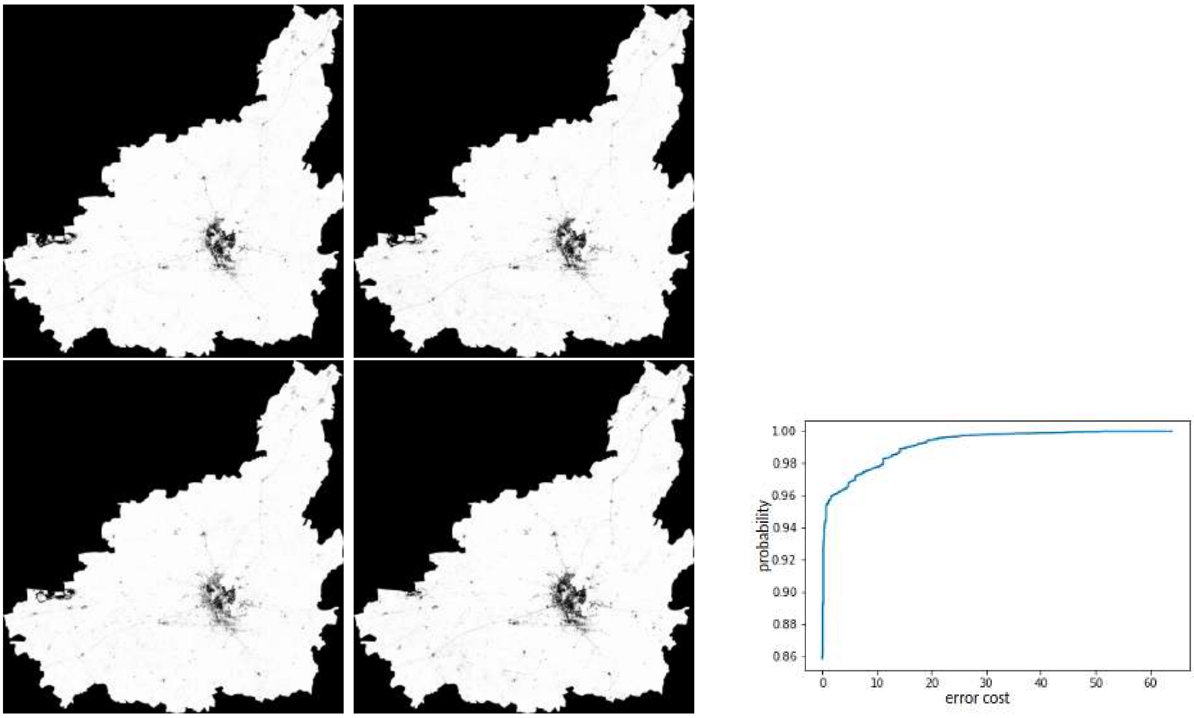


Figure 4.12: Jaipur: After spatial smoothing using 1-level Weighted Neighbourhood Score (a) Classified results (b) CDF plot

- **2 Level Adjacent Neighbours**

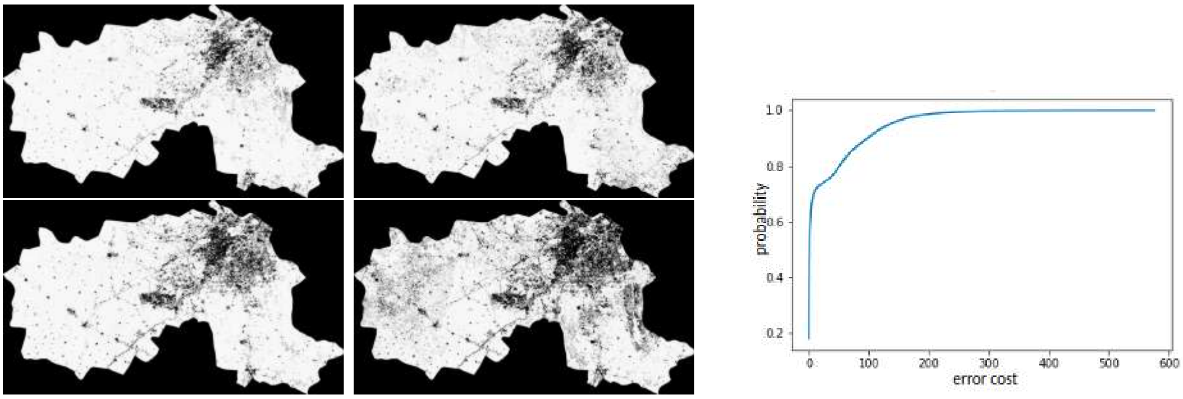


Figure 4.13: Gurgaon: After spatial smoothing using 2-level Weighted Neighbourhood Score (a) Classified results (b) CDF plot

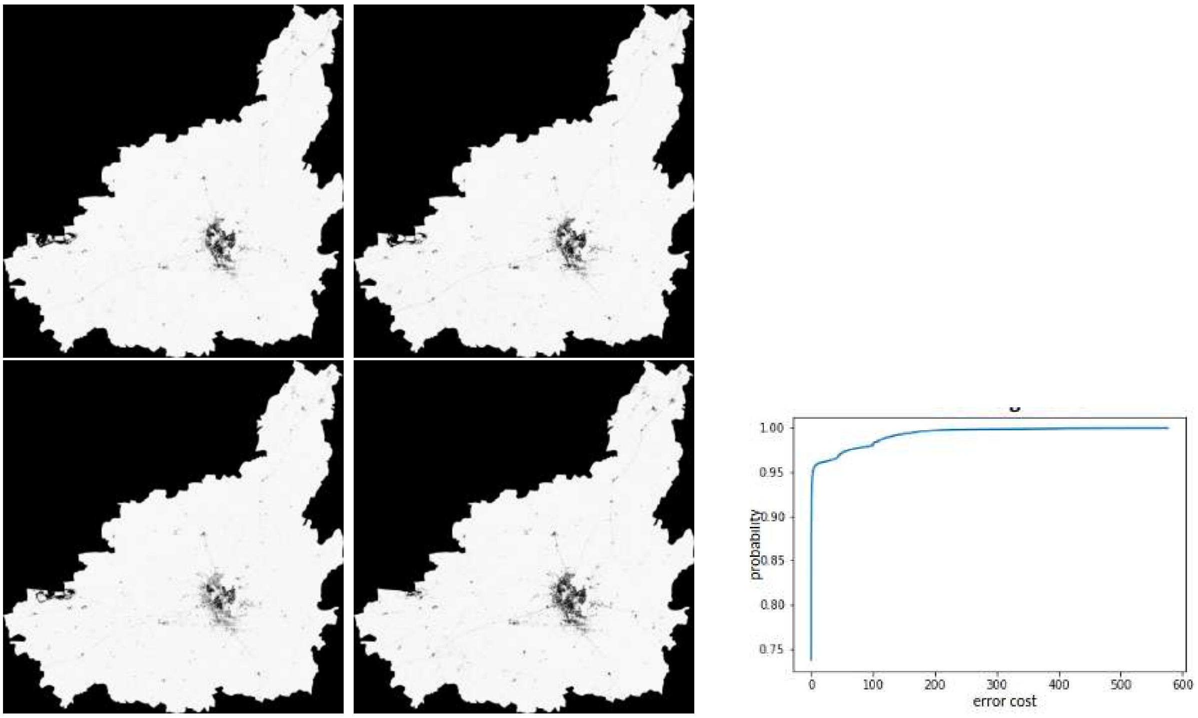


Figure 4.14: Jaipur: After spatial smoothing using 2-level Weighted Neighbourhood Score (a) Classified results (b) CDF plot

- **3 Level Adjacent Neighbours**

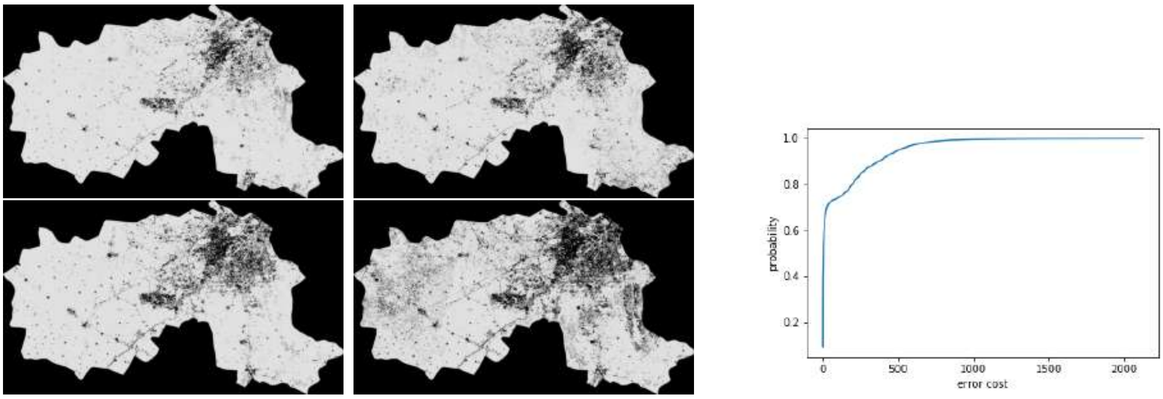


Figure 4.15: Gurgaon: After spatial smoothing using 3-level Weighted Neighbourhood Score (a) Classified results (b) CDF plot

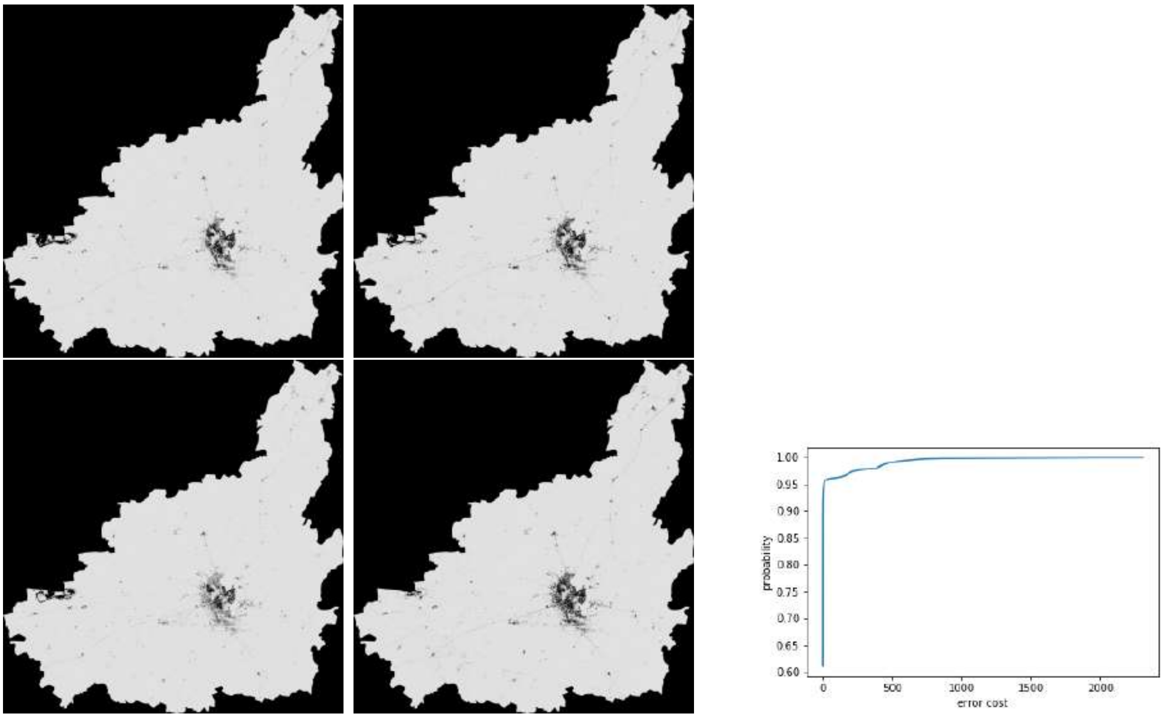


Figure 4.16: Jaipur: After spatial smoothing using 3-level Weighted Neighbourhood Score (a) Classified results (b) CDF plot

### Simple Box Blur followed by Gaussian filter

In this method, we replace each pixel by the value obtained after convolving it with its n level neighbours and then applying a gaussian filter to it. The value post convolution can be calculated as:

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

where:
N is the set of all n level adjacent neighbours of $x_{ij}$

We again apply this scoring for 1,2 and 3 levels of adjacent neighbours.

- **1 Level Adjacent Neighbours**

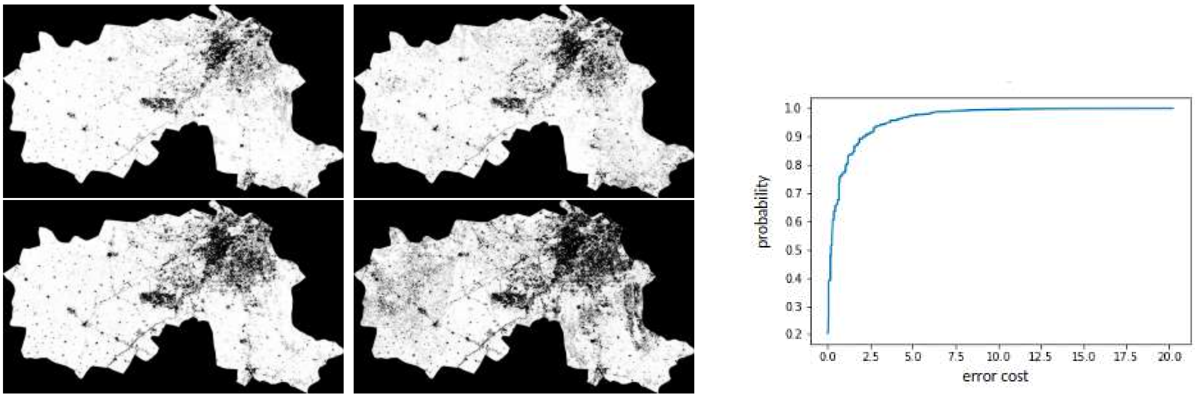


Figure 4.17: Gurgaon: After spatial smoothing using 1-level Simple Box Blur followed by Gaussian filter (a) Classified results (b) CDF plot

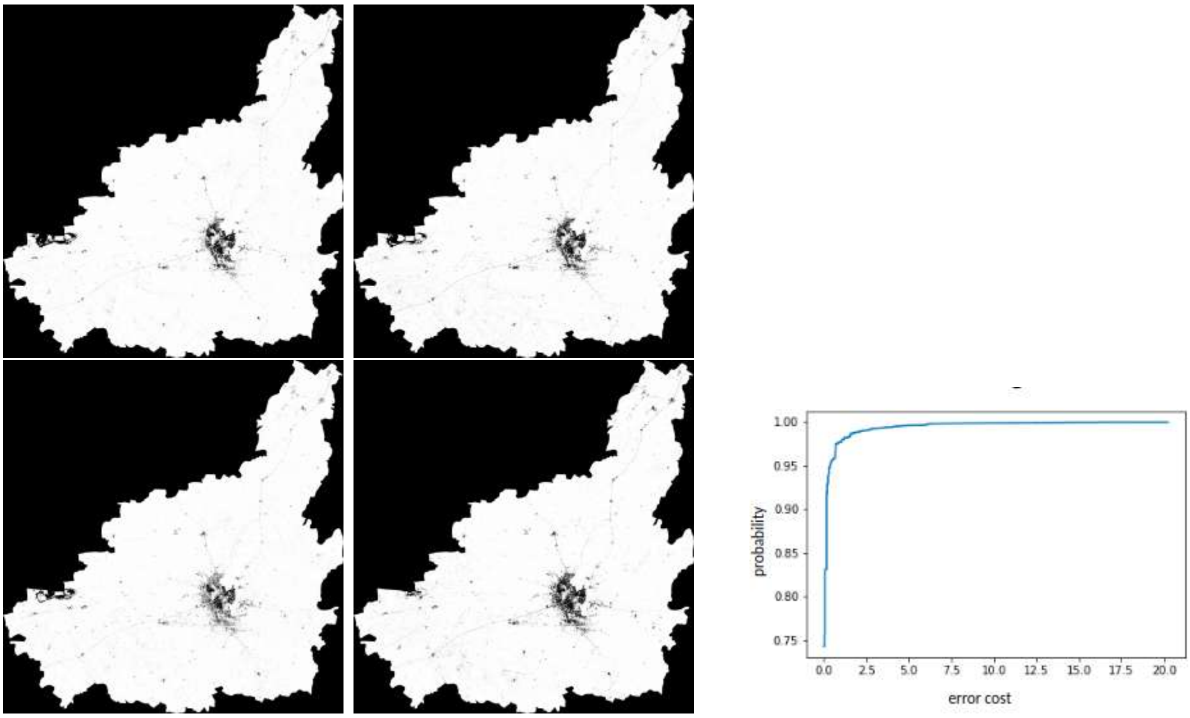


Figure 4.18: Jaipur: After spatial smoothing using 1-level Simple Box Blur followed by Gaussian filter (a) Classified results (b) CDF plot

- **2 Level Adjacent Neighbours**

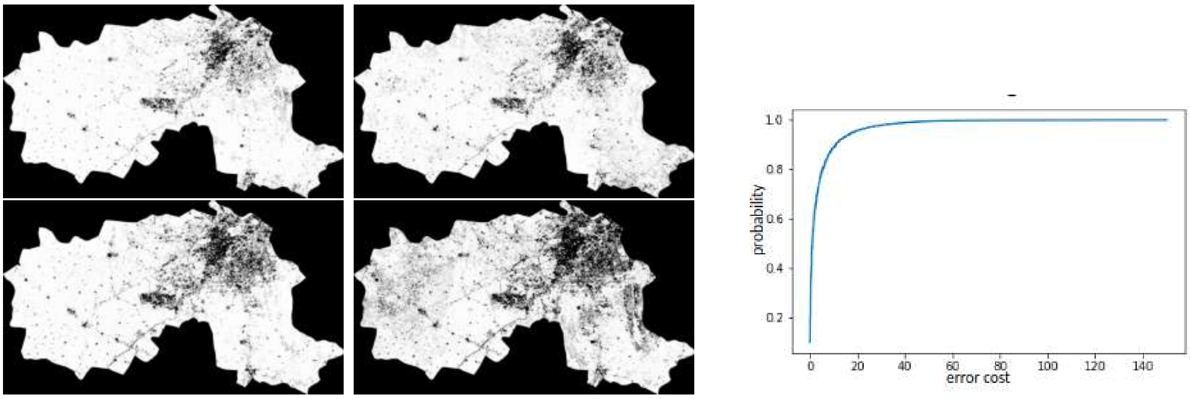


Figure 4.19: Gurgaon: After spatial smoothing using 2-level Simple Box Blur followed by Gaussian filter (a) Classified results (b) CDF plot

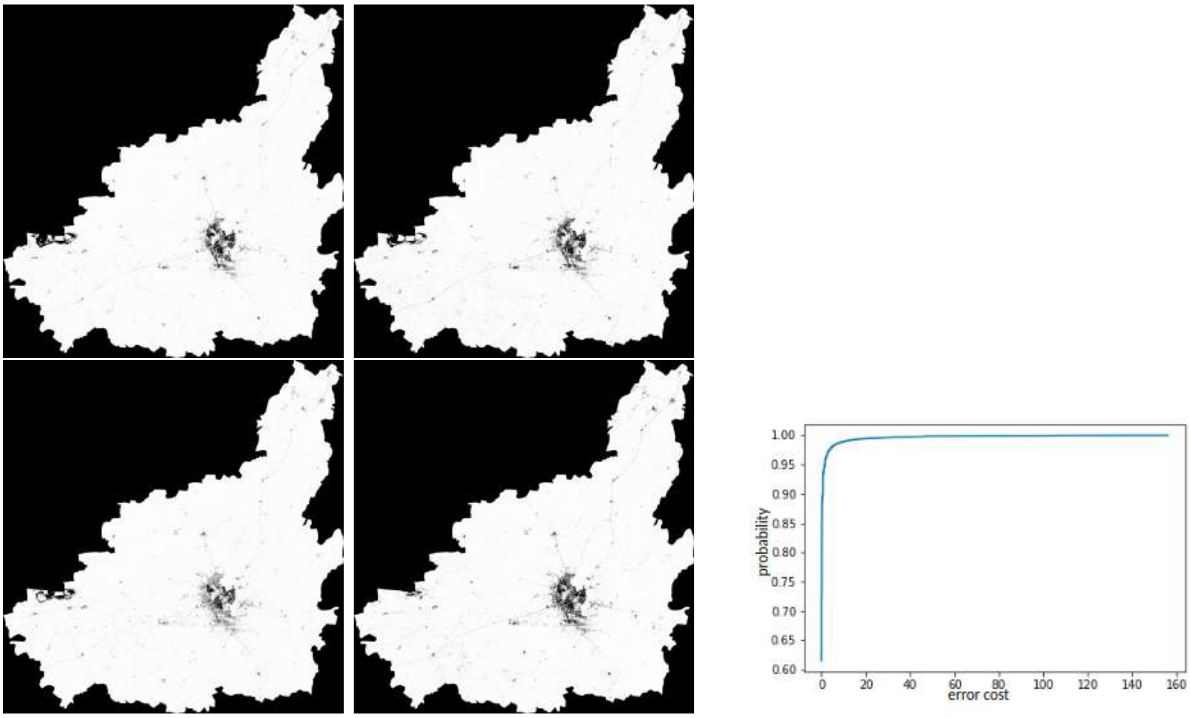


Figure 4.20: Jaipur: After spatial smoothing using 2-level Simple Box Blur followed by Gaussian filter (a) Classified results (b) CDF plot

- **3 Level Adjacent Neighbours**

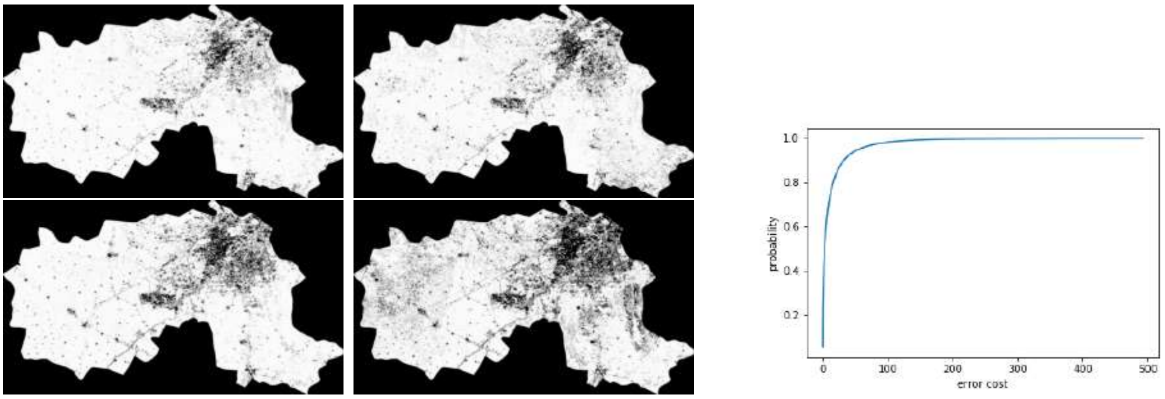


Figure 4.21: Gurgaon: After spatial smoothing using 3-level Simple Box Blur followed by Gaussian filter (a) Classified results (b) CDF plot

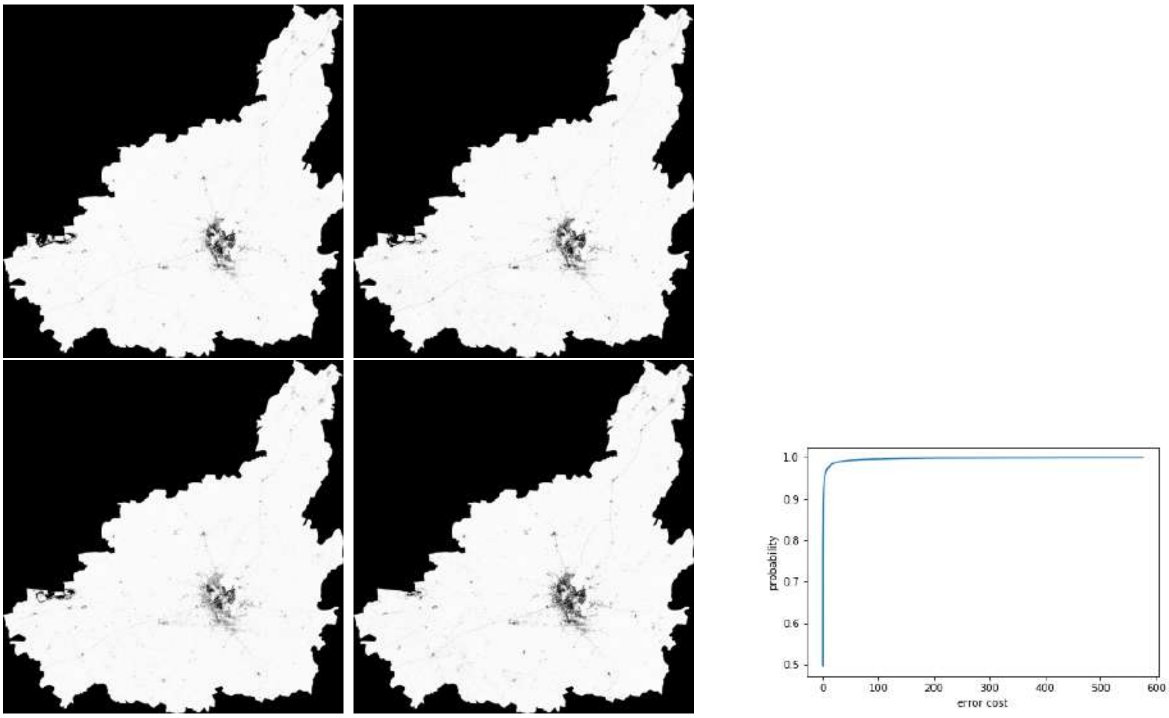


Figure 4.22: Jaipur: After spatial smoothing using 3-level Simple Box Blur followed by Gaussian filter (a) Classified results (b) CDF plot

#### 4.2.1.2 Temporal Analysis

Now we observe the patterns of change of the measure of built-up values for each pixel across the period of the four years. For this, we use Linear Regression to fit a line on the 4 values measure of built-up for each pixel and calculate the error in prediction. After applying spatial smoothing to the district, we apply linear regression on the set of values of each pixel across 2014-2018 (except 2016) and get a value of slope and intercept. We then calculate the mean squared error of prediction for all the 4 values of pixel across these years on their corresponding regression lines. We then plot the cdf curves for all the values of mean squared error obtained for each district and use this curve to observe patters of urbanisation in the district.

From the cdf plot we observe 3 regions as follows:

**Region 1:** Points very close to 0 error. These are generally the pixels that have been constantly BU or NBU throughout the years.

**Region 2:** Points in the curved region. These are the pixels that have changed throughout the years.

**Region 3:** Points in the high cost region that make the asymptote. These are the pixels that are erratically changing between NBU and BU values throughout the years. We can these points as unreliable and should not be considered for deriving indices.

We observe that while for 1 level neighbours, the cdf plots are quite rough which shows that the smoothing method is not very successful in creating uniform regions. With the increase in the levels of neighbours in consideration we get smoother curves, which are shifted closer to 0 error. For 3 level neighbours the CDF plots indicate that more and more pixels become closer to 0 with the points in region 2 becoming lesser. This means due to more rigorous smoothing, we are losing the changing points and thus moving all the pixels to constantly BU or constantly NBU states. In order to preserve the changing points as well as achieve uniform smooth regions, we thus take 2-level neighbours for further analysis. Also we observe that the cdf plots for simple box blur followed by gaussian filter is monotonically increasing with more favourable points, thus in the following discussion we work with 2-level box blur followed by gaussian spatial smoothing, the CDF plot for which can be found in Fig. 4.23. We move on to classify each pixel in three classes: Constantly Built-up, Constantly Non Built-up and Changing.

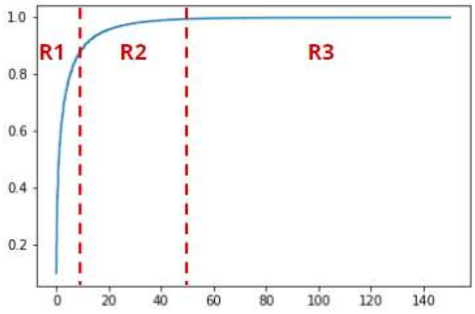


Figure 4.23: CDF plot with R1, R2 and R3

#### 4.2.1.3 Determining thresholds and Classification

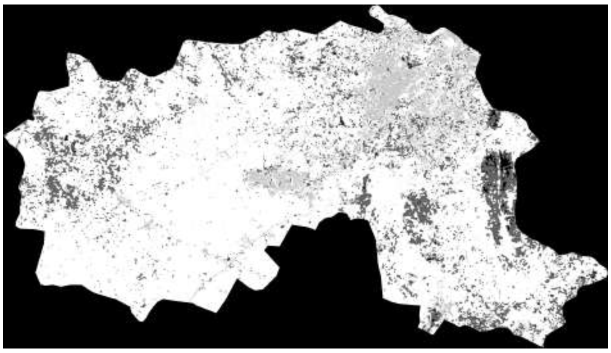
In order to classify our pixels in the above mentioned categories, we need to find the boundaries of the changing region. For this, we fit a curve to the discrete cdf values and find the double derivative of the fitted curve. We observe that the double derivative is a negative impulse curve and the region around the 'impulse' is essentially the Region 2 that is required. On observing the absolute value of double derivative, the value suddenly shoots up around the curvature region and soon it drops down. Thus on finding the values at which the double derivate shoots up and then becomes close to 0, we mark the thresholds for the three region. Once we have the thresholds, we classify pixels in the three categories as follows:

1. The pixels that lie in Region 1 and are Non Built-up in 2014 (initial year) are classified as Constantly NBU.
2. The pixels that lie in Region 1 and are Built-up in 2014 (initial year) are classified as Constantly BU.
3. The pixels that lie in Region 2 are classified as Changing.

Upon classification, we finally generate the following indices determining the extent of change of urbanisation:

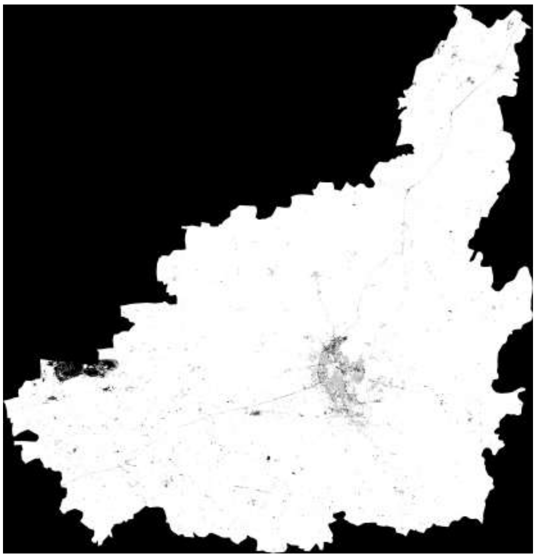
- % pixels that are constantly non-built up
- % pixels that are constantly built up
- % pixels that have changed throughout the years

Marking the points in these regions us a spatial map, figures for which can be seen in Fig. 4.24.



(a) Classification Map for Gurgaon

Figure 4.24: Classification Map



(b) Classification Map for Jaipur

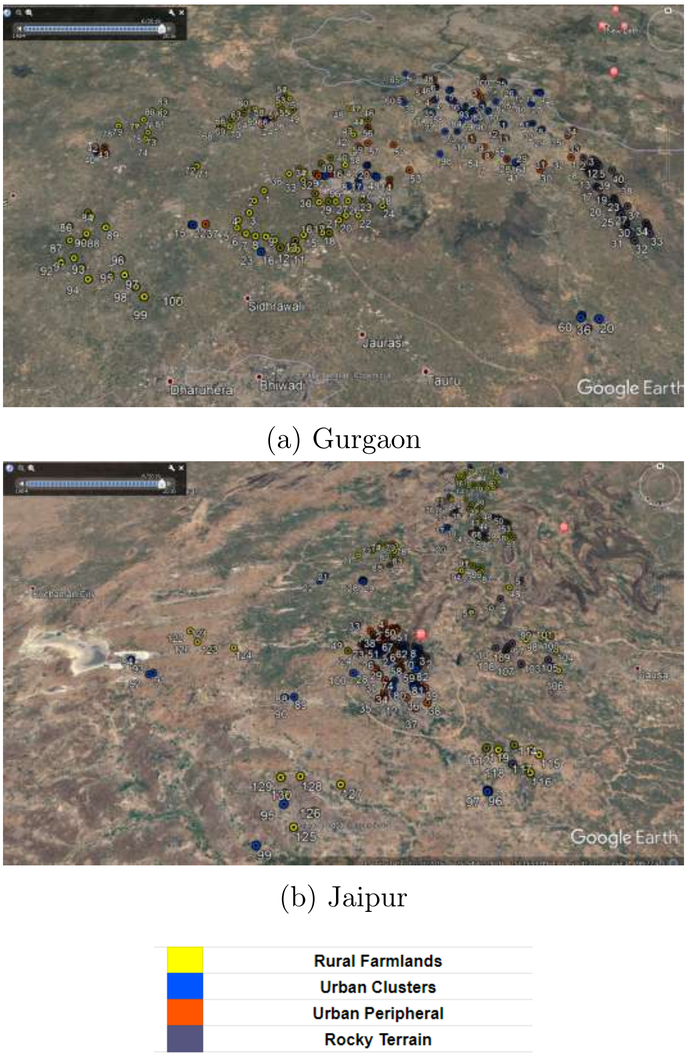
| | |
|---|---|
| □ | Constantly Non-builtup |
| ░ | Constantly Builtup |
| ▓ | Changing |
| ■ | Unreliable |

Figure 4.24: Classification Map (contd.)

#### 4.2.1.4 Validation of results

Once we have obtained the parameters, it is important to validate them against the ground truth. We therefore generate a ground truth dataset of 300 points (pixels of 30m X 30m), manually labelled as constantly non-builtup, constantly built-up and changing using historical imagery. This is done using the following steps:

1. **Marking points on the map of Gurgaon and Jaipur :** We mark 300 points each on the maps of Gurgaon and Jaipur using MATLAB. Which choosing the points, the following distribution of the points is ensured:
    - 100 points from rural farmlands
    - 100 points from urban clusters
    - 60 points from urban peripheral areas
    - 40 points from rocky terrains
2. **Getting shapefiles :** The Longitude and Latitude values of the points obtained from the marked points in the above process are used to convert these set of points into shapefiles. This is done using QGIS.
3. **Overlaying points and developing groundtruth :** The shapefiles are then imported into Google Earth Pro (Fig. 4.25) and using historical imagery, we manually mark the points as Constantly NBU, Constantly BU and Changing.



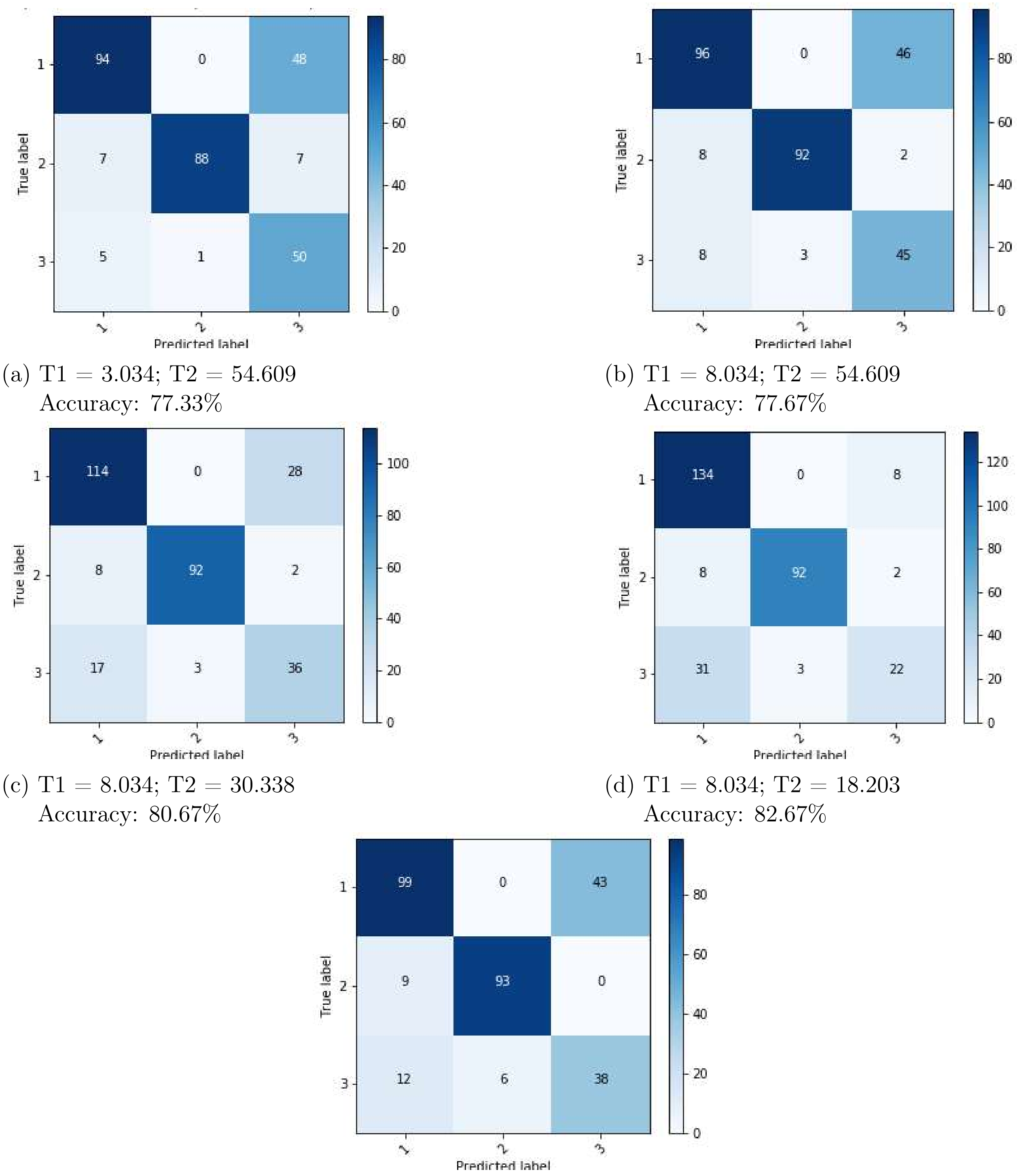
(a) Gurgaon

(b) Jaipur

Figure 4.25: Distribution of Groundtruth Dataset on Google Earth Pro

Once we have developed the Groundtruth, we then estimate the accuracy of our classification and how it changes with moving both the thresholds by certain differences for both Gurgaon and Jaipur.

For Gurgaon, we get the following changes in accuracies on moving thresholds (Fig. 4.26):



(a) T1 = 3.034; T2 = 54.609
Accuracy: 77.33%

(b) T1 = 8.034; T2 = 54.609
Accuracy: 77.67%

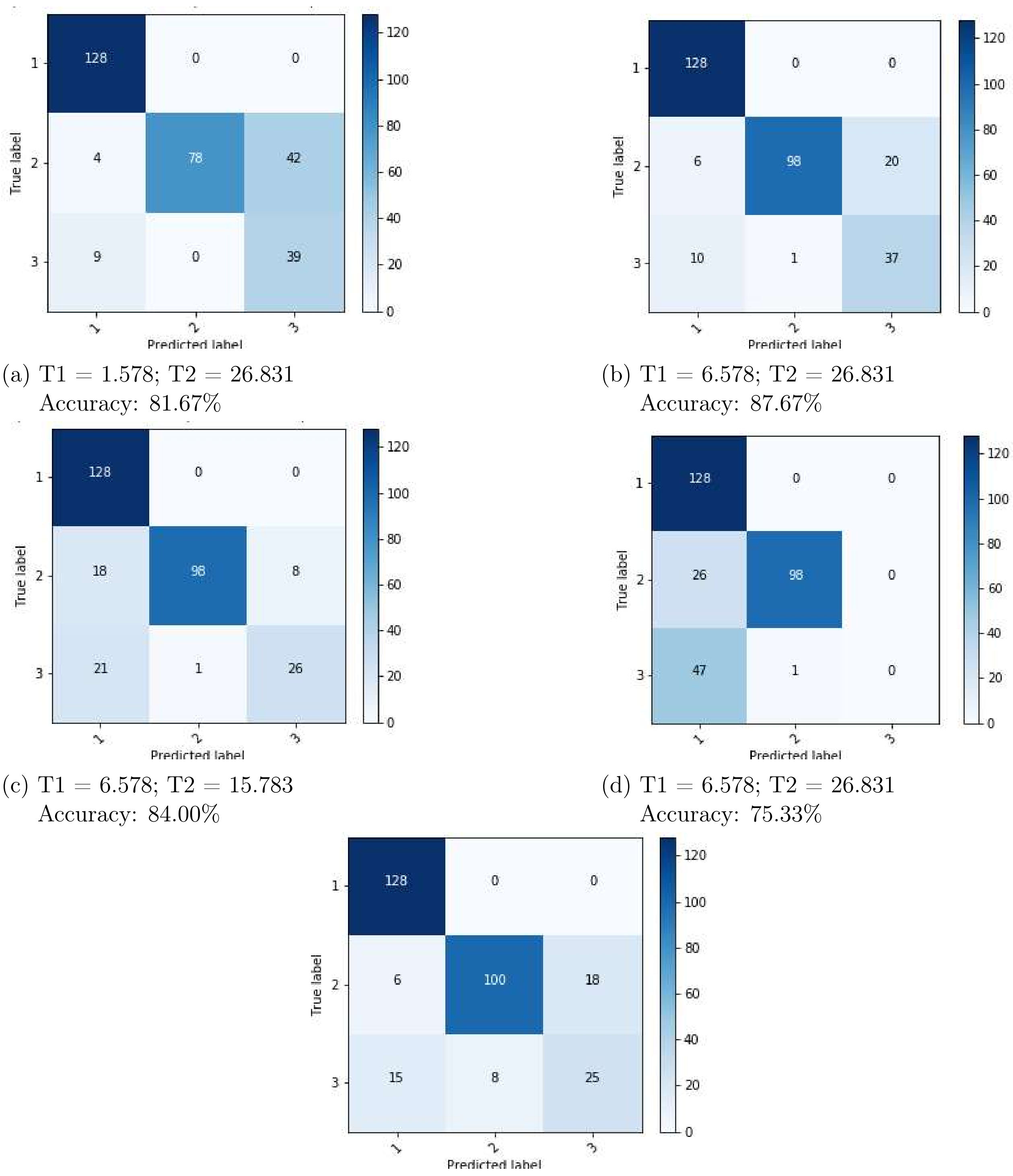
(c) T1 = 8.034; T2 = 30.338
Accuracy: 80.67%

(d) T1 = 8.034; T2 = 18.203
Accuracy: 82.67%

(e) T1 = 13.034; T2 = 54.609
Accuracy: 76.67%

Figure 4.26: Gurgaon: Accuracies & Confusion matrices for different thresholds

For Jaipur, we get the following changes in accuracies on moving thresholds (Fig. 4.27:



(a) T1 = 1.578; T2 = 26.831
Accuracy: 81.67%

(b) T1 = 6.578; T2 = 26.831
Accuracy: 87.67%

(c) T1 = 6.578; T2 = 15.783
Accuracy: 84.00%

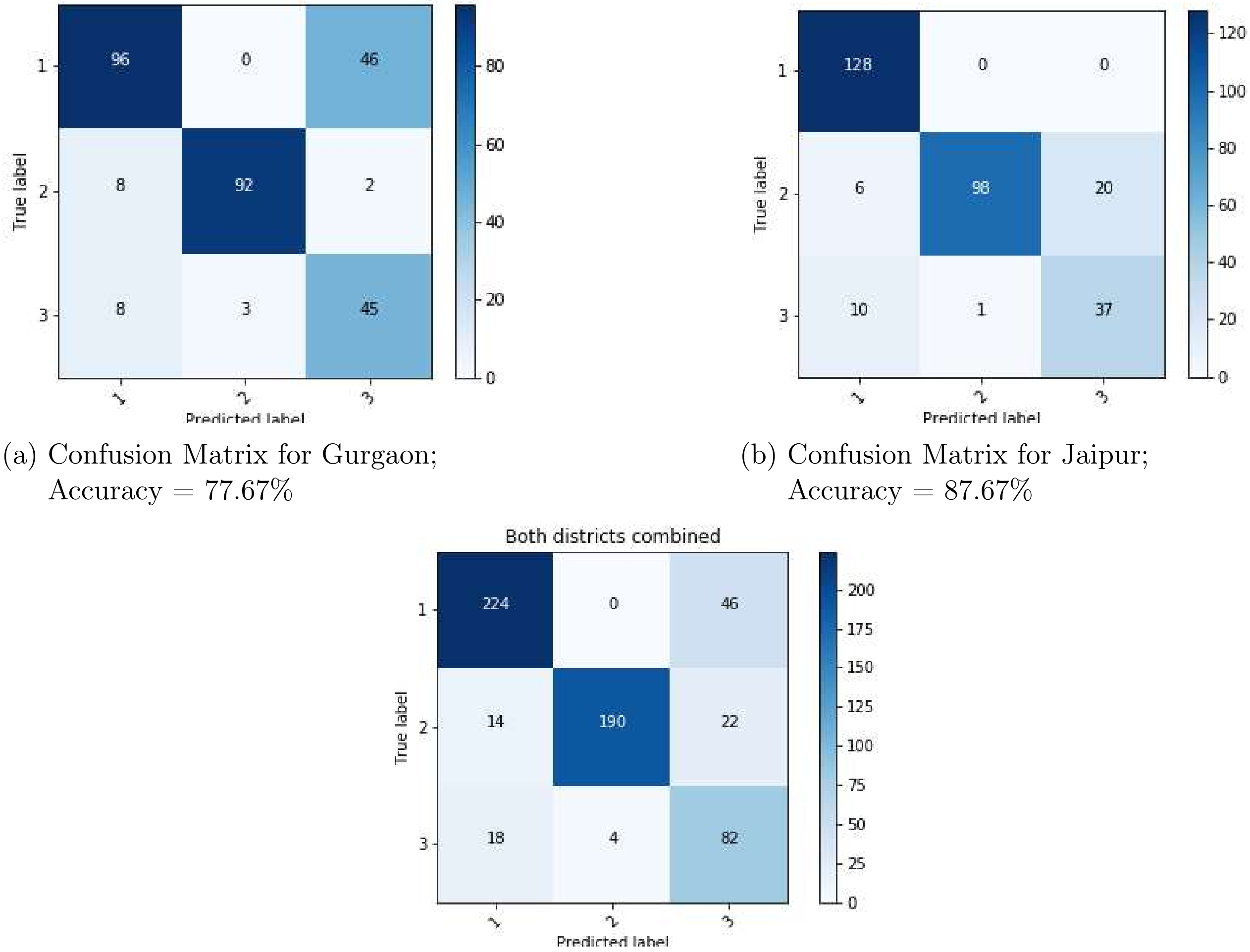
(d) T1 = 6.578; T2 = 26.831
Accuracy: 75.33%

(e) T1 = 11.578; T2 = 26.831
Accuracy: 84.33%

Figure 4.27: Jaipur: Accuracies & Confusion matrices for different thresholds

This helps us determine the optimal cut-offs for thresholds while automatic thresholding using double derivative of the cdf curves. For this choice of cutoffs, the values of threshold1 and threshold2 for Gurgaon are 8.034 and 54.609 respectively, and for Jaipur are 6.578 and

26.831 respectively. On combining the dataset for the two districts, we get an optimal overall accuracy of 82.67%. The confusion matrices for Gurgaon, Jaipur and both combined for these values can be found in Fig. 4.28.



(a) Confusion Matrix for Gurgaon; Accuracy = 77.67%

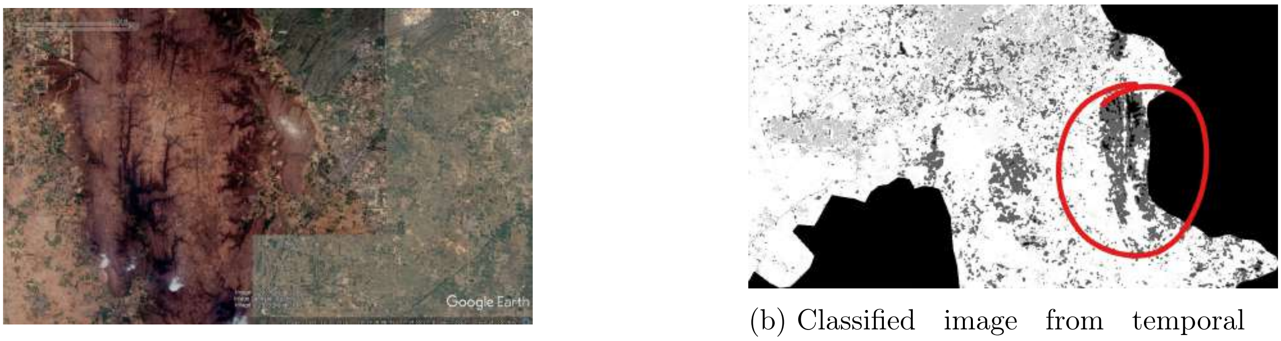
(b) Confusion Matrix for Jaipur; Accuracy = 87.67%

(c) Confusion Matrix for Gurgaon & Jaipur combined; Accuracy = 82.67%

Figure 4.28: Confusion Matrices for optimal thresholds

### Analysis of unreliable points in Gurgaon

We observe that in Gurgaon, there is a prominent area which is like a rocky terrain (Fig. 4.29a). The pixels in this area are classified erratically over the span of years because our training dataset does not include these kind of points. Due to erratic patterns, these points lie in majorly high cost region of the cdf plots and are classified as unreliable or changing depending on the choice of thresholds (Fig. 4.29b). These points in the rocky terrain are marked as Constantly Non-Builtup in the groundtruth.



(a) Google earth image

(b) Classified image from temporal analysis

Figure 4.29: Patch of rocky terrain in Gurgaon

We proceed with the analysis of how capable our classifier is in identifying the rocky terrains by looking at the accuracy of classification in the following ways:

1. Mapping all the points lying in the unreliable region to Constantly Non-Builtup during prediction
    - For threshold 1 as 8.034 and threshold 2 as 54.609, we get an overall accuracy of 77.67% with the following confusion matrix:

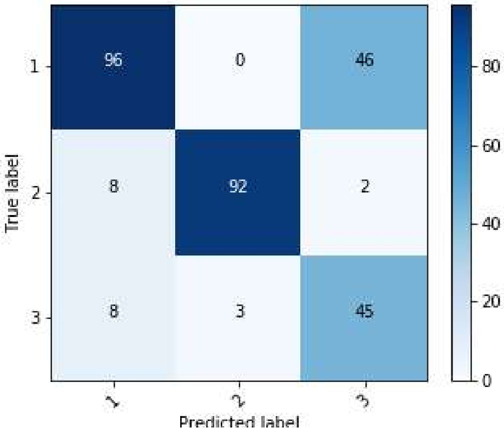


Figure 4.30: Confusion Matrix: Accuracy = 77.67%

- For threshold 1 as 8.034 and threshold 2 as 21.237, we get an overall accuracy of 82.67% with the following confusion matrix

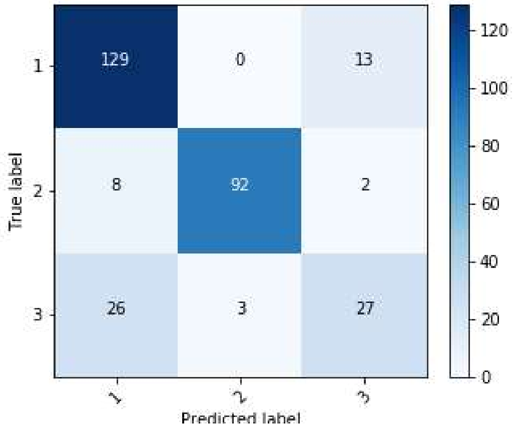


Figure 4.31: Confusion Matrix: Accuracy = 82.67%

2. Ignoring all the points lying in the unreliable region from accuracy calculations
   - For threshold 1 as 8.034 and threshold 2 as 54.609, we get an overall accuracy of 77.85% with the following confusion matrix

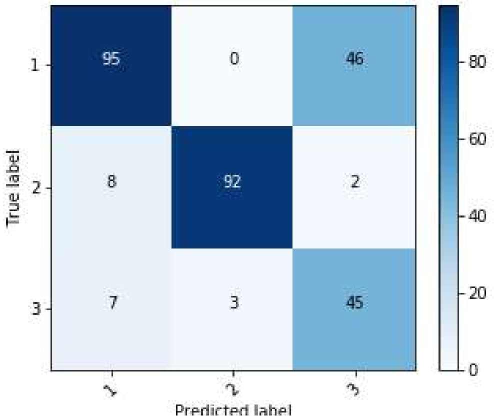


Figure 4.32: Confusion Matrix: Accuracy = 77.85%

Although we observe that there is not much change in accuracy, there is a large chunk (46 to be precise) of points that are actually Constantly Non-Builtup but are being predicted as Changing. This is because these points lie in the rocky terrain region.

   - For threshold 1 as 8.034 and threshold 2 as 21.237, we get an overall accuracy of 86.64% with the following confusion matrix

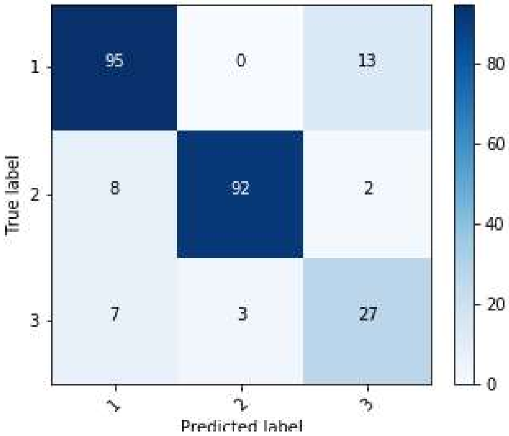


Figure 4.33: Confusion Matrix: Accuracy = 86.64%

> This is because on reducing the value of threshold 2, we shrink the Changing region as well. So although there is a better prediction of rocky terrain as unreliable, it comes along with an underestimation of the points that are Changing in a district.

### 4.2.2 Patterns in change using Blob Detection

As discussed before, urbanisation cannot be just represented by the percentage change of builtup area (i.e. fraction of area urbanised). The spatial distribution and its growth patterns also convey an important aspect, like whether the growth in a district is unicentered or multicentered. The urbanisation could be because of expansion around the existing urban clusters or emergence of new settlement areas altogether. To understand the clustering in the district, we need to identify the blobs of settlement areas and then monitor the changes over the given duration of time.

We use the technique of blob detection on the smoothened classified images obtained from the previous temporal analysis. For this, we identify the connected builtup components which can be interpreted as the blobs, by doing a simple breadth first search. Once identified, we generate a few parameters once on regions which stay constant (BU or NBU) throughout, i.e. don't change in the duration of 5 years and once on the regions which remain constant or change over this duration. These regions can be identified from the temporal analysis done for the extent of change. To understand the distribution and concentration of the urban clusters, we use the Herfindahl-Hirschman Index (HHI), which here is used to measure the concentration for a district, based on the number of clusters and their share in the urbanisation. HHI Index is defined as follows:

$$HHI = s_1^2 + s_2^2 + s_3^2 + ....s_n^2$$

where:

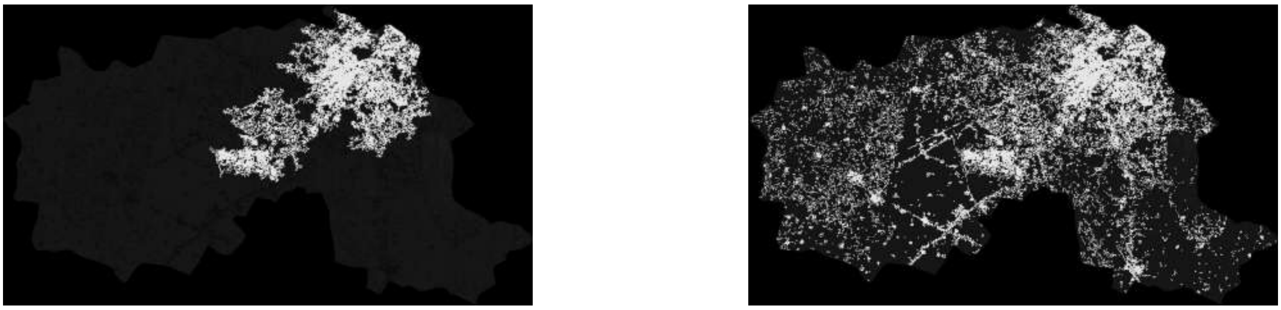
$s_n$ is the market share percentage of firm, n expressed as a whole number, not a decimal

In each of these, the unreliable points obtained from the previous analysis are skipped. The parameters generated from this analysis are:

- Constant regions:
  1. Number of blobs in the district
  2. % of BU pixels (out of the total BU pixels) present in the largest blob
  3. HHI Index
- Constant & changing regions:
  1. Number of blobs in the district
  2. % of BU pixels (out of the total BU pixels) present in the largest blob
  3. HHI Index

For Gurgaon, the results can be found in Fig. 4.34 and Fig. 4.35.

**Regions which are constant (no unreliable or changing points):**



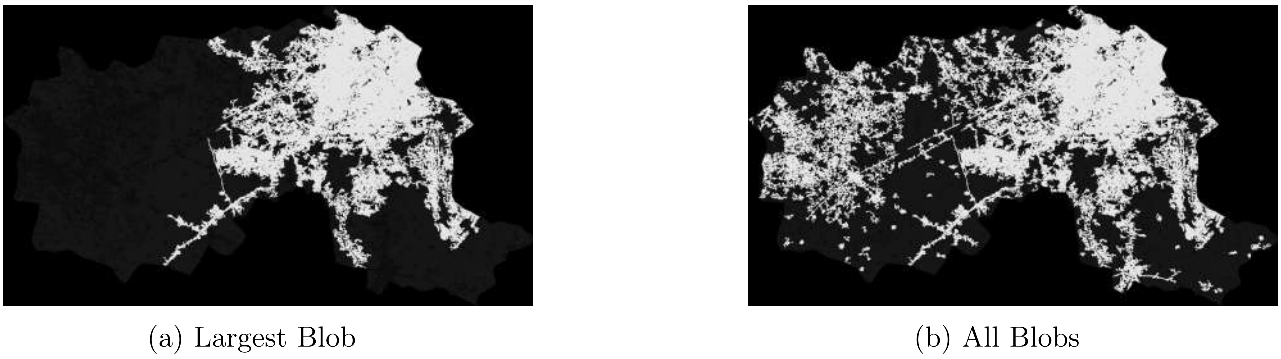
(a) Largest Blob

(b) All Blobs

Figure 4.34: Blobs: For constant regions

- Num of blobs : 522
- %BU in the largest blob : 56.2209
- HHI Index : 3240.73

**Regions which are constant + changing (no unreliable points):**



(a) Largest Blob (b) All Blobs

Figure 4.35: Blobs: For constant & changing regions

- Num of blobs : 127
- %BU in the largest blob : 66.9556
- HHI Index : 4839.49

We therefore obtain 6 parameters to represent the patterns of change for a district, by analysing the spatial distribution of the urban clusters and the changes over the given time period.

# Chapter 5

# RESULTS

Using our analysis and validation, we hereby have a pipeline to generate the 9 parametes to estimate urbanisation in a district using spatial analysis on remotely sensed imagery. For our complete analysis on Gurgaon and Jaipur, the parameters generated can be found in Table. 5.1.

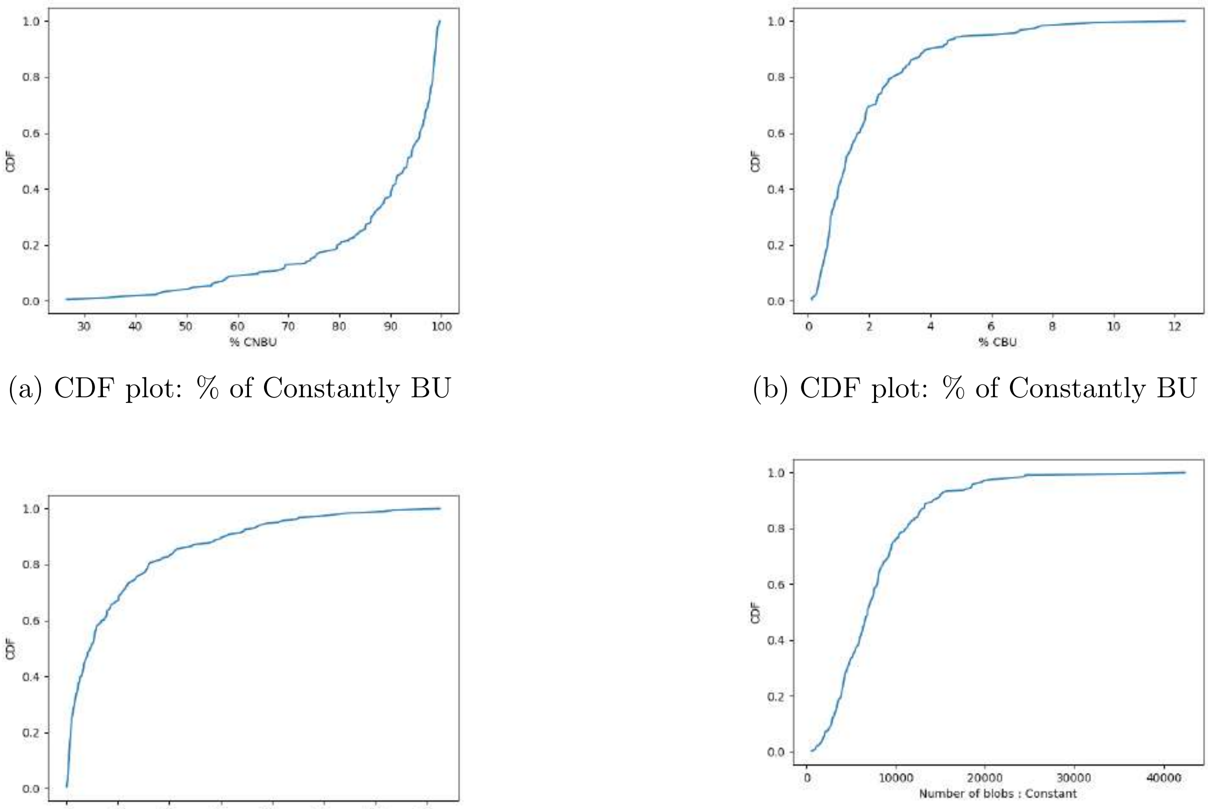
| **Paramter** | **Gurgaon** | **Jaipur** |
|---|---|---|
| % of Constantly NBU | 78.1858 | 97.36003 |
| % of Constantly BU | 8.2863 | 1.5895 |
| % of Changing | 13.5279 | 1.0505 |
| Num of blobs for constant regions | 522 | 5182 |
| % BU in largest blob for constant regions | 56.2209 | 43.5371 |
| HHI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

Table 5.1: Parameters generated for Urbanisation Index

The first three parameters represent the extent of urbanisation by discussing the changing and not changing percentages over the given time period. Since, most of the region in Jaipur is non-settlement area, we can see that the percentage of constantly NBU pixels for it is way higher than that for Gurgaon. Also, Gurgaon being a part of the NCR has seen a rapid development and hence rapid urbanisation. This is also validated by the observation that the percentage of constantly BU pixels is less than the percentage of changing pixels, which is remarkable in terms of urbanisation.

The last six parameters together represent the patterns of change in urban clusters. Here, we observe a visible change in the duration of 5 years, where urbanisation occurs around the peripherals of the urban cluster center. We also need to understand that although the largest blob has considerably expanded, it doesn't mean that the areas in the periphery were non-builtup earlier. It just means that these smaller hamlets/clusters got merged with the bigger cluster. We also observe that the number of blobs has reduced and now we have larger and lesser blobs. This is quantitatively conveyed by the increasing HHI index. Comparing Gurgaon and Jaipur, we observe that Jaipur has much higher number of hamlets and remote settlement areas than Jaipur, represented by the number of blobs. Apart from this, although Jaipur has a significantly dense urban cluster at the center, the HHI index for it is lower than Gurgaon. This can be accounted for by the extremely large number of smaller clusters/hamlets.

We extend this analysis on a larger set of 186 districts spread across 9 states. The distribution of these 9 parameters over these districts, calculated as the cumulative distribution (CDF) can be seen in Fig. 5.1.



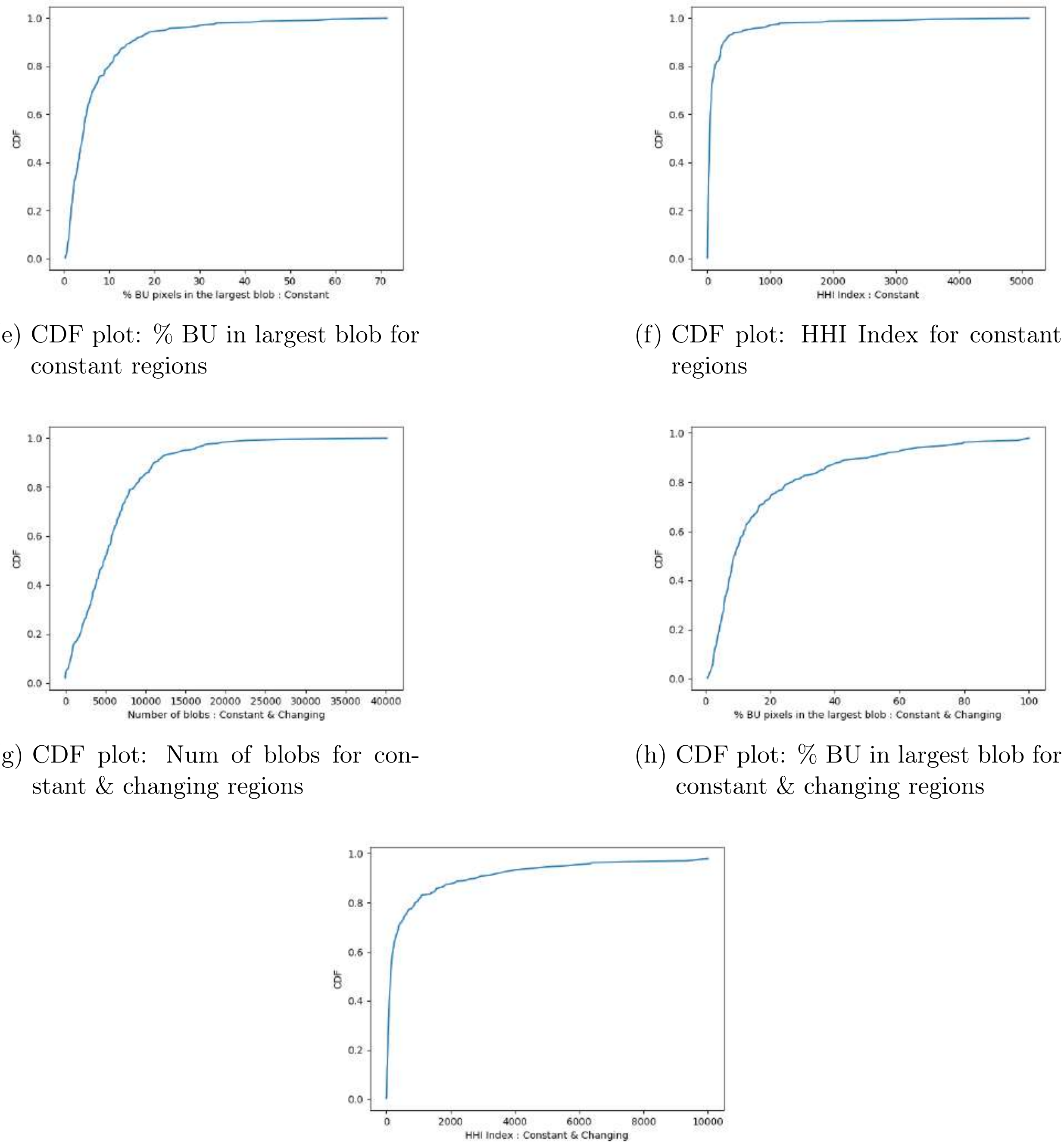
(a) CDF plot: % of Constantly BU

(b) CDF plot: % of Constantly BU

(c) CDF plot: % of Changing

(d) CDF plot: Num of blobs for constant regions

Figure 5.1: Distribution plots of 9 parameters for 186 districts



(e) CDF plot: % BU in largest blob for constant regions

(f) CDF plot: HHI Index for constant regions

(g) CDF plot: Num of blobs for constant & changing regions

(h) CDF plot: % BU in largest blob for constant & changing regions

(i) CDF plot: HHI Index for constant & changing regions

Figure 5.1: Distribution plots of 9 parameters for 186 districts (cont.)

# Chapter 6

# CONCLUSION

It is essential to note that our analysis does not merely give parameters for urbanisation, but is also able to identify the main urban clusters in Indian districts. This knowledge can help the government streamline the focus of their policies and narrow it down to either these urban clusters or remote rural settlements. The validation set that has been generated for Gurgaon and Jaipur can find many applications in different fields, and if extended further to other districts can then be useful for training and validating various models and analyses.

Finally we have tried to understand urbanisation based on the presence of buildings and infrastructure in a region. While this is a very significant factor in deciding whether the region in concern is urbanised or not, we understand that there are many other socio-economic factors that play a role in deciding if the population in an area is urbanised. To be able to clearly define the urbanisation process, it is important to understand the significance of all these factors and the correlation between them. Since there is always a scope of improving on the way we define urbanisation, we understand that there is a long way to go and a lot of other factors that need to be taken into consideration to obtain a thorough and vivd picture of changing cities.

# Bibliography

[BFSO84] L. Breiman, J. Friedman, C.J. Stone, and R.A. Olshen. *Classification and Regression Trees*. The Wadsworth and Brooks-Cole statistics-probability series. Taylor & Francis, 1984.

[DAT] Goldblatt et al. 20k dataset for image classification of built up areas in india. `https://fusiontables.google.com/DataSource?docid=1fWY4IyYiV-BA5HsAKi2V9LdoQgsbFtKK2BoQiHbO#rows:id=1`.

[FW18] Mst Ilme Faridatul and Bo Wu. Automatic classification of major urban land covers based on novel spectral indices. *ISPRS International Journal of Geo-Information*, 7(12), 2018.

[GDH18] Ran Goldblatt, Klaus Deininger, and Gordon Hanson. Utilizing publicly available satellite data for urban research: Mapping built-up land cover and land use in ho chi minh city, vietnam. *Development Engineering*, 3:83–99, 2018.

[GYHK16] Ran Goldblatt, Wei You, Gordon Hanson, and Amit Khandelwal. Detecting the boundaries of urban areas in india: A dataset for pixel-based image classification in google earth engine. *Remote Sensing*, 8(8):634, 2016.

[LDS$^+$15] Erzhu Li, Peijun Du, Alim Samat, Junshi Xia, and Meiqin Che. An automatic approach for urban land-cover classification from landsat-8 oli data. *International Journal of Remote Sensing*, 36(24):5983–6007, 2015.

[LS8] Usgs landsat 8 collection 1 tier 1 toa reflectance. `https://developers.google.com/earth-engine/datasets/catalog/LANDSAT_LC08_C01_T1_TOA`.

[MS17] Paul Macarof and Florian Statescu. Comparasion of ndbi and ndvi as indicators of surface urban heat island effect in landsat 8 imagery: A case study of iasi. *Present Environment and Sustainable Development*, 11, 10 2017.

[PM17] Darius Phiri and Justin Morgenroth. Developments in landsat land cover classification methods: A review. *Remote Sensing*, 9(9), 2017.

[RPM] Rodrigo e. principe cloud masking module. `https://github.com/fitoprincipe/geetools-code-editor/wiki/Cloud-Masks`.

[SGBB16] Szilard Szabo, Zoltán Gácsi, and Boglarka Bertalan-Balazs. Specific features of ndvi, ndwi and mndwi as reflected in land cover categories. *Landscape & Environment*, 10:194–202, 10 2016.

[YJS03] Y.Zha, J.Gao, and S.Ni. Use of normalized difference built-up index in automatically mapping urban areas from tm imagery. *International Journal of Remote Sensing*, 24(3):583–594, 2003.