IMPROVING PIXEL-BASED ACCURACY FOR APPLYING SHANNON ENTROPY AND URBAN EXPANSION INTENSITY INDEX MODEL FOR LULC CLASSIFICATION

Submitted by

JADHAV SANGAM ANANT ASMITA

ABSTRACT

Monitoring and assessing changes in land use and land cover (LULC) over large areas is crucial in a variety of study fields, including climate change, sustainable development, and the preservation of natural resources. However, because of the widespread landscape heterogeneities, it is challenging to effectively extract LULC using simply the spectral properties of satellite pictures. Numerous research have added additional auxiliary variables to the classification model in an effort to increase the accuracy of LULC classification. In addition to offering strong processing capabilities, the Google Earth Engine (GEE) also offers a significant amount of remote sensing data and a number of auxiliary datasets. However, approaches that can optimise combinations of auxiliary datasets for classification along with the varied effects of various auxiliary datasets in the GEE need to be clarified in order to increase the LULC classification accuracy. Here, a thorough analysis of the effectiveness of several auxiliary features in enhancing the precision of medium-resolution pixel-based LULC classification models is presented. Sentinel-2 MSI data is chosen as the primary dataset, with Mumbai City as the study area. Auxiliary open-source datasets in GEE are used to extract three different types of features, including spectral features, remote sensing multi-indices, and topographic features. The accuracy of four pixels-based was then improved, and the impact of auxiliary datasets on this improvement was examined. The results demonstrate that the classification's overall accuracy can be raised regardless of the kind of auxiliary characteristics used. The pixel-based categorization model's best overall accuracy is 96.49% for Mumbai City. The topographic features are crucial for increasing the classification accuracy across the board in pixel-based models using all data.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
INTRODUCTION

1.1 General

In multiple fields, including natural resource protection, sustainable development, and climate change, detailed land use/land cover (LULC) information at global and regional scales is crucial. It is common practise to get LULC information using remote sensing data. However, it is challenging to classify huge areas with geographic variation with high accuracy using simply satellite data. Researchers have made an effort to increase the LULC classification's accuracy by combining satellite data with other auxiliary datasets that are now accessible.

For the past many years, there has been interest in the value of merging auxiliary features with remote sensing data to increase the classification accuracy at both the regional and global levels. The accuracy of different remote sensing indices for different land use types varies greatly due to the complexity of the actual ground surface. As a result, various remote sensing indices are frequently used as supporting features to identify particular LULC types.

Although auxiliary characteristics can improve LULC classification accuracy, there are two main challenges preventing open-source geographical datasets from being used more widely. There aren't many free and open spatial datasets available. These datasets must be requested from several sources and acquired. There are a finite number of features that can be employed in classification models due to computational and storage constraints. For planetary-scale geospatial analysis, Google Earth Engine (GEE) offers a potent cloud-based platform that can directly call multi-petabyte satellite imagery and a variety of geospatial datasets. For high-precision land use classification at regional, governmental, and even international levels, GEE is frequently utilised.

1.2 Objectives

The objectives of this study are as follows;

1. To compare different pixel-based model on the basis of Overall accuracy, Confusion Matrix, kappa coefficient
2. To identify which type of pixel-based model with auxiliary dataset gives higher accuracy which can further used for applying ‘Shannon Entropy’ and ‘Urban Expansion Intensity Index’ model for LULC classification of Mumbai city.

CHAPTER 2
LITERATURE REVIEW

2.1 General

This chapter discusses the review of the literature conducted for the study. It includes the literature that helps achieve a better understanding of the purpose of this particular research and improve the gaps present in the past literature.

2.2 Literature Review

Norouzi's (2022) paper "Measuring Land Use Changes and Quantifying Urban Expansion Using Remote Sensing and GIS Techniques - A Case Study of Qom" presents a case study on the use of remote sensing and GIS techniques to detect land use changes and quantify urban expansion in Qom, Iran. The study employs various methods, including visual interpretation and object-based image analysis, to provide valuable insights into the drivers and impacts of urban growth. The paper highlights the importance of accurate spatial data and the role of satellite imagery in monitoring urbanization, making it a valuable contribution to the field of remote sensing and GIS. The study's findings will be of interest to urban planners, policymakers, and researchers seeking to understand the dynamics of urbanization and its impacts on land use in rapidly growing cities.

Klouček et al. (2018) aimed to identify suitable variables for detecting grassland to cropland changes using high-resolution satellite data. They used object-based image analysis to classify satellite images into grassland and cropland classes and calculated several spectral, textural, and vegetation indices as potential variables. The study found that the normalized difference vegetation index (NDVI) was the most important variable for detecting grassland to cropland changes, followed by texture measures and other vegetation indices. This research provides valuable guidance for selecting appropriate variables and using advanced machine learning algorithms to accurately detect and monitor land cover changes. The findings of this study can be applied to other regions and land cover types, providing valuable insights for land management and conservation efforts.

Huang et al. (2017) used Landsat images and a random forest algorithm to map and monitor major land cover dynamics in Beijing. The authors achieved high accuracy in land cover classification and quantification of land cover changes over time using ground truth data. The study demonstrated the potential of remote sensing data and machine learning algorithms, specifically Google Earth Engine, in accurately monitoring land cover changes in urban areas. This research provides valuable insights into the effective use of remote sensing and machine learning for land cover monitoring and management in urban areas, making it a useful reference for researchers and practitioners in this field.

CHAPTER 3
MATERIALS AND METHODOLOGY

3.1 General

The chapter briefly describes the data used in this study. It mentions all the satellite datasets, software and in-situ data involved.

3.2 Study Area

Mumbai City, also known as Bombay, is a district of Maharashtra state in India, with its headquarters in the bustling metropolis of Mumbai. The city is located on the west coast of India, at a latitude of 19.0760° N and longitude of 72.8777° E. The district is bordered by the Arabian Sea to the west, Thane district to the north, Mumbai Suburban district to the east, and Raigad district to the south. Mumbai City has a mean sea level of approximately 14 meters.

Mumbai is the financial capital of India and a major center of commerce, trade, and entertainment. Mumbai is the financial capital of India and a major center of commerce, trade, and entertainment. It is home to the country's largest stock exchange, the Bombay Stock Exchange, and numerous major corporations. The city is served by an extensive network of public transportation, including buses, trains, and taxis. Mumbai is also home to India's largest airport, the Chhatrapati Shivaji Maharaj International Airport, which connects the city to destinations across India and around the world.

The district covers an area of 603.4 square kilometers and is bordered by the Arabian Sea to its west. Mumbai also faces the challenge of unplanned urban expansion, leading to issues like congestion, inadequate infrastructure, and environmental degradation. In recent years, the city has been facing issues related to water scarcity and pollution, as well as the growing threat of climate change.

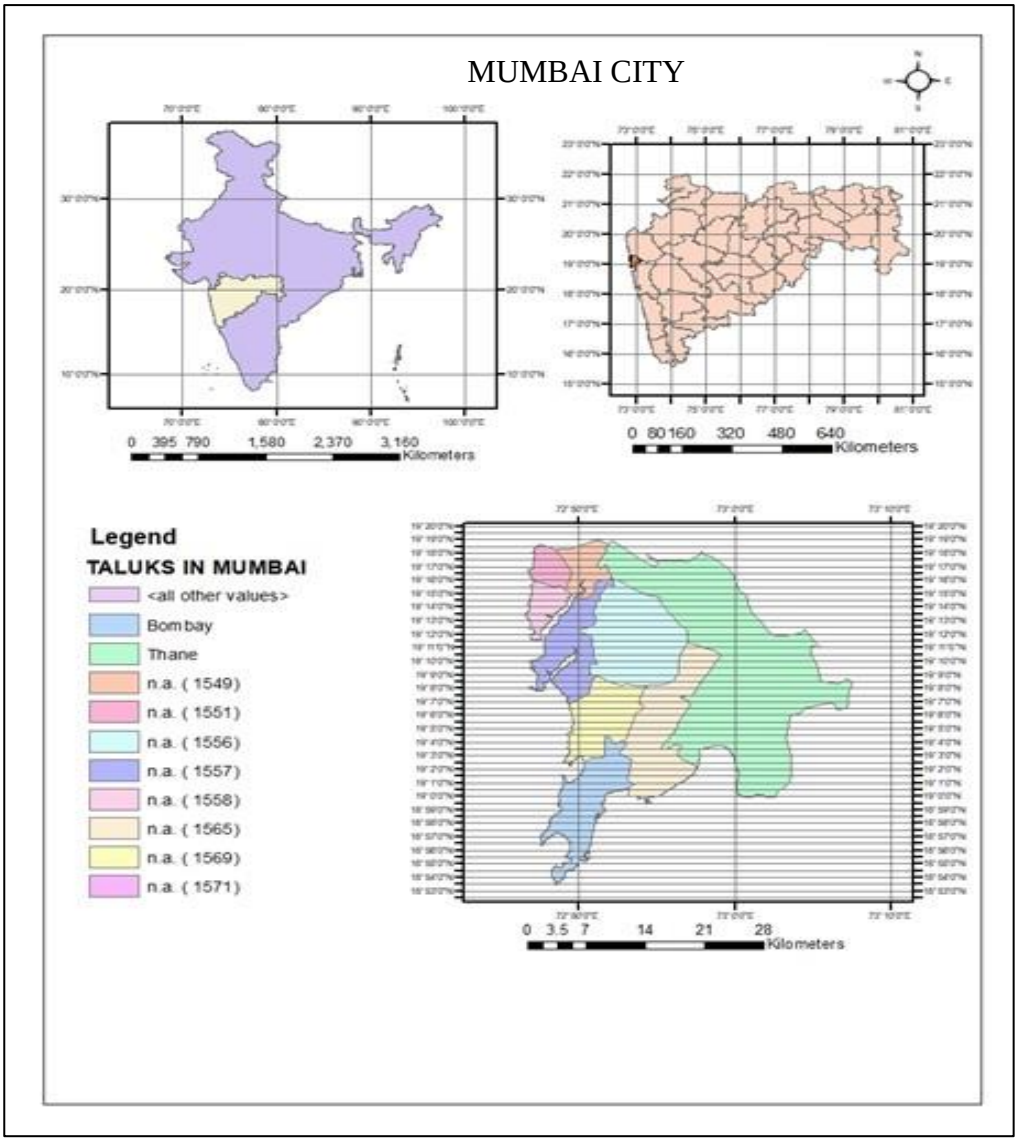

Figure 3.1: Study Area.

3.3 Data Collection

Sentinel-2 MSI: Multispectral Instrument, Level-2A data was used. The Google Earth Engine was used for classification. The medians of the Red, Green, Blue, Near-infrared (NIR), and Short-wave infrared (SWIR-1 and SWIR-2) spectral bands of the Sentinel-2 MSI image were considered in the analysis. The administrative boundary of Mumbai City district was downloaded from DIVA-GIS and all the data was clipped to the Mumbai City district boundary.

Table 3.1. Datasets used.

DATA	SPATIAL RESOLUTION	DATA FORMAT	USAGE
Sentinel-2 MSI,Level-2A	10m,20m	JPEG2000	Land use Classification
SRTM DEM	Approximately 30m	GeoTiff	Elevation
Administrative Boundary	-	Shapefile	Mumbai City Boundary.

3.4 Auxiliary Data

The medium-resolution geospatial datasets available on GEE, which are as close as possible to Sentinel-2 MSI. The time coverage of these datasets was used to extract auxiliary features.

The median values for five different spectral indices were calculated using the pre-processed Sentinel-2 images . Different remote sensing indices are susceptible to various kinds of LULC, according to prior research. As a result, there isn't a single index that covers all LULC kinds. Previous studies have demonstrated that the Normalised Difference Vegetation Index (NDVI), the Normalised Difference Water Index (NDWI), and the Normalised Difference Built-up Index (NDBI) are sensitive to vegetation features, water bodies, and built-up regions, respectively. Research on forests has shown that spectral indices such as the near infrared wavelength, among others, display weaker connections with LULC than the shortwave infrared wavelength, despite the fact that vegetation indices, such as NDVI, have been recommended in earlier studies to discriminate LULC kinds. Therefore, we also included the Normalized Burn Ratio (NBR) and Normalized Difference Moisture Index (NDMI) spectral indices to examine their contributions to LULC classification. In order examine their contributions to LULC categorization, inclusion of the Normalised Burn Ratio (NBR) and Normalised Difference Moisture Index (NDMI) spectral indices was done.

Table 3.2. Features in the land use/land cover (LULC) classification.

Features	Description	Data Sources
Spectral Features	Median of bands 2, 3, 4, 8, 11, 12	Sentinel-2 MSI
Spectral Indices	NDVI, NDWI, NDBI, NBR, NDMI	Sentinel-2 MSI
Topographic Features	Elevation, Slope, and Aspect	SRTM

Topographic features generated from Digital Elevation Models (DEMs) include elevation (affecting the temperature and precipitation level), slope, and aspect (affecting the solar radiation and vegetation growth). To describe the terrain features, 30 m DEM generated from the Shuttle Radar Topography Mission (SRTM) was used. This DEM is a post-processed elevation dataset widely used due to its high accuracy and extensive coverage.

3.5 LULC Classification Method

A decision tree-based classifier called random forest is a machine learning method. Each tree contributes one vote, and the classification or prediction outcomes are determined by voting. Numerous studies have demonstrated that RF results in LULC classification with reasonably high classification accuracy. The RF approach also offers the benefits of simple parameterization, collinear feature management, and high-dimensional data handling.

Designing of 4 RF classification models is done. Model P1 only for spectral features. Model P2 include spectral features combined with 5 different vegetation indices. P3 include spectral features combined with terrain features i.e DEM. Finally, model P4 includes all the features i.e. spectral features, vegetation indices, DEM data .

The supervised classification was implemented in the GEE platform. The training data were used for the training, and validation data were used to evaluate the classification error.

Table 3.3. Models used in this study for LULC classification.

Model	Description
P1	Pixel-based spectral features
P2	Pixel-based spectral features + vegetation indices classification model
P3	Pixel-based spectral features + topographic features (DEM) classification model
P4	Pixel-based all features classification model (spectral features, 5 vegetation indices, DEM data)

3.6 Classification Accuracy Assessment

Assigning the training sample points and verification sample points to calculate the respective confusion matrix. To quantitatively analyse the accuracy of the LULC classification, following measurements obtained

- The producer's accuracy (PA)
- User's accuracy (UA)
- Overall accuracy (OA)
- Kappa coefficient.

CHAPTER 4
RESULTS AND DISCUSSIONS

4.1 General

This chapter discusses the results obtained for objectives and comparing Kappa coefficient and overall accuracy of all models to state the best model for LULC classification.

4.2 Classification Results

Figure 4.1, 4.2, 4.3, and 4.4 shows the results of the LULC maps generated by the pixel-based models. The main LULC types include Water Bodies, Vegetation, Forest, Built-up Land, and Barren Land. The LULC maps generated by the pixel-based classification models.

The P1 pixel-based models misclassified in barren land and water, also areas where water bodies are mixed with the mangroves situated on the East coast of Mumbai city. The P2 model which includes vegetation indices masks some of the built-up area and barren land by vegetation (i.e misclassification occurred which results in decreasing overall accuracy).The P3 model which includes spectral features along with DEM data shows good built-up classification, but introduces errors by misclassifying some barren or vegetation area in the built-up class. The P4 model applied all the features, showing the best accuracy with kappa coefficient and overall accuracy as 0.955 and 96.49% respectively

Table 4.1: Kappa Coefficient and Overall accuracy of each model.

Model	Kappa Coefficient	Overall Accuracy (%)
P1	0.774	83.46
P2	0.830	87.40
P3	0.866	90.24
P4	0.955	96.49

From above Table, the classification model using the spectral features (P1) was used as the base model with overall accuracy 83.46%. The overall accuracy can be increased by almost 13% by importing spectral features along with vegetation indices and DEM data.

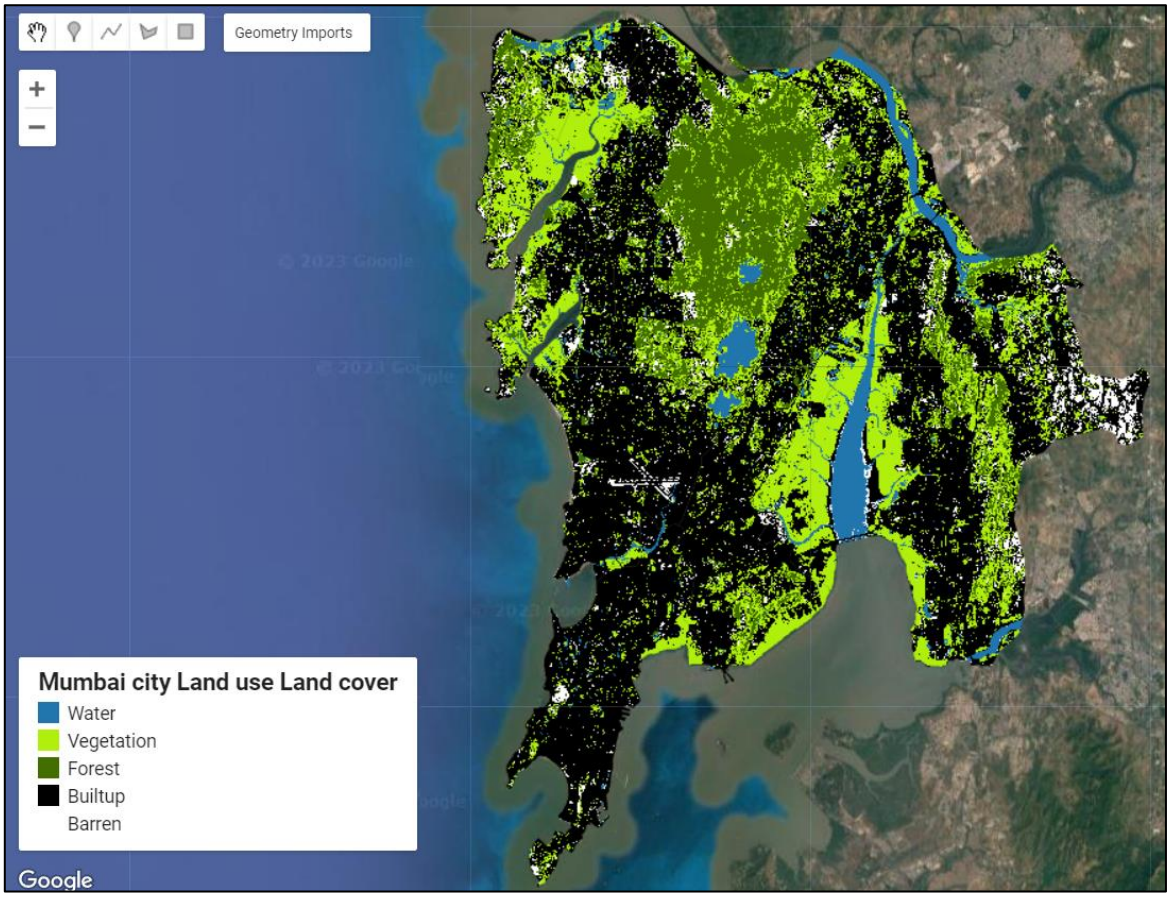

Figure 4.1. Pixel-based spectral features classification model (P1).

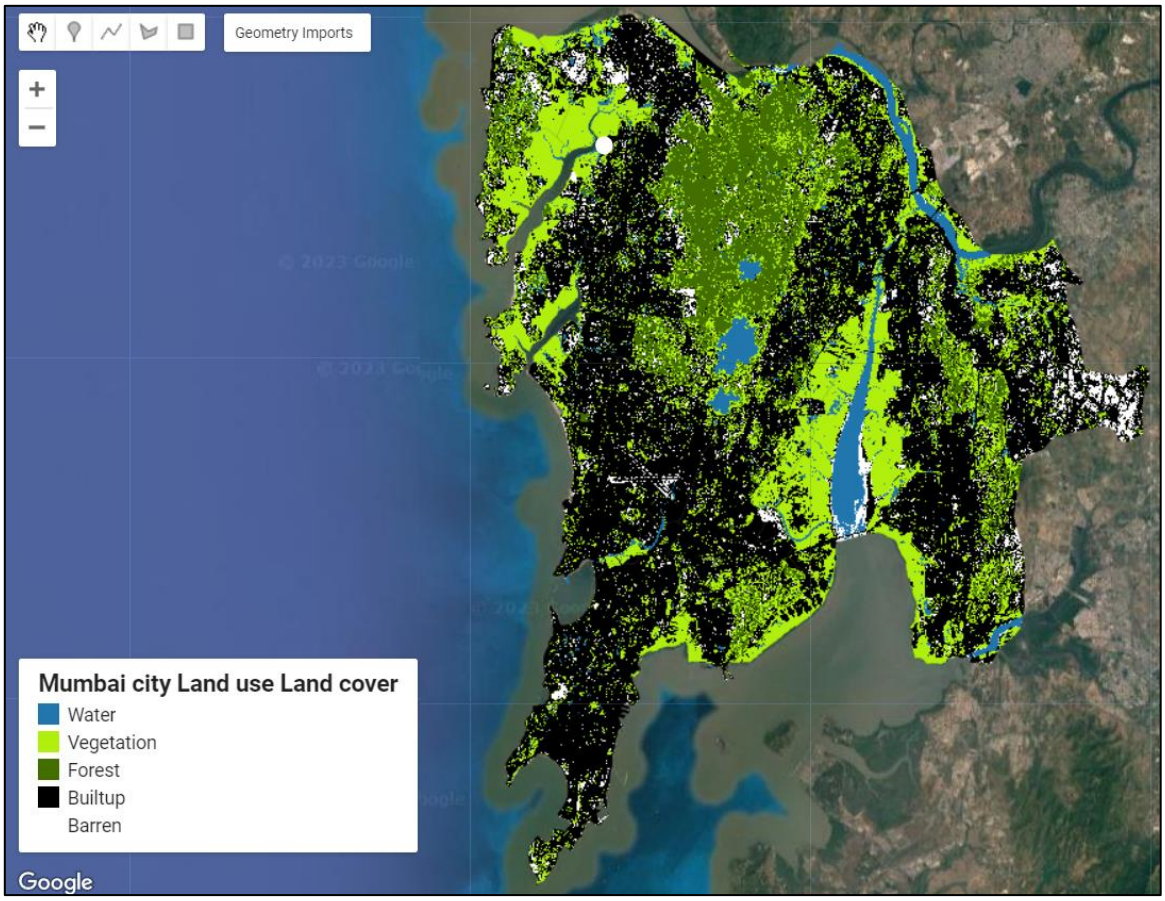

Figure 4.2. Pixel-based spectral features and vegetation indices classification model (P2).

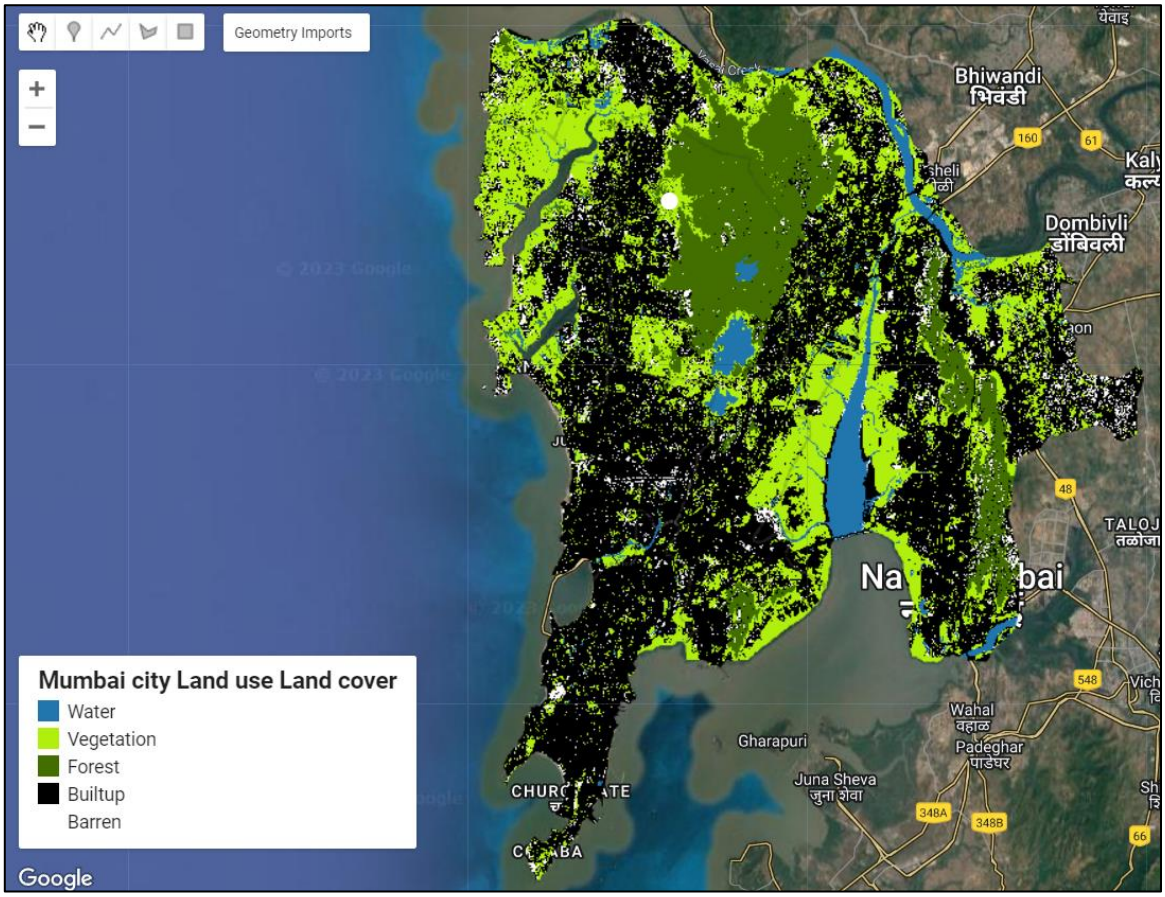

Figure 4.3. Pixel-based spectral features + topographic features classification model (P3).

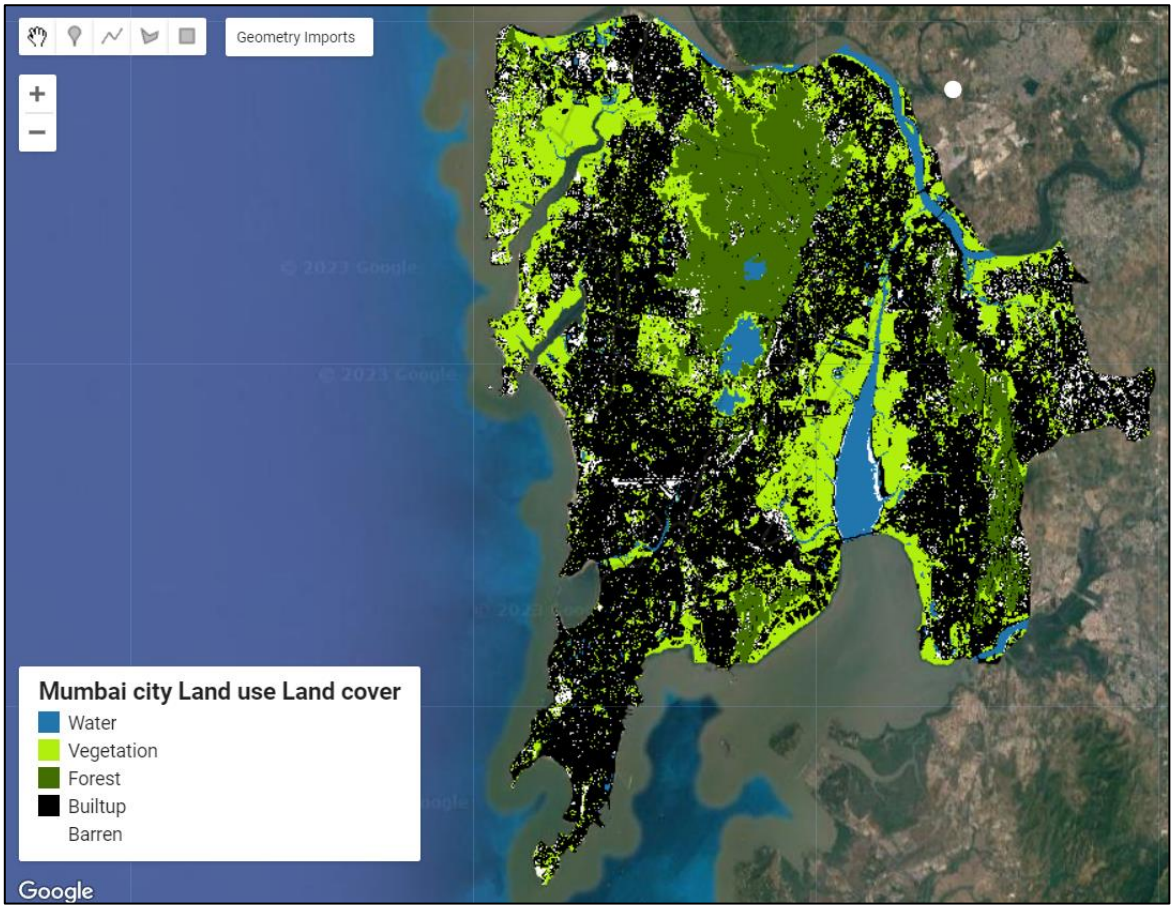

Figure 4.4. Pixel-based all features classification model (P4).

CHAPTER 5

FUTURE SCOPE

5.1 Shannon Entropy

Shannon introduced the term "entropy" in 1948 as a way to measure randomness and chaos. Shannon's entropy is frequently used by academics to gauge urban sprawl data and describe how dispersed or compact a particular characteristic is geographically in a particular area. The entropy model may analyse the dispersion of geographical phenomena, which features the configuration and direction of spatial patterns and indicators. The computation of urban sprawl can effectively use the entropy idea because to the integration of GIS and remote sensing technology. The area must be split into 'n' geographical zones in order to calculate the concentration and dispersion of urban sprawl. The value of the Shannon's entropy lies between 0 and Log(n). A value close to 0 denotes focused urban growth (greater concentration), whereas a value close to Log(n) denotes extensively distributed urban sprawl. Thus, urban sprawl from 1985 to 2021 was calculated using Shannon's entropy and GIS technology. Shannon's entropy is calculated using equation (1)

$$\boldsymbol{H_n = \sum_{i}^{n} P_i \times log\left[\frac{1}{P_i}\right]}$$

Equation-1

Where n is the total number of zones, and Pi is the prospect or portion of the variable happening in the ith zone. We have applied relative entropy to make the entropy value range Between 0 and 1. Equation (2) is used to compute the relative entropy:

$$\boldsymbol{H_n = \frac{\{\sum_{i}^{n} P_i \times log\left[\frac{1}{P_i}\right]\}}{log_e[n]}}$$

Equation-2

The land use/land cover maps were classed into five classes in order to apply the entropy model. The concentration of built-up areas in each zone was then calculated using them GEE.

Table 5.1: User's accuracy of the Classified Images

Model	User's accuracy				
	Vegetation	Built-up	Forest	Barren	Water
P1	0.894	0.826	0.810	0.666	1
P2	0.85	0.869	0.886	0.75	1
P3	0.904	1	0.875	0.727	1
P4	1	1	1	1	0.818

Table 5.2: Producer's accuracy of the Classified Images

Model	Producer accuracy				
	Vegetation	Built-up	Forest	Barren	Water
P1	0.85	0.904	0.921	0.444	0.882
P2	0.85	0.952	0.921	0.666	0.882
P3	0.95	1	0.960	0.571	0.823
P4	1	1	0.846	1	1

Table 5.2: Producer's accuracy of the Classified Images

5.2 Urban Expansion Intensity Index

Using the Urban Expansion Intensity Index, the urban spatial expansion disparity is statistically assessed. (UEII). Additionally, the UEII can be used to determine the tendencies of urban expansion as well as to evaluate the rate or intensity of changes in urban land use over a certain time period. Equation (3) shows the calculation for UEII.

$$\mathbf{UEII}_{it} = \frac{[\frac{\boldsymbol{ULA_{i,b} - ULA_{i,a}}}{\boldsymbol{t}}]}{\boldsymbol{TLA_i \times 100}}$$

Equation-3

Where: UEIIit is the annual average urban expansion intensity index of the (ith) zone in period (t) ULAi,a and ULAi,b are the amount of built-up area during time spans a and b in the (ith) spatial zone, respectively. TLAi is the total area of the (ith) spatial zone The UEII has been classified into several classes, as shown in Table

Table 5.3: The Division Standard Of UEII

UEII	SPEED
>1.92	High-speed development
1.05-1.92	Fast development
0.59-1.05	Medium-speed development
0.28-0.59	Low-speed development
0-0.28	Slow development

CHAPTER 6

CONCLUSION

In this study, improvement of pixel-based accuracy was done using 4 different classification models. This improved accuracy can be used in applying further ‘Shannon Entropy’ and ‘Urban Expansion Intensity Index’ for LULC classification. The conclusions are as follows:

1. Auxiliary features, such as multiple vegetation indices, topographic features (DEM), can improve the Overall Accuracy in non-uniform landscapes such as Mumbai city. Sentinel-2 MSI remote sensing image data were combined with the various auxiliary features used in this study, and showed that they effectively improve the accuracy of LULC classification. The OA of the pixel-based method increased from 83.46 to 96.49%.
2. The performance of the auxiliary features is not the same. The topographical model performed more accurately than model including vegetation indices model for the City of Mumbai.

3. Best Overall Accuracy can be produced by combining all the models which depicting all classes present in the area.

4. This overall accuracy obtained by P4 model can be used for applying ‘Shannon Entropy’ and ‘Urban Expansion Intensity Model’ which can result in analysis of unplanned urban expansion happened in Mumbai city. Also, necessary suggestions can be provided to the respective authority for slowing down Unplanned Urban Expansion.

REFERENCES

- Al-sharif, A. A. A. and Pradhan, B.: Spatio-temporal Prediction of Urban Expansion Using Bivariate Statistical Models: Assessment of the Efficacy of Evidential Belief Functions and Frequency Ratio Models, Applied Spatial Analysis and Policy,9,213-231,10.1007/s12061-015-9147-1,2016.
- Al-shalabi, M., Billa, L., Pradhan, B., Mansor, S., and Al- Sharif, A. A. A.: Modelling urban growth evolution and land-use changes using GIS based cellular automata and SLEUTH models: the case of Sana'a metropolitan city, Yemen, Environmental Earth Sciences,70,425-437,10.1007/s12665-012-2137-6, 2013
- Cabral, P., Augusto, G., Tewolde, M., and Araya, Y.: Entropy in Urban Systems, Entropy, 15, 10.3390/e15125223, 2013.
- Corcoran, J.M., Knight, J.F., and Gallant, A.L. (2013). "Influence of Multi-Source and Multi-Temporal Remotely Sensed and Ancillary Data on the Accuracy of Random Forest Classification of Wetlands in Northern Minnesota." *Remote Sens.,* 5(7), 3212–3238.
- He, C., Shi, P. J., and Chen, J.: A study on landuse/cover change in Beijing area, Geogr. Res, 20, 679-687, 2000.
- Huang, H., Chen, Y., Clinton, N., Wang, J., Wang, X., Liu, C., Gong, P., Yang, J., Bai, Y., and Zheng, Y. (2017). "Mapping Major Land Cover Dynamics in Beijing Using All Landsat Images in Google Earth Engine." *Remote Sens. Environ.,* 202, 166–176.
- Klouˇcek, T., Moravec, D., Komárek, J., Lagner, O., and Štych, P. (2018). "Selecting Appropriate Variables for Detecting Grassland to Cropland Changes Using High Resolution Satellite Data." *PeerJ.,* 6, e5487.
- Maxwell, A.E., Strager, M.P., Warner, T.A., Ramezan, C.A., Morgan, A.N., and Pauley, C.E. (2019). "Large-Area, High Spatial Resolution Land Cover Mapping Using Random Forests, GEOBIA, and NAIP Orthophotography: Findings and Recommendations." *Remote Sens.,* 11(12), 1409.
- Qu, L., Chen Z., Li, M., Zhi, J., and Wang, H. (2021). "Accuracy Improvements to PixelBased and Object-Based LULC Classification with Auxiliary Datasets from Google Earth Engine." *Remote Sens.,* 13(3), 453.
- Y. Norouzi 1,:Measuring Land Use Changes And Quantifying Urban Expansion Using Remote Sensing And Gis Techniques - A Case Study Of Qom ,SPRS Annals of the