Predicting Urban Expansion in Kamrup Metropolitan District using CA-Markov Model and XGBoost XAI Model: A Data-Driven Approach

Faculty In-Charge: Prof. D Sriram
Mentor: Mr. Nilay Nishant, Scientist SE, NESAC

Prepared by: Sai Vara Prasad Kotha (2021B5A71541P)
Adi Mukherjee (2021B5A70776P)
Harshit Nagpal (2021B5A72792P)
Krish Mantri (2021B3A71732H)

1. Acknowledgement

First and foremost, we would like to express our heartfelt thanks to Dr. D. Sriram, the Faculty in Charge, for his unwavering support and mentorship during our internship. His vast knowledge, expertise, and dedication have been instrumental in shaping our understanding of the subject matter. Dr. Sriram's constant encouragement, insightful feedback, and willingness to address our queries have been crucial in enhancing our skills and knowledge in the field. We are truly grateful for his invaluable guidance and mentorship.

We would also like to extend our deep gratitude to Mr. Nilay Nishant, our mentor, for his exceptional guidance and support throughout the internship. Mr. Nishant's expertise in the field, along with his patience and willingness to share knowledge, has been truly inspiring. His valuable insights, constructive feedback, and suggestions have been instrumental in refining our understanding of the project and ensuring its successful execution. We are immensely thankful to Mr. Nishant for his constant motivation and belief in our abilities.

We are indebted to the North Eastern Space Applications Center (NESAC) for providing us with this incredible opportunity to intern with them. We are grateful to the entire team at NESAC for their warm welcome, support, and guidance throughout our internship. The stimulating environment, rich resources, and state-of-the-art facilities at NESAC have greatly contributed to our learning experience and professional growth. We extend our sincere appreciation to NESAC for giving us the platform to apply our skills and knowledge in a real-world setting.

Furthermore, we would like to acknowledge our PS Division at Birla Institute of Technology and Science (BITS), Pilani, for giving us this opportunity.

We would also like to express our gratitude to our fellow interns and colleagues at NESAC. Their camaraderie, support, and shared experiences have made this internship an enriching and enjoyable journey. Their collective wisdom, collaborative spirit, and willingness to assist each other have fostered a conducive learning environment, and we are grateful for their friendship and support.

Last but not least, we would like to thank our families and friends for their unwavering support, love, and encouragement throughout our internship. Their belief in our abilities and constant motivation have been the driving forces behind our success. We are deeply grateful for their presence in our lives.

In conclusion, this internship has been a transformative experience, and we are immensely grateful to everyone who has been a part of our journey. The knowledge, skills, and experiences gained during this internship will undoubtedly shape our future endeavors. We are truly thankful for the guidance, support, and opportunities provided by Dr. D. Sriram, Mr. Nilay Nishant, NESAC, and BITS Pilani.

Thank you.

2. Table of Contents

3. Abstract

Urbanization, as a global phenomenon, necessitates accurate prediction and monitoring tools to support evidence-based urban planning and policy decisions. This project presents a scientific approach that utilizes satellite data and combines machine learning techniques, specifically XGBoost, with statistical methods such as the Markov chain model to forecast urbanization patterns.

Leveraging the rich information embedded in satellite imagery, including land cover classifications, vegetation indices, and spectral data, this study develops predictive models for urban growth. Through the extraction of pertinent features from satellite images and the integration of auxiliary datasets encompassing population density and socioeconomic indicators, the aim is to capture the intricate dynamics of urban expansion.

The proposed methodology employs the XGBoost algorithm, renowned for its capability to handle high-dimensional datasets and capture nonlinear relationships. By training the model on historical urbanization data paired with satellite-derived features, accurate forecasts of future urban growth patterns can be achieved.

Additionally, the integration of a Markov chain model facilitates the examination of temporal dependencies and transition probabilities pertaining to urban land use changes. This statistical technique enables the assessment of likelihoods for specific urbanization scenarios and provides insights into associated uncertainties.

The proposed approach exhibits substantial potential for urban planning and policy-making, empowering decision-makers with robust predictions and valuable insights into forthcoming urbanization patterns. By seamlessly integrating satellite data, machine learning algorithms, and statistical techniques, this framework offers a comprehensive scientific foundation for comprehending and forecasting the dynamics of urban growth, thereby facilitating sustainable and informed decision-making processes.

Keywords: satellite data, urbanization prediction, machine learning, XGBoost, Markov chain model, urban planning

4. Introduction

The rate of global urbanization presents urban planners and policymakers with significant challenges. Effective management and planning of urban areas require accurate projections of urban expansion so that proactive measures can be taken to address potential issues relating to infrastructure, land use, and resource management. In this context, the combination of computational modeling techniques with sophisticated machine learning algorithms offers a promising strategy for predicting urban growth patterns.

This study proposes a novel framework for predicting urban growth in Kamrup Metropolitan District, Assam, India, a rapidly developing region. We employ the Cellular Automata-Markov (CA-Markov) model, a well-known method for simulating land-use changes, and the eXtreme Gradient Boosting (XGBoost) algorithm, a potent machine learning algorithm renowned for its exceptional predictive abilities. In addition, we employ the XGBoost eXplainable Artificial Intelligence (XAI) model to provide interpretability and insight into the process of prediction.

Based on cellular automata, the CA-Markov model simulates urban growth by capturing the dynamic interactions between different land-use categories. Using transition probability matrices derived from historical land-use data and geographic information system (GIS) data, it is possible to generate projections of future urban expansion within the study area.

To enhance the accuracy and interpretability of our forecasts, we've incorporated the XGBoost algorithm into our framework. XGBoost is a method of ensemble learning that combines numerous decision trees to produce a robust predictive model. Its ability to capture complex relationships and interactions between multiple predictors makes it ideal for analyzing the multidimensional factors that affect urban expansion. In addition, the XGBoost XAI model provides interpretability via feature importance analysis, allowing us to identify the most significant urban growth drivers in the Kamrup Metropolitan District.

5. Literature Survey

Urban expansion and its predictions are important topics in spatial analysis. Several studies have contributed valuable insights to this field, focusing on different regions and employing various methodologies.

Kim and Kim (2022) conducted a study on urban expansion in South Korea, utilizing an Explainable Artificial Intelligence (XAI) model. They collected data on land cover, population density, Gross Regional Domestic Product (GRDP) per capita, and environmental assessments. Their findings revealed that urbanized areas had higher population density and GRDP per capita compared to non-urbanized areas. Additionally, urbanized areas exhibited closer proximity to various land cover types, except for forests. These results provided a comprehensive understanding of the socioeconomic, topographic, land-cover, and environmental features associated with urban expansion in South Korea.

In Vietnam, Hoang et al. (2018) conducted a study on the evaluation of tourism potential in the Central Highlands. They employed the Geographic Information System (GIS), the Analytic Hierarchy Process (AHP), and Principal Component Analysis (PCA) to assess the region. The study found that the Central Highlands had emerged as a promising tourism region with significant potential for further development. The evaluation indicated that internal potential outweighed external potential. The paper contributed to the understanding of GIS, AHP, and PCA techniques in assessing tourism potential and provided valuable insights for similar spatial prediction projects.

Aburas et al. (2021) focused on the spatio-temporal simulation of urban growth trends in Seremban, Malaysia, using an integrated CA-Markov model. Their study incorporated multiple classified and unclassified land-use maps obtained from the Department of Agriculture. The simulation results indicated that Seremban was experiencing urban sprawl in a disaggregated manner, with a projected worsening of the urban growth scenario in the future. This study demonstrated the application of the integrated CA-Markov model in predicting urban growth patterns and highlighted the significance of incorporating relevant factors, such as accessibility, for accurate predictions.

Overall, these studies contribute to the understanding of urban expansion and prediction in different regions. They employ advanced methodologies, such as XAI, GIS, AHP, PCA, and integrated CA-Markov models, to analyze the socioeconomic, topographic, land-cover, and environmental factors influencing urbanization. By providing insights into spatial patterns and future trends, these studies offer valuable guidance for urban planning and development strategies.

6. Objective

This study aims to contribute to the field of urban planning and management by providing a comprehensive approach that combines the CA-Markov model and the XGBoost XAI model's strengths. The most accurate results can be used to draw conclusions based on a comparison of the two models' efficacy. Local authorities, policymakers, and urban planners can make informed decisions regarding infrastructure development, resource allocation, and land-use regulations with the help of the proposed framework's predictive capabilities

7. Study Area and Data Used

7.1. Study Area

The Kamrup Metropolitan District is one of Assam State's 35 districts in northeastern India. Guwahati, the capital of Kamrup Metropolitan District, is the largest city in Assam and serves as its commercial and cultural hub. Geographically, the Kamrup Metropolitan District is situated in the western portion of Assam, bounded to the north by the Brahmaputra River and to the south by the Kamrup Rural District. The district encompasses roughly 1,528 square kilometers of land.

Over the years, the district has experienced significant urbanization and economic expansion. As the primary urban center, Guwahati attracts a large number of individuals from surrounding areas in search of employment, education, and better opportunities. This rapid urbanization has resulted in the district's expansion into residential, commercial, and industrial areas. Numerous industries, markets, shopping malls, educational institutions, and medical facilities are located there. The district is well-connected by road, rail, and air, with Lokpriya Gopinath Bordoloi International Airport serving as the region's primary gateway.

The district's population density is 2,010 per square kilometer (5,200 per square mile). The fact that its population grew by 37.34 percent between 1991 and 2001 and by 18.95 percent between 2001 and 2011 is one of the primary reasons we are studying this region. From 1991 to 2011, the per capita gross domestic product in this region increased significantly.

7.2. Data Used

We have collected land-use land-cover data for the years 2005-06, 2011-12, and 2015-16 in order to analyze and predict the urban expansion in the Kamrup Metropolitan District. We have generated proximity maps of roads, railways, waterways, transportation hubs, and buildings for the year 2023 using OpenStreetMap data. We have also collected topographic information, including elevation, slope, friction, and flood-risk zones. Socio-economic data like GDP was also taken into account.

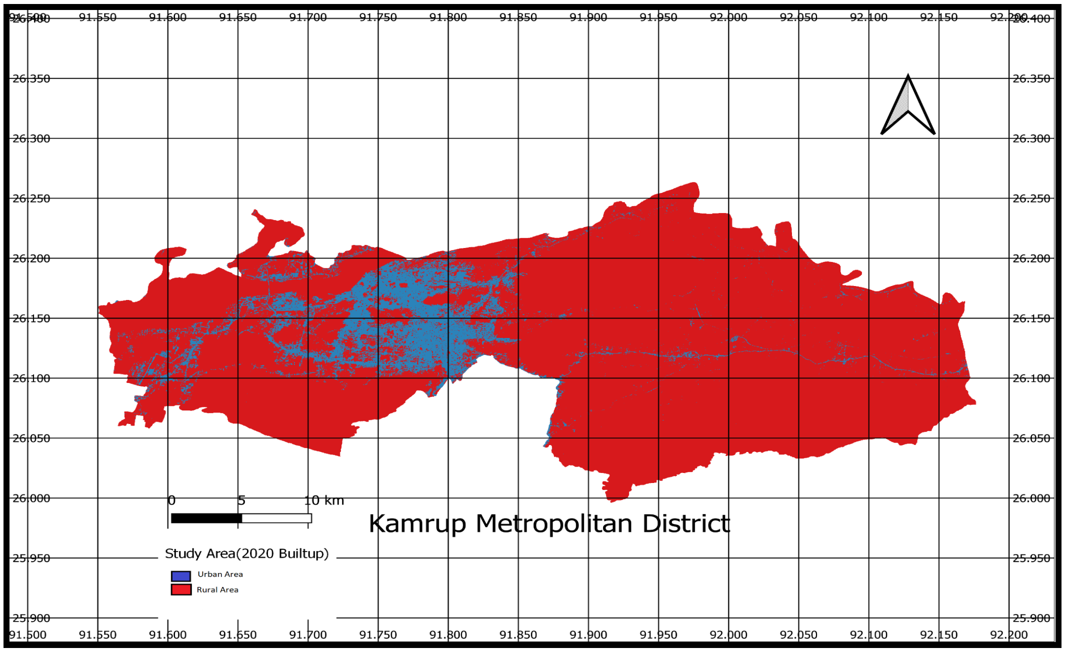


Figure 1: Study Area

	Data	
Dependent Variable	Dummy variables for urbanization from 2005 to 2020 (0: non-urbanized area, 1: urbanized area) Temperature (used in XGBoost only)	
Independent Variable	Socioeconomic Features	Population Density GDP (used in XGBoost only)
	Topographic Features	Elevation Slope Friction Flood Hazard Zones
	Landcover Features (Proximity to:)	Roads Railways Waterways Buildings Restricted Zones

Table 1: Description of variables

8.Methodology

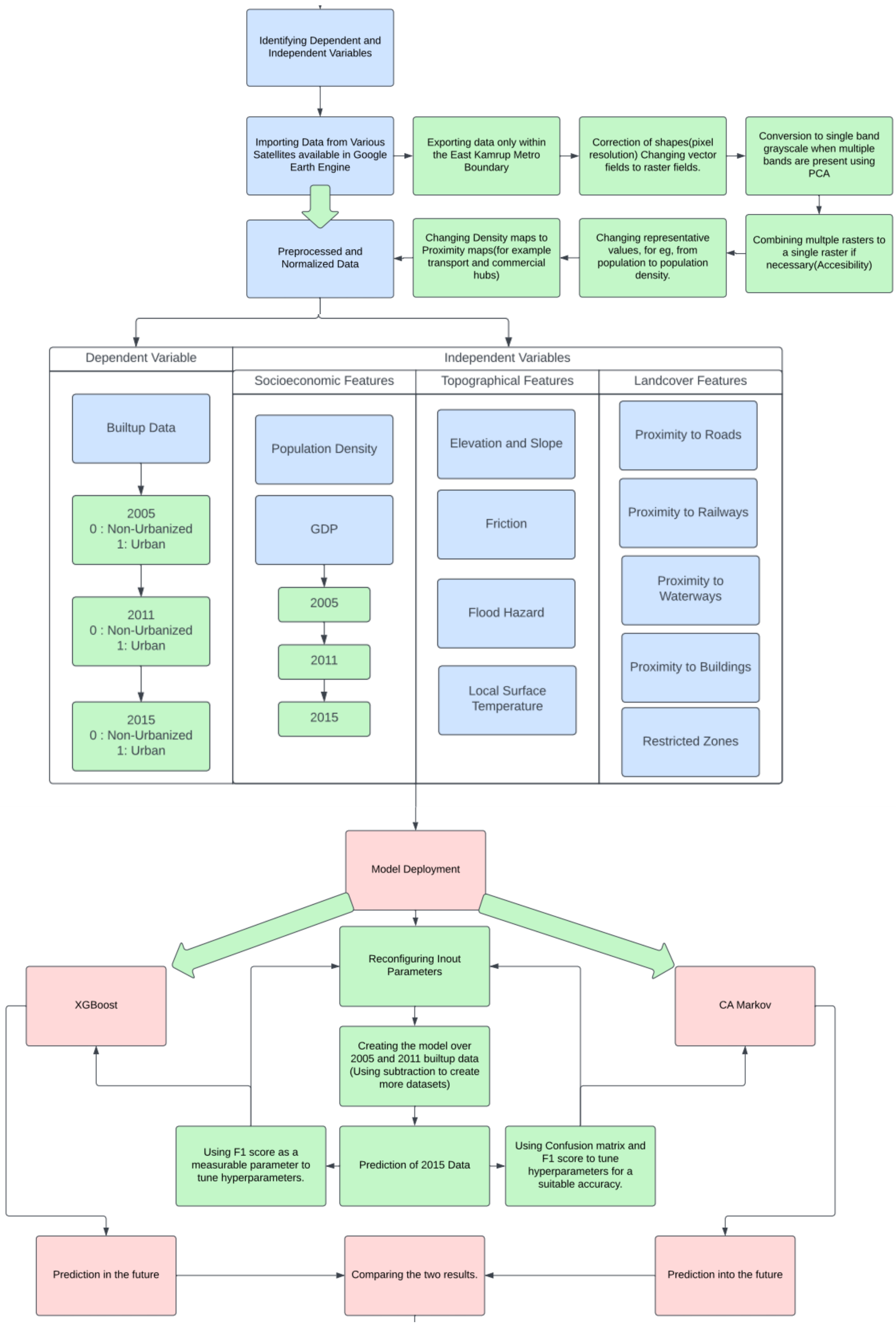


Figure 2: Research Procedure

8.1. Data Preprocessing

To study the urbanization process, it is necessary to collect pertinent data, which enables us to make more precise predictions. We have compiled OpenStreetMap data on road networks, railway networks, waterways, and transport hubs. Google Earth Engine elevation and slope (computed using DEM from Catrosat data) data have also been collected. As the Kamrup Metropolitan District is located on hilly and rugged terrain, we have also collected data on the friction (collected from Global Friction Surface) encountered when accessing land data from Google Earth Engine. The district is located near the Brahmaputra River, whose flood dynamics are highly variable. As we have done here, we must consider the Flood Hazard Zones (collected from Bhuvan Flood Hazard Zones) for better urban planning.

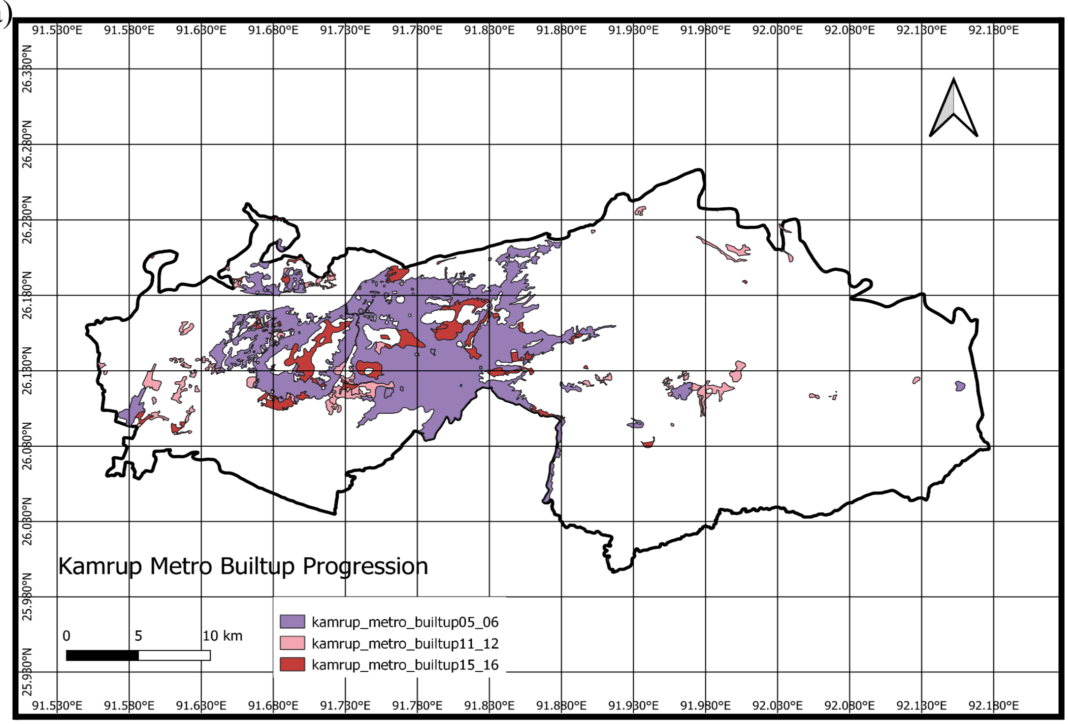
The collected transportation and building data were used to rasterize the vector data. The proximity between these zones was then computed using the Raster Proximity function in QGIS 3. We rescaled the proximity legend from 0 to 1 to ensure uniformity across all data sets. Using QGIS 3's raster slope functions, we extracted the slope map from the collected elevation data.

Eventually, we defined a new metric called Accessibility, which combines all of the transport data, friction data, and excluded zone data onto a single map, thereby enhancing data representation and providing us with a useful metric for our model. By assigning weights to each proximity map and combining them, accessibility was determined. Using the Raster Calculator function of QGIS 3, additional operations were performed. Using the Analytic Hierarchy Process (AHP) method, the weights were determined.

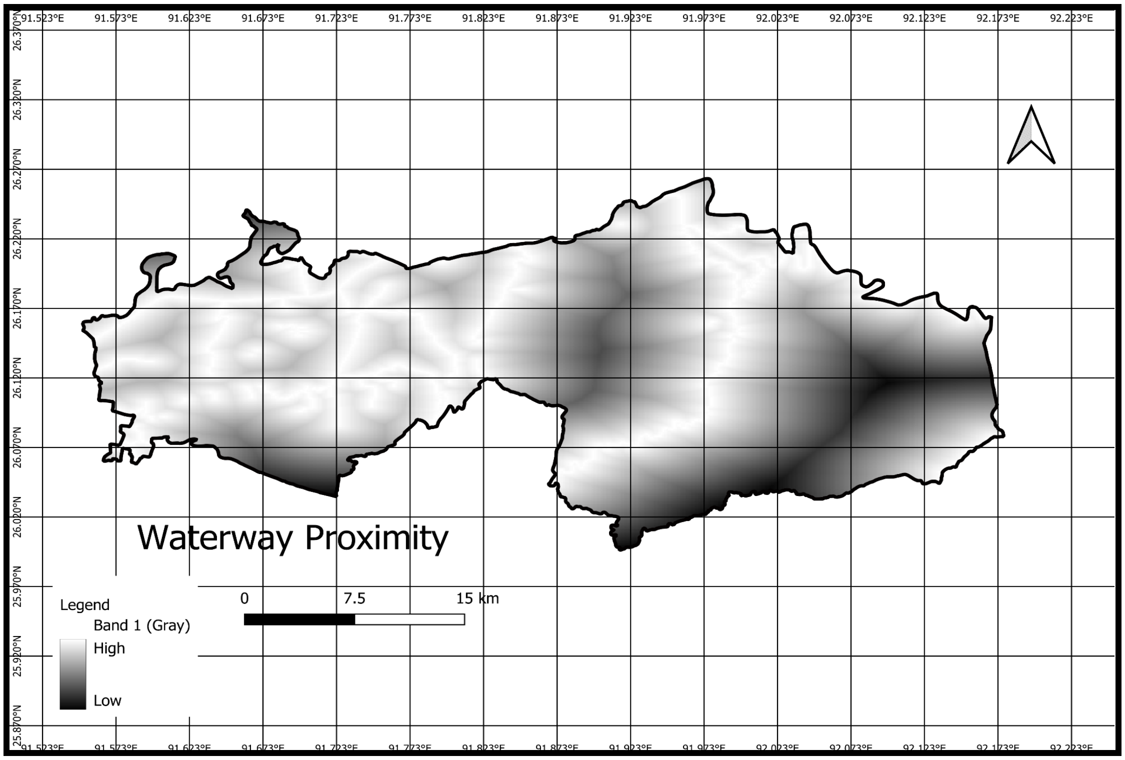
The data we received regarding the Flood Zones had three dimensions (three bands). Principal Component Analysis (PCA) was utilized to reduce this to a single dimension while retaining the majority of the data. The received population data was distributed and presented in point format. Before creating a map, we therefore created 2500x2500m grids in QGIS 3 and divided the population by the grid size using attribute table operations.

Following are the maps that were generated through these processes.

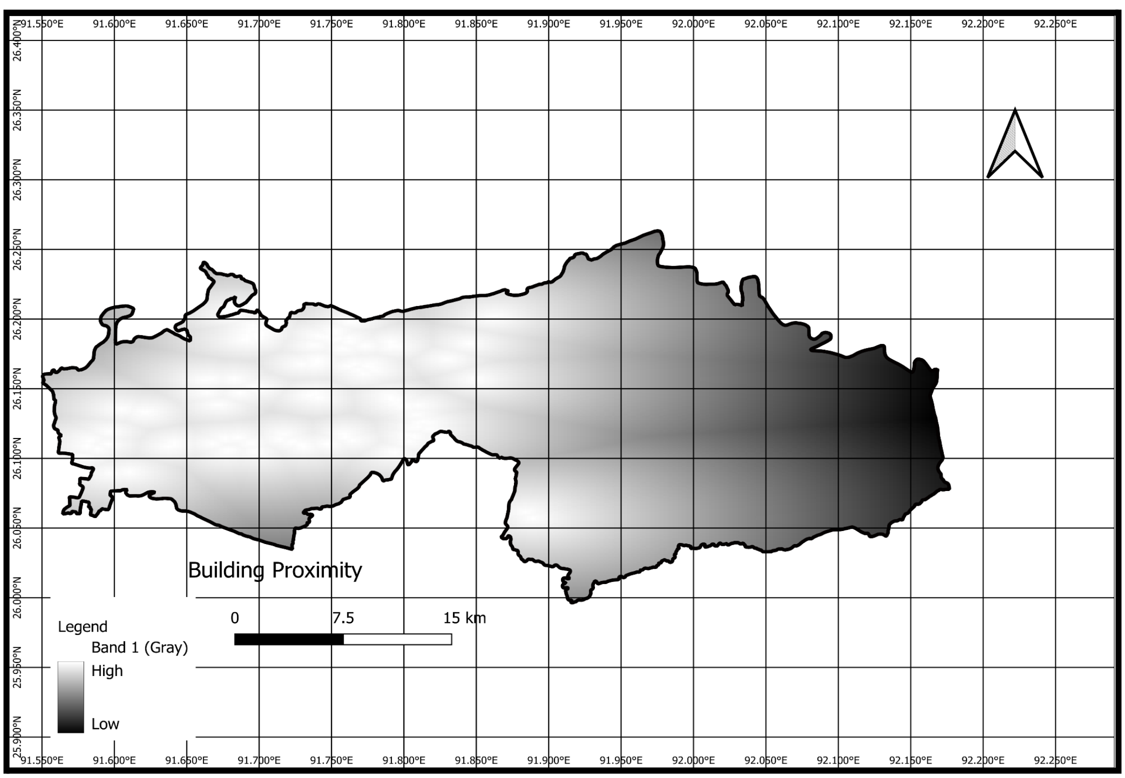
a)



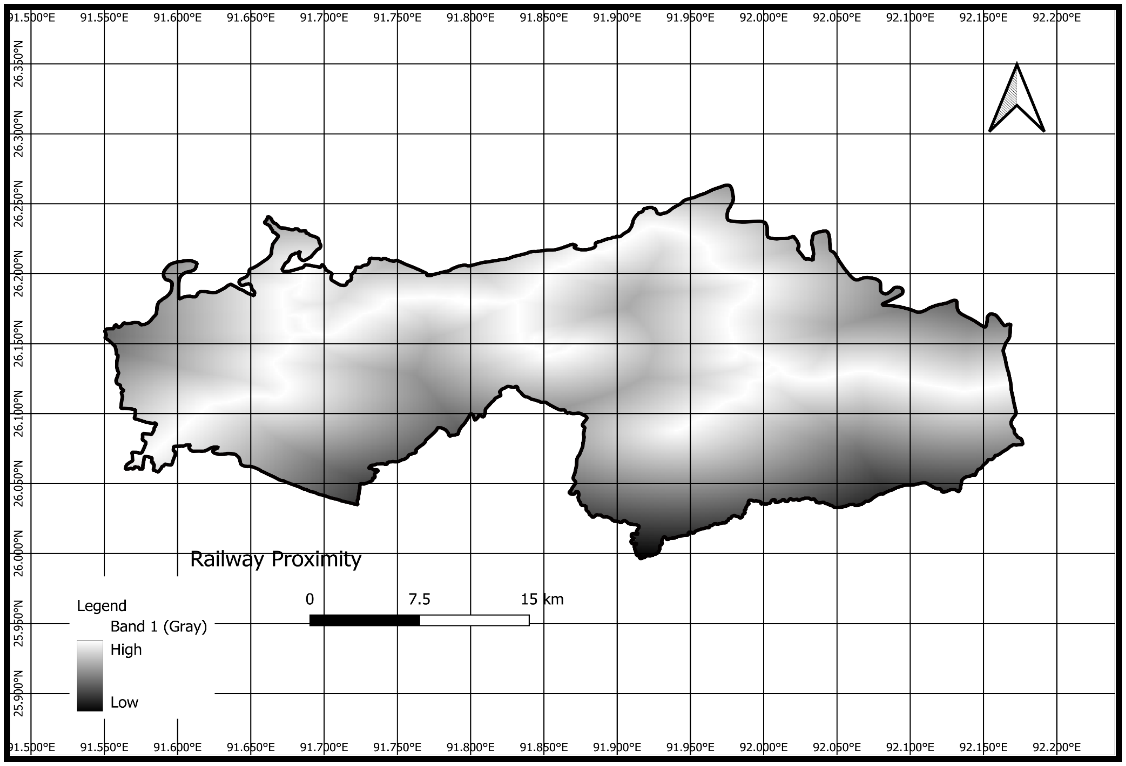
b)


Waterway Proximity
Legend
Band 1 (Gray)
High
Low
0
7.5
15 km

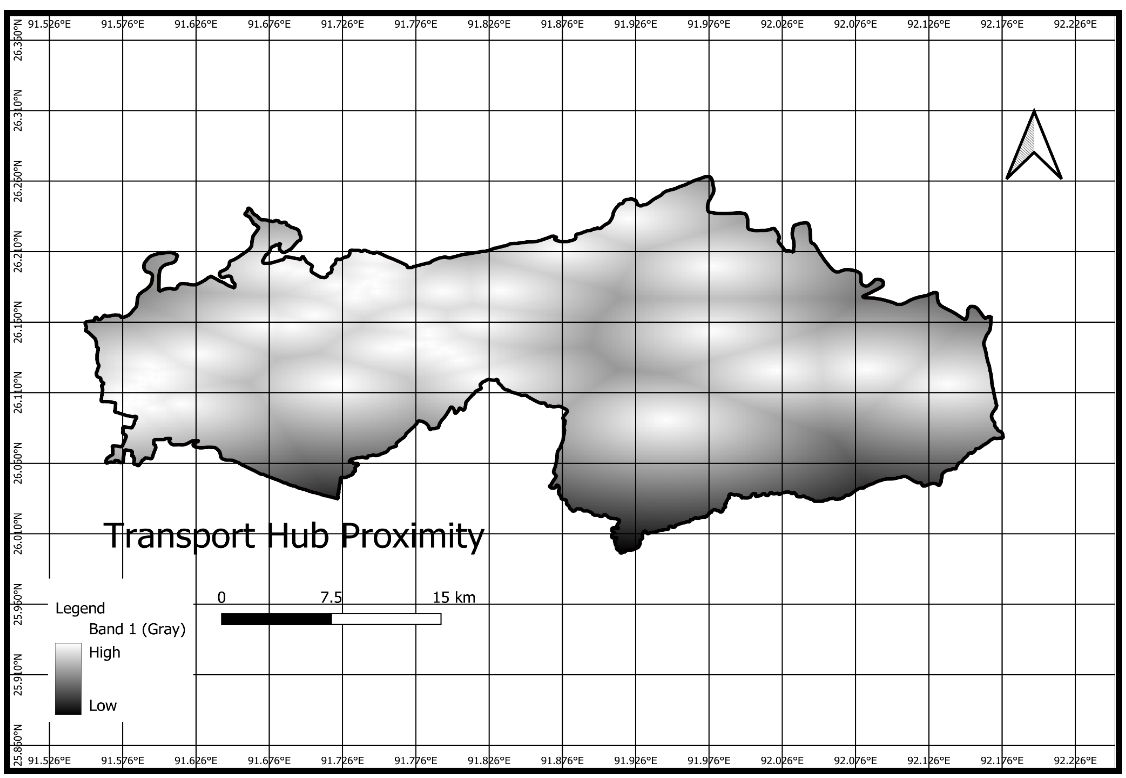
c)


Building Proximity
Legend
Band 1 (Gray)
High
Low
0
7.5
15 km

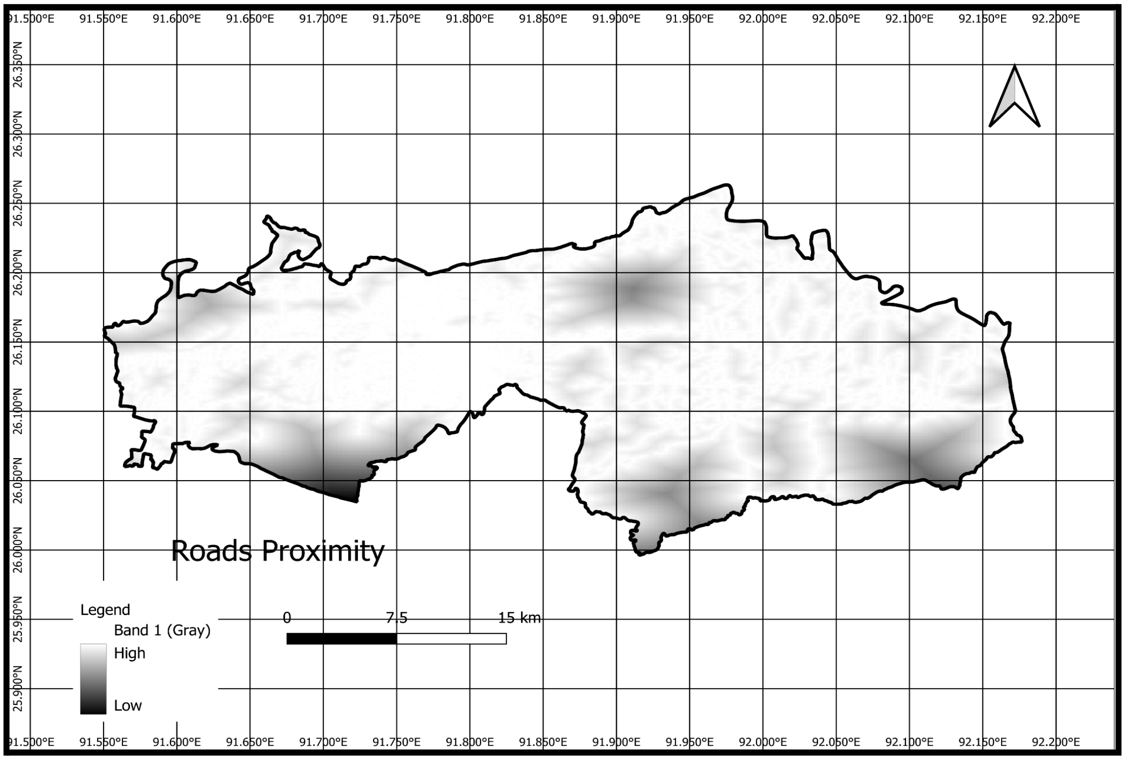
d)


91.500°E
91.550°E
91.600°E
91.650°E
91.700°E
91.750°E
91.800°E
91.850°E
91.900°E
91.950°E
92.000°E
92.050°E
92.100°E
92.150°E
92.200°E
26.350°N
26.300°N
26.250°N
26.200°N
26.150°N
26.100°N
26.050°N
26.000°N
25.950°N
25.900°N
Railway Proximity
Legend
Band 1 (Gray)
High
Low
0
7.5
15 km

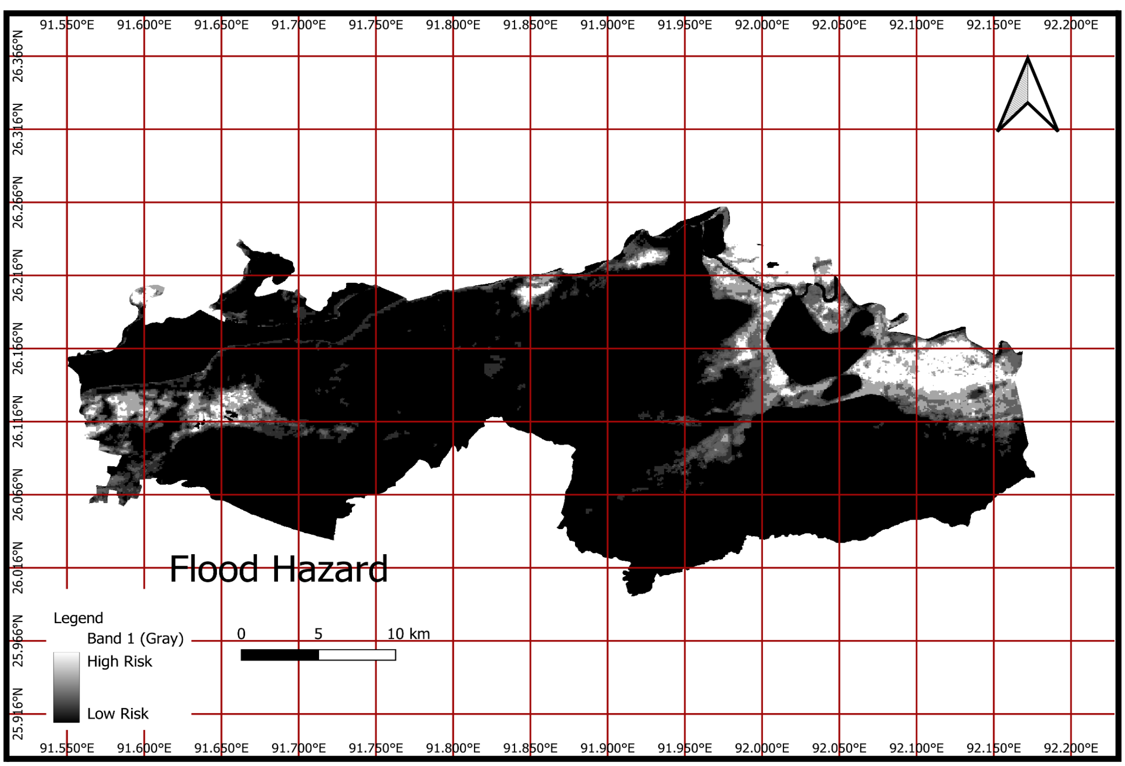
e)


91.526°E
91.576°E
91.626°E
91.676°E
91.726°E
91.776°E
91.826°E
91.876°E
91.926°E
91.976°E
92.026°E
92.076°E
92.126°E
92.176°E
92.226°E
26.360°N
26.310°N
26.260°N
26.210°N
26.160°N
26.110°N
26.060°N
26.010°N
25.960°N
25.910°N
25.860°N
Transport Hub Proximity
Legend
Band 1 (Gray)
High
Low
0
7.5
15 km

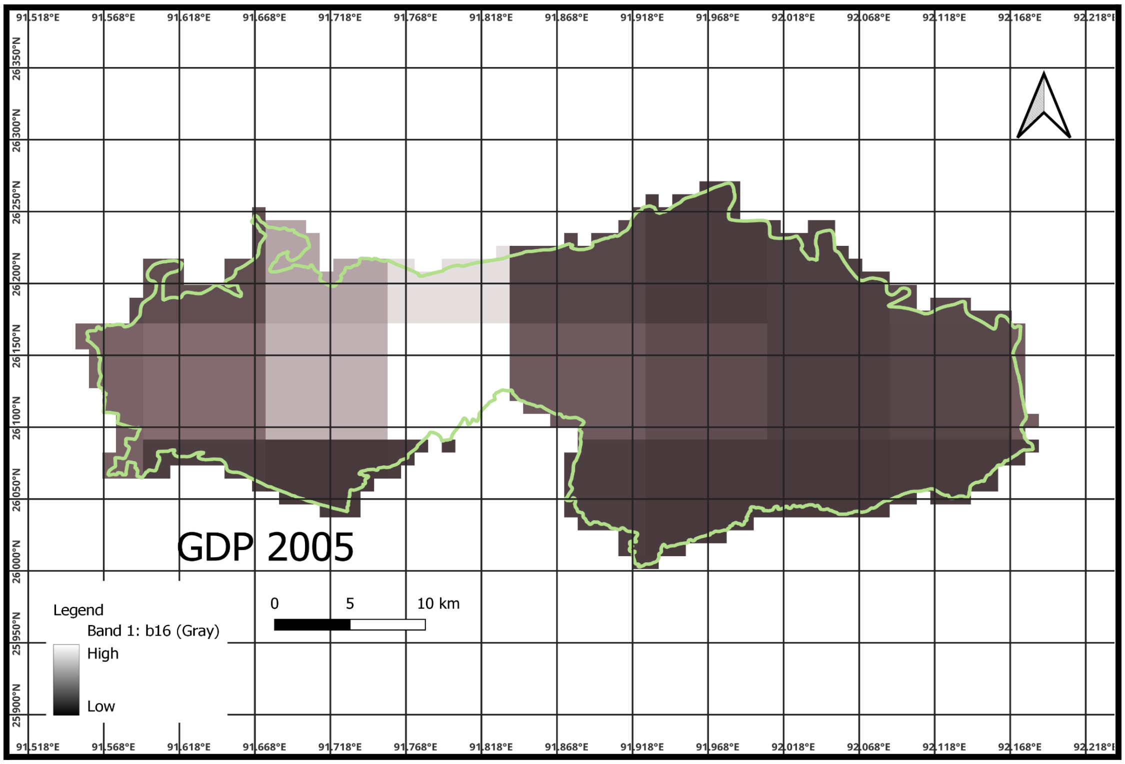
f)


Roads Proximity
Legend
Band 1 (Gray)
High
Low
0
7.5
15 km

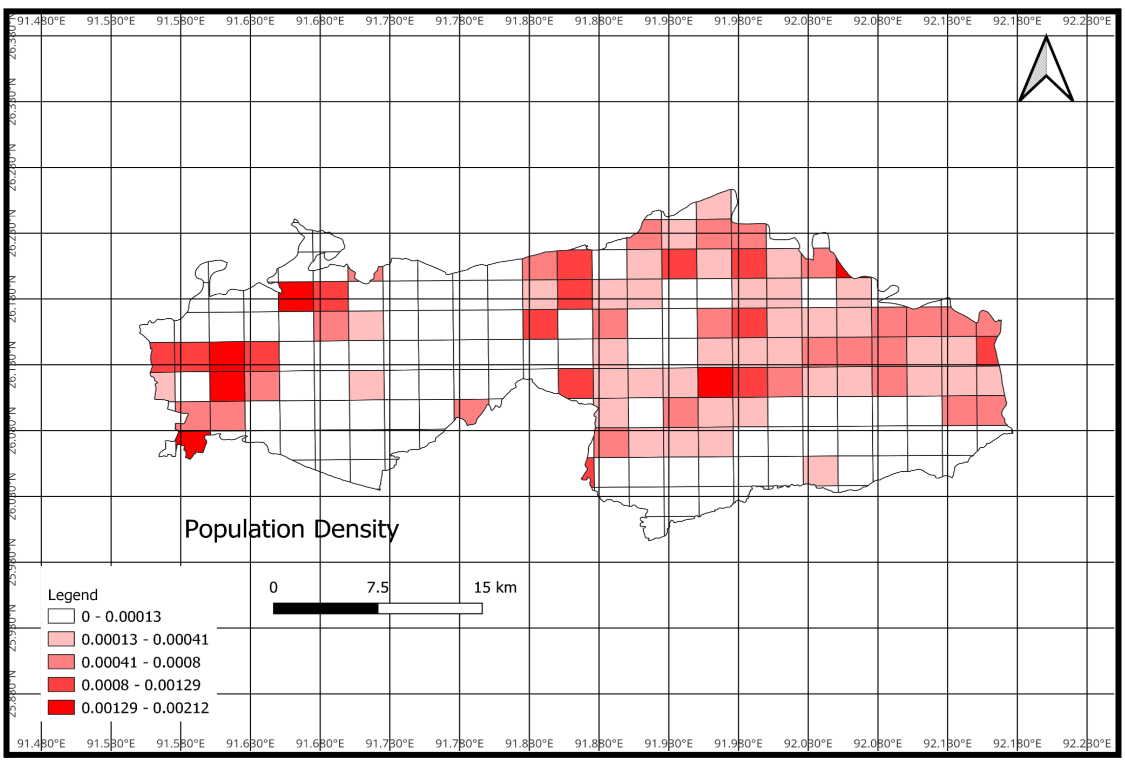
g)


Flood Hazard
Legend
Band 1 (Gray)
High Risk
Low Risk
0
5
10 km

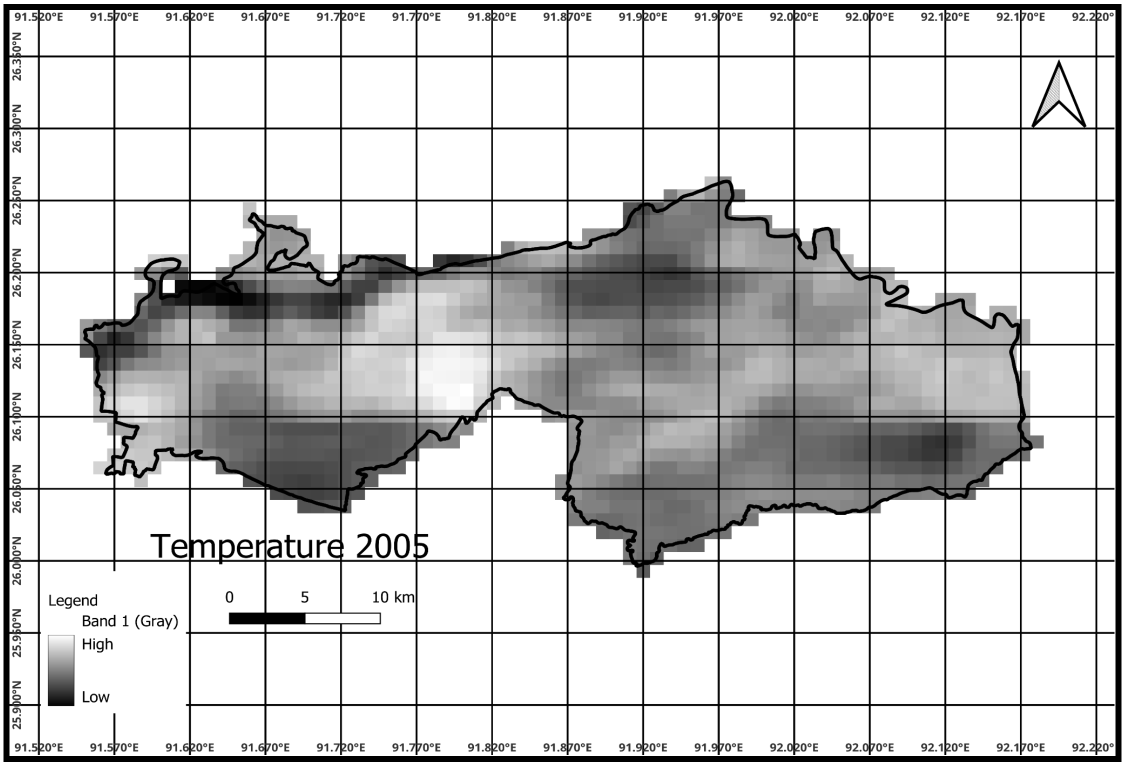
h)


GDP 2005
Legend
Band 1: b16 (Gray)
High
Low
0
5
10 km

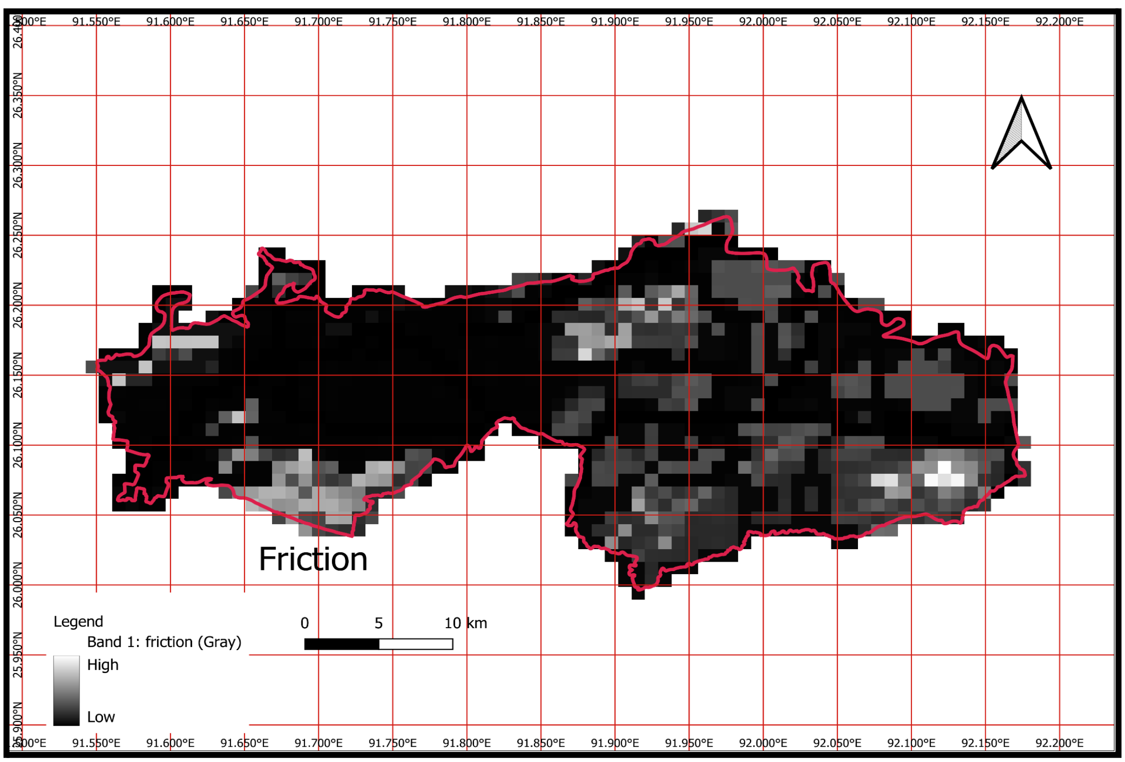
i)


Population Density
Legend
0 - 0.00013
0.00013 - 0.00041
0.00041 - 0.0008
0.0008 - 0.00129
0.00129 - 0.00212
0
7.5
15 km

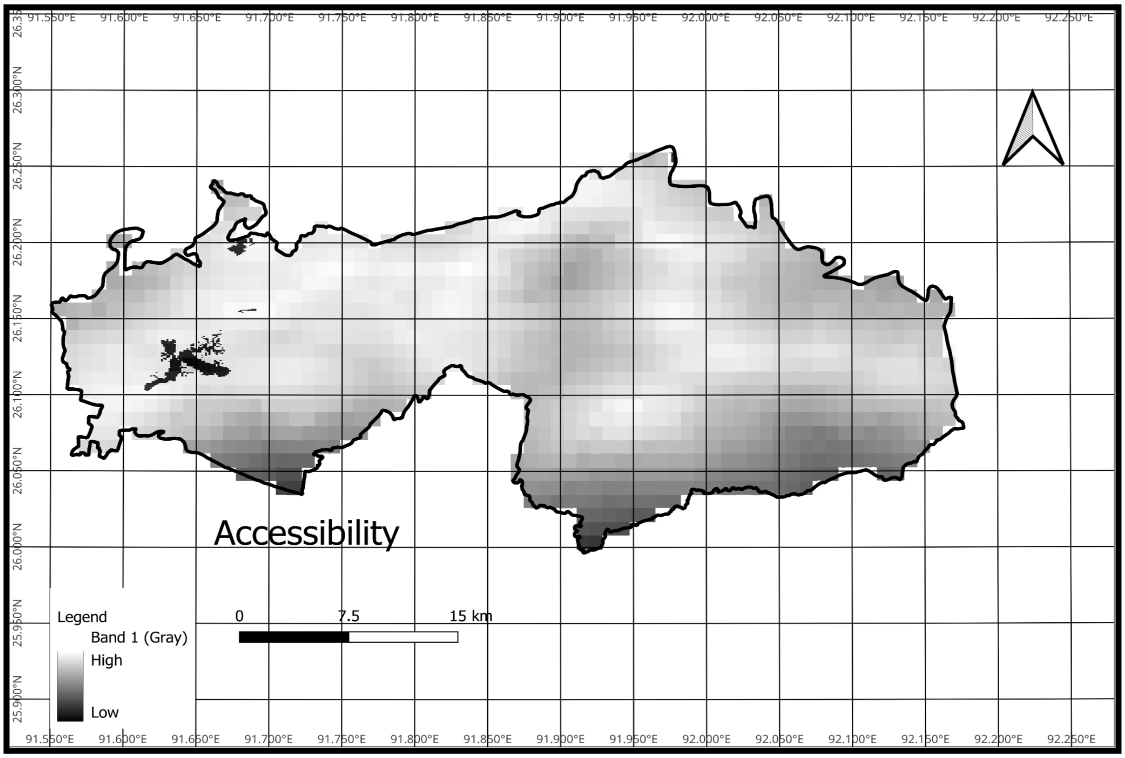
j)

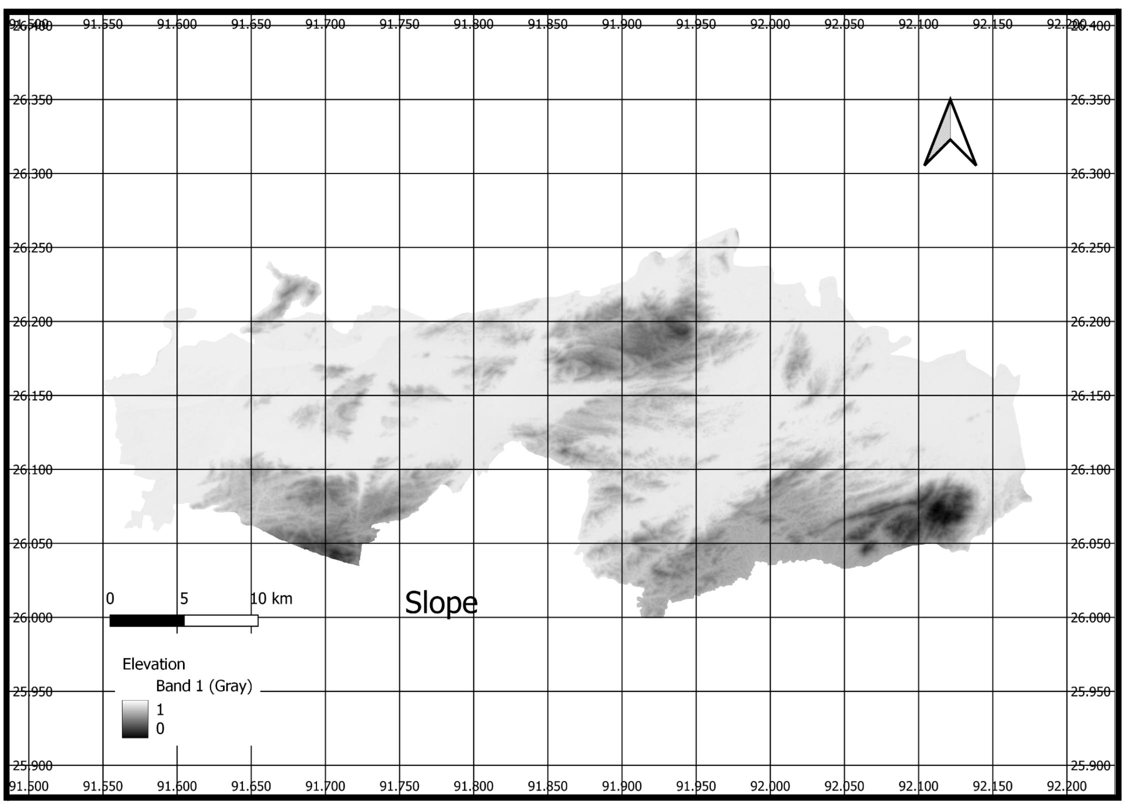

Temperature 2005
Legend
Band 1 (Gray)
High
Low
0 5 10 km

(k)


Friction
Legend
Band 1: friction (Gray)
High
Low
0 5 10 km

l)


Accessibility
Legend
Band 1 (Gray)
High
Low
0
7.5
15 km


Slope
0
5
10 km
Elevation
Band 1 (Gray)
1
0

m)

n)

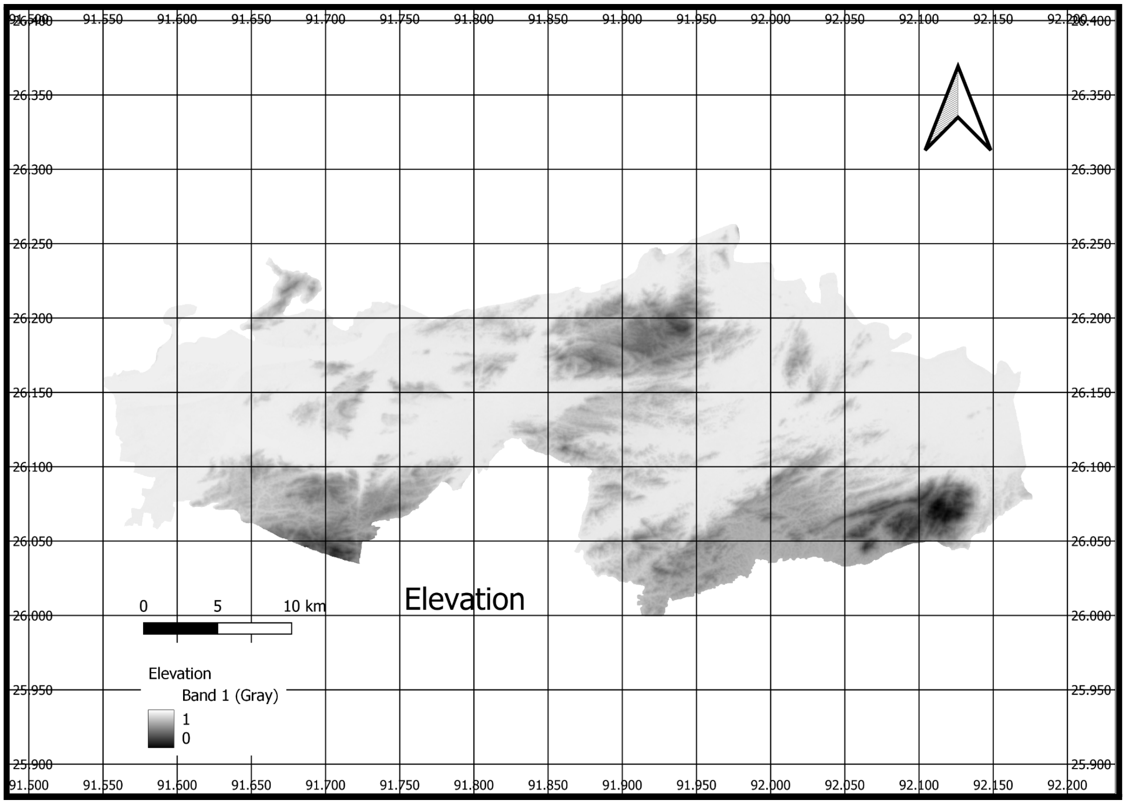


Figure 3: Variables used in the study (a) Kamrup Metro Builtup Progression **(b)** Waterways Proximity **(c)** Building Proximity **(d)** Railways Proximity **(e)** Transporthubs Proximity **(f)** Roads Proximity **(g)** Flood Hazard Zones **(h)** GDP 2005 **(i)** Population Density **(j)** Temperature 2005 **(k)** Friction **(l)** Accessibility **(m)** Slope **(n)** Elevation

9. Results and Discussion

9.1. CA Markov Model

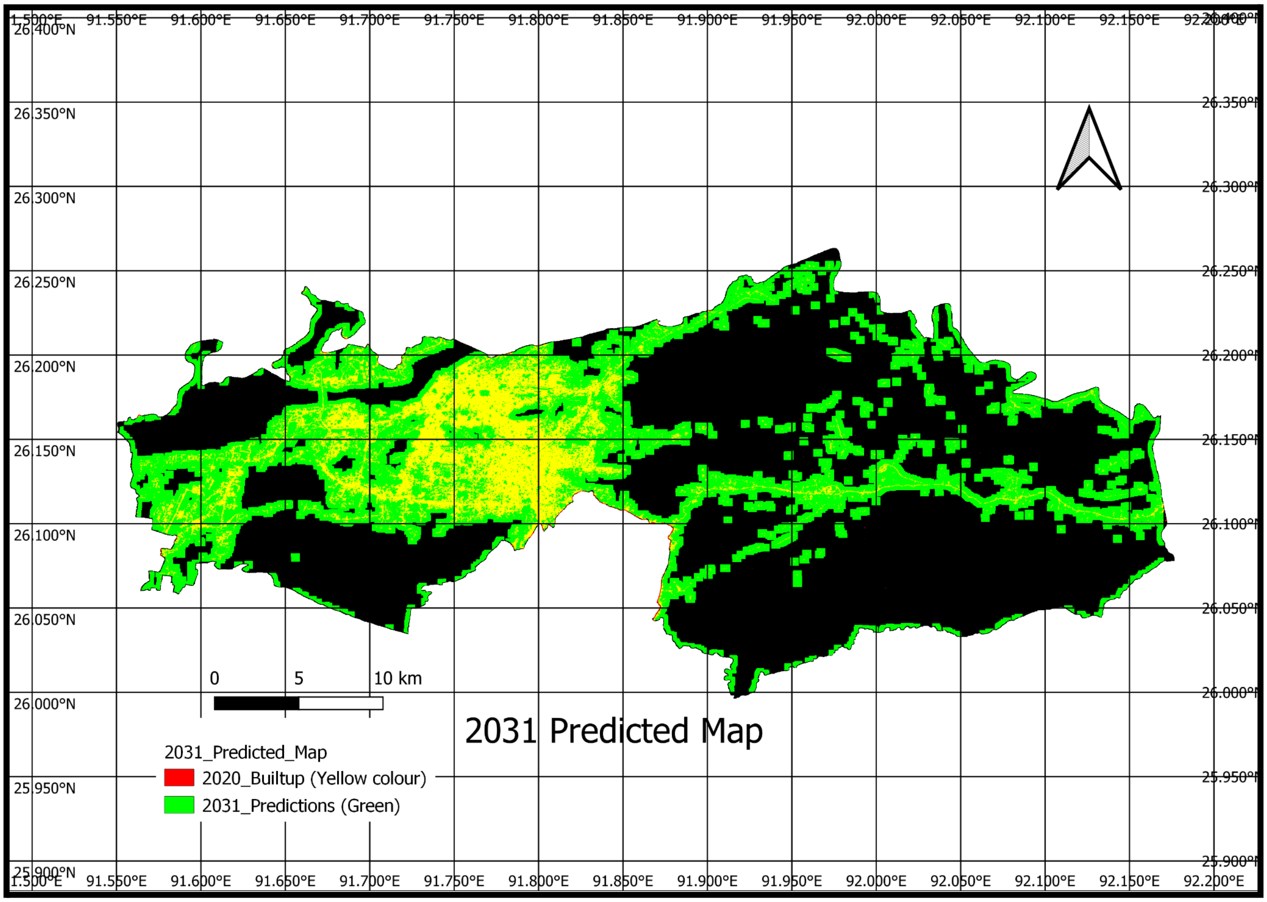


Figure 4: CA Markov Prediction Map. The green areas are the newly predicted areas and the yellow areas are existing urban areas(as of 2020)

The CA Markov model, also known as the Cellular Automata Markov model, is a potent computational tool used for simulating and predicting changes in land use and land cover, including urbanization. It combines concepts from cellular automata and Markov chain modeling to capture the spatial patterns and transition probabilities associated with time-varying land use changes.

The CA Markov model can provide valuable insights into the expansion and transformation of urban areas in the context of urbanization detection. It simulates potential changes based on the probabilities of transitioning from one land use class to another, taking into account the characteristics of the current land cover.

Using the CA Markov model for urbanization detection typically entails the steps outlined below:

The initial step is to collect pertinent data, such as satellite imagery, land cover maps, and socioeconomic data. Typically, these datasets are divided into various land use categories, such as urban, agricultural, forest, and water bodies. Additionally, the data may contain historical records of land use changes, which are restricted areas in our use case.

In this step, the model is calibrated using historical data to estimate the probabilities of transition between various land use categories. On the basis of observed land use changes in the past, Markov chain analysis is utilized to compute these transition probabilities. The calibration procedure guarantees that the model accurately represents urbanization's historical trends and patterns.

Initialization: To initiate the simulation, the CA Markov model requires an initial land use map. This map depicts the distribution of land use at the beginning of the simulation. It can be derived from actual land use data or created using expert knowledge and current urbanization patterns.
We use the 2005 built up data for this model.

Iterative Simulation: The model simulates land use changes iteratively by updating the land use states of individual cells in accordance with the transition probabilities calculated during the calibration step. The simulation takes into account the local neighborhood characteristics and the probabilities of changing land use categories. The state of each cell is modified based on its neighboring cells, and the process is repeated multiple times over multiple time steps. There has to be a specific kernel that is iterated through to obtain the necessary results. We used the simplest 3x3 kernel. We trained our model to predict 2011 built up from 2005, predict 2015 built up from 2011, and predict 2020 built up from 2015. Finally, we predicted 2031 urban cover from 2020 built up.

Validation and Accuracy Evaluation: After the simulation is complete, the resulting land use maps are compared to actual land use data or ground truth information to evaluate the model's accuracy. Validation typically employs statistical metrics such as overall accuracy, Kappa coefficient, and user and producer accuracies. Since there are only two states in this predictive cycle, we use the well known confusion matrix to calculate the accuracy of our model.

Visualization and Interpretation: The final step involves visualizing and interpreting the simulated land use maps. The maps can highlight areas of urban expansion, pinpoint hotspots of land use changes, and offer insights into urbanization's spatial patterns and trends.

Researchers and urban planners can gain a deeper understanding of urbanization dynamics and make more informed decisions regarding land use management, infrastructure planning, and sustainable development by utilizing the CA Markov model. The ability of the model to simulate and predict urbanization patterns is a valuable asset for evaluating future scenarios and guiding policy interventions to promote efficient and sustainable urban growth. This model gives us a powerful tool to use.

Following were the hyperparameters we set after several iterations:
Weights from AHP:

1. Accessibility: 0.48103461
2. Population Density: 0.05658601
3. Elevation: 0.23956212
4. Flood zones: 0.04366571
5. Slope: 0.10001182
6. Friction: 0.07913973

AHP Statistics:
Consistency Index (CI): 0.03749613775072511
Random Index (RI): 1.24
Consistency Ratio (CR): 0.0302388207667138
Threshold for conversion of state:
0.95
Accuracy obtained:
91%

9.2. XGBoost in Predicting Urban Expansion in Kamrup Metropolitan

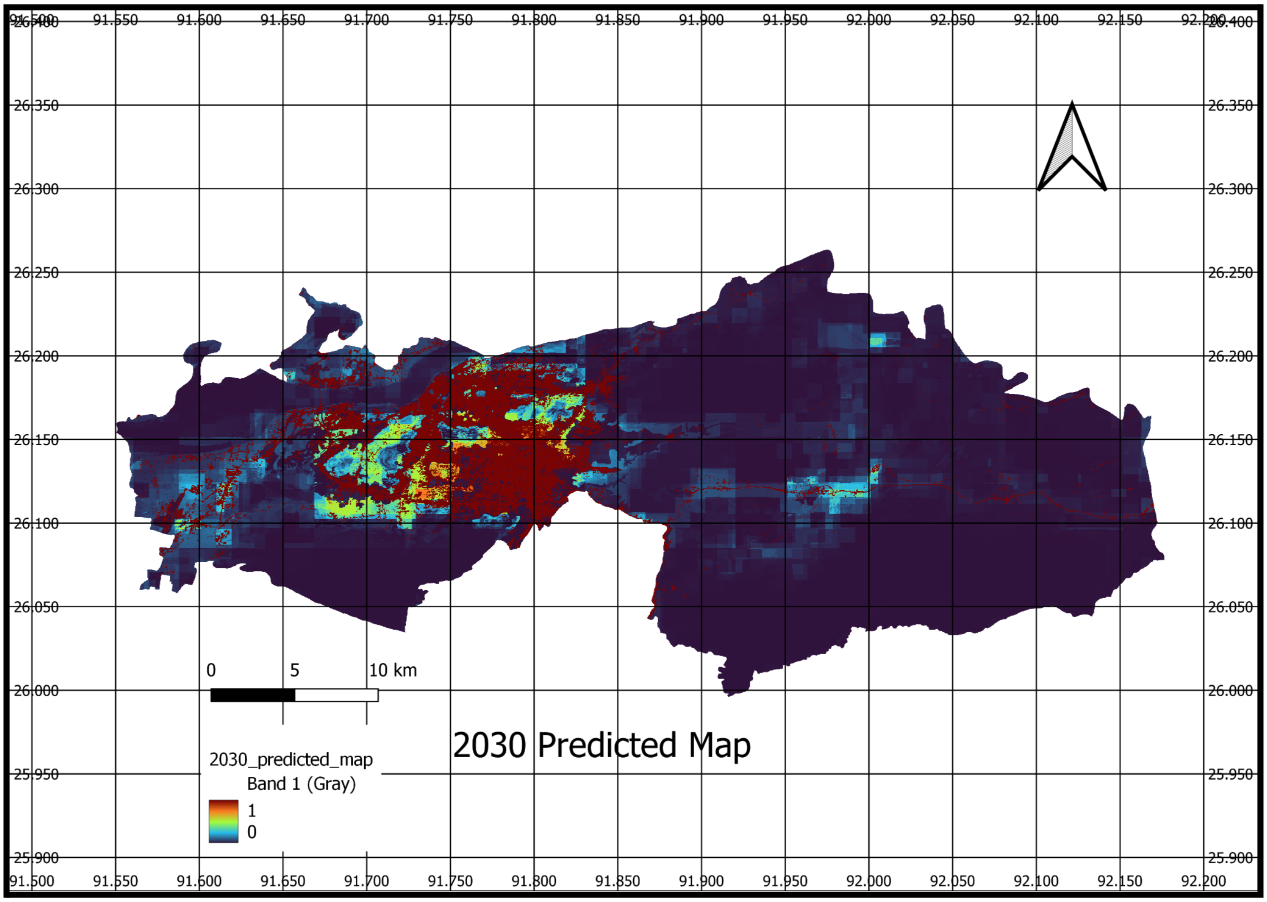


Figure 5: 2030 Predicted Map from XGBoost Model

In our study, we employed XGBoost, an optimized gradient boosting algorithm, as a regression model to predict urban expansion in the Kamrup Metropolitan District. XGBoost is a powerful machine learning algorithm that has gained popularity for its excellent predictive performance and scalability.

Data Preprocessing

Before training the XGBoost model, we performed data preprocessing on the input data. We had a total of nine layers, each representing a different geographic aspect. To ensure consistency, we rescaled and converted each layer to have the same shape and pixel resolution. Additionally, we normalized each layer, assigning specific NaN (Not a Number) values for missing or invalid data.

Model Training

For model training, we selected seven independent features: Accessibility, Elevation, Slope, Floods, GDP, Population Density, and Friction. The two dependent features were the Temperature (data collected from MODIS LST, Google Earth Engine) for the corresponding year and the built-up area for that year. We trained the XGBoost regression model using the dataset from 2005 to 2015. The independent features (X) consisted of the seven aforementioned factors, and the dependent variable (Y) represented the new built-up regions in 2015 that were not present in 2005.

XGBoost Regression Model

XGBoost is an ensemble learning method that combines multiple weak predictive models (decision trees) to create a strong predictive model. It builds trees in a sequential manner, optimizing a specific objective function that measures the model's performance. XGBoost employs gradient boosting, where subsequent trees are trained to correct the mistakes made by previous trees, resulting in improved predictions.

Evaluation and Error Metric

To evaluate the performance of our model, we used the 2010-2020 set of maps. We employed the root mean square error (RMSE) as our error metric. RMSE is a commonly used measure for regression tasks, as it quantifies the average difference between the predicted and actual values. It provides a comprehensive assessment of the model's accuracy and helps identify the magnitude of the prediction errors.

Feature Importance

We obtained a feature importance map from XGBoost, which revealed the relative contribution of each feature in predicting urban expansion. The most influential factors were found to be Accessibility and Temperature. Hypothesizing on this observation, we can speculate that accessibility plays a crucial role in determining urban expansion, as areas with better connectivity and transportation networks are more likely to witness development. Moreover, temperature could be indicative of environmental conditions and local preferences, influencing the choice of urban areas.

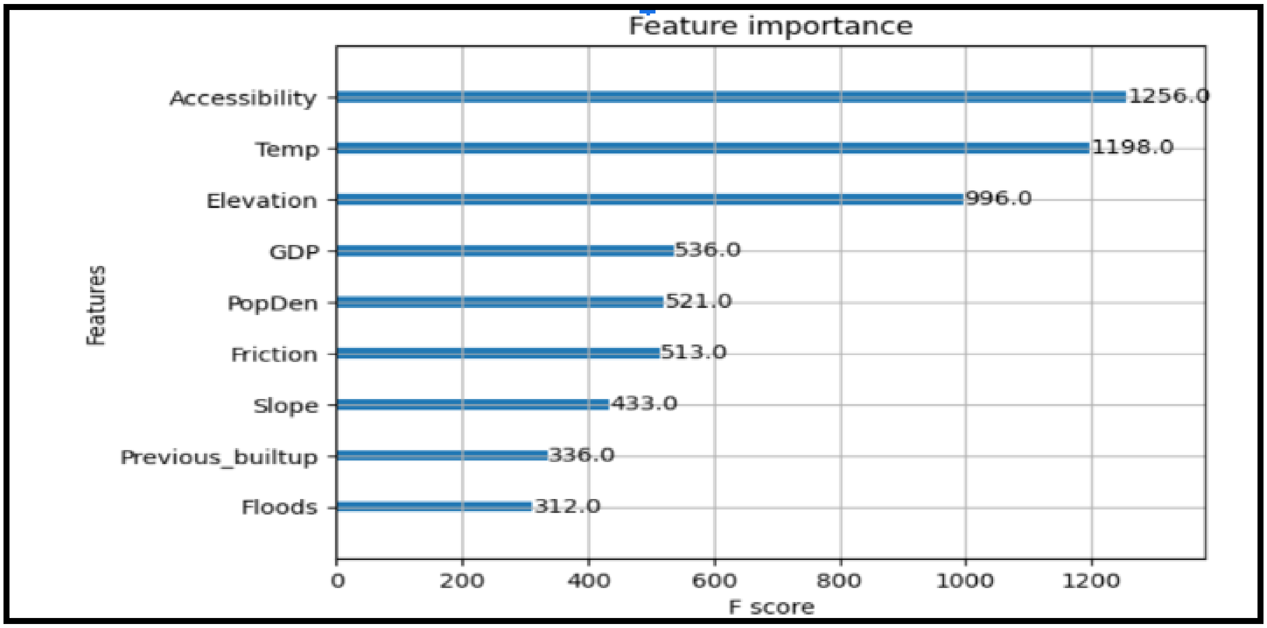


Figure 6: Feature Importance graph of the parameters used in XGBoost Model

Grid Search CV and Best Parameters

Grid Search CV is a technique used to systematically search for the best combination of hyperparameters for a machine learning model. In our study, we employed Grid Search CV to optimize the performance of our XGBoost model. It involved exploring a predefined set of hyperparameters and evaluating the model's performance for each combination.

To ensure robustness and reliability, we performed a 3-fold cross-validation during the Grid Search CV process. This means that we divided our dataset into three subsets: two subsets were used for training the model, and the remaining subset was used for validation. We repeated this process three times, each time using a different subset as the validation set. By doing so, we obtained a more accurate estimation of the model's performance.

Furthermore, to speed up the parameter search process, we utilized five parallel jobs. This allowed us to simultaneously evaluate different parameter combinations, reducing the overall time required to find the optimal parameters.

After conducting the Grid Search CV, we identified the best parameters for our XGBoost model as {'eta': 0.05, 'max_depth': 5}. The 'eta' parameter represents the learning rate, which controls the step size at each boosting iteration. A lower learning rate generally results in more conservative and accurate predictions. The 'max_depth' parameter specifies the maximum depth of the decision trees in the ensemble. A higher value allows the model to capture more complex relationships but may increase the risk of overfitting.

By fine-tuning these hyperparameters through Grid Search CV, we aimed to improve the predictive power and generalization ability of our XGBoost model, enabling it to capture the underlying patterns and dynamics of urban expansion more effectively.

Model Performance

After training and evaluating our model, we obtained the following performance metrics:

F1 Score = 0.6009109740783735
Precision = 0.6790141686401914
Recall = 0.5714270524493812
Accuracy = 0.983405079064589
RMSE = 0.12882127516606515

Interpretation of Performance Metrics

Considering the context of our study, where the positive test cases (newly built-up areas) accounted for only approximately 1.5% of the total cases, an F1 score of 0.6 is considered good. The F1 score balances precision and recall, and in this case, it indicates that our model performs well in correctly identifying and classifying the positive cases, despite their low prevalence. It demonstrates the effectiveness of the XGBoost model in predicting urban expansion in the Kamrup Metropolitan District.

The recall of 0.5714270524493812 suggests that the model successfully captured around 57% of the actual urban expansion areas. While it may not capture all instances of urban expansion, it still identifies a substantial proportion of them.

With a precision of 0.6790141686401914, the model correctly predicted approximately 68% of the urban expansion areas out of all the predicted positive cases. This indicates a relatively low rate of false positives.

The accuracy of 0.983405079064589 indicates a high level of correctness in the overall predictions made by the model. It suggests that the model achieved a high percentage of correct classifications for both positive and negative cases.

Additionally, the low RMSE value of 0.12882127516606515 indicates that the model's predictions were close to the actual values, on average.

Overall, XGBoost proved to be a valuable tool in our data-driven approach, providing accurate predictions and insightful feature importance analysis for understanding urban expansion patterns.

10. Conclusion

In predicting urbanization, XGBoost and CA Markov are two distinct modeling approaches, each with its own advantages and limitations. XGBoost frequently outperforms CA Markov in terms of urbanization prediction due to its superior accuracy and adaptability. XGBoost's ability to handle nonlinear relationships between input variables and the target variable is a significant advantage. Urbanization processes are complex and influenced by a variety of variables, including population growth, economic development, and infrastructure expansion. The capability of XGBoost to capture and model complex nonlinear interactions between these variables makes it well-suited for a more accurate prediction of urbanization patterns.

In addition, XGBoost is highly adaptable and can easily accommodate a variety of input data types, such as satellite imagery, socioeconomic data, and historical urbanization records. It can effectively integrate multiple data sources and extract pertinent characteristics, enabling a thorough analysis of urbanization's drivers and dynamics. CA Markov, on the other hand, relies primarily on historical land use data and transition probabilities to simulate land use changes, which may limit its adaptability to include diverse data types.

In addition to enhanced interpretability, XGBoost provides feature importance rankings that enable researchers and planners to identify the most influential factors contributing to urbanization. This information can more effectively guide policy interventions and urban planning strategies. CA Markov, on the other hand, focuses on simulating land use changes based on historical patterns and may not provide detailed insights into the fundamental drivers of urbanization.

While CA Markov has its merits, particularly in capturing spatial dependencies and providing a cellular-level analysis, XGBoost's improved predictive power, flexibility, and interpretability make it the best option for urbanization prediction. Its ability to handle nonlinear relationships, incorporate diverse data sources, and provide insights into influential factors significantly contributes to more accurate and informed decision-making in urban planning and policy formulation.

11. Future Scope

In some respects, the predictions that the CA Markov model has made are not entirely accurate. We were unable to determine the reason for the results, which were not at all what we had anticipated. Few areas have a large pixel resolution, and even fewer areas have what would be considered normal. Finding a solution to the problem is going to be one of our primary focuses in the future.

The model for predicting urbanization has proven its ability to capture and predict the complex dynamics of urban growth. As we look to the future, several opportunities arise for expanding and refining the existing model's capabilities:

Integration of High-Resolution Data: The urbanization prediction model's future scope includes the incorporation of high-resolution satellite imagery and other remote sensing data. With the growing availability of high-resolution data, incorporating this information into the model can provide more precise insights into urbanization patterns and facilitate the identification of smaller-scale changes within urban areas.

Real-Time Predictions: Technological advancements permit the model to make real-time predictions of urbanization trends. By leveraging near-real-time data streams and cloud computing, the model can continuously update its predictions, enabling decision-makers to swiftly respond to emerging urbanization challenges and opportunities.

Climate Change Impact Assessment: The future scope of the model will include integrating climate change variables in order to evaluate their impact on urbanization patterns. Understanding how climate change affects urbanization can inform climate-resilient urban planning and help identify areas at risk of environmental degradation.

Incorporation of Socioeconomic Drivers: Future improvements to the model should include the incorporation of socioeconomic drivers such as population dynamics, economic indicators, and migration patterns. Taking these factors into account can provide a more complete understanding of the forces driving urbanization and aid in the design of policies that promote inclusive and sustainable development.

The expansion of the model to include simulations of the development of urban infrastructure can be extremely beneficial. The model can contribute to efficient urban planning and maximize resource allocation by predicting the locations and types of future infrastructure projects, such as roads, utilities, and public facilities.

Integrating advanced machine learning algorithms and hybrid modeling techniques can further improve the model's predictive accuracy and adaptability. Ensemble methods, deep learning architectures, and hybrid models can capture complex interactions between various urbanization drivers, resulting in more accurate forecasts.

Interactive Visualization and Scenario Analysis: The future scope of the model includes the development of interactive visualization tools that allow users to explore various policy interventions and scenarios. Using the model's outputs, decision-makers can evaluate the efficacy of various urban planning strategies.

Application on a Global Scale: While the current model may concentrate on specific regions, its future scope includes expanding its application to encompass urbanization forecasts on a global scale. Expanding the scope of the model to include additional countries and continents can support global efforts in sustainable urban development and inform international policy initiatives.

Integrating the urbanization prediction model with decision support systems can provide policymakers with empirical insights. The incorporation of the model's predictions into decision support tools can facilitate informed decision-making and aid in the formulation of effective urban policies.

In conclusion, the future scope of the model for predicting urbanization includes a number of enhancements and expansions. By incorporating high-resolution data, real-time predictions, an assessment of the impact of climate change, socioeconomic drivers, and simulations of urban infrastructure, the model can provide a more comprehensive understanding of urbanization dynamics. Integration with sophisticated machine learning algorithms, interactive visualization, and decision support systems increases its value to policymakers and urban planners. The ultimate goal of the model's future development is to promote sustainable and resilient urban growth in response to evolving global challenges.

12. Appendix

1. Principal Component Analysis (PCA) Code:

```
# Install required packages
install.packages(c("raster", "rasterVis"))

# Load required libraries
library(raster)
library(rasterVis)

# Set the path to the input geotiff file
input_file <- "C:\\Users\\Adi\\Downloads\\FloodHZ\\FloodHZ\\Kamrup_FHZ.tif"

# Load
r <- brick(input_file)

# Reshape raster stack into a matrix
mat <- as.matrix(r)

# Perform PCA on the preprocessed matrix
pca <- prcomp(mat, scale. = TRUE)

# Extract the first principal component as a single-band raster
pc1 <- pca$x[, 1]

# Create a raster object from the single-band data
pc1_raster <- raster(r)
values(pc1_raster) <- pc1
#While taking the single band output after the PCA, techically, more intense is white-
#-pixel wise but if you wish it to be inverted, we can simply deduct the value from-
#-255
#pc1_raster <- 255 - pc1_raster

#Plot
plotRGB(r, r = 1, g = 2, b = 3)
plot(pc1_raster)

#Output
output_file <- "C:\\Users\\Adi\\Downloads\\FloodHZ\\FloodHZ\\Kamrup_FHZ_pca.tif"
writeRaster(pc1_raster, filename = output_file, format = "GTiff", overwrite = TRUE)
```

2. Analytical Hierarchy Process (AHP) Code:

```
# Define the pairwise comparison matrix
pairwise_matrix <- matrix(c(
  1, 3, 2, 4, 3,
  1/3, 1, 1/2, 3, 2,
  1/2, 2/3, 1, 4, 3,
  1/4, 1/3, 1/4, 1, 1/2,
  1/3, 1/2, 1/3, 2, 1
), nrow = 5, byrow = TRUE)

# Calculate the initial weights as the column-wise geometric mean
weights <- apply(pairwise_matrix, 2, FUN = function(col) prod(col) ^ (1 / length(col)))

# Iterate until the consistency index is less than 1
consistency_index <- Inf
tolerance <- 0.0001  # Define the desired tolerance
max_iterations <- 100  # Define the maximum number of iterations

for (i in 1:max_iterations) {
  # Normalize the pairwise matrix by dividing each element by its respective column sum
  normalized_matrix <- pairwise_matrix / colSums(pairwise_matrix)

  # Calculate the updated weights as the column-wise geometric mean of the normalized matrix
  updated_weights <- apply(normalized_matrix, 2, FUN = function(col) prod(col) ^ (1 / length(col)))

  # Calculate the consistency index (CI)
  consistency_index <- (sum(updated_weights * colSums(pairwise_matrix))) / sum(updated_weights)

  # Calculate the consistency ratio (CR)
  random_index <- c(0, 0, 0.58, 0.9, 1.12, 1.24, 1.32, 1.41, 1.45, 1.49)
  consistency_ratio <- consistency_index / random_index[length(pairwise_matrix)]

  # Check if the consistency index is less than the tolerance
  if (consistency_index < 1) {
    break
  }

  # Update the weights for the next iteration
  weights <- updated_weights
}
```

3. [CA Markov Code]()
4. [XGBoost Code]()

13. References

1. Modeling and Predicting Urban Expansion in South Korea Using Explainable Artificial Intelligence (XAI) Model (link)
2. Spatio-temporal simulation of future urban growth trends using an integrated CA-Markov model (link)
3. Multicriteria Evaluation of Tourism Potential in the Central Highlands of Vietnam: Combining Geographic Information System (GIS), Analytic Hierarchy Process (AHP) and Principal Component Analysis (PCA) (link)
4. CA Markov Approach in Dynamic Modelling of LULCC Using ESA CCI Products over Zambia (link)