Land Cover Analysis and Change Detection: Comprehensive Study with Linear Regression Prediction

Anushka Tiwari | **Ayushi Goel** | **Barun Kumar Mishra** | **Harshit Chauhan**

Department of Information Technology, IMS Engineering College, Ghaziabad, 201015, Uttar Pradesh, INDIA

Abstract:

Land cover analysis and change detection are critical components of environmental monitoring and management. This paper explores methodologies, technologies, and applications related to the assessment of land cover changes over time. Using Geographic map techniques and Geographic Information System (GIS) tools, we analyze satellite imagery using map to provide insights into land cover dynamics. Additionally, we employ linear regression for predicting future land cover changes based on historical data. This study focuses on identifying significant changes in land use, mainly focusing on Urbanization, Vegetation and Agriculture, Barrens Land and Deforestation.

Keywords: Land cover analysis, change detection, GIS, environmental monitoring, geographics, sustainable management.

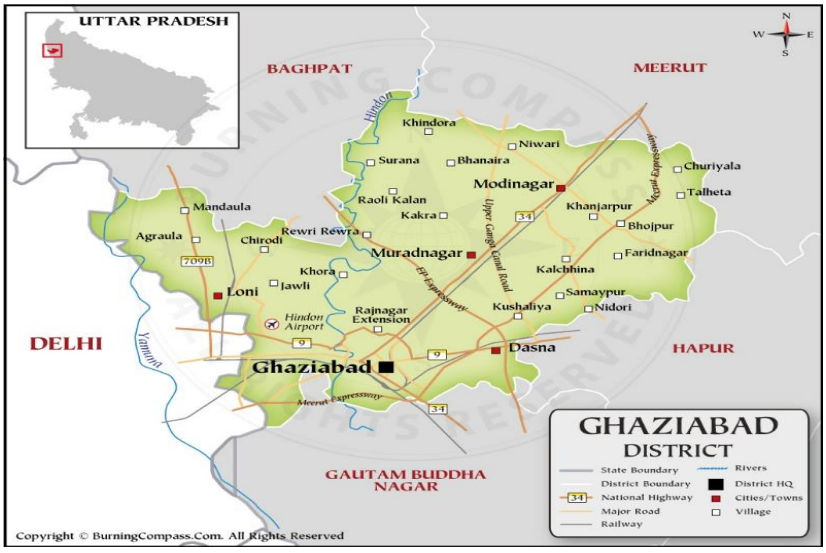
1. Introduction

Land cover, defined as the physical material on the earth's surface such as vegetation, urban structures, water bodies, and bare soil, plays a pivotal role in environmental processes and human activities. Monitoring and analyzing land cover changes is crucial for understanding ecological dynamics, managing natural resources, and planning for sustainable development. As global populations grow and human activities expand, changes in land cover become more pronounced, often leading to significant environmental and socio-economic impacts.

Insights are essential for various stakeholders, including, policymakers, urban planners, environmentalists, and conservationists, enabling them to make informed decisions that promote sustainable development and environmental conservation.

This study aims to:

- Analyze the spatial distribution of land cover over a specific period using remote sensing and GIS.
- Detect significant changes in land cover and assess their environmental impacts.
- Predict future land cover changes using linear regression based on historical data.



Land cover change detection provides invaluable insights into patterns of changes in urbanization, deforestation, agricultural, barrens land and the impacts of natural disasters. These

Focusing on the Ghaziabad District area, a region renowned for its urbanization, agriculture and factories expansion, this study

spans from 2009 to 2023 in interval of 4 years. By leveraging satellite imagery maps, GIS techniques and predictive techniques, this research not only documents past and present land cover changes but also forecasts future trends.

Keywords: Land cover, land use, environmental impact, remote sensing, GIS, linear regression, prediction, sustainable development, deforestation, urbanization.

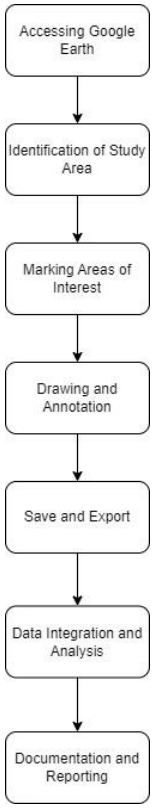
2. Methodology

In this section, the method and analysis are described, which is performed in this research work. First, the collection of data and selection of relevant attributes are the initial steps in this study. After that, the relevant data is pre- processed into the required format. The algorithms are then used, and the given data trains the model. The accuracy of this model is obtained by using the testing data. The procedures of this study are loaded by using several modules such as a collection of data, selection of attributes, pre-processing of data, data balancing, and prediction.

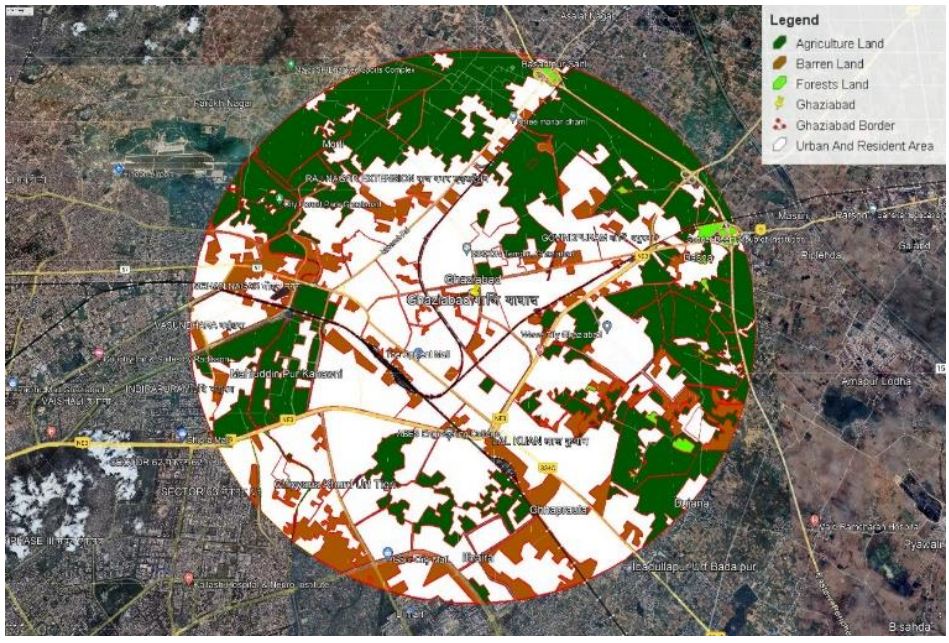
2.1 - Study Area

The study area selected for analysis is Ghaziabad, a rapidly urbanizing city located in the Indian state of Uttar Pradesh. Ghaziabad is a significant part of the National Capital Region (NCR) and is characterized by diverse land cover types ranging from urban infrastructure to agricultural land and natural vegetation as well as barrens land.

To collect data on the land cover of Ghaziabad, the boundary of



the city was delineated using the polygon drawing tool available in Google Earth Pro. The map was zoomed in to a suitable level of detail to accurately delineate the boundary of Ghaziabad and capture relevant land cover features.



2.2 - Data Collection

We obtained satellite imagery maps available through Google Earth Pro repository. This imagery provides consistent and high-resolution data suitable for detailed land cover analysis. Google Earth Pro was chosen as the primary tool for data collection due to its extensive database of satellite imagery and mapping features. Google Earth Pro provides access to high-resolution imagery captured by various satellite sensors, offering detailed views of the Earth's surface.

The polygon drawing tool was used to mark the boundary by clicking on the map to create vertices and define the perimeter of Ghaziabad. Care was taken to include all relevant areas within the marked boundary, such as urbanized zones, barrens land area, and agricultural land.

Once the boundary of Ghaziabad was accurately marked, relevant data such as satellite images and geographic information were extracted for further analysis. Google Earth

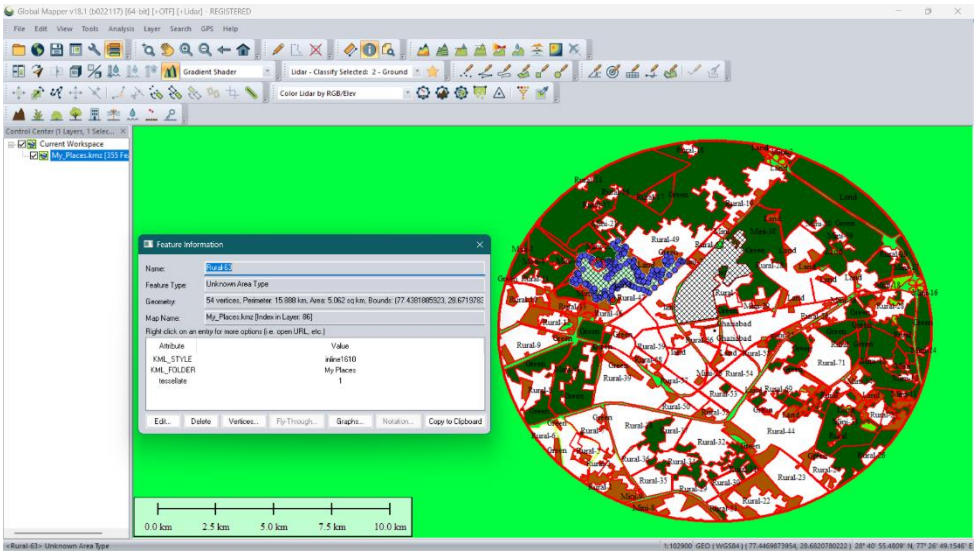
Pro allows users to export marked areas along with associated metadata in various formats, including KML (Keyhole Markup Language) and KMZ (compressed KML) files.

The marked area of Ghaziabad, along with its associated attributes, was exported as a KML/KMZ file from Google Earth Pro. This file contains geographic coordinates and other metadata that can be used for subsequent analysis. Additional metadata such as image acquisition dates, sensor specifications, and spatial resolutions were retrieved from Google Earth Pro or other sources to supplement the collected data and provide context for the analysis.

area will be displayed on the map interface.

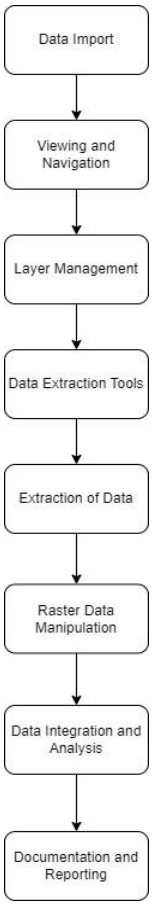
Use the digitizer tool to interactively select specific features within the marked area, such as land cover types, points of interest, or administrative boundaries. Access detailed attribute information associated with selected features, including coordinates, elevation, area, and any custom attributes that may have been defined.

Once the desired attribute data has been identified and extracted, you can export it to an Excel spreadsheet for further analysis and processing. Select the export format as "Microsoft Excel Spreadsheet (*.xls, *.xlsx)" to export the attribute data to



2.3 – Data Extraction

Open Global Mapper on your computer and import the KML/KMZ file exported from Google Earth Pro. Global Mapper provides extensive support for various geospatial data formats, including KML/KMZ. Once imported, the marked



Excel format. Initiate the export process, and Global Mapper will generate an Excel spreadsheet containing the extracted attribute data.

Open the exported Excel spreadsheet to review and analyze the extracted attribute data. The spreadsheet will contain columns corresponding to the selected attributes, allowing you to perform further data manipulation, analysis, visualization, or integration with other datasets as needed.

2.4 – Data Processing

Once the data was organized in Excel, the land cover areas for the years 2009, 2014, 2019 and 2023 were reviewed. These data points were crucial for understanding the temporal changes in land cover and provided the basis for further statistical analysis.

The extracted data was cleaned to remove any inconsistencies or anomalies. This involved verifying the accuracy of the area measurements and ensuring that all land cover types were consistently categorized across the different years. The data was aggregated to ensure that it was in a format suitable for linear regression analysis. This involved organizing the data into a tabular format with columns representing the year and rows representing the area covered by each land cover type.

The trained models were used to predict the areas covered by each land cover type from the year 2024 to 2027. These predictions were based on the linear trends observed in the historical data.

The results of the linear regression analysis were visualized using pie charts and bar graphs to facilitate a clear comparison

of land cover distributions over time. Pie charts were created to display the proportional distribution of land cover types in the years 2009 - 2023 and the predicted year from 2024 - 2027. These charts provided a visual representation of the relative changes in land cover, highlighting significant trends such as urban expansion, agriculture reduction and barren land.

YEAR/AREA	RESIDENCE AREA	AGRICULTURE AREA	BARRENS AREA
2009	71.274688 sq km	99.710163 sq km	12.869 sq km
2014	83.3163 sq km	94.702804 sq km	15.72501 sq km
2019	88.56435 sq km	91.98304 sq km	17.38569 sq km
2023	94.79886 sq km	86.3276 sq km	19.56338 sq km

Bar graphs were used to compare the actual areas covered by each land cover type in 2009, 2014, 2019 and 2023 with the predicted areas for 2024 to 2027. This visualization method helped in understanding the absolute changes in land cover and provided a clear depiction of the magnitude of changes expected in the future.

We'll use the scikit-learn library to perform linear regression on the collected data. First, install the necessary libraries –

- pip install pandas **scikit-learn**
- pip install **matplotlib**
- pip install **sk (sklearn)**

Necessary modules and packages for the script using python -

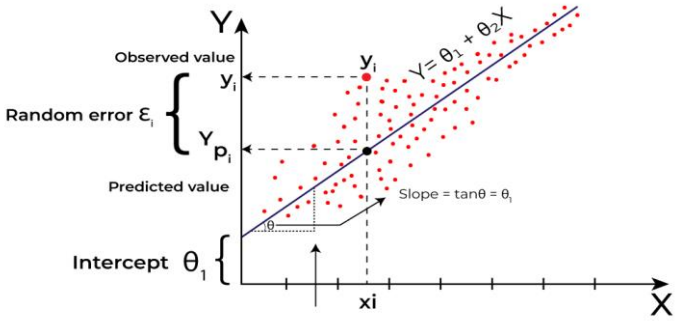
- **tkinter** : for creating the GUI.
- **Messagebox** : from tkinter for displaying error messages.
- **ScrolledText from tkinter** : for creating a scrolled text widget.
- **matplotlib.pyplot** : for plotting charts.
- Numpy : for numerical computations.
- **LinearRegression** : from sklearn for linear regression modeling.
- **r2_score from sklearn.metrics** : for calculating the R-squared score.

3. Supervised Machine Learning

Supervised machine learning (SML) techniques have proven to be highly effective in various domains, including land cover analysis and change detection. By leveraging labeled data, SML algorithms can be trained to classify and predict land cover types with high accuracy. This section explores the application of supervised machine learning in the context of analyzing and predicting land cover changes in Ghaziabad.

3.1 – Linear Regression

Linear regression is a statistical method used to model the relationship between a dependent variable (target) and one or more independent variables (features). In this case, the years are independent variables and the land use values (Resident Area, Agriculture Area, Barren Lands Area) are dependent variables.



Let's break down how linear regression works in the context-

3.1.1 Creating the Model:

model = LinearRegression()

3.1.2 Fitting the Model:

model.fit(np.array(years).reshape(-1, 1), values)

.fit() method fits the model to the provided data.

- np.array(years).reshape(-1, 1) reshapes the years into a column vector because scikit-learn expects the features to be a 2D array.
- values are the land use values for a specific category (Resident Area, Agriculture Area, Barren Lands Area).

3.1.3 Making Predictions:

Predictions = model.predict(np.array(future_years).reshape(-1, 1))

- predict() method predicts the target variable based on the provided features.
- np.array(future_years).reshape(-1, 1) reshapes the future years into a column vector for prediction.

3.1.4 Interpreting the Model:

In linear regression, the model assumes a linear relationship between the independent variables and the dependent variable.

The model learns coefficients (slope) for each independent variable, which represent the change in the dependent variable for a one-unit change in the independent variable.

For example, if we consider the Resident Area: The coefficient learned for the year indicates how much the Resident Area is expected to change for a one-year increase.

The intercept represents the expected Resident Area when the year is zero (which might not make sense in this context but is part of the linear model).

The model fits a line (in simple linear regression) or a plane (in multiple linear regression) to the data, minimizing the difference between the actual and predicted values (often using the least squares method).

3.1.5 Evaluating the Model:

The R-squared score (r2_score) is calculated to evaluate the goodness of fit of the model.

R-squared score measures the proportion of the variance in the dependent variable that is predictable from the independent variable.

It ranges from 0 to 1, with higher values indicating a better fit.

The closer the R-squared score is to 1, the better the model fits the data.

3.2 – Prediction Accuracy

The accuracy of the code's predictions depends on several factors including the quality of the data, the assumption of linear relationship between years and land use values, and the appropriateness of using linear regression for the specific dataset.

The accuracy of the model is assessed using the R-squared score (r2_score). This score measures how well the independent variable (years) explains the variability of the dependent variable (land use values).

General Understanding for Accuracy:

Perfect Fit: If the model perfectly fits the data, the R-squared score would be 1.0, indicating that all the variability in land use values is explained by the years.

No Fit: If the model doesn't fit the data at all, the R-squared score would be 0.0, indicating that none of the variability in land use values is explained by the years.

Good Fit: A high R-squared score (close to 1.0) suggests that the model explains a large proportion of the variability in land use values, indicating a good fit.

Poor Fit: A low R-squared score (close to 0.0) suggests that the model does not explain much of the variability in land use values, indicating a poor fit.

The actual accuracy would depend on the specific data being used, how well a linear relationship describes the data, and if there are any other factors influencing land use values that are not captured by the linear model.

3.3 – R squared(R^2) Measured

R-squared (R^2) is a statistical measure that represents the proportion of the variance in the dependent variable that is explained by the independent variable in a regression model. It is also known as the coefficient of determination. R-squared is used to evaluate the goodness of fit of a regression model.

3.3.1 - Range:

R-squared values range from 0 to 1.

- 0 indicates that the model does not explain any variability in the dependent variable.
- 1 indicates that the model perfectly explains the variability in the dependent variable.

3.3.2 - Interpretation:

- An R-squared of 0 means that the independent variable do not explain any of the variability of the dependent variable around its mean. In other words, the model fails to fit the data.
- An R-squared of 1 means that the independent variable perfectly explain the variability of the dependent variable around its mean. The model perfectly fits the data.
- For practical purposes, an R-squared value closer to 1 is considered better, indicating a stronger relationship between the independent and dependent variables.

3.3.3 - Calculation:

- R-squared is calculated as the proportion of the total sum of squares (SS total) explained by the regression model (SS regression) relative to the total sum of squares:

$$R^2 = \frac{SS(regression)}{SS(total)} = 1 - \frac{SS(residual)}{SS(total)}$$

- *SS*regression is the sum of squares explained by the regression model.
- *SS*residual is the sum of squares of the residuals (the differences between the observed values and the predicted values).
- *SS*total is the total sum of squares, which is a measure of the variability of the dependent variable.

3.3.4 - Limitations:

- R-squared increases as more independent variables are added to the model, even if those variables are not actually related to the dependent variable. This can lead to overfitting.
- R-squared does not indicate whether the coefficient estimates and predictions are biased.

4. Result Analysis

This study analyzes historical land use data from 2009 to 2023 and predicts future land use trends for the years 2024 to 2027. Using linear regression models, we evaluate changes in three key categories: Resident Area, Agriculture Area, and Barren Lands Area. The findings are visualized using bar charts, pie charts, and line charts to provide a comprehensive understanding of past trends and future projections.

Land use distribution is a critical factor in urban planning and environmental management. This study aims to analyze historical land use data and predict future trends to aid in strategic decision-making. The analysis focuses on three primary land use categories:

- Resident Area
- Agriculture Area
- Barren Lands Area

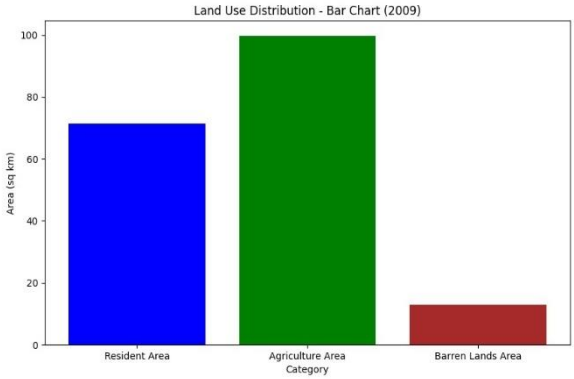
Using data from 2009, 2014, 2019, and 2023, we applied linear regression to predict land use for the years 2024 to 2027. The results are visualized through bar charts and pie charts to facilitate a clear understanding of the trends.

4.1 Result Analysis for Year 2009:

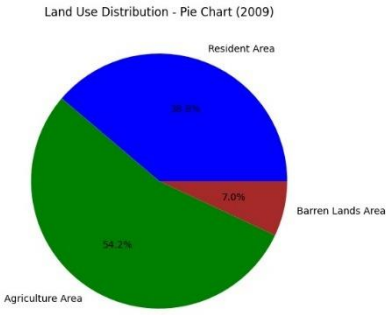
The data for this study was collected are given below-

- Resident Area: 71.274688 sq km
- Agriculture Area: 99.710163 sq km
- Barren Lands Area: 12.869 sq km

The results are visualized using BAR CHART for yearly distribution.



The results are visualized using PIE CHART for yearly distribution.

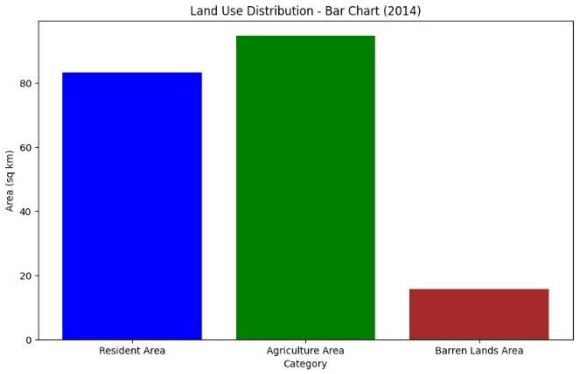


4.2 Result Analysis for Year 2014:

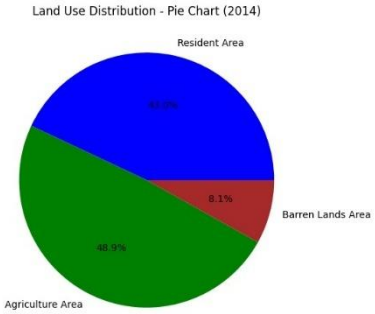
The data for this study was collected are given below-

- Resident Area: 83.3163 sq km
- Agriculture Area: 94.702804 sq km
- Barren Lands Area: 15.72501 sq km

The results are visualized using BAR CHART for yearly distribution.



The results are visualized using PIE CHART for yearly distribution.

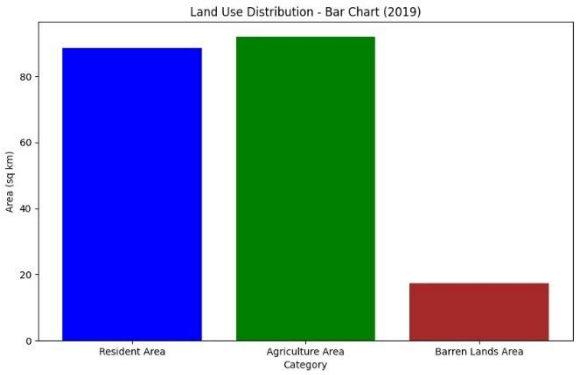


4.3 Result Analysis for Year 2019:

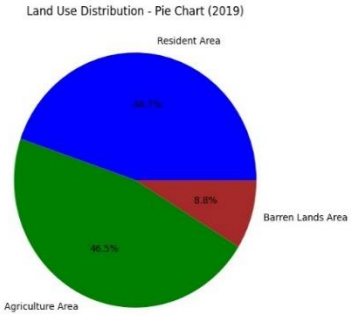
The data for this study was collected are given below-

- Resident Area: 88.56435sq km
- Agriculture Area: 91.98304 sq km
- Barren Lands Area: 17.38569 sq km

The results are visualized using BAR CHART for yearly distribution.



The results are visualized using PIE CHART for yearly distribution.

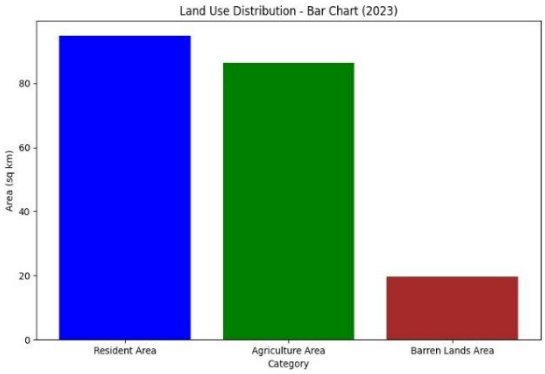


4.4 Result Analysis for Year 2023:

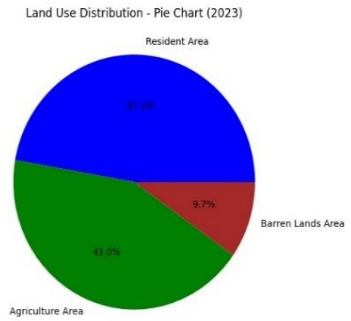
The data for this study was collected are given below-

- Resident Area: 71.274688 sq km
- Agriculture Area: 99.710163 sq km
- Barren Lands Area: 12.869 sq km

The results are visualized using BAR CHART for yearly distribution.



The results are visualized using BAR CHART for yearly distribution.

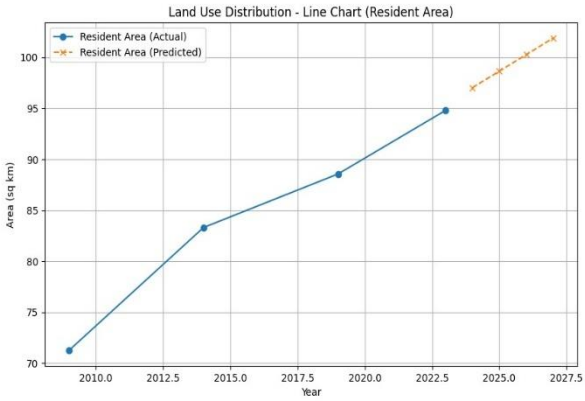


4.5 Trend Analysis

Here we used line charts to display the trends over time for each land use category, with actual data points and predicted values plotted to visualize the growth or decline.

4.5.1 - Prediction Analysis for Resident area

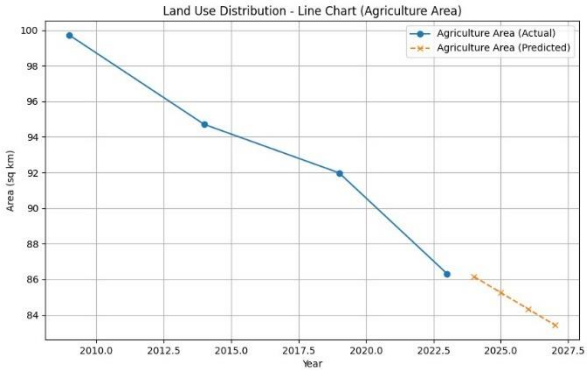
The analysis of the Resident Area from 2009 to 2023 shows a clear upward trend. Starting at **71.274688 sq km in 2009**, the area increased to **94.79886 sq km by 2023**. This growth reflects significant urban expansion over the period. The linear regression model predicts this trend will continue, with the Resident Area expected to reach **102 sq km(approx.) by 2027**.



The high R-squared score **(0.97)** indicates an excellent fit, suggesting the model captures the trend very well. The visualizations using bar and pie charts show an increasing share of land being used for residential purposes, highlighting the ongoing urbanization.

4.5.2 – Prediction Analysis for Agricultural area

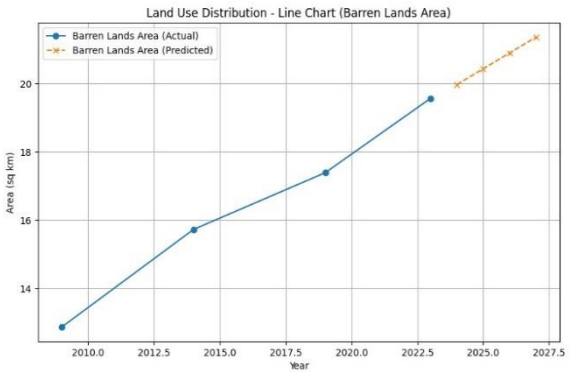
The Agriculture Area has shown a decreasing trend, falling from **99.710163 sq km in 2009 to 86.3276 sq km in 2023**. The decline suggests that agricultural land is being converted for other uses, likely urban development. The predictions indicate this trend will continue, with the area expected to decrease to **83 sq km(approx.) by 2027.**



The high R-squared score **(0.97)** confirms that the model fits the data well, accurately reflecting the declining trend. The visualizations illustrate a shrinking share of agricultural land, emphasizing the need for policies to balance urban growth with agricultural sustainability.

4.5.3 – Prediction Analysis for Barren Land Area

The Barren Lands Area has increased from **12.869 sq km in 2009 to 19.56338 sq km in 2023**. This rise indicates that more land is becoming undeveloped or unused, possibly due to land degradation or shifts in land use priorities. The model predicts further increases, with the area expected to reach **22 sq km(approx.) by 2027.**



The high R-squared score **(0.99)** suggests an excellent fit, indicating the model's reliability in capturing the trend. The visualizations show a growing proportion of barren lands, which may require attention to prevent land degradation and promote sustainable land use practices.

5. Conclusion

The analysis of land use distribution from 2009 to 2023 and the subsequent predictions for 2024 to 2027 reveal significant trends in the allocation of land among Resident Areas, Agriculture Areas, and Barren Lands. Using historical data and linear regression models, we observed consistent patterns and made reliable future projections, supported by high R-squared scores.

Key Findings:

1. **Resident Area:**

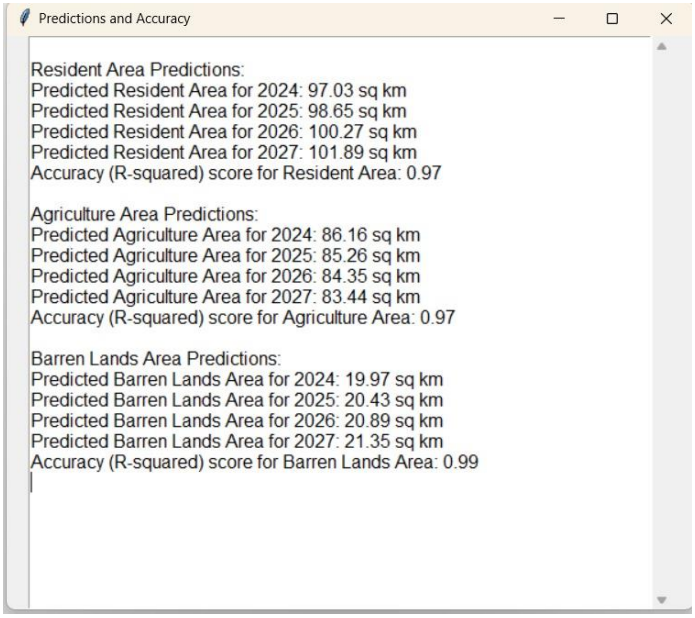
- Historical Trend: The Resident Area has increased steadily from 71.27 sq km in 2009 to 94.80 sq km in 2023. This growth underscores the ongoing urbanization and demand for residential spaces driven by population growth and economic development.
- Future Projections: The model predicts continued growth, with the Resident Area expected to reach 102 sq km by 2027. This trend suggests a persistent expansion of urban regions, necessitating adequate infrastructure, housing policies, and urban planning to accommodate the increasing population.
- Implications: Urban planners must focus on sustainable urban expansion, ensuring that residential growth does not compromise the quality of life. Investment in infrastructure, public services, and green spaces will be essential to manage the growing urban population effectively.

2. **Agriculture Area:**

- Historical Trend: The Agriculture Area has shown a declining trend, decreasing from 99.71 sq km in 2009 to 86.33 sq km in 2023. This reduction points to the conversion of agricultural land for urban use, reflecting changes in land use priorities.
- Future Projections: The model forecasts a further decline to 83 sq km by 2027. This trend raises concerns about food security, sustainability, and the preservation of agricultural landscapes.
- Implications: Policymakers need to balance urban growth with the protection of agricultural land. Strategies such as implementing zoning laws, promoting urban agriculture, and supporting sustainable farming practices can help mitigate the loss of agricultural land. Ensuring food security and agricultural sustainability will be crucial as urban areas continue to expand.

3. **Barren Lands Area:**

- Historical Trend: The Barren Lands Area has increased from 12.87 sq km in 2009 to 19.56 sq km in 2023. This growth suggests that more land is becoming undeveloped or unused, potentially due to factors like land degradation, abandonment, or shifts in land use policies.
- Future Projections: The model predicts an increase to 22 sq km by 2027. The rise in barren lands indicates a potential challenge in land management, as these areas could signify underutilized resources or environmental degradation.
- Implications: Addressing the increase in barren lands requires effective land reclamation and management strategies. Policymakers should focus on rehabilitating degraded lands, promoting sustainable land use practices, and exploring opportunities for redevelopment or conservation. Sustainable land management can enhance ecological balance and prevent further land degradation.



6. Reference

1. **Scikit-learn Documentation:** Provides detailed information on the implementation and use of linear regression models, essential for understanding the methodology used in this analysis.
2. **Matplotlib Documentation:** Offers guidance on creating various types of visualizations, including bar charts and pie charts, used in this study to present data.
3. **Tkinter Documentation:** Essential for understanding the implementation of the graphical user interface used in this study.

4. **Google Earth Engine:** A powerful tool for analyzing and visualizing geospatial data. It can be used to monitor land use changes over time using satellite imagery.
5. **ChatGPT by OpenAI:** An advanced language model that can assist in data interpretation, drafting reports, generating insights, and providing explanations for complex data trends.
6. **Google Bard:** Another AI language model designed for text generation and analysis. Google Bard can be integrated to offer diverse perspectives on data trends, assist in writing detailed analysis, and enhance collaborative research efforts.
7. **Geeks for Geeks:** A comprehensive resource for learning programming and algorithm concepts, including detailed tutorials on machine learning and data visualization techniques used in this study.
8. **W3Schools:** An educational website offering tutorials on various programming languages and tools, including Python and data visualization libraries such as Matplotlib.

Acknowledgement

The authors would like to express sincere thanks for the encouragement and constant support provided by Dr SN Rajan Sir (Professor), Department of Information Technology, IMS Engineering College.