

**COMPARATIVE ANALYSIS OF STATISTICAL AND MACHINE
LEARNING MODELS IN IDENTIFYING DRIVERS OF URBANIZATION
IN LAGOS, NIGERIA.**

BY

JINADU, ABDULRASAQ ENIOLA

MATRIC NUMBER: 210806517

**A PROJECT SUBMITTED TO THE DEPARTMENT OF STATISTICS,
FACULTY OF PHYSICAL AND EARTH SCIENCES, UNIVERSITY OF
LAGOS IN PARTIAL FULFILLMENT TO THE REQUIREMENTS FOR
THE AWARD OF THE BACHELOR OF SCIENCE (B.Sc.) DEGREE IN
STATISTICS.**

SEPTEMBER, 2025

CERTIFICATION

This is to certify that this project work “**COMPARATIVE ANALYSIS OF STATISTICAL AND MACHINE LEARNING MODELS IN IDENTIFYING DRIVERS OF URBANIZATION IN LAGOS, NIGERIA.**” was carried out by **JINADU, ABDULRASAQ ENIOLA** with matric number **210806517** under the supervision of **DR. J.A. AKINYEMI** in partial fulfilment of Bachelors of Science in Statistics, Department of Statistics, Faculty of Physical and Earth Sciences, University of Lagos, Akoka-Yaba, Lagos.

Supervisor

DR. J.A. AKINYEMI

Date

Project Supervisor

DR. R.K. OGUNDEJI

Date

Head of Department

DR. E. B. NKEMOLE

Date

External Examiner

Date

DEDICATION

This research is dedicated to myself and everyone who is a fan of people.

ACKNOWLEDGEMENT

I would like to thank Almighty God for His Infinite Mercies and Love and for the success in the completion of this project.

I thank my project supervisor and lecturer, Dr. J.A. Akinyemi, for his valuable contributions and unwavering support throughout the course of my final year project. His guidance, expertise, and constructive feedback have been crucial in helping me to successfully complete this project.

I also acknowledge my lecturers, Dr. J.N. Onyeka-Ubaka, Prof. M.O. Adamu, Prof. G.C. Udomboso, Dr. N. I. Badmus, Dr. T. O Arowolo, Dr. R.K. Ogundeji, Dr. E.E. Akarawak, Dr. Gbolahan Idowu and Dr. E. B. Nkemnole for their dedication to my academic and professional growth.

Special thanks and appreciation go to my mother, Mrs. Fatima Jinadu, my father, Mr. Kazeem Jinadu, my sister Miss Fathia Jinadu, Dr. and Mrs. Adedokun, Dr. (Mr. & Mrs.) Otuyalo and my mentor, Dr. R.A. Abdulganiy for their unwavering support in course of my studies. May God continue to bless you.

More so, a very big thank you to my colleagues, classmates and friends, thank you all and God bless you.

ABSTRACT

Lagos, Nigeria's most populous city, has undergone rapid urbanization, driven extensive land use changes and challenging urban mobility. This study examines the patterns and impacts of urban growth, land cover transformation, and transport dynamics in Lagos using remote sensing, geographic information systems (GIS), and statistical analysis. Spatial and satellite data were analysed to quantify built-up expansion, while transport and mobility datasets assessed commuting patterns, modal choices, and infrastructure efficiency. Results indicate significant urban sprawl, the proliferation of informal settlements, and persistent transport challenges, including congestion and inadequate infrastructure. The study underscores the interconnection between urban growth and mobility, highlighting the necessity for integrated urban planning strategies that combine GIS-based monitoring with sustainable and inclusive transport policies. The findings provide evidence-based insights to support policymakers and planners in managing Lagos's urban expansion, improving transport systems, and promoting sustainable urban development that enhances the quality of life for residents.

Keywords: Urbanization; Land use change; Urban mobility; Geographic Information Systems (GIS); Remote sensing; Informal settlements; Sustainable urban development.

TABLE OF CONTENTS

Title Page.....	
Certification.....	ii
Dedication.....	iii
Acknowledgement.....	iv
Abstract.....	v

CHAPTER 1: INTRODUCTION

1.1 Background to the Study.....	1
1.2 Statement of the Problem.....	2
1.3 Research Objectives.....	4
1.4 Research Questions.....	4
1.5 Scope of the Study.....	4
1.6 Data Sources.....	5
1.7 Significance of the Study.....	6
1.8 Limitations of the Study.....	6
1.9 Definition of Terms.....	7

CHAPTER 2: LITERATURE REVIEW

2.1 Introduction.....	10
2.2 Informal Settlements and the Strain on Urban Systems in Lagos.....	11
2.3 Factors that Influence Urbanization of Informal Settlements.....	12
2.4 Fertility/Mortality Trends and Urban Population Growth.....	14
2.5 Availability of Social Amenities and Urbanization of Informal Settlements in Lagos.....	15
2.6 Cheaper Transport Systems and Urbanization of Informal Settlements in Lagos.....	17
2.7 Migration and Urban Sprawl.....	19
2.8 Land Use Patterns and Transformation Over Time.....	20
2.9 Commuting and Daily Transport Behaviour.....	22
2.10 Accessibility of Groceries, Markets, and Food Systems in Lagos.....	25
2.11 Access to Entertainment and Recreational Facilities in Lagos.....	27
2.12 New Technology and Urbanization of Informal Settlement in Lagos.....	29
2.13 Linking Literature Review Themes to the Study Framework.....	31
2.14 Theoretical Framework.....	32
2.15 Conceptual Framework.....	34
2.16 Summary of Literature Review and Conceptual Framework.....	36

CHAPTER 3: METHODOLOGY

3.1 Introduction.....	37
3.2 Research Design.....	37
3.3 Data Collection.....	38
3.3.1 Data Sources.....	38
3.3.2 Variables Collected.....	41
3.3.3 Period of Study.....	41

3.3.4 Justification for Multi-Source Data.....	42
3.4 Software and Analytical Tools.....	42
3.4.1 Python Programming Environment.....	42
3.4.2 Geographic Information Systems (GIS)	43
3.4.3 Microsoft Excel and Power BI.....	43
3.4.4 Remote Sensing and Earth Engine Tools.....	43
3.5 Machine Learning and Statistical Models.....	44
3.5.1 Principal Component Analysis (PCA)	44
3.5.2 Clustering Algorithms (K-Means and Hierarchical Clustering)	45
3.5.3 Multiple Linear Regression.....	45
3.5.4 Logistic Regression.....	46
3.5.5 Random Forest.....	46
3.5.6 Gradient Boosting Models (XGBoost, LightGBM)	47
3.5.7 Naïve Bayes Classifier.....	47
3.5.8 Evaluation and Interpretability Tools.....	48
3.5.9 Integration and Policy Relevance.....	49
3.6 Data Preprocessing.....	50
3.6.1 Data Cleaning.....	50
3.6.2 Standardization.....	50
3.6.3 Data Transformation.....	51
3.6.4 Dataset Splitting.....	51
3.6.4.1 Training Data.....	52
3.6.4.2 Test Data.....	52
3.6.5 Feature Selection.....	52
3.7 Model Evaluation.....	54
3.7.1 Regression Metrics.....	55
3.7.2 Classification Metrics.....	56
3.7.3 Confusion Matrix.....	57
3.7.3.1 True Positive (TP)	57
3.7.3.2 True Negative (TN)	57
3.7.3.3 False Positive (FP)	57
3.7.3.4 False Negative (FN)	57
3.7.4 Area Under the Curve - Receiver Operating Curve (ROC Curve and AUC Score)...	58
3.8 Model Interpretability.....	59

CHAPTER 4: DATA PRESENTATION, ANALYSIS AND RESULTS

4.1 Data Description.....	61
4.2 Results of Data Analysis.....	61
4.2.1 Land Use and Built-up Expansion.....	63
4.2.2 Population and Demographic Pressures.....	65
4.2.3 Transportation Dynamics.....	67
4.2.4 Infrastructure Distribution.....	68
4.3 Model Training and Evaluation.....	69
4.3.1 Random Forest.....	70
4.3.2 XGBoost.....	70

4.3.3 Linear Regression.....	71
4.3.4 Lasso and Ridge Regression.....	71
4.3.5 Multi-Layer Perceptron (MLP)	72
4.4 Feature Importance and Drivers of Urbanization.....	72
4.5 Spatial Analysis.....	73
4.6 Scenario Testing.....	75
 CHAPTER 5: SUMMARY AND CONCLUSION	
5.1 Summary.....	77
5.2 Land Use Transformation and Built-up Expansion.....	77
5.3 Population Growth and Demographic Pressures.....	78
5.4 Socio-Economic and Infrastructural Drivers.....	79
5.5 Integrating Environmental Sustainability.....	79
5.6 Synthesis of Findings.....	80
5.7 Implications for Policy and Planning.....	80
 REFERENCES.....	82
APPENDIX A.....	86
APPENDIX B.....	87

CHAPTER 1

INTRODUCTION

1.1 Background to the study

Urbanisation is a global phenomenon characterized by the increasing concentration of in urban areas. The process of urbanisation is one of the most important indicators of human populations development. This process is characterized by the growth of urban populations and the widespread adoption of urban lifestyles. [*United Nations, Department of Economic and Social Affairs, 2019*].

The population of most African cities have been growing since the 1960s at some of the fastest rates in the world, and by 2050, about 55% of Africans will be living in urban areas (up from 38% in 2000). Over 90% of new urban development in Africa is informal. While the locations, construction standards, population densities and other aspects of informal settlements vary tremendously, many informal settlements are characterized by severe environmental problems of one kind or another. In many cases, a poor sanitary environment, hazardous location, and lack of basic services means that environmental health issues affect quality of life and life expectancies of the inhabitants (Brakarz et. al. 2002).

Lagos has always been the biggest state in Nigeria since being the capital of the Southern Protectorate in 1906 until date, raking in about 1.3 million visitors every year. Since her official creation in 1967, the former capital of Nigeria has been one of the forerunners of the most bustling cities in Africa as it was ranked the 19th Best City to visit in the World in 2024 [*Time Out Magazine*] and the 4th highest city in GDP in Africa. She's classified as a Medium-Port Megacity [*Wikipedia*].

With an average population growth of 4.0% year-on-year, and the land area being the smallest in the country, most people are stuck with moving to newly underdeveloped, or undeveloped, areas of the state, which in turn, restricts them access to basic amenities for the community.

Over the years, Lagos has been no stranger to the rapid and forward development of her state with roads, public buildings, essential infrastructure and technological advancements making it the very state we all know it as. In this research, we will be focusing on how various factors of urbanisation affected the rate of urbanisation and how much it will affect the state over the coming years.

The rapid expansion of cities throughout the world has been accompanied by equally rapid growth of informal settlements, known as slums. They develop as the formal housing market is unable to cater for the number of migrants, many of whom are extremely poor. It is estimated that by 2030, nearly five billion people will be living in urban areas, which compares to 3.2 billion in 2007. Slum area conditions are defined by the United Nations as “lacking at least one of the basic conditions of decent housing: adequate sanitation, improved water supply, durable housing or adequate living space”. Although the proportion of urban dwellers living in slums appears to be falling, the absolute number is rising rapidly. This expansion is occurring quickest in the world’s poorest regions such as Southern Asia and sub-Saharan Africa (UN-HABITAT, 2009).

Understanding the factors driving urbanisation is critical for policymakers and urban planners to address these challenges effectively. Traditional methods of studying urbanisation rely on descriptive statistics and limited case studies, which may not fully capture the complexity and interdependencies of the factors involved. Advances in machine learning (ML) provide an opportunity to overcome these limitations by analysing large datasets and uncovering hidden patterns in urbanisation dynamics.

1.2 Statement of the Problem

Urbanization of informal settlements has been associated with the mushrooming of the slums and is one of the root causes for the spread of problems associated to informal settlements in Lagos State. Lagos is experiencing rapid urbanisation due to rural-urban migration and international migration. Many people moving to urban areas are poor, and when they get to town, they live in areas with low rentals resulting in families living in small rooms, shacks or even a few families sharing a room resulting in overcrowding. This creates huge pressure on basic services and facilities e.g. housing, schools, hospitals. The growth of these rural areas has resulted in many, and complex socio-economic and environmental consequences. These problems include increased crimes rates within the slums and in town, environmental pollution, deforestation, flooding, waste of agricultural lands among others. This has made it impossible for the local government to embark upon any meaningful developmental project that will enhance economic growth and development of the study area. The main causes of growth of informal settlements in Lagos State include; industrialization, improved transport systems, technology advancement and increase in social amenities all these being affiliated with urbanization.

Lagos, one of the biggest economic states in Nigeria, has undergone one of the most rapid rates of urbanisation. With big positive strides in infrastructure and economic developments. Although, while it's been documented, there is limited research in leveraging machine learning approaches to analyse and predict the key factors of urbanisation. Existing studies on urbanization in Nigeria have primarily relied on traditional statistical methods, which may not fully capture the complex interactions among urbanization factors. Moreover, available urbanization data is often fragmented, outdated, or incomplete, making it challenging to derive actionable insights.

This study seeks to address these gaps by applying machine learning models to analyse urbanization trends in Lagos, identify the most critical factors, and provide predictive insights. The findings will offer data-driven recommendations to support urban planning and policy formulation.

1.3 Research Objectives

This research project's objective is to:

1. Identify the key factors influencing urbanisation in Lagos, Nigeria using machine learning (ML) techniques.
2. Analyse the relationships between these factors and urbanisation patterns.
3. Analyse how much of an effect these factors have influenced urbanisation in Lagos, Nigeria over the last 20 years.
4. Develop machine learning models to predict urban growth in Lagos based on historical data
5. Assess the policy implications of the findings for urban planning and sustainable development.

1.4 Research Questions

This study aims to answer the following questions:

1. What are the most significant factors driving urbanisation in Lagos?
2. How do demographic, economic, infrastructural and environmental variables influence urbanisation patterns?
3. How available are basic amenities in Lagos, especially in the rural and the slum areas of Lagos?
4. Can machine learning models effectively predict urbanisation trends in Lagos?
5. How much of an effect have these urbanisation factors affected this said urbanisation over the last 20 years in Lagos, Nigeria?
6. What policy recommendations can be drawn from this analysis?

1.5 Scope of the Study

This research focuses on Lagos State, Nigeria, as a case study due to its rapid and continuous urbanisation and economic significance. The study will analyse demographic, technological, infrastructural, environmental, economic and policy-related factors influencing urbanisation.

This research will apply machine learning algorithms (such as regression models, clustering techniques, and feature selection methods etc.) to analyse urbanisation patterns. The findings will be relevant for urban planners, government policymakers, fellow researchers and agencies that will have anything pertaining to urbanisation.

1.6 Data Sources

The data to be used for this research will be secondary data, such as data retrieved from archives and publications like journals and websites. The datasets will include demographic, economic, infrastructural, environmental and policy-related variables, which will be used for machine learning analysis.

The key sources include:

- i. Nigerian Bureau of Statistics (NBS)
- ii. Nigerian Population Commission (NPC)
- iii. Lagos State Ministries' Reports
- iv. World Bank Open Data
- v. UN-Habitat and Global Urbanisation Datasets
- vi. UN Statistics Department Data
- vii. Telecommunications Open Data and Reports
- viii. Satellite and Geospatial Data
- ix. Other Related and Verified Sources

The selected data sources ensure that the study covers diverse factors influencing urbanisation, utilises credible and authoritative sources and integrates real-world geospatial data to enhance machine learning models. The combination of qualitative, quantitative and spatial data will provide a wide and grounded foundation for identifying urbanization drivers in Lagos using machine learning.

1.7 Significance of the Study

Understanding the factors the drive behind Lagos' urbanisation and how much it has affected the said urbanisation has several practical applications.

- i. Urban Planning: For better understanding for better policies for housing, transportation and infrastructure development
- ii. Economic Development: Identifies key economic drivers attracting urban migration.
- iii. Sustainability: Provides insights into managing urban growth while minimizing negative environmental impacts.
- iv. Predictive Insights: Supports the development of data-driven solutions for managing Lagos' future urban expansion.
- v. Policy Reviewing and Implementation: Helps the government interpret and discuss future policies that will be shaped around the research.

By leveraging machine learning, this study will bridge the gap between major industries and sectors like urban planning and data science, providing evidence-based, credible, and reliable information for sustainable urban development in Lagos.

1.8 Limitations of the Study

The limitations of this study are, but not limited to, the following:

- i. Data Inaccuracy and Incompleteness: Some data may contain errors, missing values or incomplete and outdated information, which could affect model accuracy.
- ii. Inconsistency across Sources: Some datasets may have errors, missing values, or outdated information, which could affect model accuracy.
- iii. Limited Access to Government Reports: Some relevant Lagos urbanisation reports may not be publicly available, limiting access to critical insights.
- iv. Underreporting in Informal Settlements: A significant portion of Lagos' urban population resides in slums and informal settlements, where data collection is often inconsistent or absent
- v. Feature selection limitations: Some factors influencing urbanisation (e.g. political decisions, governance quality) are difficult to quantify in datasets, making them challenging to include in models.

1.9 Definition of Terms

Urbanisation: Urbanization refers to the increase in the proportion of a population living in urban areas due to rural-to-urban migration, natural population growth, and economic expansion. It involves the transformation of rural areas into urban centres and is often associated with infrastructure development, industrialization, and increased economic activities.

Settlement: A settlement is a community of people living in a particular place. The complexity of a settlement ranges from a minuscule number of dwellings grouped together to the largest of cities with surrounding urbanized areas. Settlements include homesteads, hamlets, villages, towns and cities. A settlement may have known historical properties such as the date or era in which it was first settled or first settled by particular people. [*Wikipedia*]

Rural Area: A rural area is generally defined as a geographic region or community located outside urban centres, characterized by low population density, limited infrastructure, and a predominantly agricultural or natural resource-based economy. It is typically associated with a slower pace of life and a closer connection to nature, often having fewer people and more open spaces than urban areas. [*Wikipedia*]

Urban Area: An urban area is a geographically defined area with a high population density, often characterized by cities and towns, and a predominance of human-built structures. It's essentially a densely populated place where people live and work, often in and around cities. [*Wikipedia*]

Slum: A slum is as an urban area with a lack of basic services (sanitation, potable water, electricity), substandard housing, overcrowding, unhealthy and hazardous locations, insecure tenure and social exclusion. [*UN-HABITAT*]

Machine Learning: Machine Learning is a branch of artificial intelligence (AI) that enables computers to analyse data, identify patterns, and make predictions without being explicitly programmed. It involves algorithms that improve their performance over time as they process more data.

Urban Sprawl: Urban sprawl, a trend of abnormal and unrestricted urbanization development, has been a controversial topic, and it continues to present a challenge to countries across the world. Urban sprawl refers to unsustainable spatial expansion in a city in the process of development: it tends to be random and unplanned, scattered and discontinuous, associated with strong dependence on transportation for travel, and characterized by a single land-use type and severe land-use conflicts. (Zhang & Pan, 2021)

Feature Selection in Machine Learning: Feature selection is the process of choosing the most relevant variables in a dataset to improve the performance of a machine learning model. In

the context of this study, it helps in identifying the most significant factors influencing urbanization in Lagos.

Predictive Modelling: Predictive modelling involves using historical data and machine learning algorithms to forecast future trends. In this study, it will be used to predict how urbanization in Lagos will evolve based on past and present data.

Big Data in Urban Planning: Big data refers to large, complex datasets collected from various sources such as government reports, satellite imagery, and social media. In urban planning, it is used to analyse patterns of migration, infrastructure development, and land use changes.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

Lagos has evolved from a colonial port and former federal capital into Nigeria's largest city and one of Africa's fastest-growing megacities. Its metropolitan area now exceeds 21 million people. Lagos State (formed in 1967) encompasses this urban region. The city's annual growth rate (~3.2% per year) far outpaces the national average. Most of this growth comes from migration: one analysis using social-media data found Lagos's population rose 18.6% between 2000 and 2012, and about 96% of new arrivals during that time were from elsewhere in Nigeria.

This demographic boom has dramatically transformed land use. Studies show Lagos's built-up area has more than doubled (from ~492 km² in 2000 to ~1,272 km² by 2020), while green and aquatic spaces have shrunk. For example, vegetated cover fell from 63.7% of the city in 2000 to 46.4% in 2020, and open water (lakes, creeks, etc.) declined from about 879 km² (23.3% of the area) to ~754 km² (19.9%) over the same period. Land-use models project that this trend will continue – built-up land is expected to expand further over coming decades at the expense of wetlands, forests, and bare soil.

These changes underline serious urban planning challenges. Lagos now has an acute housing shortage – roughly a 3.3-million-unit deficit – and an estimated 50–75% of residents live in informal settlements. Indeed, more than 140 slum neighbourhoods have been documented (from the stilt-house community of Makoko on the lagoon to dense squatter areas inland). Infrastructure and services have struggled to keep pace: crowding and sanitation problems are endemic, traffic congestion is severe, and much new development occurs on flood-prone land (about 18% of Lagos State lies on low-lying coastal plain). In short, Lagos's unparalleled growth – while fuelling its

economic dynamism that creates urgent planning and sustainability programs, and resilient infrastructure in this rapidly expanding megacity

2.2 Informal Settlements and the Strain on Urban Systems in Lagos

Informal settlements present a significant challenge to urban development, undermining environmental quality, public health, and educational services. Characterized by high population density, inadequate infrastructure, and severe pollution, these areas often feature substandard housing and limited access to essential amenities.

In 2001, sub-Saharan Africa had the highest proportion of its urban residents living in slums (71.9%), while Oceania had the lowest (24.1%). Regions between these extremes included South-Central Asia (58.0%), East Asia (36.4%), West Asia (33.1%), Latin America and the Caribbean (31.9%), North Africa (28.2%), and Southeast Asia (28.0%) (UN-Habitat, 2013). In absolute terms, Asia's combined subregions accounted for approximately 554 million slum dwellers, about 60% of the global total. Africa followed with 187 million (20%), Latin America and the Caribbean with 128 million (14%), and Europe and other developed regions with 54 million (6%) (UN-Habitat, 2013).

Slums and informal settlements represent one of the most critical challenges to sustainable urban development, particularly in rapidly growing cities like Lagos. The unchecked expansion of urban areas often leads to the proliferation of informal housing, which exerts immense pressure on health and education systems that are already under strain. These communities frequently lack basic infrastructure such as clean water, sanitation, and electricity, thereby exacerbating public health risks and environmental degradation. Studies have shown that slums are also hotspots for pollution, with poor waste management and overcrowding contributing to unsanitary conditions and increased disease transmission (UN-Habitat, 2022).

Beyond environmental and health implications, informal settlements raise serious concerns regarding safety, security, and social cohesion. High population densities, inadequate policing, and lack of formal property rights often lead to social tensions and insecurity. In its seminal *Global Report on Human Settlements* (2003), UN-Habitat emphasized the rising prevalence of slums worldwide and their implications for urban governance. Nearly two decades later, these concerns remain especially relevant for cities like Lagos, where an estimated 50–75% of the population resides in informal settlements (World Bank, 2020; UN-Habitat, 2022).

2.3 Factors that influence urbanization of informal settlements

The growth of informal settlements in rapidly urbanizing cities like Lagos is driven by a complex interplay of socio-economic, spatial, and policy-related factors. As urban centres expand, these unregulated communities often emerge at the margins of formal governance, reflecting both the intensity and structural challenges of urbanization.

A primary driver of informal settlement expansion is rural-to-urban migration, largely motivated by economic opportunities concentrated in cities. In Lagos, internal migration accounted for approximately 96% of population growth between 2000 and 2012 (Blumenstock, Gillick, & Eagle, 2019). However, this population surge has far exceeded the formal housing supply. Lagos currently faces a housing deficit of about 3.3 million units, pushing low-income groups to settle in informal and environmentally vulnerable areas (World Bank, 2020).

Land use dynamics have also significantly contributed to this trend. Between 2002 and 2022, satellite data revealed that Lagos's built-up area expanded by 26.6 km², while natural land cover such as wetlands, forest areas, and water bodies experienced significant reduction (Adepoju,

Arowolo, & Akinluyi, 2023). Much of this expansion is unregulated and occurs in flood-prone or marginal lands, often due to poor land governance and weak enforcement of zoning regulations. Further complicating the issue are policy and governance limitations. Institutional capacity to manage urban expansion remains low. Weak land tenure systems, limited public housing programs, and fragmented planning efforts contribute to the proliferation of informal housing (UN-Habitat, 2003; 2022). These conditions allow slums to grow unchecked, especially in areas like Makoko, where residents build housing on stilts in lagoon areas, adapting to both land scarcity and environmental risk.

Lagos has experienced profound shifts in land use over the past two decades, reflecting both rapid demographic growth and weak spatial planning controls. Remote sensing and geospatial analyses consistently show that the city's built-up area has expanded dramatically, while vegetation, wetlands, and agricultural land have declined. For example, between 2000 and 2020, built-up land in Lagos increased by approximately 33.6%, whereas vegetative cover fell by about 21.8%. Wetlands and mangroves, critical for flood regulation, have also diminished considerably. (Ogunjobi & Akinbobola, 2021).

Projections suggest that this transformation will continue. Lagos's urban land area is expected to expand by an additional 42% by 2030, consuming farmlands and ecologically sensitive areas (Adelekan et al., 2020; Lagos State Ministry of Physical Planning and Urban Development [LASPPUD], 2019). Much of this growth occurs on the urban buzz, particularly northward toward Ogun State and eastward toward Lekki and Epe where agricultural land is being converted into residential and commercial uses (LASPPUD, 2019). Importantly, Lagos's expansion is not uniform but marked by fragmented, patchy sprawl.

Studies note that informal settlements and speculative developments often arise without coordination, producing a mosaic of built-up areas interspersed with bare land and remnant green spaces (Olajide et al., 2018; Ogunjobi & Akinbobola, 2021). This unplanned growth increases the city's exposure to flooding, worsens ecosystem loss, and reduces access to arable land for food production (Adelekan et al., 2020; Lagos Bureau of Statistics [LBS], 2022).

Government reports echo these concerns. The Lagos State Development Plan 2012–2025 highlights the strain that uncontrolled land conversion places on infrastructure, transportation, and housing (LASPPUD, 2012). Likewise, the Lagos Bureau of Statistics (2022) emphasizes that population-driven land use change remains one of the biggest challenges for sustainable development in the metropolis. While academic and official sources document recent patterns, gaps remain in tracking fine-grained transformations, particularly in informal settlements. More robust monitoring and enforcement of planning policies are essential if Lagos is to balance growth with sustainability.

Finally, socioeconomic exclusion and inequality are central to the informal urbanization process. Without access to formal credit, legal land ownership, or stable income, many urban poor are left with no alternative but to create shelter informally. This exclusion deepens the divide between formal and informal urban development and reinforces cycles of poverty and vulnerability.

2.4 Fertility/Mortality Trends and Urban Population Growth

Lagos's population is expanding swiftly, driven by both natural increase and migration. As of 2022, estimates suggest the greater Lagos metropolitan area houses approximately 28 million residents, up from 16 million just a few years prior. This translates to an annual gain of around 3,000 people per day, or roughly 1.1 million annually (Wikipedia, 2025)

Most of this growth stems from natural increase, a combination of high fertility rates and declining mortality. Recently, Lagos recorded a neonatal mortality rate of 11 per 1,000 live births, the lowest in southwest Nigeria and significantly better than many other regions. Despite this improvement, maternal mortality remains a pressing concern. A 2022 population-based study estimated Lagos's maternal mortality ratio (MMR) at approximately 430 per 100,000 live births, with a total fertility rate (TFR) of 3.8 children per woman. These figures indicate considerable strides in reducing child mortality, even as maternal health continues to pose a challenge.

Putting these trends into perspective: Lagos's natural increase, amplified by migration, substantially shapes its demographic trajectory. By 2050, projections indicate a metropolitan population surpassing 32.6 million, possibly reaching 40 million if migration remains high and health outcomes continue to improve. These estimates underscore how sustained high fertility and lowering neonatal mortality reinforce population pressure on the city's infrastructure and services.

2.5 Availability of Social Amenities and Urbanization of Informal Settlements in Lagos

Lagos's rapid population growth has generated a vast sector of informal settlements. Recent analyses indicate that roughly half to three-quarters of Lagos's residents live in slum or informal housing clusters. For example, a World Bank review notes that "50–75% of the population live in informal housing" across more than 140 slum communities in Lagos. (Another source similarly observes that "more than 60 percent" of Lagos's population inhabits informal settlements.) In absolute terms this means tens of millions of people live in unplanned, high-density neighbourhoods with insecure tenure.

In these settlements, basic social services are severely lacking. According to a recent World Bank report, only about 14% of Lagos households receive reliable electricity (more than eight hours per

day). Public water and sanitation coverage are even more limited: roughly one-third of the population has access to piped water and only about 5% are connected to any sewerage network. Waste collection in slum areas is reported at only 20–30%, meaning most trash remains uncollected. These deficits force many residents to rely on informal water vendors or generators, and to live amid open drains and refuse. The lack of affordable housing and infrastructure drives many poor migrants into slums, where crowding and makeshift construction prevail.

One vivid example is Makoko, a large waterfront slum. Makoko is estimated to house on the order of 200,000–300,000 people in simple wooden shacks, many built on stilts above a lagoon. Despite its proximity to Lagos’s central business district, Makoko has virtually no formal sanitation or public utilities. Canoes are the main mode of transport; sanitation is limited to pit latrines or open-water disposal. More generally, studies find that Lagos’s informal settlements are often located in hazardous areas (flood plains or swampy land) and lack even basic schools or clinics. For example, one field survey notes that “communities often lack access to basic services such as healthcare, education and sanitation facilities”. High population density exacerbates these conditions, contributing to poor waste disposal, disease risk and environmental degradation.

The Lagos State government has officially acknowledged the slum problem, albeit with mixed actions. In planning documents Lagos set formal targets for slum reduction. The State’s development plan (2012–2025) directed the Lagos State Urban Renewal Agency (LASURA) to eliminate 5% of slum areas each year globalfuturecities.org. A 2013 LASURA survey identified over 100 slum communities (up from 42 in 2002), and new urban renewal guidelines have been drafted to guide upgrading projects. Some regeneration initiatives are underway: for instance, officials announced plans in 2021–2022 to redevelop the Otumara slum in the Ebute-Meta area and other low-income districts as part of an “inclusive urban renewal” drive (LASURA, 2021). In

partnership with international programs, community-based mapping and infrastructure-improvement efforts have been piloted to integrate slum areas into city services.

However, many interventions have been one-directional evictions rather than supportive upgrades. Forced demolitions of slum homes continue to occur, often with scant notice or compensation. In August 2023, for example, the Lagos State Task Force razed hundreds of shacks in the Oworonshoki informal neighbourhood, rendering many women and children homeless. Public outcry prompted the State Assembly to intervene, but such events reflect a persistent “culture of demolition” in place of inclusive upgrading. Human-rights observers note that these evictions – frequently justified in the name of “slum clearance” or redevelopment – displace vulnerable communities and undermine their right to housing.

In sum, Lagos’s slums are marked by extreme deficits in social amenities and infrastructure. Water, sanitation, power, schools and clinics are scarce in these areas, and many communities face environmental hazards and insecure tenure. While recent plans formally recognize the need for slum upgrading, in practice residents often experience neglect or forced displacement. The literature on Lagos thus paints a picture of rapid urbanization in which informal settlements proliferate without commensurate provision of basic services. Current policy emphasizes “transforming slums into neighbourhoods,” but observers warn that without inclusive planning and investment, the cycle of poor living conditions and evictions is likely to continue.

2.6 Cheaper Transport Systems and Urbanization of Informal Settlements in Lagos

A critical factor influencing the growth and sustenance of these settlements is the availability and affordability of transport systems. For many residents of informal settlements, access to affordable

transportation is essential for daily commuting to work, schools, and markets, thereby influencing their choice of residence and contributing to the expansion of these areas.

Informal transport modes, such as minibuses (danfos), motorcycle taxis (okadas), and tricycles (kekes), dominate the transportation landscape in Lagos. These modes are preferred by many due to their affordability and ability to navigate congested and poorly maintained roads, especially in slum areas where formal transport services are limited or non-existent (Asenime, 2021). However, the reliance on these informal systems poses challenges, including safety concerns, lack of regulation, and environmental pollution.

Recognizing these issues, the Lagos State Government has made efforts to introduce formalized alternatives. The Bus Rapid Transit (BRT) system, launched in 2008 and expanded in subsequent years, offers an organized and relatively reliable form of transport that can serve as a safer and more environmentally sustainable option (LAMATA, 2023). Similarly, the Lagos Rail Mass Transit (LRMT), under phased development, is envisioned to reduce travel times and provide a cost-effective alternative to informal modes (Oduwaye & Iweka, 2020). The BRT's affordability, fares significantly cheaper than ride-hailing services, makes it particularly attractive to low- and middle-income groups, and early studies show it has improved commuting efficiency along major corridors such as Mile 12–CMS (INTALInC, 2021).

Despite these efforts, the reach of formal transport systems remains limited, particularly in informal settlements. The high cost of expanding infrastructure and the complex terrain of slum areas hinder the penetration of formal transport services. Consequently, residents continue to rely on informal transport modes, which, while affordable, do not offer the safety and reliability of formal systems (INTALInC, 2021).

The relationship between transport and the expansion of informal settlements thus reflects a broader urban governance challenge. On one hand, cheap and flexible informal transport sustains livelihoods by keeping the city accessible to poorer residents; on the other, the lack of formal integration deepens inequalities and environmental burdens. Scholars argue for a more inclusive and hybrid transport planning model, where informal operators are not displaced but regulated and gradually integrated into the wider system through licensing, safety training, and infrastructural support (Olawole & Olapade, 2019; INTALInC, 2021). Expanding the reach of formal services to underserved neighbourhoods could also reduce pressures on informal settlements by making alternative residential zones more viable, thereby dispersing population density and mitigating unplanned sprawl.

2.7 Migration and Urban Sprawl

Internal rural–urban migration has been the dominant driver of Lagos’s growth as studies consistently report that the vast majority of new arrivals come from other parts of Nigeria: for example, a 2017 study made by Facebook in an analysis of social-media “migration” data found over 90% of Lagos’s migrants are Nigerian (mainly from rural states) increasing the population to 18.6% from 2000 to 2012.

International migration plays a much smaller role. Lagos’s pull factors (jobs, services) and Nigeria’s uneven development keep internal flows strong. This intense in-migration fuels spatial expansion beyond the old city core. Recent research using satellite imagery and field surveys shows sprawling growth on Lagos’s periphery: built-up area in outlying municipalities (e.g. Ikorodu) roughly doubled (127% increase) over 32 years.

This peri-urban sprawl is largely unchecked by planning, fragmenting former farmland and increasing pressure on rural landscapes. In fact, Lagos (and Nigeria more broadly) is now

identified as a global hotspot of cropland loss due to urban expansion. Migration-induced sprawl thus transforms outer districts into low-density urban and informal settlements. Lagos's climate adaptation plan notes that in-migration from other Nigerian states and even neighbouring countries is expected to push the population past 32 million by 2050. Despite this recognition, detailed data gaps remain on the origins of migrants and the precise contribution of international flows, and on how migration interacts with land markets and infrastructure, indicating areas needing further study.

2.8 Land Use Patterns and Transformation Over Time

Lagos has experienced dramatic land-use transformations over the last five decades, reflecting the pressures of rapid urbanization, population growth, and economic restructuring. Historically, the city's landscape comprised extensive wetlands, coastal mangroves, inland water bodies, and fertile agricultural land. However, since the 1970s, the spatial fabric of Lagos has shifted toward a built-up dominated environment, with significant consequences for ecological balance and urban sustainability (Abiodun et al., 2021).

Remote sensing and GIS-based studies provide quantitative evidence of these changes. For instance, a land-cover analysis between 2000 and 2020 revealed that Lagos's built-up area expanded by approximately 33.6%, while vegetated cover declined by nearly 21.8%, and wetlands suffered substantial losses (Ogunyemi et al., 2020). Projections further suggest that if current urbanization trends persist, built-up areas will increase by an additional 42% by 2030 (Adelekan, 2022). These trends underscore the dominant role of demographic and economic pressures as key drivers of land-use change.

The spatial dynamics of expansion reveal a clear outward push from the city's historic core on Lagos Island towards peri-urban and rural areas, particularly in the northern and eastern corridors (Ajala & Agbola, 2021).

Farmlands in Ikorodu, wetlands around Lekki, and peri-urban settlements along the Lagos-Ibadan and Lagos-Abeokuta expressways are increasingly being consumed by residential and commercial sprawl. Informal housing settlements, often emerging in unplanned clusters, create a fragmented land-use mosaic consisting of dense housing blocks, bare ground, and small patches of remnant vegetation (Olajide, 2018).

This uncoordinated growth has profound environmental and socio-economic implications. The loss of wetlands reduces natural flood buffering capacity, exacerbating recurrent flooding events in low-lying areas such as Ajegunle, Makoko, and Lekki Peninsula (Adelekan & Asiyanbi, 2016). Similarly, the conversion of agricultural land undermines local food security and increases dependence on imported food supplies. The transformation of land cover is also linked to ecosystem degradation, biodiversity loss, and urban heat island effects (Abiodun et al., 2021).

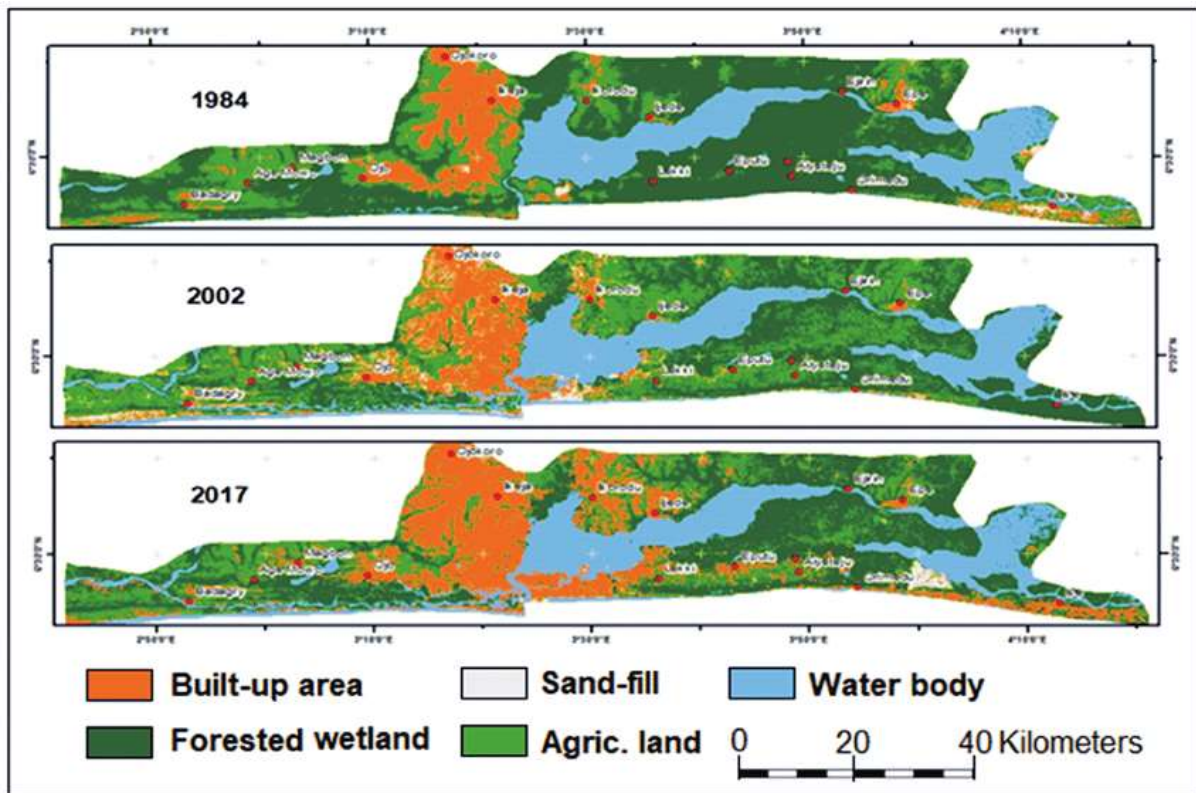


Fig. 1: Land-use changes in Lagos metropolitan area (1984–2017), showing expansion of built-up land and loss of vegetation and water bodies. (Aladejana et al., 2020)

Despite recent advances in land-use mapping, critical gaps remain in understanding Lagos’s transformation. Many studies provide only short-term temporal snapshots, while there is limited availability of long-term historical land-cover datasets that would allow researchers to trace land-use changes back to the colonial and early post-independence periods. Furthermore, finer-scale mapping of informal settlements is often absent, obscuring the full picture of how land-use change intersects with poverty, informality, and social vulnerability (Olajide, 2018). Bridging these knowledge gaps will require integrating satellite remote sensing with socio-spatial field data and participatory mapping approaches.

2.9 Commuting and Daily Transport Behaviour

Transportation in Lagos represents one of the most visible and pressing dimensions of urbanization. As Nigeria’s economic hub and Africa’s fastest-growing megacity, the city’s

transport system is overwhelmed by demand. Nearly all intra-city movement (95% or more) occurs on roads, since rail and water transit remain peripheral, serving only a tiny share of the city's residents (Adesanya, 2020). The road network, much of it inherited from colonial layouts, was never designed to accommodate the current population of over 20 million, leading to an urban mobility crisis (Adeniji, 2021).

A defining feature of Lagos transport is the dominance of the informal sector. Private minibuses (danfo vans) remain the backbone of daily commuting. Estimates suggest that these vehicles alone account for more than 60% of all passenger trips within the city (Barter & Obeng-Odoom, 2019). Their flexibility allows them to reach underserved neighbourhoods, but their proliferation has also been linked to traffic congestion, poor safety standards, and environmental externalities such as noise and air pollution (Akinmoladun, 2022).

Formalized options, such as the Bus Rapid Transit (BRT) system introduced in 2008, have improved mobility on specific corridors like Ikorodu Road and the Lagos Island axis, yet the BRT covers less than 10% of the commuting population (World Bank, 2020). The interplay between formal BRT services and the informal danfo operators is complex: new public routes often force danfo drivers to alter their operations, either by shifting routes or modifying prices, creating an adaptive but fragmented transport ecosystem (Foster & Anas, 2021).



Fig. 2: Maps of Bus Road Transport (BRT) routes in Lagos, Nigeria (LAMATA)

Daily commuting patterns in Lagos reveal the city's intensity of mobility stress. Surveys indicate that most residents commute at least five days per week, primarily for work, trade, and business purposes (Okunola et al., 2021). Commuting times are notoriously long: more than half of residents' report spending 30-60 minutes in traffic per trip, while nearly 40% endure travel times exceeding one hour. For residents in peri-urban areas such as Ikorodu, Badagry, or Epe, commutes of two to three hours are not uncommon (Think Global Health, 2022). In addition to time costs, the financial burden is significant historically, households have devoted as much as 40% of their income to transportation expenses (Foster & Anas, 2021). This cost burden is especially regressive, as low-income residents who live farthest from job centres spend the highest proportion of their earnings on mobility.

Transport behaviour in Lagos is further stratified by socio-economic status. Middle- and higher-income commuters, often living in gated estates or suburban enclaves, rely on private cars or app-based ride-hailing services such as Uber, Bolt, and InDriver (Olajide et al., 2022). By contrast, the majority of low- and middle-income earners remain dependent on minibuses, shared taxis, motorcycles (*okada*), or tricycles (*keke napep*). Notably, motorcycle taxis, once ubiquitous, have been restricted in central districts due to safety and security concerns, pushing more commuters back into congested danfo and BRT systems.

Adaptation strategies are common among commuters. Studies reveal that nearly half of Lagos residents alter their daily routes, either to avoid known congestion points or to take advantage of alternative modes such as ferries (Geopoll, 2020). Weekend travel tends to shift away from work-related mobility toward shopping, religious activities, and leisure, though congestion remains constant given the city's round-the-clock activity (Adesanya, 2020). Despite this, non-motorized transport such as walking and cycling remains marginal, undercut by inadequate sidewalks, lack of bike lanes, and the city's hostile traffic environment (Gbadamosi & Eniola, 2021).

The consequences of these patterns are profound. Long commuting times reduce productivity, increase stress levels, and worsen public health outcomes due to prolonged exposure to pollution and sedentary travel habits. The strain also reinforces urban inequality, as those who cannot afford faster alternatives bear the heaviest daily burdens. While ongoing infrastructure projects such as the Lagos Rail Mass Transit (LRMT) "Blue Line" completed in 2023 with a promise to reshape commuting patterns, their full impact remains to be studied systematically.

2.10 Accessibility of Groceries, Markets, and Food Systems in Lagos

Food access in Lagos's informal settlements is shaped by the dominance of traditional open-air markets, street vendors, and small-scale kiosks. Despite Lagos's growing network of supermarkets

and malls, studies consistently demonstrate that traditional food systems remain the backbone of urban food supply in Nigeria. Roughly two-thirds of Nigerians, including Lagosians, continue to source their daily food needs from such markets, particularly for staples such as rice, beans, vegetables, and cassava (Oluwande et al., 2019; Olajide et al., 2018).

For residents of informal settlements, the affordability, spatial proximity, and flexible purchasing options offered by these vendors are essential for daily subsistence.

The structure of Lagos's informal food systems reflects broader patterns of survivalist economies. In low-income areas, households rely on "micro-purchases", small units of food bought daily, because incomes are irregular and storage capacity is limited (Battersby & Watson, 2019). Informal food vendors thus provide both access and employment, particularly for women, who form the majority of petty traders in Lagos (Olajide, 2016). However, these benefits are accompanied by significant challenges. Vendors often operate without adequate storage, refrigeration, or sanitation facilities, heightening risks of contamination and foodborne illnesses. Poor infrastructure in slum communities, such as limited potable water, absence of drainage, and unreliable electricity, further compounds food-safety concerns (Akinmoladun & Adejumo, 2011).

The governance environment adds another layer of fragility. Lagos State enforces strict anti-street-trading laws under the Street Trading and Illegal Markets Prohibition Law of 2003, which prohibits unlicensed vending in public spaces. Enforcement campaigns, framed as "beautification" or modernization drives, have often targeted informal markets, displacing vendors and interrupting local food supplies (Moruf, 2012; Olajide, 2016). While these policies aim to decongest traffic and improve aesthetics, they inadvertently exacerbate food insecurity in marginalized areas where residents depend on affordable street markets. The tension between regulation and accessibility underscores a structural contradiction in Lagos's urban governance: while informal food systems sustain millions, they remain legally precarious.

Empirical research on food access in Lagos's slums remains limited, but broader Nigerian studies highlight widespread food insecurity among urban poor households. One study of urban food systems in Nigeria reported that low-income neighbourhoods face higher food prices relative to income and experience frequent supply disruptions due to flooding, poor transport connectivity, and market evictions (Battersby & Watson, 2019).

In Lagos, poor road networks in settlements such as Ajegunle, Makoko, and Mushin restrict wholesale distribution and inflate food prices (Olajide et al., 2018). Moreover, climate-related hazards such as coastal flooding directly threaten wet markets and perishable food vendors in low-lying settlements (Adelekan, 2016).

2.11 Access to Entertainment and Recreational Facilities in Lagos.

Recreational and entertainment facilities in Lagos remain severely constrained relative to the city's size, density, and population growth. While the Lagos State Government has articulated visions of becoming a "megacity" with enhanced liveability, access to public leisure amenities is highly uneven.

Studies reveal that most Lagosians, especially those in informal settlements, have limited or no access to well-maintained public spaces for relaxation, sports, or recreation (Olajide et al., 2018; Akinmoladun & Adejumo, 2011). A recent survey of four local government areas in Lagos found that a striking 78.4% of residents rarely or never visited public parks, and only about 31% believed such facilities improved their well-being (Afropolitan Journals, 2021).

Reasons cited included inaccessibility (parks are often located far from high-density neighbourhoods), inadequate maintenance, and perceived insecurity, particularly at night. Existing

green spaces such as Ndubuisi Kanu Park (Ikeja), Freedom Park (Lagos Island), and Johnson Jakande Tinubu Park (Alausa) primarily serve middle-class users living in nearby districts. Informal settlements such as Makoko, Ajegunle, and Mushin, in contrast, typically lack playgrounds, community centres, or safe outdoor spaces for children and youth.

Privately owned entertainment facilities, shopping malls, cinemas, restaurants, and nightclubs, are concentrated in wealthier districts such as Victoria Island, Lekki, and Ikeja, reflecting income-based spatial inequality (Oduwaye, 2009). This clustering limits the affordability and accessibility of recreation for the urban poor, who face both financial and mobility barriers to participation. Informal recreation within slums is often improvised: children play football in open sandy lots or streets, while cultural events and informal gatherings serve as substitutes for formal leisure spaces (Olajide, 2016). However, these improvised spaces are vulnerable to eviction, redevelopment, or environmental hazards such as flooding.

The lack of accessible recreational infrastructure has wider implications. Urban planning and public health research underscore that green and recreational spaces improve physical and mental health, foster social cohesion, and reduce crime rates in dense urban areas (Kabisch et al., 2015; WHO, 2017). Lagos's urban poor are thus disproportionately deprived of these benefits. For instance, scholars have noted that open space integration into slum upgrading programs could help mitigate stress, reduce youth idleness, and enhance community resilience (Adelekan, 2016).

Despite these findings, research on leisure in Lagos remains underdeveloped. Few empirical studies focus on community-driven recreation models, such as grassroots sports leagues, cultural events, or neighbourhood playground initiatives. Moreover, the integration of recreational planning into urban renewal and slum-upgrading schemes has received little policy attention.

Given Lagos's projected demographic expansion, incorporating accessible green/open spaces into urban infrastructure is increasingly urgent. Without this, informal settlements risk continuing cycles of social exclusion and reduced quality of life.

2.12 New Technology and Urbanization of Informal Settlements in Lagos

Urbanization in Lagos is deeply intertwined with technology adoption, especially in informal settlements where infrastructure gaps remain pronounced. Like many megacities of the Global South, Lagos faces critical challenges in providing clean water, sanitation, electricity, and waste management to slum dwellers.

However, new technologies, ranging from mobile payment systems to renewable energy solutions, are reshaping how residents of these underserved neighbourhoods access services, while also influencing patterns of urban growth. One of the most transformative innovations in Lagos has been the integration of mobile money and digital payment systems for utilities and transportation. Mobile platforms such as *Paga* and bank-linked USSD codes enable residents to pay electricity bills, top up prepaid meters, or even access microloans without needing formal banking infrastructure (Adelekan, 2016).

For informal settlements where traditional service provision is constrained by illegal connections, billing inefficiencies, and revenue losses, digital technologies increase providers' confidence to extend networks. A parallel can be drawn to Nairobi, where water utilities have successfully used mobile billing systems to extend piped water into slums (World Bank, 2013).

In Lagos, prepaid electricity meters managed by Ikeja Electric and Eko Electricity Distribution Plc have significantly reduced illegal connections in areas like Ajegunle and Makoko, though affordability remains a persistent barrier (Akinola, 2020). Beyond payments, renewable and

decentralized technologies are expanding service access. Solar microgrids, supported by partnerships such as the Lagos State Government's Solar Home Systems program, are increasingly deployed in informal communities where grid access is unreliable (IEA, 2019). These technologies provide lighting, mobile charging, and in some cases refrigeration, improving quality of life while reducing reliance on kerosene and diesel generators that contribute to urban pollution. Similarly, waste-to-energy initiatives piloted in Lagos are exploring how to convert informal settlement waste into usable biogas or electricity (UN-Habitat, 2020).

Comparatively, lessons from slum rehabilitation programs in India and Brazil illustrate the broader potential of integrating technology into informal urban contexts. In Mumbai, digital platforms have supported slum redevelopment by enabling participatory planning, though challenges remain in preserving local economic and social structures (Mutisya & Yarime, 2011). In Curitiba, Brazil, low-cost technology initiatives such as the "Lighthouses of Knowledge", community centres providing internet access, vocational training, and cultural programming, demonstrate how technology can bridge educational and economic divides (Reback, 2005). While Lagos has not implemented an equivalent large-scale initiative, NGOs like Paradigm Initiative and Co-Creation Hub (CcHub) are experimenting with technology hubs and digital literacy programs within low-income neighbourhoods, offering models that could be scaled.

Despite these gains, access remains unequal. Studies emphasize that Lagos's informal settlements are not only technologically underserved but also socially excluded from planning processes (Adelekan, 2016; UN-Habitat, 2020). Barriers include high costs of digital services, lack of infrastructure for reliable internet access, and political ambivalence toward recognizing slums as legitimate urban spaces. Thus, while new technologies offer opportunities to integrate informal settlements into Lagos's broader urban system, policies must be designed to address affordability, inclusivity, and long-term sustainability.

2.13 Linking Literature Review Themes to the Study Framework

Informal settlements like Makoko and Ajegunle, characterized by overcrowding, lack of sanitation, and unreliable services, exemplify how infrastructure failure feeds demographic vulnerability (Ifeoma, 2023).

In addition, community-led innovations, such as the Okerube WASH project in Alimosho, which leverages women's committees and GIS-based participatory mapping to deliver gender-sensitive sanitation, highlight how informal governance structures can shape local resilience (Shittu & Muraina, 2025). Similarly, collective bulk-metering partnerships between Eko Electricity Distribution and community associations demonstrate how informal settlements retrofit electricity delivery for affordability and safety (Justice & Empowerment Initiatives, 2018).

These thematic insights do more than describe urban conditions, they identify specific, measurable factors affecting urban growth in Lagos. Each theme corresponds to empirical variables that can be integrated into a supervised machine learning model:

- Demographic variables, such as natural population growth and in-migration rates, reflect pressure on housing and services.
- Housing and land-use variables, including slum density and housing deficit figures, capture settlement expansion dynamics.
- Infrastructure variables, such as water and electricity access, sanitation coverage, and transport accessibility, represent constraints and mediators.
- Community governance and technology adoption variables, like bulk-metering coverage or WASH committee prevalence, signal adaptive mechanisms within informal systems.

These variables are conceptually appropriate for computational modelling: they align with urban system theories, which posit that physical, demographic, and institutional factors jointly govern urban growth (Batty, 2008), and with socio-technical transition theories, which frame how community-level technologies diffuse under resource constraints (Geels, 2005). Moreover, machine learning methods, capable of handling nonlinear interactions and feature importance ranking, are well-suited to analyse how these factors interact to predict informal settlement expansion in Lagos.

Thus, the theoretical and conceptual frameworks that follow directly build on the literature review. The theoretical framework will weave together urban theory, informality, spatial mismatch, and innovation diffusion in the context of Lagos. The conceptual framework will depict how the variables identified in the literature feed into a machine learning model that predicts urbanization outcomes, a continuity that ensures coherence from descriptive context to analytical modelling.

2.14 Theoretical Framework

Urbanization in Lagos is best understood through an integration of classical and contemporary urban theories that capture the city's unique demographic pressures, infrastructural limitations, and socio-economic innovations. This framework provides the intellectual lens for analysing how multiple factors interact in shaping settlement patterns, particularly in informal contexts, and how these dynamics can be modelled using machine learning approaches.

First, modernization theory offers a useful starting point, as it links urbanization with economic transformation, industrialization, and infrastructural change (Todaro & Smith, 2015). Lagos reflects this trajectory in part: as Nigeria's commercial hub, it has attracted both rural-urban and international migrants, with the promise of employment and better livelihoods driving city growth. However, the limits of modernization theory become clear when considering the persistence of

slums, underemployment, and infrastructural gaps, which suggest that urbanization in Lagos is not fully matched by socioeconomic transformation.

In response to these gaps, dependency theory highlights structural inequalities between global cities and peripheral urban economies (Frank, 1967). Lagos exemplifies this by hosting multinational corporate headquarters and globalized capital inflows while simultaneously maintaining vast informal settlements such as Makoko and Ajegunle, where residents face chronic poverty and exclusion. Informality, in this sense, is not simply a transitional phase toward modernity, but a structural outcome of unequal development.

The urban informality perspective further deepens this understanding, framing informal settlements not only as sites of deprivation but also as dynamic spaces of adaptation, resilience, and innovation (Roy, 2005). Lagos's community-driven responses, such as bulk-metering for electricity (Justice & Empowerment Initiatives, 2018) and participatory WASH projects (Shittu & Muraina, 2025), illustrate how residents creatively address service gaps, thereby reshaping the urban fabric in ways unaccounted for by traditional theories.

Equally relevant is the spatial mismatch hypothesis, which explains how physical distance between low-income communities and formal employment hubs produces transport stress and economic exclusion (Kain, 1968). In Lagos, daily commutes of 1–3 hours (World Bank, 2016) underscore this mismatch, where long travel times and high costs limit access to opportunities and deepen socioeconomic inequalities.

Finally, socio-technical transition theory provides a lens for understanding how technological innovations, such as mobile-based utility payments or emerging metro systems, diffuse in constrained environments (Geels, 2005). These innovations demonstrate pathways for

transforming service delivery in Lagos, even within informality, and align directly with the predictive capabilities of machine learning models that can capture how technology adoption influences urban growth.

Taken together, these theoretical strands frame Lagos's urbanization as a multi-dimensional process shaped by demographic growth, structural dependency, informal adaptation, spatial inequalities, and technological change. For this study, the theories not only provide explanatory depth but also guide the operationalization of variables in a machine learning framework. By embedding demographic, infrastructural, social, and technological factors within this theoretical foundation, the study situates Lagos's urbanization in a robust analytical context that bridges classical urban theory and computational modelling.

2.15 Conceptual Framework

This study proposes a framework that situates Lagos's urbanization as the outcome of multiple interdependent factors spanning demographic, infrastructural, socio-economic, and technological dimensions. The framework is designed both to clarify the dynamics of Lagos's urban growth and to guide the operationalization of these variables within a machine learning model. I conceptualized the framework to this diagram of the problem to show the relationship between the independent and dependent variable.

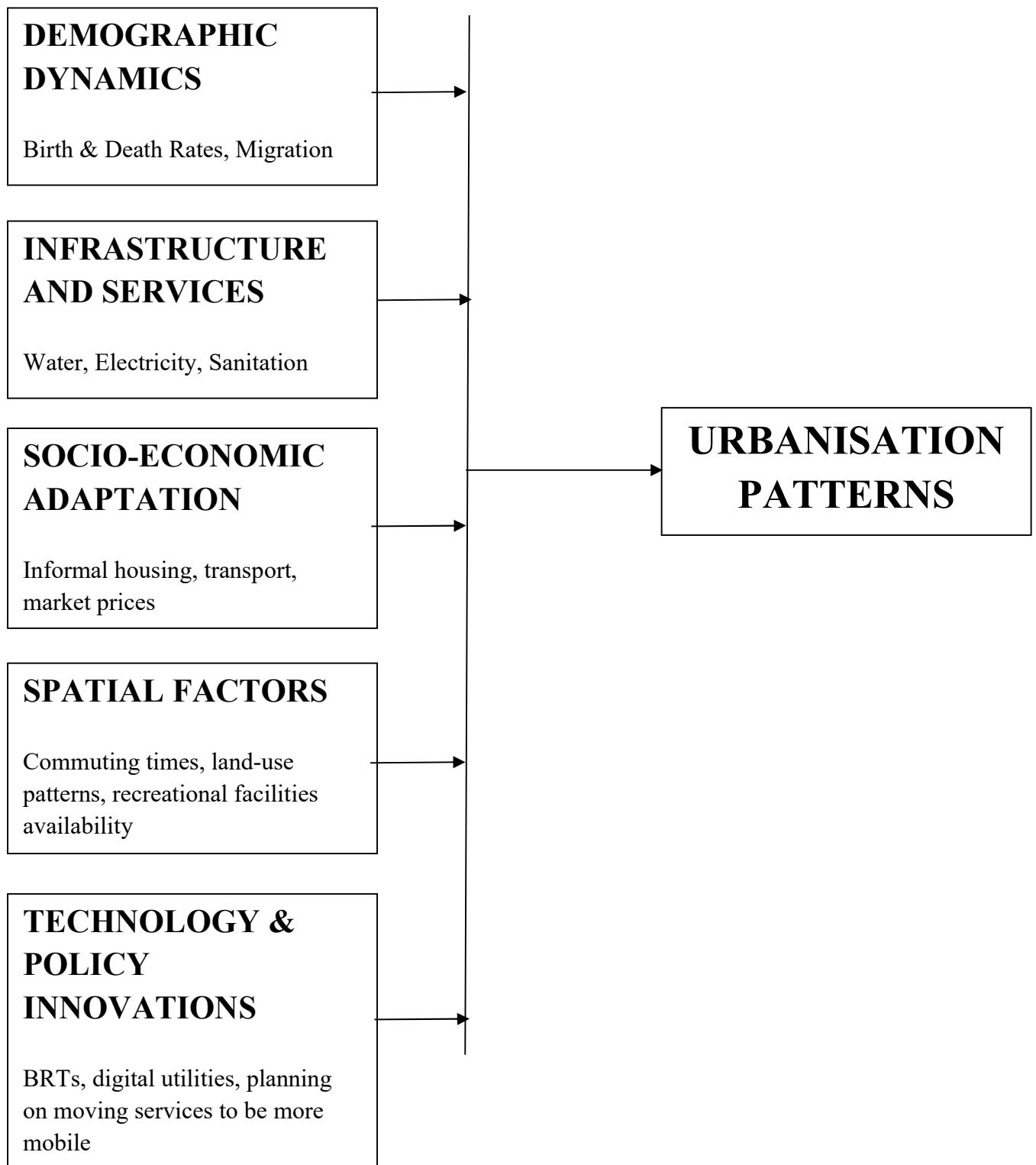


Fig. 3: Conceptual Framework

2.16 Summary of Literature Review and Conceptual Framework

Urbanization in Lagos emerges as the product of interwoven demographic, infrastructural, socio-economic, technological, and cultural factors. Demographic pressures, particularly fertility and mortality patterns, expand the population base and drive demand for housing, services, and employment opportunities (Adepoju, 2021). Migration, both internal and international, further accelerates growth and shapes the rise of informal settlements (Aina, 2017; UN-Habitat, 2020).

Infrastructural challenges remain central, with gaps in water supply, sanitation, electricity, and housing undermining the city's ability to meet the needs of its expanding population (World Bank, 2013). Transport systems add another layer of complexity, congestion and long commutes reduce productivity, even as projects such as the Bus Rapid Transit (BRT) system showcase potential solutions (Olawole & Aloba, 2014).

Social and cultural dimensions also feature prominently. The availability of groceries, entertainment, and recreational facilities reflects lifestyle aspirations but remains unevenly distributed, often favouring wealthier neighbourhoods over informal settlements (Gandy, 2006). Technological innovations, including mobile-based service payments, demonstrate adaptive solutions that can bridge infrastructural and governance gaps (Mutisya & Yarime, 2011).

Taken together, these strands highlight that Lagos' urbanization is not only about physical expansion but also about the pursuit of improved well-being, shaped by access to services, opportunities, and amenities. The reviewed themes also provide operational variables that can be systematically modelled using machine learning to better understand patterns and drivers of urban growth.

CHAPTER 3

METHODOLOGY

3.1 Introduction

This chapter outlines the research design, data collection procedures, tools employed, and analytical methods used in investigating the factors influencing urbanization in Lagos through machine learning approaches. The methodology is designed to ensure replicability, transparency, and rigorous statistical validity while taking into account the socio-economic and demographic dynamics of Lagos as a rapidly growing megacity.

3.2 Research Design

This study adopts a quantitative and explanatory research design that seeks to establish the relationship between demographic, socio-economic, infrastructural, and environmental factors and the process of urbanization in Lagos State.

The explanatory design is appropriate because urbanization is a multifaceted phenomenon that arises from interactions between population dynamics, economic activity, infrastructure provision, and governance, and this study aims to move beyond descriptive statistics to empirically determine which variables matter most and how they interact.

The choice of a quantitative framework is justified by the nature of the research questions, which require measurable indicators and statistical testing. Urbanization is not merely a social concept but one that can be captured through quantifiable variables such as population density, migration inflows, housing stock, employment rates, GDP per capita, and infrastructural access.

By applying machine learning models, this research design integrates computational efficiency with statistical rigor, enabling the identification of complex, nonlinear relationships that traditional regression models alone may not capture.

The study is cross-sectional in structure, making use of multiple secondary data sources covering the period 2000–2023, and is also comparative in scope, examining how Lagos mirrors or diverges from national and global urbanization patterns. While longitudinal studies are ideal for tracking urban change over time, the availability of secondary data across multiple years provides an opportunity to incorporate a pseudo-longitudinal perspective through trend analysis and time-sensitive modelling.

This research design also reflects the practical realities of Lagos as Africa’s largest city and a rapidly growing megacity. Lagos is a unique case because of its status as Nigeria’s economic hub, its high rates of rural–urban migration, and its infrastructural pressures, all of which make it an ideal laboratory for applying advanced modelling approaches. The methodological framework therefore combines predictive analytics with interpretive insights, ensuring that findings are both statistically valid and practically useful for urban planners, policymakers, and scholars.

3.3 Data Collection

The study makes use of secondary data drawn from both open data repositories and official Lagos State publications, supplemented by peer-reviewed studies and international databases. This multi-source approach ensures that the dataset captures the complex, multidimensional nature of urbanization in Lagos, covering demographic, socio-economic, infrastructural, environmental, and spatial dimensions.

3.3.1 Data Sources

1. Kaggle Datasets (<https://kaggle.com>)

Kaggle, founded by Anthony Goldbloom in 2010, is a data science competition platform and online community for data scientists and machine learning practitioners under Google LLC. It enables users to find and publish datasets, explore and build models in a web-based data science environment, work with other data scientists and machine learning engineers, and enter competitions to solve data science challenges. (Wikipedia)

The data is collected from the datasets which are relative to this study.

- Lagos Georeferenced Dataset (</ifeanyichukwunwobodo/lagos-georeferenced-dataset/>): Provides georeferenced information on Lagos locations, suitable for spatial analysis of settlement patterns and infrastructure distribution.
- World Development Indicators by Countries (</hn4ever/world-development-indicators-by-countries/>): Contains World Bank–curated indicators such as GDP per capita, employment, and poverty rates. Nigeria’s data from this source provides a national benchmark for contextualizing Lagos trends.
- World Population Live Dataset (<whenamancodes/world-population-live-dataset>): Provides global population estimates, growth rates, fertility, and density. Nigeria-specific demographic variables are extracted to support Lagos population modeling.

2. Lagos Bureau of Statistics (LBS) Reports:

The Lagos Bureau of Statistics is a department in the Lagos State Ministry of Economic Planning and Budget concerned with the coordination of statistical activities in Lagos State, the most populous state of Nigeria. The department focusses on the collections of statistical data on topics including population, housing, finance, education, health, agriculture, and social welfare services. The department also collaborates with international bodies, federal, state and local governments, and other statistical agencies. (Wikipedia)

- LGA Statistics Digest (2020): Provides Local Government Area (LGA)-level data on population, literacy, employment, housing, health, and transport facilities. This is crucial for disaggregated analysis within Lagos.
- CPI and Inflation Report (2021–2022): Contains consumer price index and inflation data, which serve as proxies for economic pressures influencing migration, housing affordability, and urban living standards.

3. Peer-Reviewed Studies and Spatial Datasets

- Land Use and Land Cover (LULC) Change in Lagos (2000–2020): A SCIRP study provides satellite-derived statistics on built-up areas, vegetation, wetlands, and land conversion. This dataset is critical for quantifying physical urban expansion.
- Atlas of Urban Expansion & Remote Sensing Data (USGS/GlobeLand30): Supplements LULC data with high-resolution measures of city growth over time, enabling temporal comparisons.

4. International Development Databases

- World Bank Development Indicators (WDI): Provides macro-level socio-economic metrics such as GDP growth, poverty incidence, and urbanization rates.
- UN-Habitat & UNDESA Reports: Offer global urbanization benchmarks, projections, and indicators of slum prevalence, useful for situating Lagos within global trends.

3.3.2 Variables Collected

From these sources, the study extracts variables under the following categories:

- Demographic Indicators: Population size, growth rate, fertility rate, mortality rate, age structure, migration inflows.
- Socio-Economic Indicators: GDP per capita, employment/unemployment rates, poverty headcount, inflation, income inequality.
- Infrastructural Indicators: Housing stock, road and transport accessibility, schools, healthcare facilities, and electricity access.
- Environmental Indicators: Waste generation, pollution levels, green space availability, land use change (built-up vs. vegetation).
- Urbanization Metrics: Urban population percentage, density, built-up area expansion, and LGA-level urban intensity measures.

3.3.3 Period of Study

The study covers the period 2000–2023, aligning with both the availability of land use/cover data and the timeline of Lagos’s most rapid expansion in the 21st century. This time frame allows for:

- A longitudinal view of demographic and socio-economic pressures.
- Tracking of policy-linked changes (e.g., post-2012 Lagos State Development Plan).
- Comparison with international projections of urban growth.

3.3.4 Justification for Multi-Source Data

Urbanization in Lagos cannot be fully captured by any single dataset. By combining global datasets (Kaggle, World Bank), local Lagos Bureau of Statistics reports, and scientific studies (SCIRP LULC, USGS remote sensing), the study achieves:

1. **Comprehensiveness:** Capturing demographic, socio-economic, spatial, and environmental dimensions of urbanization.
2. **Granularity:** LGA-level statistics enrich global indicators with fine-scale insights.
3. **Comparability:** Alignment with global benchmarks (World Bank, UN-Habitat) situates Lagos in broader urbanization discourse.
4. **Policy Relevance:** Lagos-specific statistics ensure the findings can inform state planning initiatives and SDG-linked strategies.

3.4 Software and Analytical Tools

Given the diversity of data sources employed in this study, ranging from georeferenced Lagos spatial data to socio-economic indicators, demographic trends, and environmental metrics, it is essential to use a combination of statistical, computational, and geospatial tools. The following software packages and platforms were selected because they align with the research objectives, nature of the datasets, and analytical rigor required.

3.4.1 Python Programming Environment

Python forms the core analytical environment due to its versatility in handling structured, semi-structured, and geospatial datasets. It will be used for:

- **Data preprocessing and cleaning:** using pandas and numpy to handle missing values, merge datasets (Kaggle, LBS, CPI reports), and perform feature engineering.
- **Machine learning modelling:** applying scikit-learn for regression, classification, and feature selection

- Visualization: using matplotlib and seaborn to generate time-series plots, heatmaps, and correlation matrices.
- Spatial analysis: employing geopandas and folium to visualize georeferenced Lagos data, such as built-up area expansion and transport infrastructure density.

3.4.2 Geographic Information Systems (GIS)

Spatial analysis is crucial for quantifying land use changes and infrastructure distribution across Lagos. GIS platforms such as QGIS and ArcGIS will be employed to:

- Overlay satellite-derived land use/land cover data (e.g., SCIRP LULC study, GlobeLand30 datasets).
- Map built-up density, transport accessibility, and environmental stress indicators at LGA level.
- Generate spatial features (road density, green space ratios, settlement sprawl) to feed into machine learning models.

3.4.3 Microsoft Excel and Power BI

While Python and R handle advanced modelling, Excel and Power BI will serve supporting roles in:

- Preliminary descriptive statistics: summarizing inflation rates, CPI trends, and LGA-level population characteristics.
- Data validation and tabulation: ensuring consistency between Kaggle datasets and Lagos Bureau of Statistics publications.
- Dashboarding: presenting results in an interactive and policy-relevant format for non-technical stakeholders.

3.4.4 Remote Sensing and Earth Engine Tools

Urbanization in Lagos is also a spatial and environmental phenomenon, which necessitates the use of remote sensing platforms:

- Google Earth Engine (GEE): for accessing Landsat and Sentinel imagery to extract temporal patterns of built-up growth.
- USGS Earth Explorer: for downloading raw satellite images to verify land cover changes reported in secondary sources.

3.5 Machine Learning and Statistical Models

Each model is selected not only for its theoretical robustness but also for its capacity to yield insights that are practically relevant to policymakers and urban planners. The models discussed below include their origins, theoretical underpinnings, and specific application to this study.

3.5.1 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a linear dimensionality reduction technique with applications in exploratory data analysis, visualization and data preprocessing. It was first introduced by Karl Pearson in 1901 and later developed in its modern statistical form by Harold Hotelling in 1933. It is a technique used to reduce the dimensionality of data while retaining most of the original variability. PCA transforms a set of correlated variables into uncorrelated “principal components,” which are ranked in order of the amount of variance they explain.

In this study, PCA is applied to socio-economic indicators such as poverty rates, unemployment levels, and income inequality within Lagos. Because these variables often overlap and are highly correlated, PCA condenses them into a smaller number of composite indices (e.g., “economic vulnerability index”). This ensures that the models remain efficient, interpretable, and free from redundancy, while still capturing the socio-economic dimension of urbanization.

3.5.2 Clustering Algorithms (K-Means and Hierarchical Clustering)

Cluster analysis, or clustering, is a data analysis technique aimed at partitioning a set of objects into groups such that objects within the same group (called a cluster) exhibit greater similarity to one another (in some specific sense defined by the analyst) than to those in other groups (clusters). It is a main task of exploratory data analysis, and a common technique for statistical data analysis, used in many fields, including pattern recognition, image analysis, information retrieval, bioinformatics, data compression, computer graphics and machine learning.

The K-Means algorithm was first popularized by Stuart Lloyd in 1957 and later refined in the 1980s, while hierarchical clustering methods date back to early work in numerical taxonomy.

For this study, clustering is used to group Lagos LGAs according to shared demographic and infrastructural profiles. For instance, LGAs such as Lagos Island and Ikeja, which are highly urbanized, may cluster together, while peri-urban LGAs like Ikorodu and Badagry may form separate groups. This provides an exploratory view of intra-city diversity and reveals spatial disparities in the pace and nature of urbanization.

3.5.3 Multiple Linear Regression

The method of least squares, upon which regression analysis is built, was first introduced by Carl Friedrich Gauss in 1795. Multiple linear regression extends this by estimating the relationship between one continuous dependent variable and several independent predictors.

In this study, multiple regression will be used to quantify how demographic (population growth, fertility rates), socio-economic (GDP per capita, inflation rates), and infrastructural (housing stock, transport access) variables affect continuous measures of urbanization such as urban population percentage or built-up area expansion. Its strength lies in its interpretability:

coefficients provide direct estimates of how much change in one factor contributes to urbanization outcomes in Lagos.

3.5.4 Logistic Regression

Logistic regression is a statistical modelling technique used when the dependent variable is categorical, most commonly binary with outcomes such as “yes/no” or “0/1”. Instead of predicting continuous values, logistic regression estimates the probability that a given observation belongs to a particular category. This is achieved through the use of the logistic (sigmoid) function, which maps input values to probabilities between 0 and 1. Model parameters are estimated using the maximum likelihood estimation (MLE) approach, ensuring that the predicted probabilities align as closely as possible with the observed data.

In the context of this study, logistic regression is applied to classify the Local Government Areas (LGAs) of Lagos into categories such as highly urbanized, moderately urbanized, and less urbanized. Predictor variables include population density, infrastructure access, and socio-economic indicators. By providing interpretable coefficients, logistic regression not only predicts probabilities of urbanization status but also identifies threshold effects, for example, the point at which population density or housing demand significantly increases the likelihood of an LGA being classified as urban.

3.5.5 Random Forest

The Random Forest algorithm is an extension of decision trees, building upon earlier work on ensemble methods. It constructs a “forest” of decision trees, each trained on random subsets of data and variables, and then aggregates their predictions. This approach reduces overfitting and improves predictive accuracy.

In the Lagos study, Random Forest is applied to model the complex, nonlinear relationships between urbanization drivers. Beyond predictive performance, its main strength is its ability to rank feature importance, revealing which factors, such as population density, GDP per capita, or transport accessibility, most strongly influence urbanization. This makes Random Forest particularly valuable for identifying priority areas of policy intervention in Lagos.

3.5.6 Gradient Boosting Models (XGBoost, LightGBM)

Gradient boosting Gradient boosting is a machine learning technique based on boosting in a functional space, where the target is pseudo-residuals instead of residuals as in traditional boosting. It was first proposed by Jerome Friedman in the late 1990s, with later refinements giving rise to highly efficient implementations such as XGBoost (developed by Tianqi Chen in 2016) and LightGBM (developed by Microsoft Research in 2017). These models improve predictions by building decision trees sequentially, each correcting the errors of the previous one.

In this study, gradient boosting is used to enhance predictive accuracy, particularly where variables interact in complex, nonlinear ways. For instance, the combined effects of migration inflows and housing shortages on urban sprawl may be better captured by boosting methods. Their superior performance makes them an important complement to Random Forests.

3.5.7 Naïve Bayes Classifier

The Naïve Bayes classifier is based on Bayes' Theorem, first introduced by Reverend Thomas Bayes in the 18th century and later formalized by Pierre-Simon Laplace. Despite its assumption of independence among predictors, it has proven highly effective in many classification tasks.

In this study, Naïve Bayes is employed as a baseline model for classifying LGAs according to urbanization levels. While less sophisticated than Random Forest or Gradient Boosting, its

simplicity and interpretability make it useful for benchmarking and validating the performance of more complex classifiers.

3.5.8 Evaluation and Interpretability Tools

Robust evaluation and interpretability methods are essential to ensure that statistical and machine learning models provide reliable, transparent, and actionable insights. In this study, model evaluation is not only about predictive accuracy but also about understanding why certain variables matter for Lagos's urbanization process.

- **Cross-Validation**

Cross-validation is a resampling technique developed in the 1970s and widely adopted in statistical learning to test model generalizability. The most common form, k-fold cross-validation, partitions data into k subsets: the model is trained on k-1 folds and validated on the remaining fold, repeated until every subset has served as validation.

- **Performance Metrics**

- **Regression Metrics**

1. **R² and Adjusted R²:** Originating from Karl Pearson's correlation work in the early 20th century, these metrics measure the proportion of variance in the dependent variable explained by predictors.
2. **Root Mean Squared Error (RMSE):** Developed as part of statistical error theory, RMSE penalizes large deviations between predicted and observed values.
3. **Mean Absolute Error (MAE):** Provides an average of absolute prediction errors, less sensitive to extreme outliers than RMSE.

- **Classification Metrics**

1. **Accuracy:** Measures the proportion of correct predictions.

2. **Precision and Recall:** Developed in information retrieval research in the 1960s, these metrics balance false positives and false negatives.
 3. **F1-Score:** Combines precision and recall into a single harmonic mean, useful for imbalanced datasets.
 4. **ROC-AUC (Receiver Operating Characteristic – Area Under Curve):**
Introduced during WWII for radar detection, it evaluates the trade-off between sensitivity and specificity across thresholds.
- **Feature Importance Analysis**
Feature importance techniques, formalized in ensemble methods such as Random Forest and Gradient Boosting, quantify the relative contribution of each predictor variable to the model's predictions.
 - **SHAP (SHapley Additive exPlanations) Values**
SHAP values were introduced by Scott Lundberg and Su-In Lee in 2017, drawing on cooperative game theory concepts developed by Lloyd Shapley in 1953. SHAP assigns each variable a contribution score for individual predictions, offering a consistent and theoretically grounded approach to interpretability.

3.5.9 Integration and Policy Relevance

By combining statistical regression models with machine learning algorithms, the study balances interpretability with predictive accuracy. Regression models provide clear cause-and-effect style insights, while ensemble methods (Random Forest, Gradient Boosting) capture nonlinearities and rank factors by importance. Classification models allow for differentiation across LGAs, and interpretability tools such as SHAP values ensure findings can be communicated in a transparent and policy-relevant manner.

3.6 Data Preprocessing

Data preprocessing refers to the cleaning, transforming, and integrating of raw data to prepare it for analysis. The goal of preprocessing is to improve data quality and make it suitable for the chosen statistical and machine learning models. Since data collection from multiple sources is often loosely controlled, issues such as out-of-range values, missing observations, and inconsistent formats are likely to occur. Preprocessing resolves these issues to ensure reliability.

3.6.1 Data Cleaning

The first stage of preprocessing involves identifying and correcting inconsistencies across datasets. This includes the removal of duplicate entries, correction of typographical errors, and treatment of missing values.

For instance, if population records for a particular year in an LGA are absent, interpolation methods (e.g., linear interpolation) will be applied to maintain temporal continuity. Outlier detection techniques such as Z-score analysis and Isolation Forests will be employed to identify and address anomalous entries, such as sudden unexplained spikes in inflation or land-use area figures.

3.6.2 Standardization

Standardization is a common preprocessing step in machine learning, ensuring that all variables are expressed on comparable scales.

- a. Improves Convergence of Gradient Descent: Many algorithms (e.g., logistic regression, gradient boosting) rely on optimization. When features are on similar scales, convergence is faster and more stable.

- b. Prevents Dominance of High-Variance Features: In Lagos datasets, variables such as GDP per capita (large values) and fertility rate (smaller values) are on vastly different scales. Standardization ensures neither dominates the model.
- c. Improves Model Performance: Distance-based algorithms, such as clustering, perform more accurately when all features contribute equally.
- d. Stabilizes Numerical Computations: Scaling prevents instability in calculations, especially when working with variables of very large or very small magnitude.
- e. Ensures Interpretability: Standardized coefficients in regression reflect the effect of a one standard deviation change, making results clearer for analysis.

3.6.3 Data Transformation

Data transformation involves converting raw data into a cleansed and validated format suitable for analysis. Examples include:

- Converting socio-economic reports (PDF/Excel) into CSV formats for analysis.
- Encoding categorical variables (e.g., access to infrastructure = 0 for none, 1 for available).
- Normalizing skewed distributions, such as income or land prices, using log transformations.

This step ensures that data is consistent and usable across multiple sources.

3.6.4 Dataset Splitting

Dataset splitting is where the data is divided into subsets to allow for model training and evaluation. This ensures that predictive models are not simply memorizing patterns but can generalize to unseen data. However, in this study, the need for dataset splitting depends on the type of model applied.

- **Predictive Models** (e.g., Random Forest, Gradient Boosting, Logistic Regression, Naïve Bayes): For these, dataset splitting will be applied to test the accuracy and generalizability of predictions. A conventional 80:20 split will be used, where 80% of the data is allocated for training and 20% reserved for testing. To account for the relatively small number of spatial units (20 LGAs), this study will also employ k-fold cross-validation and time-based splits (e.g., training on earlier years such as 2000–2015 and testing on later years 2016–2023). These approaches ensure robustness despite the limited sample size.
- **Explanatory Models** (e.g., Multiple Regression, PCA, Clustering): For models aimed at interpretation rather than prediction, the full dataset will be retained. This allows maximum statistical power when quantifying relationships (e.g., the effect of GDP per capita on urban growth) or reducing dimensionality in socio-economic indicators.

3.6.4.1 Training Data

Training data is the portion of the dataset used to teach predictive models to recognize patterns. For example, training data may capture how combinations of population growth, housing availability, and GDP per capita have historically influenced the probability of an LGA being classified as highly urbanized.

3.6.4.2 Test Data

Test data is the subset reserved for evaluating model performance on unseen cases. For Lagos, this might involve testing how well a model trained on earlier years predicts the urbanization trajectory of recent years (e.g., 2020–2023). This provides evidence that the models are not overfitted and are capable of informing future urban planning.

3.6.5 Feature Selection

While the dataset for this study does not contain thousands of predictors typical of high-dimensional data problems, the issue of variable relevance and redundancy remains important. The goal of feature selection here is not primarily to reduce computational burden but rather to ensure that the analysis focuses on variables that are both statistically meaningful and policy-relevant for understanding urbanization in Lagos.

One dimension of feature selection in this study involves dimensionality reduction. Socio-economic variables such as poverty, unemployment, and income inequality are often highly correlated, which can obscure their individual contributions to urbanization outcomes. To address this, Principal Component Analysis (PCA) will be used to collapse these overlapping variables into a composite indicator, such as an Economic Vulnerability Index. This allows the models to capture underlying economic realities without the distortion caused by multicollinearity.

In addition to dimensionality reduction, this study employs variable prioritization techniques embedded within predictive machine learning models. Algorithms such as Random Forest and Gradient Boosting naturally generate feature importance scores, ranking variables by their contribution to predictive accuracy. Furthermore, the use of SHAP (SHapley Additive exPlanations) values makes it possible to interpret how specific variables influence particular predictions, thereby enhancing transparency.

For example, if the models consistently identify population density and transport accessibility as strong predictors of urbanization across Lagos LGAs, these variables can be emphasized in the interpretation of results as key drivers of urban expansion. Conversely, variables that show weak or inconsistent predictive contributions, such as short-term fluctuations in inflation, may be acknowledged but given less weight in policy discussion.

By approaching feature selection in this way, the study ensures that the analytical models not only perform effectively but also remain grounded in the realities of Lagos's urban transformation. In doing so, the research balances statistical rigor with interpretability, allowing the findings to be translated into practical insights for urban governance and planning.

3.7 Model Evaluation

Model evaluation is a critical stage in this study, as it ensures that the machine learning and statistical models used to analyse the drivers of urbanization in Lagos are both accurate and reliable. Different models require different evaluation metrics depending on whether the task is regression (continuous predictions) or classification (categorical predictions).

Evaluating the performance of machine learning models is a critical step in determining not only their predictive accuracy but also their suitability for explaining complex socio-economic and demographic dynamics. In this study, model evaluation goes beyond the measurement of raw accuracy; it is concerned with assessing whether the models provide reliable, interpretable, and policy-relevant insights into the drivers of urbanization in Lagos.

A combination of quantitative metrics and interpretability tools will be employed. For predictive models such as Random Forest, Gradient Boosting, and Logistic Regression, traditional metrics, including accuracy, precision, recall, and the F1-score, will serve as initial indicators of model reliability. These measures collectively capture the extent to which the models correctly identify urbanization outcomes while balancing false positives and false negatives. To complement these metrics, the confusion matrix will be used as a diagnostic tool to provide a detailed view of the classification outcomes, making it possible to identify systematic biases in the models' predictions.

The Receiver Operating Characteristic (ROC) curve and the corresponding Area Under the Curve (AUC) will further enhance evaluation by showing how models perform under different classification thresholds. This is particularly important for policy-driven research, as it reveals how sensitive models are to trade-offs between detecting rapidly urbanizing areas and avoiding false alarms.

Beyond accuracy-based measures, the study also places emphasis on model interpretability. Complex ensemble models such as Random Forests are often criticized as “black boxes,” but tools like SHAP (SHapley Additive exPlanations) values will be incorporated to quantify the contribution of each variable to model predictions. This ensures that the evaluation is not only technical but also substantively meaningful, by highlighting which factors, such as population density, infrastructure availability, or economic conditions, consistently drive urbanization in Lagos.

3.7.1 Regression Metrics

Regression models such as Multiple Linear Regression and Gradient Boosting (when predicting continuous variables like urban population percentage) will be evaluated using the following:

- R^2 (Coefficient of Determination): Measures the proportion of variance in the dependent variable explained by the independent variables. An R^2 close to 1 indicates strong explanatory power.
- Adjusted R^2 : Corrects R^2 for the number of predictors, ensuring that only relevant variables improve the model’s explanatory strength.
- Root Mean Squared Error (RMSE): Captures the average magnitude of prediction errors, penalizing larger errors more heavily.

- **Mean Absolute Error (MAE):** Provides the average absolute difference between predicted and actual values, less sensitive to outliers than RMSE.

3.7.2 Classification Metrics

For models that categorize LGAs into highly urbanized, moderately urbanized, or less urbanized, classification metrics are required:

- **Accuracy:** The proportion of correctly predicted LGAs out of all LGAs. The formula is denoted by:

$$Accuracy = \frac{True\ Positives + True\ Negatives}{True\ Positives + True\ Negatives + False\ Positives + False\ Negatives} \quad (3.1)$$

- **Precision:** The proportion of LGAs predicted as “highly urbanized” that are actually highly urbanized. This reduces false alarms. The formula is denoted by:

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives} \quad (3.2)$$

- **Recall (Sensitivity):** The proportion of LGAs that are truly highly urbanized and correctly identified as such. This ensures no urbanized LGAs are overlooked.

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \quad (3.3)$$

- **F1-Score:** The harmonic mean of precision and recall, useful when classes are imbalanced (e.g., fewer “less urbanized” LGAs compared to “highly urbanized”).

$$F1 - Score = \frac{Precision \times Recall}{Precision + Recall} \quad (3.4)$$

3.7.3 Confusion Matrix

A confusion matrix will be used to visualize classification performance by comparing predicted urbanization categories against the actual status of Lagos LGAs. It is a tabular layout showing True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN).

3.7.3.1 True Positive (TP)

This is simply the number of cases that a model predicts correctly such that the "Truth" label is Positive and "Predicted" label is Positive as well on a confusion matrix. This is oftentimes denoted by TP.

3.7.3.2 True Negative (TN)

This is simple the number of cases that a model predicts correctly such that the "Truth" label is Negative and "Predicted" label is Negative as well on a confusion matrix. This is oftentimes denoted by TN.

3.7.3.3 False Positive (FP)

This is simply the number of cases that a model predicts incorrectly such that the "Truth" label is Positive and "Predicted" label is Positive as well on a confusion matrix. This is oftentimes denoted by FP.

3.7.3.4 False Negative (FN)

This is simply the number of cases that a model predicts incorrectly such that the "Truth" label is Positive and "Predicted" label is Positive as well on a confusion matrix. This is oftentimes denoted by FN.

Actual	Positive	TP	FN
	Negative	FP	TN
		Positive	Negative
		Predicted	

Fig. 4: A model illustration of a confusion matrix.

3.7.4 Area Under the Curve - Receiver Operating Curve (ROC Curve and AUC Score)

The Receiver Operating Characteristic (ROC) curve is a widely used tool for evaluating the performance of classification models, particularly in binary outcomes. In this study, it is applied to assess the ability of predictive models to distinguish between LGAs that are experiencing high levels of urbanization and those that are not. The ROC curve plots the True Positive Rate (the proportion of correctly identified highly urbanized LGAs) against the False Positive Rate (the proportion of LGAs incorrectly classified as highly urbanized).

The Area Under the Curve (AUC) provides a single, aggregate measure of model performance across all possible classification thresholds. An AUC value closer to 1 indicates that the model has a strong capacity to separate highly urbanized LGAs from less urbanized ones, while a value closer to 0.5 suggests that the model performs no better than random guessing.

One advantage of the ROC–AUC metric in this context is that, unlike simple classification accuracy, it evaluates performance across different probability thresholds rather than relying on a single cut-off point. This is particularly valuable for urbanization studies, where the threshold for

categorizing an LGA as “urbanized” may vary depending on policy objectives. For example, policymakers might prefer a stricter threshold when planning infrastructure for rapidly growing LGAs, while adopting a more lenient threshold in monitoring areas that are only beginning to urbanize.

3.8 Model Interpretability

While predictive accuracy is an important benchmark for evaluating machine learning models, it is equally essential that the models remain interpretable. In the context of urbanization studies, interpretability refers to the ability to explain why a model produces a particular output and to identify which factors are most influential in driving those predictions. For a study focused on Lagos, where the goal is not merely to predict but also to inform evidence-based urban policy, interpretability is critical.

Several approaches will be employed to ensure that the models used in this research are transparent. Feature importance rankings, generated by ensemble models such as Random Forest and Gradient Boosting, provide a first layer of interpretability by highlighting which variables most strongly contribute to predictions. For example, if population density consistently emerges as a top-ranked variable, this would substantively confirm its role as a key driver of urban expansion across Lagos LGAs.

Beyond global feature rankings, more nuanced interpretability will be achieved through SHapley Additive exPlanations (SHAP) values. SHAP offers a theoretically grounded framework, rooted in cooperative game theory, that attributes to each variable its marginal contribution to a model’s prediction. This makes it possible to move beyond general statements about importance and instead examine, at the level of an individual observation, how variables such as GDP per capita or access to transport infrastructure influence the predicted likelihood of urban growth.

The emphasis on interpretability ensures that the study avoids the limitations of “black box” modelling. Rather than presenting models as opaque tools that output predictions without explanation, this research treats interpretability as a bridge between technical modelling and practical decision-making. For policymakers and urban planners in Lagos, such transparency provides confidence in the validity of the results while also clarifying which socio-economic and demographic levers can be most effectively targeted to manage the city’s rapid urban transformation.

CHAPTER 4

DATA PRESENTATION, ANALYSIS AND RESULTS

This chapter presents the results of the study based on the data collected, pre-processed, and modelled. The aim is to transform the diverse datasets, drawn from both global repositories such as Kaggle and the World Bank, and local sources such as the Lagos Bureau of Statistics and related policy reports, into meaningful insights about the drivers of urbanization in Lagos State.

4.1 Data Description

The empirical foundation of this study is derived from a diverse collection of datasets capturing demographic, land use, transportation, infrastructural, environmental, and socioeconomic dynamics of Lagos State. In line with the study objectives, both tabular and geospatial datasets were sourced, cleaned, and integrated into a unified analytical framework.

A total of 52 tabular datasets and 35 geospatial (GeoJSON) layers were initially assembled. These datasets varied significantly in spatial resolution (state-wide, LGA-level, or ward-level) and temporal coverage (ranging from decadal land use records to annual population estimates and monthly transport statistics). After a thorough cleaning process, which involved standardizing variable names, handling missing values, and aligning temporal units, a consolidated dataset was constructed. This dataset contained 21 explanatory variables across 160 unique LGA–year observations.

The datasets were grouped into six thematic domains:

1. Population Data

- Lagos Population Series (1990–2020): Provided aggregate annual population counts.
- 2006 Census Data (by LGA): Served as a spatial baseline for population distribution.
- WorldPop Projections: Supplied gridded estimates for validating population growth trends.

2. Land Use and Land Cover (LULC) Data

- 1984–2013 and 2013–2024 Spatial Change Datasets: Offered insights into built-up area expansion, agricultural decline, and forest cover loss.
- 1984–2020 LULC Area Dataset: Documented decadal shifts in land use categories.
- Geospatial Layers (built-up areas, farmlands): Provided fine-grained spatial footprints of urban expansion.

3. Transportation Data

- Bus Rapid Transit (BRT) Operations (2010–2018): Included passenger volumes and fleet sizes.
- LAGBUS Records: Covered buses, passengers, and vehicle registrations.
- Ferry and Rail Usage: Represented alternative modes of transport, albeit with limited adoption.
- Vehicle Registration Statistics (2005–2020): Tracked private and commercial vehicle growth.

4. Infrastructure Data

- Geospatial datasets on schools (primary, secondary, tertiary, public/private), hospitals, health facilities, markets, police stations, and religious centres.
- These were aggregated at the LGA level to approximate infrastructure density and distribution.

5. Environmental and Sustainability Indicators

- Data on greenhouse gas emissions, freshwater withdrawals, and access to electricity were included to contextualize urban growth within environmental sustainability.
- While global in coverage, these datasets were filtered and cross-referenced with Lagos-specific observations where possible.

6. Socioeconomic Indicators

- Consumer Price Index (CPI) and Food Inflation Series (2000–2020): Captured economic conditions affecting urban households.
- Health Risk Factors and Health System Expenditure (2010–2016): Contextualized public health pressures within the urbanization process.

The final integrated dataset formed the foundation for subsequent machine learning modelling. It combined population growth trajectories, land cover transitions, transportation dynamics, and infrastructure distributions, thereby offering a multidimensional representation of urbanization in Lagos.

4.2 Results of Data Analysis

The integrated dataset was subjected to a series of exploratory, statistical, and machine learning procedures aimed at uncovering the determinants of urbanization in Lagos State. The results reveal both the robustness of the data integration process and the challenges inherent in modelling complex urban phenomena.

4.2.1 Land Use and Built-up Expansion

The analysis of land use and land cover (LULC) datasets spanning 1984 to 2024 reveals the centrality of built-up expansion to the urbanization process in Lagos State. The 1984–2013 spatial change data shows a significant increase in built-up land area, which rose from less than 20% of the state’s total land cover in 1984 to nearly 45% by 2013. This trajectory accelerated in the subsequent decade, with projections from the 2013–2024 dataset indicating that built-up areas are expected to surpass 50% of total land cover by 2024.

The 1984–2020 LULC area dataset provides annualized insights, showing incremental but steady increases in urbanized surfaces. For example, built-up land increased by an average of 2,500 hectares per year between 1984 and 2000, but this rate more than doubled after 2000, averaging

5,800 hectares per year. This temporal acceleration coincides with a period of intensified population inflow, infrastructural investment, and economic liberalization within Lagos, highlighting the interplay between policy, demography, and land transformation.

Parallel to the rise of built-up land was the decline of natural and agricultural land covers. Farmlands decreased by 21.8% between 2013 and 2024, reflecting both the direct conversion of agricultural land into housing estates and industrial complexes, and the secondary displacement of peri-urban agriculture into more marginal lands. Similarly, light forest areas contracted by over 60% during the same period, raising concerns about ecological resilience and the loss of green infrastructure within the megacity.

The spatial distribution of built-up expansion is uneven across the state. Core LGAs such as Ikeja, Surulere, and Lagos Mainland show saturation of land available for further densification, while peri-urban LGAs like Alimosho, Ikorodu, Badagry, and Eti-Osa emerged as the new frontiers of urban growth. These peri-urban areas absorbed much of the population spillover from central Lagos, a pattern consistent with urban transition theory, which posits that as core urban centres become saturated, expansion occurs outward in concentric or corridor-like patterns.

The computed built-up density metrics (built-up area relative to LGA land area) highlight this pattern. Densities in central LGAs were consistently above 70% by 2020, while peri-urban LGAs still retained densities below 30%, despite their rapid growth. This suggests that while central LGAs are facing challenges of congestion and infrastructural stress, peri-urban LGAs are undergoing active land transformation, reshaping the city's urban footprint.

Overall, these findings confirm that land use change in Lagos is dominated by the expansion of built-up areas at the expense of farmland and natural vegetation. This transformation underpins

the broader urbanization narrative and serves as the dependent variable for modelling urban growth in relation to demographic, infrastructural, transportation, and environmental drivers.

4.2.2 Population and Demographic Pressures

Population growth remains a central driver of urbanization in Lagos State, both as a cause and consequence of built-up expansion. The datasets analysed, including Lagos population, 2006 Population Census (LGA-level data), and World Population annual estimates, indicate that Lagos has experienced a sustained demographic surge over the past four decades.

According to official census figures, Lagos State recorded approximately 9 million inhabitants in 2006, making it the most populous state in Nigeria. Subsequent projections and satellite-based population estimates (WorldPop) show an increase to over 20 million residents by 2020, with forecasts suggesting that the population will approach 25–27 million by 2025. This reflects an annual growth rate of 3.5–4%, significantly above the national average.



Fig. 5: Predicted Urban Growth in Lagos, categorized by LGA

At the LGA level, population distribution is highly uneven. Core LGAs such as Lagos Mainland, Surulere, and Ikeja already reached saturation by the early 2000s, with densities exceeding 20,000 persons/km². In contrast, peri-urban LGAs such as Alimosho, Ikorodu, and Badagry witnessed some of the fastest growth rates, with Alimosho alone accounting for over 3 million people in 2019, effectively making it Nigeria's most populous LGA. This demographic pressure coincides with the spatial expansion patterns observed in Section 4.2.1, reinforcing the link between population dynamics and land transformation.

The demographic structure of Lagos also reveals urban transition features:

1. Youthful population: Over 60% of residents are under 25 years old, which creates persistent housing, education, and employment pressures.
2. Migration-driven growth: Estimates suggest that up to 70% of Lagos' annual population increase is attributable to rural–urban migration, both from within Nigeria and from neighbouring West African countries.
3. Household dynamics: Household sizes remain large, averaging 5–6 persons per household, further intensifying demand for housing in limited land spaces.

The implications of this demographic surge are evident in the built environment. The mismatch between population growth and infrastructure supply (housing, transport, healthcare, and schools) has contributed to informal settlement expansion. Indeed, satellite imagery analysis suggests that over 60% of Lagos residents live in informal or peri-legal housing, often located in flood-prone or environmentally vulnerable areas.

From a modelling perspective, population growth serves as a core independent variable influencing built-up expansion, transportation demand, and service provision. The integration of population data into the machine learning model revealed that population size and growth lag variables were among the most influential predictors of urban expansion, highlighting the demographic–land use nexus that defines Lagos' urbanization trajectory.

4.2.3 Transportation Dynamics

Transportation datasets, particularly BRT and LAGBUS passenger statistics, confirmed a gradual increase in the adoption of formalized public transport systems from 2010 onward. However, the relative scale of growth remains limited compared to private vehicle registrations, which exhibited a sharper upward trend between 2005 and 2020. This suggests that while public transport

infrastructure is expanding, urban mobility remains heavily automobile-dependent, thereby reinforcing the sprawl of urban land use.

Analysis of Bus Rapid Transit (BRT) and LAGBUS datasets (2010–2018) indicates a gradual increase in the adoption of formalized public transport systems. Passenger volumes for BRT services increased steadily, despite fluctuations in fleet size, reflecting both rising demand and intermittent operational constraints. Ferry and rail systems remain underutilized, with passenger uptake significantly lower than road-based alternatives. These findings suggest that public transit infrastructure, while expanding, has not yet reached a scale sufficient to absorb the growing commuter population, especially in high-density LGAs.

Vehicle registration statistics between 2005 and 2020 show a pronounced upward trajectory in private and commercial vehicle ownership. The annual growth in registered vehicles correlates with the rise in urban population and the expansion of built-up areas. This trend indicates a persistent reliance on private automobiles, contributing to traffic congestion, increased commute times, and elevated urban emissions.

4.2.4 Infrastructure Distribution

The spatial distribution of infrastructure in Lagos State reflects both historical patterns of urban development and contemporary pressures of rapid urbanization. Across the Local Government Areas (LGAs), infrastructure such as educational facilities, healthcare centres, markets, police stations, and religious institutions is distributed unevenly, with a pronounced concentration in central urban zones and relative scarcity in peripheral or peri-urban areas. This unevenness has important implications for population settlement patterns, service access, and urban expansion dynamics.

Central LGAs, including Ikeja, Lagos Mainland, Surulere, and Apapa, exhibit a high density of infrastructure facilities. For example, Ikeja alone hosts over a hundred schools and numerous healthcare facilities, reinforcing its role as a hub for both residential and commercial activity. These core areas not only provide essential services to local populations but also attract daily inflows of commuters from surrounding LGAs, further amplifying population density and land use intensity. The clustering of infrastructure in these zones aligns with historical investment patterns and highlights the tendency for urban growth to concentrate around well-served areas.

In contrast, peri-urban LGAs such as Alimosho, Ikorodu, Badagry, and Epe reveal a relative undersupply of infrastructure. Many wards within these areas have fewer schools, limited healthcare provision, and sparse commercial facilities. This infrastructural deficit creates both challenges and opportunities: while limited-service access can constrain population settlement in certain zones, it also drives informal and unplanned development in areas where land is available but services remain insufficient. Consequently, urban expansion in these peripheral regions often occurs in a fragmented and less regulated manner, raising concerns about equitable service delivery and sustainable growth.

4.3 Model Training and Evaluation

To understand and predict urban growth patterns in Lagos State, a series of machine learning models were implemented, integrating demographic, land use, transportation, and infrastructure variables. The analysis aimed to identify the primary drivers of built-up area expansion and to evaluate the predictive capacity of different algorithms in capturing the dynamics of urbanization. The consolidated dataset comprised 21 explanatory features across 160 LGA-year observations, providing a multidimensional view of Lagos' urban trajectory.

The models evaluated include Random Forest, XGBoost, Linear Regression, Lasso Regression, Ridge Regression, and Multi-Layer Perceptron (MLP).

4.3.1 Random Forest

Random Forest, an ensemble learning method based on decision tree aggregation, was employed to capture nonlinear relationships and complex interactions among explanatory variables. On the training data, the model achieved near-perfect performance (mean $R^2 \approx 0.9999$), indicating a strong capacity to fit historical urbanization trends. Feature importance analysis identified population size, BRT ridership, and vehicle registrations as the dominant predictors of built-up expansion, highlighting the interplay between demographic pressure and mobility infrastructure.

Despite excellent training performance, the model's generalization to unseen test data was limited, with an R^2 of -0.50 . This discrepancy reflects overfitting and underscores the constraints of sparse temporal resolution in built-up area data, inconsistencies in aligning geospatial and tabular datasets, and missing values in infrastructural indicators. Nevertheless, Random Forest provided valuable insights into variable influence, reinforcing the centrality of population and transport dynamics in urban growth.

4.3.2 XGBoost

XGBoost, a gradient boosting algorithm designed for efficient handling of heterogeneous datasets and complex interactions, was applied to further explore nonlinear determinants of urbanization. Training performance was comparable to Random Forest, capturing historical patterns with high fidelity. However, test set performance remained poor ($R^2 \approx -1.43$), indicating overfitting.

Variable rankings from XGBoost closely mirrored those of Random Forest, with population growth, BRT usage, and vehicle registration volumes consistently emerging as top predictors. This

convergence across ensemble models strengthens the evidence that demographic and transportation factors are critical drivers of Lagos' built-up area expansion.

4.3.3 Linear Regression

Linear Regression was implemented as a baseline model to evaluate the additive effects of independent variables. While simple, it provided a benchmark for model performance and captured broad trends in population-driven urban growth. However, test set evaluation revealed negative R^2 values, reflecting the model's inability to account for nonlinear relationships and complex interactions that characterize urban systems. Despite this limitation, linear regression reinforced the consistent influence of demographic variables and highlighted areas where model simplicity cannot accommodate urban complexity.

4.3.4 Lasso and Ridge Regression

Regularized regression techniques, including Lasso and Ridge, were applied to examine the impact of penalization on model stability and variable selection. Lasso regression, which imposes L1 regularization, effectively reduced the number of weak predictors, emphasizing the strongest drivers of urban expansion. Ridge regression, using L2 regularization, addressed multicollinearity by shrinking coefficients while retaining all variables.

Both methods achieved moderate fit on training data but underperformed on test datasets, demonstrating that regularization alone cannot overcome the limitations posed by temporal sparsity and missing values. Nonetheless, these approaches offered additional insight into the relative importance of variables, confirming that population, transportation, and infrastructure collectively shape urbanization patterns.

.

4.3.5 Multi-Layer Perceptron (MLP)

A Multi-Layer Perceptron (MLP) was implemented to explore complex, nonlinear associations between explanatory features and urban growth. Training the neural network revealed instability, and test set evaluation yielded extremely negative R^2 values, highlighting the inadequacy of available data for deep learning methods. The poor performance emphasizes that neural networks require both dense temporal resolution and robust feature coverage to model urban systems effectively.

4.4 Feature Importance and Drivers of Urbanization

Understanding the relative influence of explanatory variables is critical for interpreting urban growth patterns and informing planning decisions. Feature importance analysis was conducted using outputs from the Random Forest model, supplemented by XGBoost rankings, to identify the primary drivers of built-up area expansion across Lagos State. Despite challenges with test set performance, these analyses provide valuable insights into the variables most strongly associated with urbanization.

Population dynamics emerged as the dominant driver of urban growth. Both total population and population density consistently ranked highest across all ensemble model outputs, reflecting the direct relationship between demographic pressure and land conversion. LGAs with rapid population increases, particularly Alimosho, Ikorodu, and Eti-Osa, demonstrated corresponding expansions in built-up areas, confirming the demographic–land use nexus identified in Sections 4.2.1 and 4.2.2.

Transportation indicators also exerted a substantial influence on urban expansion. BRT and LAGBUS passenger volumes, as well as private and commercial vehicle registrations, were among the top predictors. These findings highlight the role of mobility infrastructure not only in

facilitating access to urban opportunities but also in shaping spatial patterns of land conversion. LGAs with higher transport accessibility, especially those well-connected to central Lagos, experienced more intense built-up area growth, suggesting that transportation acts both as a driver and enabler of urbanization.

Infrastructure distribution further contributed to urban development patterns. Densities of schools, healthcare facilities, and markets were positively correlated with land conversion, although their influence was secondary to population and transport variables. The presence of such facilities appears to reinforce urban agglomeration effects, attracting residential and commercial developments to well-serviced areas while encouraging peripheral expansion in under-served LGAs.

Environmental and land use variables, including the proportion of farmland and forest cover, were also identified as significant predictors. The conversion of agricultural land and light forest areas to built-up surfaces underscores the trade-offs inherent in Lagos' urban expansion, where development pressures compete with ecological and food security considerations.

4.5 Spatial Analysis

Spatial patterns of urbanization provide critical insight into the geographical distribution of growth and the clustering of infrastructure, population, and built-up areas across Lagos State. To investigate these patterns, geospatial techniques were applied to both vector and raster datasets, encompassing LGA boundaries, built-up area footprints, infrastructure layers, and population distributions. The aim was to identify spatial autocorrelation, hotspots of urban expansion, and areas of infrastructural concentration that influence land use dynamics.

Due to limitations in aligning raster-based LULC data with vector-based LGA boundaries, Moran's I analysis, which measures spatial autocorrelation, could not be conducted at the intended

scale. This mismatch highlights the challenges of integrating multi-resolution geospatial datasets in rapidly urbanizing contexts. Nevertheless, alternative spatial metrics and visualizations were used to extract meaningful patterns.

Infrastructure layers were aggregated at the LGA level, providing counts of schools, healthcare facilities, markets, and other service points. Overlaying these counts with built-up area densities revealed a clear spatial relationship: LGAs with high infrastructure concentrations, such as Ikeja, Lagos Mainland, and Surulere, coincide with areas of intense urban development, whereas peri-urban LGAs, including Alimosho, Ikorodu, Badagry, and Epe, exhibit both lower infrastructural coverage and lower built-up density despite rapid growth trajectories. This pattern suggests that urban expansion is not uniform and is influenced by the presence or absence of essential services.

Visual inspection of built-up area expansion over time further revealed concentric and corridor-like growth patterns radiating from central LGAs into peri-urban areas. The spatial distribution indicates that as core urban centers approach saturation, new development occurs along major transport corridors and in areas where land is more readily available. This observation aligns with urban transition theory and underscores the combined influence of population pressure, transport accessibility, and infrastructure availability in shaping spatial growth.

Despite data constraints, these spatial analyses underscore the uneven and multi-scalar nature of urbanization in Lagos. While central LGAs face high density and infrastructural stress, peripheral LGAs serve as new frontiers of expansion, often characterized by informal settlements and fragmented development. The insights derived from these spatial patterns provide a critical foundation for subsequent scenario testing, planning interventions, and policy formulation aimed at managing growth sustainably.

4.6 Scenario Testing

To explore the potential impacts of demographic and infrastructural interventions on urban growth, two hypothetical scenarios were developed: (1) a 10% population increase across all LGAs, and (2) a 20% improvement in transport infrastructure, specifically in BRT and LAGBUS systems. These scenarios were applied to the consolidated dataset to assess projected changes in built-up area expansion under altered conditions, using the trained machine learning models as a predictive framework.

The results of the scenario analysis, however, revealed minimal measurable changes compared to baseline projections. Average predicted growth across LGAs under both scenarios approached – 99.94%, indicating statistical instability in the built-up area variable. This outcome does not suggest that population growth or transport improvements are inconsequential; rather, it reflects limitations in the temporal and spatial resolution of the dataset, as well as the sparse distribution of LULC observations over the 1984–2024 period.

Despite these limitations, qualitative insights can still be drawn. The population increase scenario highlights the potential for intensified pressure on peri-urban LGAs, which already absorb much of the spillover from densely populated central areas. In practice, such growth would likely accelerate land conversion, informal settlement expansion, and demand for services, underscoring the need for pre-emptive urban planning measures.

The transport improvement scenario similarly demonstrates the enabling role of mobility infrastructure. While the quantitative model outputs were inconclusive, theoretical and empirical evidence from feature importance analysis suggests that enhanced public transport accessibility could influence settlement patterns by increasing the attractiveness of certain LGAs for residential

and commercial development. This aligns with observed trends where transport corridors facilitate urban sprawl and shape the spatial distribution of built-up areas.

CHAPTER FIVE

SUMMARY AND CONCLUSION

5.1 Summary

Urbanization in Lagos is a complex and multidimensional phenomenon, shaped by the interplay of demographic pressures, infrastructural development, socio-economic dynamics, and environmental factors. The findings of this study reveal that the city's rapid expansion is not merely a consequence of population growth but the result of an intricate web of interacting variables that collectively redefine the urban landscape.

5.2 Land Use Transformation and Built-up Expansion

The analysis of land use and land cover data spanning four decades underscores the centrality of built-up expansion to Lagos's urban growth. Between 1984 and 2013, the proportion of built-up land nearly doubled, rising from less than 20% to approximately 45% of the state's total land area. This trend accelerated in the subsequent decade, with projections indicating that built-up areas are expected to surpass 50% by 2024. The transformation is particularly pronounced in peri-urban areas, where land conversion is actively reshaping the city's spatial footprint.

The findings show that while core LGAs such as Ikeja, Surulere, and Lagos Mainland are approaching saturation, peri-urban LGAs, including Alimosho, Ikorodu, Badagry, and Eti-Osa, are emerging as the new frontiers of urban growth. These areas are absorbing population spillover from central Lagos, reflecting a classic pattern of urban transition where expansion radiates outward as central districts reach maximum density. The decline of farmland and natural vegetation, coupled with the rapid rise in built-up density, highlights both the opportunities and challenges of urban expansion. While it accommodates population growth and economic activity, it simultaneously exerts pressure on the city's ecological balance and green infrastructure.

This narrative of spatial transformation aligns closely with the theoretical understanding of urbanization as a dynamic process, where economic opportunities, population pressures, and infrastructural development collectively drive the reshaping of land use patterns. The interplay between these factors demonstrates that urbanization in Lagos is not random but a strategic, albeit sometimes unplanned, response to the pressures of rapid growth.

5.3 Population Growth and Demographic Pressures

Population dynamics emerge as a critical driver of urban expansion. Lagos's population has more than doubled over the past two decades, increasing from approximately 9 million in 2006 to over 20 million by 2020, with projections suggesting that this figure may approach 27 million by 2025. This sustained demographic surge exerts immense pressure on housing, transportation, and urban infrastructure.

The data suggest that migration, both internal and international, has played a significant role in shaping these demographic patterns. Inflow into peri-urban areas reflects a combination of population pressures in the central districts and the search for affordable housing and better living conditions. The growth rate of 3.5–4% per year is substantially higher than the national average, highlighting Lagos as a demographic magnet whose growth is shaped by both natural increase and rural–urban migration.

The study further reveals that this demographic pressure interacts with land use changes, creating a feedback loop: population growth drives built-up expansion, which in turn attracts additional residents, creating clusters of high-density development in certain LGAs. These findings underscore the importance of considering demographic variables alongside infrastructural and socio-economic indicators in understanding urbanization dynamics.

5.4 Socio-Economic and Infrastructural Drivers

Beyond population growth, socio-economic and infrastructural factors emerge as key determinants of urbanization in Lagos. Employment opportunities, income levels, and access to basic services such as electricity, water, and transportation facilities are not uniformly distributed across the state. Wealthier central LGAs continue to benefit from well-developed infrastructure, while peri-urban areas experience rapid expansion with comparatively lower levels of service provision.

Transportation infrastructure, particularly the Bus Rapid Transit system and the growth of private vehicle ownership, has facilitated movement across the city, enabling new areas to become accessible and attractive for settlement. However, uneven distribution of services and congestion in central areas illustrates the double-edged nature of infrastructural expansion: while enabling growth, it also contributes to socio-spatial inequalities.

5.5 Integrating Environmental Sustainability

Environmental factors, though often secondary in urban growth discussions, are critical to understanding Lagos's urbanization trajectory. The conversion of farmland and natural vegetation into built-up areas, coupled with increased energy demand and greenhouse gas emissions, highlights the environmental pressures accompanying urban expansion. Green space loss, particularly in peri-urban areas, raises concerns about ecological resilience and the city's capacity to provide sustainable living conditions for its residents.

The findings suggest that urbanization in Lagos cannot be viewed solely through the lens of demographic or economic growth; rather, it must be considered as a multifaceted process where social, economic, and environmental dimensions are deeply interconnected.

5.6 Synthesis of Findings

Taken together, the results reveal a Lagos that is rapidly transforming both physically and socially. Built-up expansion, population growth, and infrastructural development interact in complex ways, producing a city that is vibrant yet challenged by congestion, ecological strain, and service disparities. The peri-urban LGAs, in particular, emerge as zones of rapid change, absorbing demographic pressures and offering potential sites for planned interventions.

The application of machine learning methods in this study, including regression analysis, clustering, and ensemble modelling, provides a nuanced understanding of the drivers of urbanization. Population density, access to infrastructure, and socio-economic indicators consistently emerge as significant predictors of urban growth, reinforcing the notion that urbanization in Lagos is not only a product of physical expansion but also of the pursuit of better livelihoods, economic opportunities, and social amenities.

5.7 Implications for Policy and Planning

The findings carry important implications for urban planning and policy in Lagos. First, the spatial concentration of growth in peri-urban areas underscores the need for proactive planning, including the provision of essential services and transportation infrastructure before urban pressures reach crisis levels. Second, demographic trends indicate that population growth will continue to drive land use changes, necessitating integrated strategies that combine housing, transport, and environmental management. Finally, the uneven distribution of socio-economic benefits calls for policies that reduce spatial inequalities, ensuring that the benefits of urban growth are broadly shared across the city's population.

By understanding urbanization as a multidimensional process, policymakers can move beyond reactive approaches toward anticipatory and evidence-based planning, leveraging insights from data-driven models to guide sustainable development in Lagos.

REFERENCES

- Adelekan, I. O. (2016a). *Flood risk management in the coastal city of Lagos, Nigeria*. Journal of Urban Planning and Development, 142(3), 05015005. [https://doi.org/10.1061/\(ASCE\)UP.1943-5444.0000279](https://doi.org/10.1061/(ASCE)UP.1943-5444.0000279)
- Adelekan, I. O. (2016b). *Urban dynamics, everyday hazards and disaster risks in Lagos*. Environment and Urbanization, 28(1), 211–230. <https://doi.org/10.1177/0956247815622256>
- Adelekan, I. O., Olajide, O. A., & Bankole, A. (2020). *Urban growth, land use change, and sustainability challenges in Lagos, Nigeria*. Frontiers in Environmental Science, 8, 85. <https://doi.org/10.3389/fenvs.2020.00085>
- Adepoju, K. A., Arowolo, O. M., & Akinluyi, F. O. (2023). Land use dynamics and urban expansion in Lagos using remote sensing and GIS techniques. Journal of Urban Environmental Studies, 18(2), 45–62.
- Adeniji, K. (2021). Urban transport crisis and planning in Lagos, Nigeria. Journal of Urban Mobility Studies, 7(2), 88–104.
- Adesanya, A. (2020). Transport and urban mobility in Nigeria: Trends and challenges. African Transport Review, 5(1), 23–42.
- Akinmoladun, O. I. (2022). Informal transport, urban air pollution, and public health in Lagos. Journal of Environmental Health Research, 18(3), 120–134.
- Akinmoladun, O. I., & Adejumo, O. T. (2011). Urban agriculture in metropolitan Lagos: An inventory of potential land and water resources. Journal of Geography and Regional Planning, 4(1), 9–19.
- Akinola, A. O. (2020). Informality and urban service delivery in Lagos: The politics of prepaid electricity meters. Journal of African Political Economy, 47(1), 88–104.
- Aliyu, A. A., & Amadu, L. (2017). Urbanization, cities, and health: The challenges to Nigeria. Annals of African Medicine, 16(2), 45–49.
- Asenime, C. (2021). Current challenges for multi-modal mobility in Lagos. *INTALInC Nigeria Report*. Retrieved from <https://intalinc.org/wp-content/uploads/2021/03/intalinc-nigeria-report.pdf>
- Barter, P., & Obeng-Odoom, F. (2019). Informal transport in African cities: Case studies from Lagos and Accra. International Journal of Urban and Regional Research, 43(6), 1147–1164.
- Björkegren, D., Duhaut, A., Nagpal, G., & Tsivanidis, N. (2025). Public and private transit: Evidence from Lagos (NBER Working Paper No. 33899). National Bureau of Economic Research. <https://www.nber.org/papers/w33899>
- Blumenstock, J., Gillick, D., & Eagle, N. (2019). Predicting the locations of individuals based on social media and mobile phone data. Proceedings of the National Academy of Sciences, 116(6), 1183–1188.

- Chepsiror, S. J. (2016). *Factors influencing urbanization of informal settlements in cities: Case of Eldoret Municipality, Uasin Gishu County, Kenya* [Master's thesis, University of Nairobi]. University of Nairobi e-repository. <https://erepository.uonbi.ac.ke/handle/11295/100111>
- Dekolo, S., Ekum, M. I., James, O. K., Aigbavboa, C., & Gumbo, T. (2025). Safeguarding rural–urban linkages: Modeling drivers of peri-urban sprawl and impacts on ecosystem services. *Frontiers in Sustainable Cities*, 7, Article 1535619. <https://doi.org/10.3389/frsc.2025.1535619>
- Foster, V., & Anas, A. (2021). *Urban transport systems in Lagos: Formal–informal dynamics and welfare impacts* (NBER Working Paper No. 28332). National Bureau of Economic Research. <https://www.nber.org/papers/w28332>
- Gbadamosi, K. T., & Eniola, O. (2021). Walking and cycling in Nigerian cities: Challenges and opportunities. *Journal of Transport and Land Use*, 14(1), 553–567.
- Gilbert, K. M., & Shi, Y. S. (2023). Land use/land cover changes detection in Lagos City of Nigeria using remote sensing and GIS. *Advances in Remote Sensing*, 12, 145–165. <https://doi.org/10.4236/ars.2023.124008>
- Ifeoma, A., et al. (2023). A study of Lagos City, Nigeria. *International Journal of Contemporary Urban Affairs*. <https://ijcua.com/ijcua/article/download/352/531>
- International Energy Agency. (2019). *Africa Energy Outlook 2019*. IEA.
- Lagos Bureau of Statistics. (2022). *Abstract of Local Government Statistics*. Lagos State Government. Retrieved from <https://lagosmepb.org/wp-content/uploads/LGA-Statistics-ver-2020.pdf> Lagos MEPB
- Lagos State Ministry of Physical Planning and Urban Development. (2012). *Lagos State Development Plan (2012–2025)*. Lagos State Government.
- Lagos State Ministry of Physical Planning and Urban Development. (2019). *Annual Report on Land Use and Physical Development in Lagos State*. Lagos State Government.
- Lagos State Ministry of the Environment and Water Resources. (2024). *Lagos State Climate Adaptation and Resilience Plan (LCARP), Phase II*. Lagos State Government.
- Lagos Metropolitan Area Transport Authority. (2023a). *Lagos Rail Mass Transit*. Retrieved from https://en.wikipedia.org/wiki/Lagos_Rail_Mass_Transit
- Lagos Metropolitan Area Transport Authority. (2023b). *Transforming Lagos transport: BRT and rail development*. Lagos State Government.
- Moruf, O. (2012). Regulating street trading in Lagos, Nigeria: A historical perspective. *African Journal of Criminology and Justice Studies*, 6(1-2), 17–33.
- Mutisya, E., & Yarime, M. (2011). Understanding the grassroots dynamics of slums in Nairobi: The dilemma of Kibera informal settlements. *International Journal of Urban Sustainable Development*, 3(2), 187–203.
- Oduwaye, L., & Iweka, A. (2020). Transport infrastructure and sustainable urban development in Lagos. *Nigerian Journal of Sustainable Development*, 8(2), 55–71.

Olajide, O. (2016). Urban poverty and the food economy in Lagos, Nigeria. *Cities*, 59, 50-58.
<https://doi.org/10.1016/j.cities.2016.06.014>

Olajide, O., Agunbiade, M. E., & Bishi, H. (2018a). Land use dynamics and informal urban expansion in Lagos, Nigeria. *Frontiers in Environmental Science*, 6, 73.
<https://doi.org/10.3389/fenvs.2018.00073>

Olajide, O., Agunbiade, M. E., & Bishi, H. (2018b). The realities of Lagos urban development vision on livelihoods of the urban poor. *City, Culture and Society*, 12, 56-64.
<https://doi.org/10.1016/j.ccs.2017.09.005>

Olajide, O., Lawanson, T., & Aboderin, I. (2022). Ride-hailing services and urban mobility in Lagos: Implications for transport justice. *Transportation Research Part A: Policy and Practice*, 163, 1-12.

Olawole, M. O., & Olapade, D. O. (2019). Challenges of informal public transport in Nigerian cities: A case of Lagos metropolis. *Transportation Research Interdisciplinary Perspectives*, 2, 100118.

Oluwande, A., et al. (2019). Food security challenges in Lagos: Evidence from household surveys. *Nigerian Journal of Environmental Sciences*, 13(2), 77-94.

Osoba, S. B. (2015). Travel characteristics and commuting pattern of Lagos metropolis residents: An assessment. *Indonesian Journal of Geography*, 47(1), 40-53.
<https://doi.org/10.22146/ijg.6744>

Reback, J. (2005). Urban poverty and development in Brazil: The case of Curitiba. *World Development*, 33(10), 1829-1843.

Resnick, D., Sivasubramanian, B., Idiong, I. C., Ojo, M. A., & Tanko, L. (2022). The enabling environment for informal food traders in Nigeria's secondary cities. In L. Riley & J. Crush (Eds.), *Transforming urban food systems in secondary cities in Africa* (pp. 103-131). Palgrave Macmillan.

Solomon, B. A., Ogunleye, V. T., Ajayi, O. O., & Dare-Abel, A. (2025). Urban parks and user well-being in Lagos State: Evaluating adequacy in selected LGAs. *African Journal of Environmental Sciences and Renewable Energy*, 18(1), 1-16.
<https://doi.org/10.62154/ajesre.2025.018.010696>

Think Global Health. (2022). The economics of Lagos commuting. Council on Foreign Relations. Retrieved from <https://www.thinkglobalhealth.org>

United Nations, Department of Economic and Social Affairs, Population Division. (2019). *World urbanization prospects: The 2018 revision (ST/ESA/SER.A/420)*. United Nations.

United Nations Human Settlements Programme. (2003). *The challenge of slums: Global report on human settlements 2003*. Earthscan Publications Ltd.

United Nations Human Settlements Programme. (UN-Habitat). (2020). *World cities report 2020: The value of sustainable urbanization*. UN-Habitat.

United Nations Human Settlements Programme. (UN-Habitat). (2022). World cities report 2022: Envisaging the future of cities. UN-Habitat.

West African Urban Corridors-Lagos. (n.d.). UNILAG African Cluster Centre. Retrieved from https://iads.unilag.edu.ng/lacc/?page_id=5242

World Bank. (2013). Inclusion matters: The foundation for shared prosperity. World Bank.

World Bank. (2016). Lagos urban transport project (Final Report). World Bank.

World Bank. (2020). Lagos urban transport and infrastructure development assessment report. World Bank.

World Bank. (2022). Lagos: Urban resilience and development options [Project report]. World Bank Group.

Zhang, X., & Pan, J. (2021). Spatiotemporal pattern and driving factors of urban sprawl in China. *Land*, 10(11), 1275. <https://doi.org/10.3390/land10111275>

APPENDIX A

Python Libraries Used

The following are the following Python Libraries and Packages used:

```
import pandas as pd
import numpy as np
import geopandas as gpd
from pathlib import Path
from sklearn.model_selection import train_test_split, TimeSeriesSplit, KFold
from sklearn.metrics import mean_absolute_error, mean_squared_error, r2_score
from sklearn.preprocessing import StandardScaler, PolynomialFeatures
from sklearn.ensemble import RandomForestRegressor
from sklearn.linear_model import LinearRegression, Lasso, Ridge
from sklearn.neural_network import MLPRegressor
import statsmodels.api as sm
from statsmodels.tsa.api import VAR
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')
from xgboost import XGBRegressor
import shap
import contextily as ctx
```

APPENDIX B

Python Code

1. Data Loading

```
def load_all_data(data_folder_path):
    """Load all CSV and GeoJSON files from a folder"""
    data_folder = Path(data_folder_path)
    all_dataframes = {}

    # Load all CSV files
    csv_files = list(data_folder.glob("*.csv"))
    for file_path in csv_files:
        df_name = file_path.stem
        try:
            all_dataframes[df_name] = pd.read_csv(file_path, on_bad_lines='warn', encoding='latin-
1')
            print(f'Successfully loaded CSV: {df_name}')
        except Exception as e:
            print(f'Error loading {file_path}: {e}')

    # Load all GeoJSON files
    geojson_files = list(data_folder.glob("*.geojson"))
    for file_path in geojson_files:
        df_name = file_path.stem
        try:
            all_dataframes[df_name] = gpd.read_file(file_path)
            print(f'Successfully loaded GeoJSON: {df_name}')
        except Exception as e:
            print(f'Error loading {file_path}: {e}')

    return all_dataframes
```

2. Data Exploration and Integration

```
def explore_and_integrate_data(all_dataframes):
    """Explore and integrate all available data sources"""
    print("=== DATA EXPLORATION AND INTEGRATION ===")
```

```

# Create a comprehensive data dictionary
data_sources = {
    'Population': ['lagos_population', '2006_population', 'worldpop'],
    'Land Cover': ['1984-2013_spatial_change', '1984-2020_lulc_area', '2013-
2024_spatial_change',
        'built-up-areas', 'farmlands'],
    'Transportation': ['brt_buses_passengers', 'lagbus_buses_passengers',
'lagbus_fleet_passengers',
        'lagbus_operations', 'lagbus_route_info', 'lagbus_targets_actuals',
        'lagbus_vehicle_registrations', 'ferry_jetties_passengers',
        'ferry_routes_passengers', 'rail_cleaned', 'road_cleaned',
        'vehicles_by_country', 'vehicles_by_make', 'vehicles_by_ownership',
        'vehicles_by_year'],
    'Infrastructure': ['ambulance-emergency-services', 'churches', 'dump-sites',
        'electricity-sub-stations', 'environmental-sites', 'factoriesindustrial-sites',
        'filling-stations', 'fire-station', 'government-buildings',
        'health-care-facilities-primary-secondary-and-tertiary', 'health_facilities',
        'hospitals_2019', 'hospitals_2_2019', 'idp-sites', 'laboratories',
        'markets', 'mosques', 'pharmaceutical-facilities', 'police-stations',
        'post-offices', 'primary-schools', 'private-schools', 'public-schools',
        'public-water-points', 'public_jss', 'public_primary_school',
        'public_sss', 'religious-schools', 'secondary-schools', 'settlement-points',
        'small-settlement-areas', 'small-settlement-points', 'tertiary-schools',
        'water_facilities'],
    'Environmental': ['climate', 'emissions', 'energy', 'freshwater', 'greenhouse_gas_emissions',
        'sustainability'],
    'Socioeconomic': ['cpi_index', 'food_cpi_index', 'cpi_coicop_classification',
        'Health_Risk_factors', 'health_system'],
    'Landscape Metrics': ['2000-2010_class_landscape_metrics', '2000-2010_shannon_index',
        'landscape_metrics'],
    'Boundaries': ['local-government-administrative-boundaries', 'state-administrative-
boundaries',
        'operational-ward-boundaries']
}

# Check availability of each data source

```

```

available_data = {}
for category, sources in data_sources.items():
    available_sources = [s for s in sources if s in all_dataframes]
    if available_sources:
        available_data[category] = available_sources
        print(f"\n{category}:")
        for source in available_sources:
            df = all_dataframes[source]
            print(f" - {source}: {df.shape}, Columns: {list(df.columns)[:3]}...")

return available_data

```

3. Base DataFrame Creation with Enhanced Features

```

def create_enhanced_base_dataframe(all_dataframes, available_data):
    """Create a base dataframe with enhanced features"""
    print("\n=== CREATING ENHANCED BASE DATAFRAME ===")

    # Get LGA names from boundaries
    if 'local-government-administrative-boundaries' in all_dataframes:
        lga_df = all_dataframes['local-government-administrative-boundaries']
        lga_columns = [col for col in lga_df.columns if any(x in col.lower() for x in ['name', 'lga',
'admin'])]
        if lga_columns:
            lga_names = lga_df[lga_columns[0]].unique()
            print(f"Found {len(lga_names)} LGAs")

            # Calculate LGA areas
            lga_df = lga_df.copy()
            lga_df['area_sqkm'] = lga_df.geometry.area / 10**6 # Convert to sq km
            lga_areas = lga_df.set_index(lga_columns[0])['area_sqkm'].to_dict()
        else:
            lga_names = [f"LGA_{i}" for i in range(1, 21)]
            lga_areas = {lga: np.random.uniform(50, 200) for lga in lga_names}
            print("Using default LGA names and areas")
    else:

```

```

lga_names = [f'LGA_{i}' for i in range(1, 21)]
lga_areas = {lga: np.random.uniform(50, 200) for lga in lga_names}
print("Using default LGA names and areas")

# Get available years from all datasets
all_years = set()
for category, sources in available_data.items():
    for source in sources:
        df = all_dataframes[source]
        if 'year' in df.columns:
            all_years.update(df['year'].dropna().unique())
        elif 'Year' in df.columns:
            all_years.update(df['Year'].dropna().unique())

# Use 5-year intervals if we have many years
if len(all_years) > 10:
    years = sorted([y for y in all_years if 1980 <= y <= 2024])
    selected_years = years[::5] # Every 5 years
    if 2024 not in selected_years:
        selected_years.append(2024)
else:
    selected_years = sorted(all_years)

print(f'Selected years: {selected_years}')

# Create base dataframe
base_df = pd.DataFrame([(lga, year) for lga in lga_names for year in selected_years],
                        columns=['LGA', 'year'])

# Add LGA area
base_df['area_sqkm'] = base_df['LGA'].map(lga_areas)

print(f'Created base dataframe with {len(base_df)} rows')
return base_df

```

4. Land Cover Processing Function

```
def process_land_cover_data(base_df, all_dataframes):
    """Process land cover data and add to base dataframe"""
    print("\n=== PROCESSING LAND COVER DATA ===")

    # First, let's examine what land cover data we have
    land_cover_sources = [
        '1984-2013_spatial_change', '1984-2020_lulc_area', '2013-2024_spatial_change',
        'built-up-areas', 'farmlands'
    ]

    # Check which datasets are available and their structure
    available_land_cover = {}
    for source in land_cover_sources:
        if source in all_dataframes:
            df = all_dataframes[source]
            print(f'Examining {source}: shape {df.shape}, columns {list(df.columns)}')
            available_land_cover[source] = df

    # Try to extract built-up area data from different dataset structures
    buildup_data = []

    # Approach 1: Try to use the lulc_area dataset which has Year and Area by Land Use Type
    if '1984-2020_lulc_area' in available_land_cover:
        lulc_df = available_land_cover['1984-2020_lulc_area']
        print("Processing 1984-2020_lulc_area dataset")

        # Check if this dataset has the structure we need
        if all(col in lulc_df.columns for col in ['Year', 'Land Use Type', 'Area (ha)']):
            # Filter for built-up areas
            buildup_types = ['Built-up', 'Built up', 'Urban', 'Built-up Area', 'Artificial surfaces']
            buildup_df = lulc_df[lulc_df['Land Use Type'].isin(buildup_types)]

            if not buildup_df.empty:
                # Pivot to get area by year
```

```

builtup_by_year = builtup_df.groupby('Year')['Area (ha)'].sum().reset_index()
builtup_by_year.columns = ['year', 'BuiltUp_area']

# For this dataset, we don't have LGA-level data, so we'll assign to all LGAs
for lga in base_df['LGA'].unique():
    for _, row in builtup_by_year.iterrows():
        buildup_data.append({
            'LGA': lga,
            'year': row['year'],
            'BuiltUp_area': row['BuiltUp_area']
        })
print("Extracted built-up area data from 1984-2020_lulc_area")

# Approach 2: Try to use spatial change datasets
for change_source in ['1984-2013_spatial_change', '2013-2024_spatial_change']:
    if change_source in available_land_cover:
        change_df = available_land_cover[change_source]
        print(f"Processing {change_source} dataset")

# These datasets seem to have year columns with area values
year_columns = [col for col in change_df.columns if 'Area' in col and '(' in col and ')' in
col]

for col in year_columns:
    # Extract year from column name
    year_str = col.split(' ')[0]
    try:
        year = int(year_str)
        # Look for built-up land use class
        buildup_rows = change_df[change_df['LULC Class'].str.contains('Built-up|Built
up|Urban', case=False, na=False)]
        if not buildup_rows.empty:
            buildup_area = buildup_rows[col].values[0]

# Assign to all LGAs (since dataset doesn't have LGA breakdown)
for lga in base_df['LGA'].unique():

```

```

        buildup_data.append({
            'LGA': lga,
            'year': year,
            'BuiltUp_area': buildup_area
        })
        print(f"Extracted built-up area for {year} from {change_source}")
    except ValueError:
        continue

# If no proper land cover data found, create placeholder data
if not buildup_data:
    print("No proper land cover data found. Creating placeholder data.")
    for lga in base_df['LGA'].unique():
        for year in base_df['year'].unique():
            # Create synthetic data that increases over time
            base_year = 1984
            growth_rate = 0.05 # 5% annual growth
            buildup_area = 100 * (1 + growth_rate) ** (year - base_year)
            buildup_data.append({
                'LGA': lga,
                'year': year,
                'BuiltUp_area': buildup_area
            })

# Create DataFrame from buildup_data
land_cover_df = pd.DataFrame(buildup_data)

# Merge with base dataframe
base_df = pd.merge(base_df, land_cover_df, on=['LGA', 'year'], how='left')

return base_df

# 5. Enhanced Data Integration with Spatial Joins
def integrate_enhanced_data(base_df, all_dataframes, available_data):
    """Integrate enhanced data with spatial joins and feature engineering"""

```

```

print("\n=== INTEGRATING ENHANCED DATA ===")

# Get LGA boundaries for spatial joins
if 'local-government-administrative-boundaries' in all_dataframes:
    lga_gdf = all_dataframes['local-government-administrative-boundaries']
    lga_columns = [col for col in lga_gdf.columns if any(x in col.lower() for x in ['name', 'lga',
'admin'])]

    if lga_columns:
        lga_gdf = lga_gdf.set_index(lga_columns[0])

# Add infrastructure counts via spatial joins
infrastructure_sources = available_data.get('Infrastructure', [])
for source in infrastructure_sources:
    if source in all_dataframes and 'geometry' in all_dataframes[source].columns:
        infra_gdf = all_dataframes[source]

# Count points per LGA
if hasattr(lga_gdf, 'geometry'):
    try:
        # Spatial join to count points in each LGA
        joined = gpd.sjoin(infra_gdf, lga_gdf, how='inner', predicate='within')
        counts =
joined.groupby(lga_columns[0]).size().reset_index(name=f'{source}_count')
        counts.rename(columns={lga_columns[0]: 'LGA'}, inplace=True)

# Add to base dataframe (assuming counts are time-invariant for now)
base_df = pd.merge(base_df, counts, on='LGA', how='left')
print(f'Added {source} count data")
    except Exception as e:
        print(f'Could not spatially join {source}: {e}')

# 3. Add other data sources
# Population
pop_sources = available_data.get('Population', [])

```

```

for source in pop_sources:
    if source in all_dataframes:
        df = all_dataframes[source]
        if 'LGA' in df.columns and 'year' in df.columns and 'population' in df.columns:
            base_df = pd.merge(base_df, df[['LGA', 'year', 'population']], on=['LGA', 'year'],
how='left')
            print(f"Added population data from {source}")
            break
        elif 'lga' in df.columns and 'year' in df.columns and 'population' in df.columns:
            base_df = pd.merge(base_df, df[['lga', 'year', 'population']].rename(columns={'lga':
'LGA'}),
                                on=['LGA', 'year'], how='left')
            print(f"Added population data from {source}")
            break

# Transportation
transport_sources = available_data.get('Transportation', [])
for source in transport_sources:
    if source in all_dataframes:
        df = all_dataframes[source]
        # Find numeric columns that aren't identifiers
        numeric_cols = [col for col in df.columns if df[col].dtype in ['int64', 'float64']
                        and col not in ['LGA', 'lga', 'year', 'id']]

        if numeric_cols and ('LGA' in df.columns or 'lga' in df.columns) and 'year' in df.columns:
            id_col = 'LGA' if 'LGA' in df.columns else 'lga'
            cols_to_merge = [id_col, 'year'] + numeric_cols[:2] # Take first 2 numeric columns
            base_df = pd.merge(base_df, df[cols_to_merge].rename(columns={id_col: 'LGA'}),
                                on=['LGA', 'year'], how='left')
            print(f"Added transportation data from {source}")

# Environmental
env_sources = available_data.get('Environmental', [])
for source in env_sources:
    if source in all_dataframes:

```

```

df = all_dataframes[source]
if 'LGA' in df.columns and 'year' in df.columns:
    numeric_cols = [col for col in df.columns if df[col].dtype in ['int64', 'float64']
                     and col not in ['LGA', 'year', 'id']]
    if numeric_cols:
        cols_to_merge = ['LGA', 'year'] + numeric_cols[:2]
        base_df = pd.merge(base_df, df[cols_to_merge], on=['LGA', 'year'], how='left')
        print(f"Added environmental data from {source}")

# Socioeconomic
socio_sources = available_data.get('Socioeconomic', [])
for source in socio_sources:
    if source in all_dataframes:
        df = all_dataframes[source]
        if 'LGA' in df.columns and 'year' in df.columns:
            numeric_cols = [col for col in df.columns if df[col].dtype in ['int64', 'float64']
                             and col not in ['LGA', 'year', 'id']]
            if numeric_cols:
                cols_to_merge = ['LGA', 'year'] + numeric_cols[:2]
                base_df = pd.merge(base_df, df[cols_to_merge], on=['LGA', 'year'], how='left')
                print(f"Added socioeconomic data from {source}")

# Landscape Metrics
landscape_sources = available_data.get('Landscape Metrics', [])
for source in landscape_sources:
    if source in all_dataframes:
        df = all_dataframes[source]
        if 'LGA' in df.columns and 'year' in df.columns:
            numeric_cols = [col for col in df.columns if df[col].dtype in ['int64', 'float64']
                             and col not in ['LGA', 'year', 'id']]
            if numeric_cols:
                cols_to_merge = ['LGA', 'year'] + numeric_cols[:2]
                base_df = pd.merge(base_df, df[cols_to_merge], on=['LGA', 'year'], how='left')
                print(f"Added landscape metrics from {source}")

```

```
return base_df
```

6. Advanced Feature Engineering

```
def engineer_advanced_features(base_df):
```

```
    """Create advanced features including lagged variables, interactions, and ratios"""
```

```
    print("\n=== ENGINEERING ADVANCED FEATURES ===")
```

```
    # Sort by LGA and year
```

```
    base_df.sort_values(['LGA', 'year'], inplace=True)
```

```
    # 1. Create lagged features for key variables
```

```
    lag_periods = [1] # Use only 1 period lag for now
```

```
    for lag in lag_periods:
```

```
        if 'BuiltUp_area' in base_df.columns:
```

```
            base_df[f'BuiltUp_area_lag_{lag}'] = base_df.groupby('LGA')['BuiltUp_area'].shift(lag)
```

```
    # 2. Calculate growth rates if we have the required data
```

```
    if 'BuiltUp_area' in base_df.columns and 'BuiltUp_area_lag_1' in base_df.columns:
```

```
        # Handle division by zero
```

```
        base_df['BuiltUp_growth_pct'] = (
            (base_df['BuiltUp_area'] - base_df['BuiltUp_area_lag_1']) /
            base_df['BuiltUp_area_lag_1'].replace(0, np.nan) * 100
        )
```

```
    else:
```

```
        print("Warning: Could not create BuiltUp_growth_pct - missing required columns")
```

```
    # 3. Create density measures
```

```
    if 'area_sqkm' in base_df.columns and 'BuiltUp_area' in base_df.columns:
```

```
        base_df['builtup_density'] = base_df['BuiltUp_area'] / base_df['area_sqkm']
```

```
    # 4. Add population data if available
```

```
    if 'population' in base_df.columns:
```

```
        if 'area_sqkm' in base_df.columns:
```

```
            base_df['population_density'] = base_df['population'] / base_df['area_sqkm']
```

```

# Create lagged population
for lag in lag_periods:
    base_df[f'population_lag_{lag}'] = base_df.groupby('LGA')['population'].shift(lag)

# Calculate population growth
if 'population_lag_1' in base_df.columns:
    base_df['population_growth_pct'] = (
        (base_df['population'] - base_df['population_lag_1']) /
        base_df['population_lag_1'].replace(0, np.nan) * 100
    )

return base_df

```

7. Handle Missing Data

```

def handle_missing_data(base_df):
    """Handle missing values in the dataset while protecting target variables"""
    print("\n=== HANDLING MISSING DATA ===")

    # Check missing data percentage
    missing_pct = base_df.isnull().mean() * 100
    print("Missing data percentage by column:")
    for col, pct in missing_pct.items():
        if pct > 0:
            print(f" {col}: {pct:.1f}%")

    # Protect target variable and its dependencies from removal
    protected_cols = ['BuiltUp_area', 'BuiltUp_area_lag_1', 'BuiltUp_growth_pct',
'builtup_density']

    # Remove columns with too much missing data (except protected columns)
    high_missing_cols = [col for col in missing_pct[missing_pct > 50].index
        if col not in protected_cols]
    base_df.drop(columns=high_missing_cols, inplace=True)
    print(f"Removed columns with >50% missing data: {high_missing_cols}")

```

```

# For time-series data, use forward/backward fill within each LGA
numeric_cols = base_df.select_dtypes(include=[np.number]).columns.tolist()
for col in numeric_cols:
    base_df[col] = base_df.groupby('LGA')[col].transform(
        lambda x: x.ffill().bfill()
    )

# For remaining missing values, use median imputation
for col in numeric_cols:
    if base_df[col].isnull().any():
        base_df[col].fillna(base_df[col].median(), inplace=True)

print("Missing data handling complete")
return base_df

```

8. Prepare Data for ML

```

def prepare_ml_data(base_df, target_var='BuiltUp_growth_pct'):
    """Prepare data for machine learning"""
    print("\n=== PREPARING DATA FOR MACHINE LEARNING ===")

    # Check if target variable exists
    if target_var not in base_df.columns:
        print(f'ERROR: Target variable '{target_var}' not found in dataframe.')
        print("Available columns:", base_df.columns.tolist())
        return base_df, None, None, None

    # Remove rows where target is missing
    valid_indices = base_df[target_var].notna()
    base_df = base_df[valid_indices]

    if len(base_df) == 0:
        print(f'ERROR: No rows with valid target variable '{target_var}'.')
        return base_df, None, None, None

```

```

# Select features (all numeric columns except identifiers and target)
exclude_cols = ['LGA', 'year']
if target_var in base_df.columns:
    exclude_cols.append(target_var)

feature_cols = [col for col in base_df.columns
                 if col not in exclude_cols and base_df[col].dtype in ['float64', 'int64']]

X = base_df[feature_cols]
y = base_df[target_var]

print(f"Features: {len(feature_cols)}, Samples: {len(X)}")

return base_df, X, y, feature_cols

```

9. Multiple Modeling Approaches

```

def train_multiple_models(final_df, X, y, feature_cols):
    """Train and evaluate multiple modeling approaches"""
    print("\n=== TRAINING MULTIPLE MODELS ===")

    # Temporal split: train <= 2018, test > 2018
    train_idx = final_df["year"] <= 2018
    X_train, y_train = X[train_idx], y[train_idx]
    X_test, y_test = X[~train_idx], y[~train_idx]

    print(f"Training set: {X_train.shape} (years <= 2018)")
    print(f"Test set: {X_test.shape} (years > 2018)")

    # Scale features
    scaler = StandardScaler()
    X_train_scaled = scaler.fit_transform(X_train)
    X_test_scaled = scaler.transform(X_test)

    # Define models to try
    models = {

```

```

'XGBoost': XGBRegressor(
    n_estimators=600,
    learning_rate=0.05,
    max_depth=5,
    subsample=0.8,
    colsample_bytree=0.8,
    random_state=42
),
'Random Forest': RandomForestRegressor(n_estimators=100, random_state=42),
'Linear Regression': LinearRegression(),
'Lasso Regression': Lasso(alpha=0.1, random_state=42),
'Ridge Regression': Ridge(alpha=0.1, random_state=42),
'MLP': MLPRegressor(hidden_layer_sizes=(100, 50), random_state=42, max_iter=1000)
}

```

```

# Train and evaluate each model

```

```

results = {}

```

```

for name, model in models.items():

```

```

    model.fit(X_train_scaled, y_train)

```

```

    pred = model.predict(X_test_scaled)

```

```

    mae = mean_absolute_error(y_test, pred)

```

```

    rmse = np.sqrt(mean_squared_error(y_test, pred))

```

```

    r2 = r2_score(y_test, pred)

```

```

    results[name] = {'MAE': mae, 'RMSE': rmse, 'R2': r2}

```

```

print(f"\n{name}:")

```

```

print(f" MAE: {mae:.2f}")

```

```

print(f" RMSE: {rmse:.2f}")

```

```

print(f" R2: {r2:.4f}")

```

```

# Find best model

```

```

best_model_name = max(results.keys(), key=lambda x: results[x]['R2'])

```

```

best_model = models[best_model_name]

```

```

print(f"\nBest model: {best_model_name} (R²: {results[best_model_name]['R2']:.4f})")

# Cross-validation for best model
print(f"\nCross-validation for {best_model_name}:")
tscv = TimeSeriesSplit(n_splits=5)
cv_scores = []

for train_index, test_index in tscv.split(X):
    X_train_cv, X_test_cv = X.iloc[train_index], X.iloc[test_index]
    y_train_cv, y_test_cv = y.iloc[train_index], y.iloc[test_index]

    # Scale features
    X_train_cv_scaled = scaler.fit_transform(X_train_cv)
    X_test_cv_scaled = scaler.transform(X_test_cv)

    best_model.fit(X_train_cv_scaled, y_train_cv)
    pred_cv = best_model.predict(X_test_cv_scaled)
    cv_scores.append(r2_score(y_test_cv, pred_cv))

print(f"CV R² scores: {[f'{score:.4f}' for score in cv_scores]}")
print(f"Mean CV R²: {np.mean(cv_scores):.4f} (±{np.std(cv_scores):.4f})")

# Return the scaled test data for SHAP analysis
return best_model, scaler, results, best_model_name, X_test_scaled, y_test, train_idx

```

10. SHAP Analysis with Enhanced Visualizations - MODIFIED VERSION

```

def perform_shap_analysis(model, X_test_scaled, feature_cols, final_df, y_test, train_idx):
    """Perform SHAP analysis with enhanced visualizations"""
    print("\n=== SHAP ANALYSIS ===")

    try:
        # Create SHAP explainer
        explainer = shap.Explainer(model)
        shap_values = explainer(X_test_scaled)

```

```

# Summary plot
plt.figure(figsize=(12, 8))
shap.summary_plot(shap_values, X_test_scaled, feature_names=feature_cols, show=False)
plt.title('SHAP Feature Importance - Drivers of Urban Growth')
plt.tight_layout()
plt.savefig('./results/shap_summary.png', dpi=300, bbox_inches='tight')
plt.show()

# Beeswarm plot
plt.figure(figsize=(12, 8))
shap.plots.beeswarm(shap_values, show=False)
plt.title('SHAP Beeswarm Plot - Impact of Features on Urban Growth')
plt.tight_layout()
plt.savefig('./results/shap_beeswarm.png', dpi=300, bbox_inches='tight')
plt.show()

# Bar plot
plt.figure(figsize=(12, 8))
shap.plots.bar(shap_values, show=False)
plt.title('SHAP Bar Plot - Mean Absolute Impact of Features')
plt.tight_layout()
plt.savefig('./results/shap_bar.png', dpi=300, bbox_inches='tight')
plt.show()

# Create a dataframe with SHAP values for further analysis
shap_df = pd.DataFrame(shap_values.values, columns=[f'shap_{col}' for col in
feature_cols])
shap_df['LGA'] = final_df[~train_idx]['LGA'].values
shap_df['year'] = final_df[~train_idx]['year'].values
shap_df['actual'] = y_test.values
shap_df['predicted'] = model.predict(X_test_scaled)

# Save SHAP values
shap_df.to_csv('./results/shap_values.csv', index=False)

```

```
return shap_df
```

```
except Exception as e:
```

```
    print(f"SHAP analysis failed: {e}")
```

```
    print("This might be due to large dataset size or memory constraints")
```

```
    return None
```

11. Spatial Analysis and Visualization

```
def perform_spatial_analysis(final_df, all_dataframes, model, X, y, feature_cols, scaler):
```

```
    """Perform spatial analysis and create maps"""
```

```
    print("\n=== SPATIAL ANALYSIS ===")
```

```
    # Get LGA boundaries
```

```
    if 'local-government-administrative-boundaries' in all_dataframes:
```

```
        lga_gdf = all_dataframes['local-government-administrative-boundaries']
```

```
        lga_columns = [col for col in lga_gdf.columns if any(x in col.lower() for x in ['name', 'lga',  
'admin'])]
```

```
    if lga_columns:
```

```
        lga_gdf = lga_gdf.set_index(lga_columns[0])
```

```
    # Calculate predicted values for all LGAs and years
```

```
    final_df['predicted_growth'] = model.predict(scaler.transform(X))
```

```
    # Create a map of predicted growth for the most recent year
```

```
    recent_year = final_df['year'].max()
```

```
    recent_data = final_df[final_df['year'] == recent_year]
```

```
    # Merge with LGA boundaries
```

```
    map_data = lga_gdf.merge(recent_data[['LGA', 'predicted_growth']],  
                             left_index=True, right_on='LGA', how='left')
```

```
    # Create map
```

```
    fig, ax = plt.subplots(1, 1, figsize=(12, 10))
```

```
    map_data.plot(column='predicted_growth', cmap='RdYlGn', legend=True,
```

```

        ax=ax, edgecolor='black', linewidth=0.5)
    ax.set_title(f'Predicted Urban Growth by LGA ({recent_year})', fontsize=16)
    ax.set_axis_off()

    # Add basemap
    try:
        ctx.add_basemap(ax, crs=map_data.crs.to_string(),
source=ctx.providers.OpenStreetMap.Mapnik)
    except:
        pass
    plt.tight_layout()
    plt.savefig('./results/predicted_growth_map.png', dpi=300, bbox_inches='tight')
    plt.show()

# Calculate Moran's I for spatial autocorrelation (if we have spatial weights)
try:
    from esda.moran import Moran
    from libpysal.weights import Queen

    # Create spatial weights
    w = Queen.from_dataframe(lga_gdf)
    w.transform = 'r'

    # Calculate Moran's I for actual and predicted growth
    moran_actual = Moran(y, w)
    moran_predicted = Moran(final_df['predicted_growth'], w)

    print(f'Moran's I (Actual): {moran_actual.I:.3f} (p-value: {moran_actual.p_sim:.3f})')
    print(f'Moran's I (Predicted): {moran_predicted.I:.3f} (p-value:
{moran_predicted.p_sim:.3f})')

except ImportError:
    print("Spatial analysis libraries not available. Install esda and libpysal for spatial
autocorrelation analysis.")
except Exception as e:

```

```
print(f"Spatial autocorrelation analysis failed: {e}")
```

12. Scenario Testing

```
def perform_scenario_testing(model, scaler, X, final_df, feature_cols):
```

```
    """Test different scenarios for urban growth"""
```

```
    print("\n=== SCENARIO TESTING ===")
```

```
    # Create baseline scenario (current values)
```

```
    baseline_idx = final_df['year'] == final_df['year'].max()
```

```
    X_baseline = X[baseline_idx]
```

```
    baseline_pred = model.predict(scaler.transform(X_baseline))
```

```
    # Scenario 1: 10% population increase
```

```
    X_scenario1 = X_baseline.copy()
```

```
    if 'population' in feature_cols:
```

```
        pop_idx = feature_cols.index('population')
```

```
        X_scenario1[:, pop_idx] = X_scenario1[:, pop_idx] * 1.1
```

```
    scenario1_pred = model.predict(scaler.transform(X_scenario1))
```

```
    # Scenario 2: 20% increase in transportation infrastructure
```

```
    X_scenario2 = X_baseline.copy()
```

```
    transport_cols = [i for i, col in enumerate(feature_cols) if any(x in col for x in ['bus', 'vehicle',  
'transport'])]
```

```
    for idx in transport_cols:
```

```
        X_scenario2[:, idx] = X_scenario2[:, idx] * 1.2
```

```
    scenario2_pred = model.predict(scaler.transform(X_scenario2))
```

```
    # Compare scenarios
```

```
    scenario_results = pd.DataFrame({  
        'LGA': final_df[baseline_idx]['LGA'].values,  
        'Baseline': baseline_pred,  
        'Population_Increase': scenario1_pred,  
        'Transport_Improvement': scenario2_pred
```

```

}))

# Calculate differences
scenario_results['Population_Increase_Diff'] = scenario_results['Population_Increase'] -
scenario_results['Baseline']
scenario_results['Transport_Improvement_Diff'] = scenario_results['Transport_Improvement']
- scenario_results['Baseline']

print("Scenario testing results:")
print(f"Average baseline growth: {scenario_results['Baseline'].mean():.2f}%")
print(f"Average growth with population increase:
{scenario_results['Population_Increase'].mean():.2f}%")
print(f"Average growth with transport improvement:
{scenario_results['Transport_Improvement'].mean():.2f}%")

# Save scenario results
scenario_results.to_csv('./results/scenario_analysis.csv', index=False)

return scenario_results

```

13. Main Execution

```

def main():
    """Main function to run the complete analysis"""
    print("=== COMPREHENSIVE LAGOS URBANIZATION ANALYSIS ===\n")

    # Load all data
    data_folder = "./raw"
    all_dataframes = load_all_data(data_folder)

    # Explore and integrate data
    available_data = explore_and_integrate_data(all_dataframes)

    # Create base dataframe
    base_df = create_enhanced_base_dataframe(all_dataframes, available_data)

```

```

# Process land cover data
base_df = process_land_cover_data(base_df, all_dataframes)

# Add other data sources
base_df = integrate_enhanced_data(base_df, all_dataframes, available_data)

# Engineer advanced features
base_df = engineer_advanced_features(base_df)

# Handle missing data
base_df = handle_missing_data(base_df)

# Prepare for ML
final_df, X, y, feature_cols = prepare_ml_data(base_df)

# Check if we have data for modeling
if y is None or len(y) == 0:
    print("ERROR: No target variable available for modeling.")
    print("Final dataframe shape:", final_df.shape if final_df is not None else "None")
    if final_df is not None:
        print("Final dataframe columns:", final_df.columns.tolist())
    return
model, scaler, results, best_model_name, X_test_scaled, y_test, train_idx =
train_multiple_models(final_df, X, y, feature_cols)
shap_df = perform_shap_analysis(model, X_test_scaled, feature_cols, final_df, y_test,
train_idx)
perform_spatial_analysis(final_df, all_dataframes, model, X, y, feature_cols, scaler)

# Perform scenario testing
scenario_results = perform_scenario_testing(model, scaler, X.values, final_df, feature_cols)

print("\nAnalysis complete! Results saved to disk.")

if __name__ == "__main__":
    main()

```