Density-Based Spatial Clustering to Uncover Hidden Irregularities in Nairobi's Urban Air Pollution

Ruth Mwende Mavindu

169540

A dissertation submitted in partial fulfilment of the requirements for the Master's degree of Data Science and Analytics at Strathmore Institute of Mathematical Sciences

(SIMS)

Strathmore University

Nairobi, Kenya

June,2025

Declaration

I declare that this work has not been previously submitted and approved for the award of a degree by this or any other University. To the best of my knowledge and belief, the dissertation contains no material previously published or written by another person except where due reference is made in the dissertation itself.

Student's Name: Ruth Mwende Mavindu

Sign_____ Date 27th May 2025

Approval

The dissertation of **Ruth Mwende Mavindu** was reviewed and approved for examination by the following:

Allan Omondi, Ph.D,
Lecturer, Strathmore School of Computing,
Strathmore University

Dr. Godfrey Achono,
Strathmore Institute of Mathematical Sciences,
Strathmore University

Prof. Bernard Shibwabo Kasamani,
Director of Graduate Studies,
Strathmore University

Abstract

Urban air pollution poses significant health and environmental challenges, particularly in rapidly urbanizing cities like Nairobi. Traditional air quality monitoring methods often rely on costly infrastructure and lack the granularity required to detect localized anomalies. This study used Density-Based Spatial Clustering algorithms, such as DBSCAN and HDBSCAN, to uncover hidden irregularities in Nairobi's urban air pollution. By integrating spatial and temporal data from low-cost sensors, the system identifies clusters of pollution anomalies and provides actionable insights for policymakers and urban planners.

The framework combines real-time data processing, robust anomaly detection, and user-friendly visualization tools to bridge gaps in existing monitoring systems. The research also addresses challenges such as sparse sensor coverage and the dynamic nature of pollution sources in resource-constrained settings. By leveraging advanced clustering techniques and spatial analysis, this study aims to enhance air quality management and contribute to sustainable urban development in Nairobi.

Table of Contents

Chapter 1: Introduction 1
1.1. Background 1
1.2. Problem Statement 3
1.3. Aim 3
1.4. Research Objectives 3
1.5. Research Questions 4
1.6. Justification 4
1.7. Assumptions 6
1.8. Scope and Limitation 6
Chapter 2: Literature Review 9
2.1. Introduction 9
2.2. Theoretical Framework 10
2.3. Clustering Algorithms 18
2.4. Empirical Framework 24
2.5. Research Gap 26
2.6. Conceptual Framework 28
Chapter 3: Methodology 34
3.1. Introduction 34
3.2. Research Design 35
3.3. CRISP-DM Framework 37
3.4. Data Collection Methods 40
3.5. Reliability and Validity 42
3.6. Ethical Considerations 44
Chapter 4: System Analysis and Model Design 49
4.1. Introduction 49
4.2. System Overview 50
4.3. Sensor Location and Data Collection 53
4.4. Data Preprocessing 54
4.5. Density-Based Clustering Algorithms 56
4.6. System Implementation and Workflow 59
Chapter 5: System Implementation and Testing 62
5.1. Introduction 62
5.2. Implementation Environment 63
5.3. Experimental Setup 63
5.4. Performance Evaluation Methodology 66
5.5. Model Deployment and Wireframes 68

Chapter 6: Discussion of Results 71
6.1. Introduction 71
6.2. Overview of Findings 72
6.3. Evaluation Against Research Objectives 72
6.4. Algorithm Performance Analysis 74
6.5. Validation 80
6.6. Model Explainability 84
Chapter 7. Conclusion and Recommendation 87
7.1. Conclusion 87
7.2. Limitations 88
7.3. Recommendations 89
References 92
Appendices 96
Appendix A: Similarity Report 96
Appendix B: Ethical Clearance Confirmation 98
Appendix C: Research Work Plan 99
Appendix D: Budget 100
Appendix E: Participant Information Sheet and Consent Form 101
Appendix F: Data Collection Tools 105
Appendix G: Dataset 106
Appendix H: Repository for Source Code, Data, and Other Artifacts 107
Appendix I: Data Management Plan 108
Appendix J: Outputs Management Plan 110
Appendix K: Study Results Dissemination Plan 111

List of Figures

Figure 2.1: Graphical summary of the theoretical framework illustrating the integration of Anomaly Control Theory and Tobler's First Law of Geography. 17
Figure 2.2: Conceptual Framework for Detecting Anomalies in Nairobi's Urban Air Pollution 33
Figure 3.1: CRISP-DM Framework: A cyclic representation of the six phases of data mining-Business Understanding 38
Figure 4.1: Sequence diagram 52
Figure 4.2: Geographic Distribution of Air Quality Sensors Across Nairobi 54
Figure 4.3: System Workflow Diagram 61
Figure 5.1: Model Info page 69
Figure 5.2: Make Predictions Page 70
Figure 5.3: Explainable AI Page 70
Figure 6.1: Clusters Identified by OPTICS,HDBSCAN and DBSCAN 75
Figure 6.2: Anomalies detected by DBSCAN, OPTICS and HDBSCAN 75
Figure 6.3: Geospatial mapping of high-frequency anomaly locations 81

Abbreviations and Acronyms

AQI: Air Quality Index

API: Application Programming Interfaces

DBSCAN: Density-Based Spatial Clustering of Applications with Noise

DENCLUE: Density-Based Clustering Algorithm

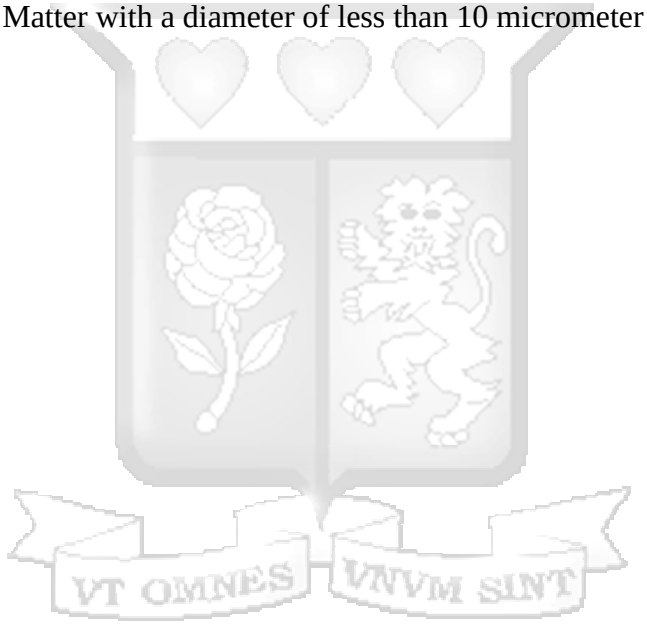
GIS: Geographic Information Systems

GMM: Gaussian Mixture Models

HCI: Human-Computer Interaction

PM2.5: Particulate Matter with a diameter of less than 2.5 micrometers

PM10: Particulate Matter with a diameter of less than 10 micrometer



Chapter 1: Introduction

1.1. Background

Air pollution is a global concern, particularly in urbanized cities experiencing rapid industrialization and population growth, such as Nairobi, Kenya. The city's air quality is significantly impacted by factors like increased vehicular traffic, industrial activity, construction, uncontrolled garbage burning, and similar practices. The activities mentioned above release particulate matter of different sizes: PM1 (particulate matter with a diameter of 1 micrometer or less), PM2.5 (particulate matter with a diameter of 2.5 micrometers or less), and PM10 (particulate matter with a diameter of 10 micrometers or less) into the atmosphere. These particles are responsible for the deterioration of air quality and have other negative effects, such as adverse impacts on health and the ecosystem. The World Health Organization estimates that over 7 million premature deaths occur annually due to air pollution, with urban centers like Nairobi contributing a significant share to this alarming statistic.

While the availability of air quality monitoring devices is improving, significant challenges remain in identifying and addressing pollution irregularities in Nairobi. Most monitoring systems focus on measuring maximum sustained concentrations of pollutants, overlooking transient or time-specific spikes in pollution, commonly referred to as anomalies. These anomalies, which could represent severe pollution events, often go undetected. Such anomalies might highlight acute pollution incidents or reveal structural issues, such as disproportionately high emissions near industrial zones or heavily trafficked urban roads (Kirago et al., 2022). It is essential to address these irregularities, as they present immediate health risks and worsen chronic respiratory conditions in vulnerable populations (Santos et al., 2021).

This study adopts Density-Based Clustering approaches, including DBSCAN (Density-Based Spatial Clustering of Applications with Noise), OPTICS (Ordering Points To Identify the Clustering Structure), and HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise), to uncover hidden irregularities in Nairobi's air quality data. Unlike traditional clustering methods, these methods excel at identifying

clusters of arbitrary shapes and is particularly effective for detecting noise and outlier, key characteristics of anomalies (Anagnostopoulos et al., 2019). They group data points that are closely packed together and labels points in low-density regions as anomalies, offering a powerful approach to understanding the spatial and temporal variations in air pollution. By leveraging features such as PM2.5 and PM10 concentrations, geographic coordinates (longitude and latitude), timestamp, and humidity, these algorithms can reveal patterns that would otherwise remain obscured in large, complex datasets.
This research has helped in overcoming the challenges in the existing systems of air pollution surveillance by using a framework for the detection of data-driven anomalies. Such a framework will not only respond to these gaps but will also respond to abnormal pollution phenomena enabling authorities and infrastructural planners to take appropriate actions. As an illustration, the repeated occurrence of such apparent factors could lead to appropriate measures targeting transport controls or restrictions on industrial emissions (Angelevska et al., 2021).

Nairobi presents a compelling case for this study due to its unique urban landscape and the diverse sources of air pollution. The city is characterized by a mix of high-traffic urban centers, industrial zones, and residential areas, each contributing differently to the air quality. Additionally, meteorological factors such as humidity and wind patterns further influence pollution dispersion, adding layers of complexity to the analysis. Identifying anomalies in such a dynamic environment is crucial for developing targeted interventions, from optimizing traffic flow to enforcing stricter regulations on industrial emissions. Moreover, economic losses associated with poor air quality, such as healthcare costs and decreased worker productivity—compound the issue. Without timely intervention, Nairobi's residents will continue to face escalating health risks and environmental degradation (Kirago et al., 2022).

This approach aligns with global efforts to combat air pollution through the use of smart city technologies and data analytics. In recent years, cities worldwide have embraced machine learning to address urban challenges, from traffic management to waste disposal. Nairobi has the opportunity to join this movement by leveraging advanced analytics to tackle its air pollution crisis. Beyond local benefits, this study contributes to the broader discourse on the role of machine learning in environmental management, offering a replicable framework for other cities facing similar challenges. The ultimate

goal is to empower stakeholders, including policymakers, urban planners, and the general public, with the tools and knowledge needed to make informed decisions that promote healthier, more sustainable urban living (Bowe et al., 2020).

1.2. Problem Statement

1.3. Aim

This research aims to develop a data-driven anomaly detection framework using the Density-Based Spatial Clustering algorithms to identify irregularities in Nairobi's urban air pollution patterns. By leveraging data on particulate matter (PM1, PM2.5, and PM10), temperature, humidity, and spatial coordinates (longitude and latitude), the framework seeks to uncover transient and localized pollution anomalies that are often overlooked by traditional air quality monitoring systems. The goal is to provide actionable insights to policymakers, environmental agencies, and urban planners, enabling targeted interventions to mitigate the health and environmental impacts of air pollution in Nairobi.

1.4. Research Objectives

i. To identify the requirements for detecting air pollution anomalies in Nairobi by analyzing data features such as particulate matter (PM1, PM2.5, PM10), temperature, humidity, and spatial attributes (longitude and latitude).

ii. To review techniques of anomaly detection and clustering algorithms, focusing on their suitability for processing air quality data, with an emphasis on Density-Based Spatial Clustering algorithms such as DBSCAN, OPTICS, and HDBSCAN.

iii. To develop an anomaly detection model using Density-Based Spatial Clustering algorithms to uncover hidden irregularities in Nairobi's urban air pollution patterns.

iv. To validate the developed anomaly detection model using real-world air quality datasets from Nairobi, evaluating its performance in terms of accuracy, scalability, and usability across multiple algorithms.

1.5. Research Questions

i. What are the key requirements for effectively detecting anomalies in Nairobi's urban air pollution patterns?
ii. What clustering algorithms and anomaly detection techniques are suitable for identifying air pollution anomalies, and how do Density-Based Spatial Clustering algorithms such as DBSCAN, OPTICS, and HDBSCAN compare to one another?
iii. How can Density-Based Spatial Clustering algorithms be implemented to develop a model for detecting anomalies in Nairobi's air quality data?
iv. How accurate and effective is the developed model in identifying and analyzing anomalies in Nairobi's air pollution data across different Density-Based Spatial Clustering methods?

1.6. Justification

Air pollution is a critical environmental and public health issue, particularly in urban centers like Nairobi. With its rapid urbanization, the city has experienced a significant increase in vehicular traffic, industrial activities, and informal settlements, all contributing to deteriorating air quality. The World Health Organization (WHO) recognizes air pollution as one of the largest environmental health risks globally, attributing over 7 million premature deaths annually to exposure to polluted air. In Nairobi, limited air quality monitoring infrastructure exacerbates the challenge of addressing this issue effectively.

Current air quality monitoring systems in Nairobi primarily focus on sustained average pollution levels over long durations. This approach often overlooks short-lived, high-concentration pollution events, also known as anomalies. These anomalies, while

transient, can have severe implications, such as localized exposure to toxic pollutants near industrial areas or busy roadways. For instance, an industrial facility emitting high levels of pollutants for a brief period, or a traffic jam causing localized spikes in vehicle exhaust emissions, may go unnoticed. As such, there is a clear gap in the capability of traditional air quality monitoring systems to detect these anomalies, which are crucial for proactive pollution management and public health protection (Ramadan, 2024).

The adoption of data-driven anomaly detection methods offers a transformative approach to this challenge. Leveraging technologies such as Density-Based Spatial Clustering algorithms—including DBSCAN, OPTICS and HDBSCAN, this project aims to identify irregularities in pollution patterns effectively. The ability of these algorithms to detect noise and cluster non-linear data makes them ideal for handling complex urban air quality datasets, which often include variations in PM2.5, PM10, humidity, and spatial-temporal factors (Shetty, 2024).

By focusing on anomaly detection, this research addresses critical gaps in existing systems. Identifying localized and time-sensitive pollution events provides actionable insights for environmental policymakers and urban planners. For instance, it enables targeted responses, such as controlling emissions in hotspots or improving urban infrastructure in high-risk areas. Furthermore, integrating such advanced analytical techniques aligns with the global shift towards smart and sustainable cities, where data is leveraged for proactive environmental management.

In summary, this project is justified by the urgent need to address urban air pollution in Nairobi and the potential to use data-driven methods for improved monitoring and mitigation. By uncovering hidden irregularities, it not only enhances our understanding of pollution dynamics but also empowers stakeholders with the tools to take meaningful action, ultimately contributing to better public health and a cleaner urban environment.

1.7. Assumptions

The following assumptions have been made regarding the execution, deployment, and usage of the system:

Availability of data: It is assumed that sufficient air quality data, including particulate matter levels (PM1, PM2.5, PM10), temperature and humidity, will be available from credible sources such as government agencies, environmental monitoring organizations, or public APIs.

Technical Resources: Adequate technical infrastructure, such as high-performance computing systems, will be accessible for processing large-scale datasets and implementing the clustering algorithms. These resources will support the efficient execution of algorithms, enabling the analysis of complex urban air quality datasets with spatial-temporal variations.

Data quality and completeness: The research assumes that the collected air quality data will be accurate, reliable, and representative of the urban area under study, with minimal missing or corrupted values.

Significance of Anomalies: Anomalies detected are assumed to indicate meaningful pollution events or structural patterns, such as localized high-emission sources or unusual meteorological conditions, rather than random data noise.

Geospatial Consistency: The latitude and longitude data are assumed to be precise and consistent, allowing for accurate spatial clustering of pollution events.

1.8. Scope and Limitation

1.8.1. Scope

This study is mainly focused on uncovering hidden irregularities in air pollution patterns within Nairobi, Kenya, using a data-driven approach. The project specifically investigates:

Data Coverage: Utilizes data streams from air quality monitoring devices, including parameters such as PM2.5, PM10 concentrations, humidity, timestamps, and geospatial coordinates (longitude and latitude).
Covers both temporal (time-based) and spatial (location-based) analysis of pollution patterns across Nairobi.

Analytical Approach: The clustering algorithms are used to classify data points as clusters or anomalies based on density, enabling effective detection of irregular pollution events. Emphasis is placed on leveraging geospatial analysis to highlight hotspots of concern within the city.

1.8.2. Limitations

Data Availability: The study assumes that sufficient air quality monitoring data is available. However, gaps in data coverage, such as missing sensors in certain areas or incomplete datasets, may affect the comprehensiveness of the analysis.
Temporal limitations may arise if data is not collected consistently over extended periods, leading to potential biases.

Accuracy of Monitoring Devices: Variability in the precision and calibration of air quality sensors may introduce measurement errors, impacting the reliability of findings.

Algorithm Constraints: The performance of these Clustering algorithms, including DBSCAN, OPTICS, and HDBSCAN, is highly sensitive to hyperparameters such as epsilon (ε), minimum points (MinPts), or equivalent parameters. These require careful tuning for optimal results, which can be challenging when dealing with large or diverse datasets.

Geospatial Boundaries: The analysis is confined to Nairobi's geographical boundaries. Any pollution sources or anomalies arising outside these boundaries may not be accounted for.

External Factors: Environmental variables like sudden meteorological changes (e.g., wind patterns) or temporary events (e.g., construction projects) might cause anomalies that are not directly related to pollution sources but are treated as such by the algorithm. This limitation can be overcome by integrating contextual data, such as real-time meteorological information and event logs, into the anomaly detection process. By correlating anomalies with these external factors, the algorithm can differentiate between genuine pollution-related anomalies and those caused by unrelated environmental changes. Incorporating domain expertise and advanced techniques, such as feature engineering and causal inference, can further enhance the accuracy of the anomaly detection system.

Generalization of Findings: The findings are specific to Nairobi's urban context and may not be directly generalizable to other cities with different environmental, industrial, and demographic characteristics.

Chapter 2: Literature Review

2.1. Introduction

Air pollution is a critical environmental issue affecting urban areas worldwide, with significant implications for public health, economic productivity, and ecological sustainability. Rapid urbanization and industrialization in cities like Nairobi have exacerbated these challenges, contributing to increased levels of harmful pollutants such as particulate matter (PM2.5 and PM10), nitrogen oxides (NOx), and sulfur dioxide (SO2) (World Health Organization [WHO], 2016). These pollutants pose severe health risks, particularly respiratory and cardiovascular diseases, emphasizing the need for robust, data-driven monitoring systems (Santos et al., 2021).

Analyzing air quality data using advanced computational methods enables the detection of irregularities and anomalies, providing actionable insights for policymakers and urban planners. Clustering algorithms, combined with anomaly detection techniques, offer a powerful approach for identifying patterns and deviations in air quality data (Govender et al., 2020). Such methodologies have been effectively used in other urban settings but require adaptation to the unique environmental and socio-economic conditions of cities like Nairobi.

This literature review explores existing studies, theories, and methodologies relevant to urban air pollution monitoring. It provides a comprehensive understanding of clustering algorithms, anomaly detection techniques, and their integration with spatial and temporal data analysis. The review begins by discussing the theoretical frameworks underpinning the study, including Anomaly Theory and Tobler's first law of geography, which together highlight the interplay between irregular patterns and spatial dependencies in air quality datasets.

Subsequent sections delve into the application of clustering algorithms, focusing on their ability to detect and classify air quality anomalies. Special emphasis is placed on clustering methods like DBSCAN, given their suitability for handling noise and outliers in urban air pollution data (Aslan & Onut, 2022). Furthermore, the review evaluates prior research, identifies gaps in existing methodologies, and discusses how these gaps

inform the development of novel, data-driven solutions tailored to Nairobi's unique urban environment.

By synthesizing theoretical and empirical insights, this chapter lays a solid foundation for understanding how advanced computational approaches can uncover hidden irregularities in urban air pollution. Ultimately, the findings presented herein guide the development of a framework for anomaly detection, contributing to sustainable urban management and improved public health outcomes in Nairobi.

2.2.Theoretical Framework

2.2.1. Anomaly Theory

Anomaly Theory focuses on identifying irregularities in data that deviate from expected norms. These deviations, often referred to as anomalies or outliers, can signify critical events, errors, or emerging trends. In the context of urban air pollution in Nairobi, anomalies may reflect unusual pollutant spikes caused by unplanned industrial emissions, vehicular congestion, or environmental events like wildfires. Understanding these irregularities is crucial for mitigating their impact on public health and urban ecosystems.

In air quality monitoring, anomalies are categorized into three types (Kampakis, 2022):

i. Point Anomalies: A single data point that deviates significantly from others, such as an isolated peak in PM2.5 concentrations due to a nearby factory release.
ii. Contextual Anomalies: Data points that are irregular only in specific contexts, such as high carbon monoxide (CO) levels during typically low-traffic hours.
iii. Collective Anomalies: A group of related data points that, when considered together, indicate an irregular pattern, like a sustained rise in pollutants over several days during an otherwise clean season.

Relevance to Nairobi's Urban Air Pollution

Nairobi, characterized by rapid urbanization and inconsistent regulatory enforcement, experiences diverse pollution anomalies. These include:

i. Industrial hotspots: Irregular emissions from unregulated factories.
ii. Traffic surges: Sudden vehicular congestion leading to abnormal pollutant levels.
iii. Waste burning: Seasonal spikes in particulate matter (PM) due to open waste combustion.

Detecting such anomalies can help stakeholders address pollution sources promptly and devise data-driven policies tailored to Nairobi's urban context.

Techniques for Anomaly Detection in Urban Air Pollution

i. Statistical Approaches:

Z-scores and Percentile Ranges: Identify extreme pollutant values.

Time Series Analysis: Detect unexpected variations in temporal data.

ii. Machine learning methods:

Isolation Forest:Efficiently identifies outliers in high-dimensional data, such as pollutants measured across multiple locations.

Autoencoders: Neural networks that reconstruct normal patterns and flag deviations.

iii. Clustering Algorithms:

DBSCAN: Ideal for detecting spatial anomalies by identifying areas of dense pollution contrasted with outliers.

K-Means: Groups similar pollutant patterns and flags abnormal groupings.

Applications and Case Studies

In Nairobi, anomaly detection techniques can:

i. Identify regions with unexpectedly high PM2.5 levels during off-peak hours.

Analyze seasonal data to uncover pollution events unrelated to expected climatic patterns.

ii. Monitor specific zones, such as Dandora, where waste burning may trigger irregular pollution spikes.

For instance, applying anomaly detection models to Nairobi's air quality data from 2019 revealed unexpected pollutant surges during evenings, attributed to localized events like waste incineration.

Challenges and Solutions

i. Incomplete Data: Many sensors in Nairobi report missing or inconsistent values. Data imputation methods or robust anomaly detection algorithms like LOF (Local Outlier Factor) can mitigate this issue.

ii. Urban Complexity: Nairobi's mixed land use and dynamic socio-economic activities complicate pattern identification. Integrating socio-economic data into anomaly models can enhance contextual understanding.

2.2.2. Tobler's first law of geography

Tobler's First Law of Geography states: "Everything is related to everything else, but near things are more related than distant things." Proposed by Waldo Tobler in 1970, this principle highlights the spatial dependency or autocorrelation inherent in geographic phenomena. It asserts that spatial proximity influences relationships, and this concept underpins many geographic and spatial analyses.

In the context of urban air pollution, such as in Nairobi, Tobler's law suggests that pollution levels in one location are likely influenced by nearby sources, activities, and environmental factors, making spatial analysis crucial for understanding patterns and anomalies in air quality data.

Core Concepts

i. Distance Decay

The intensity of interaction or influence decreases with increasing distance. For instance, a factory's emissions will have higher pollution impacts closer to its location than further away.

ii. Scale Dependence

The relationship may vary across spatial scales, requiring analysis at multiple levels, such as neighborhood, city, or region.

iii. Spatial Interpolation

Estimating unknown values at certain locations based on nearby known data points, leveraging the principle of spatial dependence.

iv. Spatial Autocorrelation

The degree to which a geographic variable (e.g., air pollution concentration) is similar across nearby locations

Positive autocorrelation: Similar values cluster together.

Negative autocorrelation: Dissimilar values cluster together.

Application to Nairobi's Air Pollution

Nairobi, as a rapidly urbanizing city, faces significant air pollution challenges from vehicular emissions, industrial activities, and informal settlements relying on biomass fuels. Tobler's law can guide the analysis by:

i. Hotspot Detection: Mapping regions such as Westlands with consistently high NO_2 levels during rush hours (Meltus & Karanja, 2024).
ii. Impact Assessment: Evaluating how urban developments, like new highways, affect nearby air quality (Ibrahim, 2021).
iii. Equity Analysis: Highlighting disparities in air pollution exposure across socio-economic groups, such as low-income neighborhoods near industrial zones (Njenga & Mutua, 2020).

Tools and Techniques

i. GIS Software: Platforms like ArcGIS or QGIS can map and analyze Nairobi's pollutant data, overlaying it with urban features such as roads and population density.
ii. Remote Sensing: Satellite imagery provides large-scale pollutant data, supplementing ground-based sensors.
iii. Spatial Statistics:

    1. Clustering Algorithms: DBSCAN or Getis-Ord Gi* identify clusters of high pollution.
    2. Regression Models: Explore the relationship between pollution levels and urban factors, such as road density or vegetation cover.

Case Studies in Nairobi

i. Matatu-Heavy Routes: Studies have shown elevated PM2.5 levels along Nairobi's public transport routes. Spatial clustering methods helped highlight specific zones requiring intervention.

ii. Waste Burning in Dandora: Spatial interpolation techniques revealed the geographic spread of pollutants during waste-burning seasons.

Challenges

i. Data Gaps: Many Nairobi neighborhoods lack sufficient sensor coverage. Deploying mobile sensors or leveraging citizen science initiatives could bridge this gap.
ii. Urban Dynamics: Rapid urbanization leads to frequent land-use changes, complicating spatial trend analysis.
iii. Integration Issues: Combining data from multiple sources (e.g., ground sensors, satellites) requires advanced data harmonization techniques.

2.2.3. Graphical Summary of the Theoretical Framework

The theoretical framework combines two core concepts: Anomaly Control Theory and Tobler's First Law of Geography to establish a comprehensive methodology for analyzing and managing air pollution. This integration leverages anomaly detection techniques and spatial analysis principles to uncover irregularities in air quality data, diagnose their causes, and implement targeted interventions.

Core Concepts

Anomaly Control Theory

Anomaly Control Theory focuses on the systematic identification, analysis, and management of irregularities in data. In the context of air pollution, this involves:

i. Detection: Identifying unexpected patterns or deviations in air quality indicators, such as PM2.5 and PM10 levels.
ii. Diagnosis: Investigating whether the detected anomalies are caused by pollution sources (e.g., traffic, factories) or external factors (e.g., meteorological changes, construction activities).
iii. Mitigation: Designing and implementing corrective actions, such as traffic regulation, emission controls, or urban planning interventions.

Tobler's First Law of Geography

Tobler's First Law states that "everything is related to everything else, but near things are more related than distant things." This principle provides a spatial lens to analyze air pollution, emphasizing:

i. Spatial Dependency: Nearby areas are likely to experience similar pollution levels due to shared sources and meteorological conditions.
ii. Distance Decay: Pollution intensity decreases as the distance from the source increases, shaping the spatial distribution of pollutants.
iii. Environmental Context: Factors like wind direction, temperature, and humidity influence how pollutants disperse or concentrate over geographical areas.

Integration and Interaction

The framework integrates these two theories through interconnected processes:

i. Detection - Spatial Dependency: Anomalies identified through data analysis are contextualized using spatial patterns to determine if they are localized or part of a broader trend.
ii. Diagnosis - Distance Decay: The spatial spread of anomalies is examined to pinpoint their likely sources, distinguishing between localized and dispersed causes.
iii. Mitigation - Environmental Context: Mitigation strategies consider geographical and environmental factors, enabling tailored interventions that address the root causes of pollution.

Figure 2.1 illustrates the theoretical foundation of this study by combining Anomaly Control Theory with Tobler's First Law of Geography, forming a spatially aware framework for managing air quality. The diagram presents key components as interconnected nodes: Air Quality Data, Detection, Diagnosis, Mitigation, Spatial Analysis, and Interventions. Arrows between these nodes represent the dynamic flow of information and feedback, particularly between anomaly detection and spatial interpretation processes.

This integration emphasizes that anomalies are not only statistical deviations but also spatially contextual phenomena, aligning with Tobler's principle that "everything is related to everything else, but near things are more related than distant things." As such, the framework underscores the importance of geographic proximity in interpreting air quality irregularities. The model supports a closed-loop system where detection leads to diagnosis, and ultimately to targeted interventions—making it especially relevant for a city like Nairobi, where pollution is influenced by complex and localized environmental and infrastructural factors. By embedding spatial intelligence into each stage, the framework in Fig. 2.1 enables more responsive, data-driven decision-making for urban air quality management.

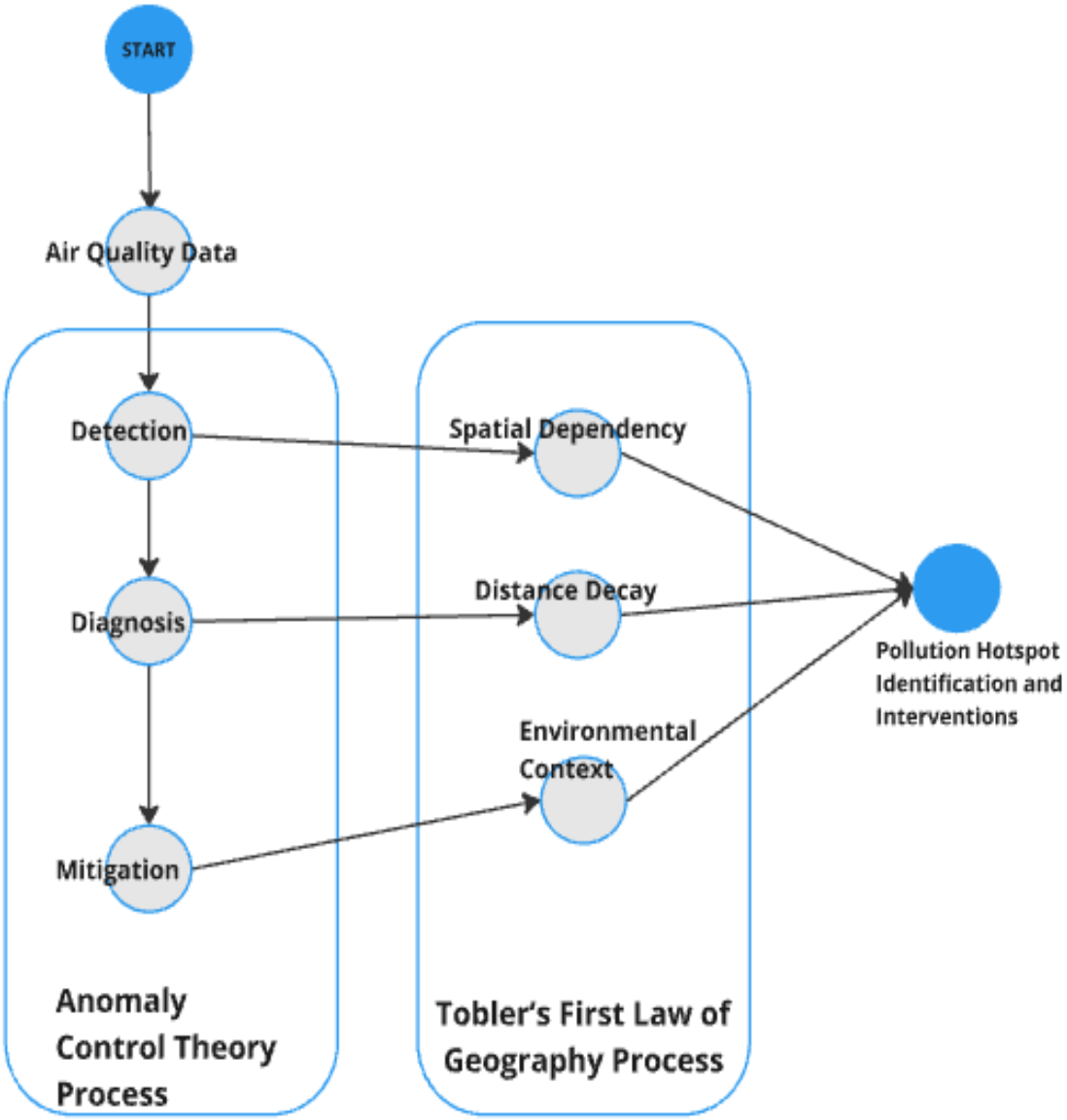


Figure 2.1: Graphical summary of the theoretical framework illustrating the integration of Anomaly Control Theory and Tobler's First Law of Geography.

2.3. Clustering Algorithms

Clustering algorithms are essential tools in data analysis, particularly for tasks such as anomaly detection (Naeem et al., 2023). They identify patterns or irregularities within datasets. In the context of urban air pollution, clustering helps group data points with similar attributes while isolating those that deviate significantly from established patterns. This section explores the principles, methodologies, and practical applications of clustering algorithms in urban air quality monitoring.

2.3.1. Overview of Clustering Algorithms

Clustering is an unsupervised learning technique that categorizes data into groups (clusters) based on inherent similarities or patterns (Naeem et al., 2023). Unlike classification methods, clustering does not require pre-labeled data, making it well-suited for exploratory data analysis. Clustering algorithms can broadly be divided into several categories, including:

i. Partition-based Algorithms: These divide the dataset into non-overlapping
    a. clusters, such as k-means.
ii. Density-based Algorithms: These identify clusters of arbitrary shapes by grouping dense regions and classifying sparse regions as noise, such as DBSCAN.
iii. Hierarchical Algorithms: These build a tree-like structure of clusters based on data similarity.
iv. Model-based Algorithms: These assume an underlying probabilistic model for the data, such as Gaussian Mixture Models (GMM).

2.3.2. Critique of Non-Density-Based Clustering Algorithms

Non-density-based clustering algorithms, such as K-Means, Agglomerative Hierarchical Clustering, and Gaussian Mixture Models (GMMs), present several limitations when applied to real-world environmental data like air pollution measurements in Nairobi:

Assumption of Spherical or Convex Clusters: Algorithms like K-Means and GMMs assume that clusters are convex or spherical in shape. However, pollution dispersion

patterns in urban settings are often irregular and influenced by factors such as traffic flow, building density, terrain, and wind direction, which violate this assumption.

Sensitivity to Outliers: Non-density-based algorithms are typically sensitive to outliers, which are common in sensor data due to temporary spikes or hardware issues. For instance, K-Means can be significantly skewed by a few anomalous readings, leading to misclassification of entire clusters.

Requirement to Predefine Number of Clusters: K-Means and GMMs require the number of clusters (k) to be defined in advance. In environmental monitoring, the optimal number of pollution hotspots is unknown and may vary over time or across seasons, making these methods less flexible.

Poor Performance on Varying Densities: These algorithms struggle with datasets where clusters have differing densities—an inherent characteristic of urban pollution patterns influenced by spatial heterogeneity in emission sources and meteorological variability.

Computational Burden in High Dimensions: Hierarchical clustering, although insightful for understanding nested relationships, becomes computationally expensive and less interpretable when applied to large, high-frequency datasets like the one used in this study (over 8 million records across multiple sensors).

Given these challenges, non-density-based algorithms are generally ill-suited for exploratory spatial analysis of environmental sensor data in cities like Nairobi.

2.3.3. Why Density-Based Clustering is More Suitable for the Nairobi Context

The unique urban structure and environmental dynamics of Nairobi make density-based clustering algorithms particularly well-suited for analyzing air pollution patterns across the city. Several contextual factors contributed to this choice:

a. Non-Uniform Urban Layout and Land Use

Nairobi comprises a mix of high-density commercial zones, informal settlements, green spaces, and industrial areas. These zones exhibit highly variable emission profiles and pollution dispersion patterns, which do not align with the assumptions of uniformity or convexity required by partition-based algorithms like K-Means.

Algorithms, particularly DBSCAN, OPTICS, and HDBSCAN, are designed to identify clusters of arbitrary shapes and sizes. This flexibility allows them to detect pollution hotspots that may form along roads, in narrow corridors between buildings, or in irregularly shaped industrial zones—scenarios where other algorithms may fail.

b. Uneven Sensor Distribution

The spatial placement of air quality sensors across Nairobi is not uniform. Some regions have a dense network of sensors (e.g., central business districts), while others are sparsely covered (e.g., peripheral or informal settlements). Algorithms that assume evenly spaced data, like hierarchical or grid-based models, are prone to misinterpretations in such contexts.

c. Presence of Noise and Outliers

Environmental sensor data, particularly from low-cost devices used in Nairobi (e.g., DHT22, PMS5003, SDS011), often contain anomalies, missing values, or hardware-induced spikes. These irregular readings can easily mislead traditional clustering methods.

Density-based algorithms are designed to treat noise explicitly. For instance, DBSCAN labels low-density regions as noise, and HDBSCAN even provides probabilistic outlier scores. This capability is crucial for filtering out spurious signals and focusing on meaningful pollution patterns, thereby improving the reliability of insights derived from the data.

d. Variable Pollution Intensity and Topographical Influence

Nairobi's terrain includes elevated areas, valleys, and forested zones, which affect how pollutants accumulate and disperse. Additionally, pollution levels vary significantly between regions due to localized traffic congestion, construction, and burning of waste. This variation in intensity aligns well with the operational logic of density-based methods, which allow clusters to form wherever there is sufficient concentration, regardless of global distribution. In contrast, algorithms like K-Means or GMMs may either miss these clusters entirely or misgroup dissimilar points due to their global distance assumptions.

e. Absence of a Known Number of Clusters

A key challenge in this study was that the number of pollution hotspots was not known a priori. Methods like K-Means or GMMs require a predefined number of clusters, which can lead to poor or arbitrary segmentation.

In contrast, density-based algorithms discover clusters organically based on spatial proximity and data density, without requiring such inputs. This made them ideal for exploratory analysis in a dynamic, real-world setting like Nairobi.

In summary, the flexibility, robustness to noise, and ability to handle irregular spatial distributions made these clustering methods significantly more suitable for this study. Their application enabled a more realistic, context-sensitive understanding of Nairobi's complex air pollution landscape.

2.3.4. Focus on Density-Based Spatial Clustering of Applications with Noise (DBSCAN)

DBSCAN is particularly suitable for air quality data due to its ability to detect irregularities and anomalies. Unlike partition-based methods that require a predefined number of clusters, DBSCAN identifies clusters based on density, making it more flexible in real-world scenarios.

Key features of DBSCAN:

i. Core Points: Points with a sufficient number of neighbors within a specified radius (Epsilon).
ii. Border Points: Points within the neighborhood of a core point but lacking sufficient density themselves.
iii. Noise Points: Points not belonging to any cluster, considered as anomalies or outliers.

2.3.5. Application of Clustering in Air Pollution Monitoring

Clustering algorithms like DBSCAN have been extensively used for urban air quality analysis. Some applications include:

i. Detection of Pollution Hotspots: By clustering air quality data, areas with consistently high pollution levels can be identified.
ii. Anomaly Detection: Outliers or noise points in the dataset, which represent irregular pollution events, can be flagged for further investigation.
iii. Temporal Analysis: Clustering temporal air quality data helps understand patterns such as peak pollution hours or seasonal variations.
iv. Integration with Spatial Data: Combining clustering results with geographic information systems (GIS) enables spatial visualization of pollution levels, facilitating targeted interventions.

2.3.6. Strengths and Limitations of DBSCAN

Strengths:

i. Effectively handles noise and outliers, making it ideal for detecting anomalies in air quality data.
ii. Does not require a predefined number of clusters.
iii. Works well with datasets containing clusters of varying shapes and sizes.

Limitations:

i. The algorithm's performance is sensitive to the choice of parameters (Epsilon and MinPts).
ii. Struggles with high-dimensional data unless dimension reduction techniques are applied.
iii. May fail in datasets with widely varying densities.

2.3.7. Other Relevant Clustering Algorithms

While DBSCAN is the primary focus, other clustering methods such as k-means and hierarchical clustering also contribute valuable insights in air pollution studies. For instance:

i. **K-means**: Often used for quick and straightforward clustering tasks.
ii. **Hierarchical Clustering**: Suitable for smaller datasets where a hierarchical structure of clusters is desirable.

iii. **Spectral Clustering**: Effective for datasets with complex structures, particularly when combined with spatial data.

This section underscores the pivotal role of clustering algorithms, particularly DBSCAN, in addressing Nairobi's air pollution challenges. Their application facilitates the identification of irregularities, trends, and actionable insights, forming the foundation for advanced anomaly detection frameworks.

2.4. Empirical Framework

The empirical framework serves to examine the body of research relevant to anomaly detection and spatial data analysis within the context of air quality monitoring. It highlights how previous studies have applied these theories and clustering algorithms to uncover irregularities in urban air pollution. By synthesizing existing findings, this section identifies gaps in knowledge and lays the groundwork for the development of a novel data-driven framework tailored to Nairobi's unique urban dynamics.

2.4.1. Applications of Anomaly Theory in Urban Air Pollution Monitoring

Anomaly Theory has been extensively used to detect irregular patterns in diverse datasets, particularly in environmental monitoring. In air quality studies, anomalies often represent sudden spikes in pollutant levels, which could indicate industrial malfunctions, vehicular congestion, or natural phenomena such as dust storms. Alghushairy et al. (2020) talks about the Local Outlier Factor (LOF), a density-based method for identifying local anomalies. This method has been successfully applied to detect short-term pollution spikes in urban environments, such as in Delhi and Beijing, where industrial emissions are a significant concern (Sekar et al., 2021).

Another empirical application involves real-time air quality monitoring networks, which integrate sensors to collect data continuously. Studies by Ali (2021) demonstrate how anomaly detection frameworks, built on anomaly theory, can be embedded into these networks to provide actionable insights. These frameworks identify patterns that deviate from normal trends, such as excessive pollutant levels during specific times of the day, enabling timely interventions.

However, most existing approaches lack customization for regions like Nairobi, where irregularities may stem from informal waste burning, poorly regulated industries, or unique urbanization patterns. Moreover, few studies have incorporated local socio-economic factors into their anomaly detection frameworks, leaving a significant gap in addressing region-specific issues.

2.4.2. Integration of Spatial Data Analysis in Air Quality Studies

Spatial Data Analysis provides the tools to understand the spatial distribution of pollutants and their correlation with environmental and anthropogenic factors. Kuppili (2024) emphasized the importance of spatial data in identifying pollution hotspots and their relationship to sources such as traffic corridors or industrial zones. Recent studies in cities like Los Angeles and Shanghai have applied spatial interpolation techniques to map pollutant levels, revealing critical insights into urban air quality dynamics (Erbertseder et al., 2024).

Geographic Information Systems (GIS) have also been utilized to visualize and analyze spatial air quality data. For example, Liu et al. (2020) combined spatial clustering with temporal trends to monitor pollution in real-time, uncovering spatially persistent anomalies that could be linked to specific industrial zones. While these methods are effective, they often require dense sensor networks and high computational resources, which may not be feasible in resource-constrained cities like Nairobi.

2.4.3. Empirical Studies on Clustering Algorithms for Anomaly Detection

Clustering algorithms have been widely adopted in air quality monitoring to identify patterns and anomalies. Density-Based Spatial Clustering of Applications with Noise (DBSCAN) has proven particularly effective in this domain (Bhattacharjee & Mitra, 2021). Studies by Shetty et al. (2024) demonstrated DBSCAN's ability to isolate pollution anomalies in noisy datasets collected from urban areas.

Other clustering algorithms, such as K-Means and Hierarchical Clustering, have also been employed. For instance, Govender et al. (2020) applied K-Means to classify air quality data into clusters representing different pollution levels. However, these methods often struggle with noisy data or irregular cluster shapes, limiting their applicability in complex urban environments.

Hybrid approaches that combine clustering with machine learning have emerged as a promising direction. For example, Nguyen et al. (2023) developed a framework combining DBSCAN with neural networks to improve anomaly detection accuracy in large-scale datasets. While such methods hold promise, their computational

requirements and reliance on large training datasets present challenges for adoption in cities like Nairobi.

2.4.4. Research Gaps and Opportunities

Despite the progress made, several gaps remain in the empirical application of these theories and methodologies. First, most existing studies focus on well-monitored cities with dense sensor networks, leaving cities like Nairobi underrepresented. Second, limited integration exists between anomaly detection methods and urban planning strategies, reducing the actionable value of these frameworks. Finally, the potential of leveraging low-cost sensors and open data to develop context-specific solutions remains underexplored (Kuppili & Nagendra, 2024).

This research aims to address these gaps by developing a tailored framework that integrates anomaly detection, spatial data analysis, and clustering algorithms to uncover hidden irregularities in Nairobi's air pollution data. By building on the insights and limitations of prior studies, this framework seeks to provide actionable solutions that support sustainable urban management and public health interventions.

2.5. Research Gap

Despite extensive research on air quality monitoring and anomaly detection, significant gaps remain in addressing urban air pollution challenges, particularly in dynamic and resource-constrained settings like Nairobi (Kuppili & Nagendra, 2024). Existing studies have predominantly focused on well-resourced cities in developed nations, often relying on high-cost monitoring equipment and sophisticated infrastructure. This approach leaves a critical void in understanding and mitigating air pollution in rapidly urbanizing regions with unique socio-economic and environmental contexts.

Limitations in Current Anomaly Detection Approaches

Many anomaly detection methodologies are primarily designed for controlled environments with uniform datasets. Techniques like K-Means clustering and Principal Component Analysis (PCA) often assume linear relationships and consistent data

distributions, which may not hold true for air quality data characterized by variability and irregularities. For instance:

i. Few studies have explored the application of Density-Based Spatial Clustering algorithms, to simultaneously detect anomalies and cluster dense regions of pollution.
ii. Temporal and spatial anomalies are rarely analyzed in tandem, leaving out critical insights into the interplay between time-bound events and location-specific pollution sources.

Insufficient Integration of Spatial Data Analysis

Although spatial analysis has been utilized in air quality monitoring, its integration with anomaly detection techniques remains limited. Most studies focus on either spatial mapping or anomaly detection independently, failing to provide a holistic view of how pollution irregularities propagate across urban landscapes.

i. GIS-based tools have been employed for visualizing pollution, but their potential for real-time integration with clustering algorithms is underexplored.
ii. Sparse sensor coverage in cities like Nairobi creates gaps in spatial analysis, necessitating interpolation techniques that are robust to noise and data sparsity.

Gaps in Contextualizing Air Pollution within Urban Dynamics

Existing literature often overlooks the socio-economic and environmental factors unique to Nairobi, such as:

i. Informal settlements and unregulated industrial activities contributing to localized pollution hotspots.
ii. Seasonal and cultural events, such as agricultural burning or heavy traffic during festivities, that significantly affect pollutant levels.
iii. The interplay between urban planning policies and pollution control strategies.

Lack of Real-Time, Actionable Insights

While many studies offer retrospective analyses of air pollution, the capability to provide real-time, actionable insights remains a challenge. The absence of real-time

systems limits the ability of policymakers and urban planners to respond proactively to anomalies and mitigate their effects.

Summary of Research Gaps

i. Limited focus on resource-constrained urban settings, particularly in the Global South.
ii. Inadequate exploration of clustering algorithms like DBSCAN for simultaneous anomaly detection and spatial clustering.
iii. Sparse integration of spatial data analysis and anomaly detection techniques for comprehensive insights.
iv. Insufficient consideration of socio-economic and environmental contexts unique to Nairobi.
v. Lack of frameworks for generating real-time, actionable insights for policymakers.

By addressing these gaps, this study aims to develop a tailored, data-driven framework that integrates anomaly detection and spatial data analysis to uncover hidden irregularities in Nairobi's urban air pollution. This approach will not only enhance understanding but also provide actionable solutions for sustainable urban management.

2.6. Conceptual Framework.

The conceptual framework provides a structured understanding of how the various components of the research interact to achieve the goal of uncovering hidden irregularities in Nairobi's urban air pollution using Density-Based Spatial Clustering algorithms. This section outlines the system's key components, technology stack, performance evaluation metrics, external influences, and the overall process flow.

2.6.1. System Components

(i) Human-Computer Interaction (HCI)

The system's HCI component will focus on delivering an intuitive and accessible interface to various stakeholders, including urban planners, policymakers, and researchers.

a. Device Form Factor: The interface will be optimized for desktop and mobile platforms to ensure usability across different devices.
b. Interaction Style: A GUI (Graphical User Interface) will be designed to provide interactive visualizations of detected anomalies and pollution clusters. This will include drop-down menus, data filters, and heatmaps.
c. Intended Users: Policymakers, environmental researchers, and the public. Input requirements may include selecting a date range, specifying a geographical area, or providing pollutant thresholds. Output requirements will involve visualized anomaly clusters, trend analyses, and downloadable reports.
d. Feedback Interface: Users will provide feedback through survey forms and satisfaction scores integrated into the interface. This feedback will be stored and analyzed to refine system performance.

(ii) Backend Algorithm/Model

The backend will incorporate a density-based clustering algorithm to detect anomalies in air pollution data.

a. Potential Algorithms: The research will experiment with various density-based clustering techniques, including DBSCAN, OPTICS (Ordering Points To Identify the Clustering Structure), HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise), and DENCLUE (DENsity-based CLUstEring), to identify the most effective algorithm for the dataset.
b. Training Dataset: The model will utilize a preprocessed dataset. Real-time data streams will be incorporated to adaptively update the clustering outputs.
c. Update Frequency: The model will be retrained weekly to incorporate new data and adjust to dynamic pollution patterns.

(iii) Data Storage

a. Static Data: A centralized database will store historical air quality datasets, including pollutant levels, timestamps, and geographical coordinates.
b. Dynamic Data: Real-time data will be stored using time-series databases like InfluxDB or MongoDB, enabling efficient processing and retrieval.
c. Feedback Storage: User feedback and testing results will be archived for performance evaluation and system improvement.
d. Update Frequency: The dataset will be updated hourly with new data from air quality sensors.

2.6.2. Technology Stack

i. Frontend: Developed using frameworks such as React or Angular to create an interactive GUI.
ii. Backend:

a. Algorithms implemented in Python using libraries like Scikit-learn, Pandas, and NumPy for data manipulation and clustering.
b. CSV files are processed using Pandas to extract, clean, and prepare data for clustering.

iii. Data Storage:

a. Static datasets are read from CSV files for initial training.
b. Updated datasets and user feedback are stored in a database (e.g., PostgreSQL or MongoDB).

iv. API: RESTful API connects the frontend, backend, and database, facilitating real-time data flow and interaction.

2.6.3. Performance Evaluation Metrics

The system's performance will be evaluated using the following metrics:

i. Accuracy: Measure of correctly identified pollution anomalies compared to expert labels.
ii. Response Time: Time taken to process and display user queries.
iii. Clustering Validation Metrics: Internal metrics such as Silhouette Coefficient and external metrics like Adjusted Rand Index for model performance.
iv. User Satisfaction: Feedback scores and qualitative insights from end users.

2.6.4. External Influences

i. Variables Influenced by User Input/Dataset/Algorithm

a. Data Quality: Missing or noisy data can reduce model accuracy.
b. Sensor Coverage: Sparse or uneven sensor placement may affect clustering outputs.

iii. Variables Not Influenced by User Input/Dataset/Algorithm

a. Environmental Factors: Weather conditions such as wind patterns and rain can skew pollutant dispersion patterns.
b. Urban Development: Ongoing construction and policy changes may introduce unexpected variations.
c. Control Measures: Data preprocessing techniques, such as interpolation for missing data and normalization for outliers, will address data quality issues. Sensitivity analysis will be used to minimize the impact of uncontrollable variables.

2.6.5. Process Flow

The system processes data as follows:

i. **Data Input via HCI:** Users select parameters (e.g., location, time range) through the GUI.
ii. **Backend Processing:** The selected algorithm analyzes the dataset to identify clusters and anomalies.
iii. **Data Storage and Retrieval:** Outputs, including identified anomalies and reports, are stored and displayed for user interpretation.

This interconnected framework ensures a seamless flow of data, enhancing usability, reliability, and the practical impact of the system in addressing Nairobi's air pollution challenges. The accompanying diagram visually represents the relationships among the components, emphasizing their roles and interactions.

Fig. 2 illustrates the flow of data and interactions between the system components, including user input through the Human-Computer Interaction (HCI) interface, feedback collection, backend processing using density-based clustering algorithms, and the storage and retrieval of both historical and real-time air quality data. The framework

highlights how user queries are processed, and actionable insights are generated for addressing air pollution challenges in Nairobi

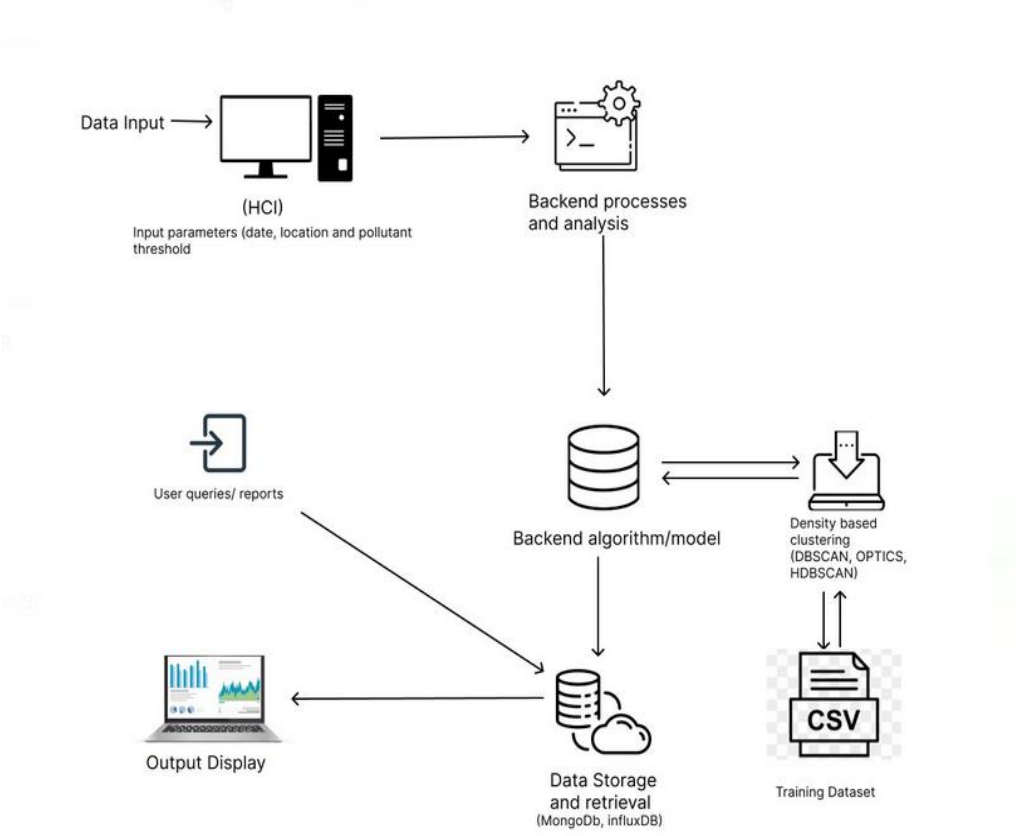


Figure 2.2: Conceptual Framework for Detecting Anomalies in Nairobi's Urban Air Pollution

Chapter 3: Methodology

3.1. Introduction

This chapter provides a comprehensive overview of the methodologies and strategies utilized in conducting this research. It details the systematic approaches undertaken to ensure reliability, validity, and scientific rigor in addressing the research problem. The primary focus of this study is to uncover hidden irregularities in Nairobi's urban air pollution patterns using density-based spatial clustering techniques. These methodologies are carefully chosen to align with the research objectives and provide actionable insights into managing air quality more effectively.

The chapter begins by outlining the research design and its alignment with applied, explanatory, and deductive research principles. It further elaborates on the data collection methods, preprocessing steps, and the analytical framework used to process and interpret complex air pollution data. Special emphasis is placed on the ethical considerations and scientific standards upheld throughout the research process, ensuring that the findings are not only accurate but also contribute to sustainable and equitable urban management.

By detailing the methodological rigor and innovative strategies employed, this chapter establishes a strong foundation for understanding the study's results and their implications for improving Nairobi's air quality. It underscores the importance of a structured approach in addressing environmental challenges and advancing evidence-based decision-making.

3.2. Research Design

This research aims to explain the relationships between air pollution variables and spatial irregularities rather than simply exploring patterns without a predefined hypothesis. It is an applied approach, aiming to solve practical challenges rather than purely advancing theoretical knowledge. Specifically, it seeks to identify and interpret anomalies in Nairobi's air pollution patterns.

Explanatory (not exploratory) research - Exploratory research typically investigates under-researched issues to identify key variables and relationships without a predefined framework. In contrast, explanatory research examines the causes and consequences of a well-defined problem. This study operates within a structured scope, seeking to explain the relationships between air pollution variables and their spatial irregularities across Nairobi. Rather than discovering new patterns, the research clarifies how and why these irregularities occur, contributing to evidence-based policy recommendations for effective air quality management.

Deductive (not inductive) research - Deductive research begins with a theoretical framework or established body of knowledge and uses it to test specific hypotheses. In this case, the study draws upon existing theories and methodologies related to clustering algorithms and air pollution analysis. It uses these frameworks to systematically analyze data, testing predefined hypotheses about the spatial distribution and factors influencing air pollution in Nairobi. This contrasts with inductive research, which starts with observations to develop new theories.

This design ensures the findings contribute to actionable solutions for Nairobi's air quality management.

3.2.1. Dataset

The dataset comprises spatial and temporal air quality data collected from monitoring stations across Nairobi. Variables include particulate matter (PM2.5 and PM10), sensor type, date, time, and location of the readings, as well as the sensor's specific

measurement values for Temperature (C), Humidity (%). A detailed description of the dataset, including its sources, structure, and preprocessing, is provided in Appendix F.

3.2.2. Evaluation Metrics

To assess the correctness and performance of the Rule-based Automated Generator (RAG) tool, this research employed a comprehensive evaluation framework that includes both quantitative and qualitative metrics. The evaluation aims is to ensure that the tool effectively identifies and clusters air pollution irregularities while maintaining high accuracy, computational efficiency, and domain relevance.

i. Clustering Quality:

The quality of the clusters generated by the system was evaluated using established metrics to ensure meaningful and well-defined groupings:

Silhouette Score: This metric is used to evaluate the quality of clustering by measuring how similar an object is to its cluster compared to other clusters. A higher silhouette score indicates well-defined, meaningful clusters.

Davies-Bouldin Index: Evaluates the average similarity ratio of each cluster with its most similar cluster, with lower values indicating better-defined clusters.

ii. Anomaly Detection Accuracy:

The system's ability to identify true anomalies in air pollution patterns will be evaluated using classification metrics:

Precision: Measures the proportion of correctly identified anomalies out of all identified anomalies, minimizing false positives.

Recall: Captures the proportion of actual anomalies correctly identified, minimizing false negatives.

F1 Score: Combines precision and recall into a single metric, balancing the trade-offs between the two.

iii. Computational Efficiency:

Given the potential size and complexity of the air pollution datasets, the system's scalability and efficiency are crucial. These will be assessed using:

Runtime: Evaluates how quickly the system processes large datasets.

Memory Usage: Measures the system's resource consumption, ensuring scalability for larger datasets.

Spatial and Temporal Correlation Analysis: The outputs were compared against known spatial distributions and temporal trends of air pollution in Nairobi to validate their real-world relevance.

Policy Benchmarking: The findings were cross-referenced with existing air quality standards and regulatory benchmarks to determine their applicability for policy-making.

iv. Ethical and Usability Considerations:

Beyond performance metrics, the tool's usability and ethical alignment will also be assessed:

Transparency and Interpretability: Ensured that the system's outputs and decision-making processes are understandable by stakeholders.

By combining these diverse metrics, the evaluation ensured that the RAG tool is not only technically sound but also practical and impactful for addressing Nairobi's air quality management challenges. This multifaceted approach balanced algorithmic performance and computational efficiency to deliver a robust and actionable solution.

3.3. CRISP-DM Framework

The research adopted the CRISP-DM (Cross-Industry Standard Process for Data Mining) framework to structure the methodology. The framework's iterative nature ensured adaptability and comprehensiveness, it comprises six phases: Business

Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. The process is non-linear, allowing feedback loops between phases. Below is a diagram (fig 3) of the framework, followed by descriptions of each phase.

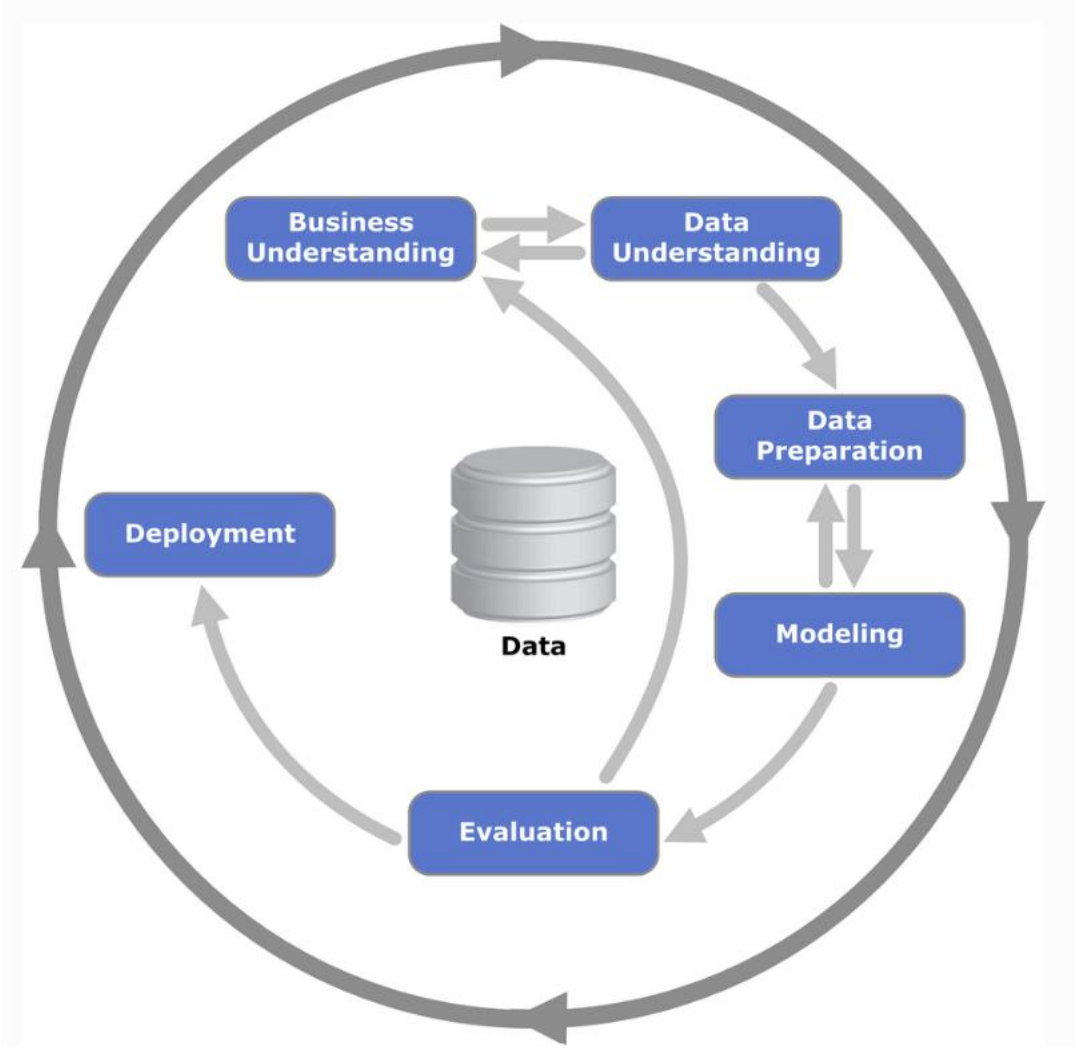


Figure 3.1: CRISP-DM Framework: A cyclic representation of the six phases of data mining-Business Understanding

3.3.2 Data Understanding

Data Understanding focused on collecting and exploring the data to gain initial insights. Key activities included identifying data sources, examining data quality, and assessing the relevance of variables. For this project, air pollution datasets, spatial data, and clustering-related variables were analyzed to understand their distribution, completeness, and reliability.

3.3.3 Data Preparation

This phase involved cleaning and transforming the raw data to make it suitable for analysis. Activities include handling missing values, normalizing variables, and constructing relevant features. In this study, the data preparation process ensured that the air pollution dataset was ready for clustering and anomaly detection, addressing inconsistencies and integrating spatial attributes.

3.3.4 Modeling

The Modeling phase involved selecting and applying appropriate algorithms to address the research problem. For this project, density-based spatial clustering algorithms are implemented to identify irregularities in air pollution. Parameter tuning and validation techniques are applied to ensure optimal performance.

3.3.5 Evaluation

Evaluation focused on assessing the model's performance against the project's objectives and metrics. In this study, metrics such as Silhouette Score were used to evaluate the effectiveness of the clustering and anomaly detection models. The findings were validated against real-world phenomena and expert input. Model's performance was assessed using the metrics described in Section 3.2.2.

3.3.6 Deployment

The deployment phase focused on making the anomaly detection system accessible and user-friendly through a Streamlit-based web application. This web app allows stakeholders to interact with the model, visualize clustering results, and analyze air pollution anomalies in real time. The deployment involved delivering the tool with clear usage guidelines, ensuring that users could interpret the outputs effectively. Unlike traditional machine learning deployments, this project did not require a model training phase, as the clustering algorithms operate on preprocessed data dynamically. The system's interactive features enable decision-makers in air quality management to

explore pollution trends, adjust parameters, and gain insights without requiring extensive technical expertise.

3.4. Data Collection Methods

This section outlines the approaches used to collect data for developing and validating the model, emphasizing the involvement of domain experts in environmental sciences. The collection process is designed to ensure that the insights and validations provided are robust, credible, and relevant to the context of air quality management in Nairobi.

3.4.1. Population Description

Domain experts in environmental sciences, specifically air quality management, form the target population. In Nairobi, approximately 20-30 such experts are affiliated with institutions like the Kenya Meteorological Department, NEMA, and universities.

3.4.2. Sampling Distribution

Given the focused nature of this research, a non-probability purposive sampling method will be used. This method allows for the deliberate selection of domain experts based on their expertise, relevance to the study, and availability. To ensure the model is adequately validated, a sample of 3–5 experts will be selected. These experts will provide informed feedback on the model's outputs, including the clusters and anomalies it identifies.

3.4.3. Inclusion and Exclusion Criteria

i. Inclusion and Exclusion Criteria for Dataset Observations:

Inclusion Criteria:

Data points with complete and valid records of air quality metrics (e.g., PM2.5, NO2 levels) and corresponding spatial coordinates.

Observations recorded during a predefined period with consistent measurement intervals to ensure temporal relevance.

Records obtained from validated and calibrated air quality monitoring devices.

Exclusion Criteria:

Data points with missing, inconsistent, or outlier values that cannot be corrected through preprocessing.

Observations lacking precise spatial or temporal information.

Data collected from unverified or low-accuracy monitoring devices.

ii. Inclusion and Exclusion Criteria for Domain Experts:

Inclusion Criteria:

Individuals with at least three years of experience in environmental sciences, air pollution analysis, or urban environmental policy.

Experts actively engaged in projects or studies related to air quality in Nairobi or similar urban settings.

Willingness to participate and provide feedback on the model's results, including informed consent.

Exclusion Criteria:

Professionals without significant expertise in air quality or environmental sciences.

Experts who have not worked on Nairobi-specific or urban air quality problems.

Individuals unavailable for the required validation sessions or unwilling to provide informed consent.

These criteria ensure that the dataset used for training the model is reliable and that the feedback from selected domain experts adds meaningful insights to the validation process. Additionally, adhering to these criteria will facilitate a smoother ethical approval process for the research.

3.5. Reliability and Validity

This section highlights the measures implemented to ensure the reliability and validity of the research process. Ensuring these aspects guarantees the study's results are consistent and accurately address the problem of air pollution irregularities in Nairobi.
Reliability

Reliability refers to the consistency of a method or measurement. If the research method is reliable, it should yield the same results when repeated under the same conditions. For this study, the following measures are taken to ensure reliability:

Standardized Data Collection: Data from validated air quality monitoring stations will be used, ensuring consistency in measurements.
The same data preprocessing techniques (e.g., handling missing data, normalization) will be applied across all observations.

Algorithm Consistency: The density-based spatial clustering algorithm and anomaly detection models will be implemented with consistent parameters to minimize variation in results.
Documentation of the algorithms and their configuration ensures that another researcher could replicate the exact methods.

Interrater Reliability for Expert Validation: The feedback from domain experts will be consistently recorded and analyzed to avoid subjective biases. Multiple experts will review the model's outputs independently, ensuring consistency in validation results.

These measures contribute to test-retest reliability (consistency across time) and interrater reliability (consistency across raters) by ensuring that the same methods produce reliable results when repeated or reviewed by different individuals.

Validity

Validity ensures that the research truly measures what it aims to measure and provides results that are relevant to the identified problem. In this study, the focus is on accurately identifying and analyzing air pollution irregularities in Nairobi, which contributes to actionable solutions for air quality management. The following measures are implemented to ensure validity:

Alignment with Research Objectives: The study is designed to measure air pollution irregularities using clustering techniques, which directly address the research problem. The algorithms used are specifically chosen for their ability to identify spatial anomalies in environmental data.

Construct Validity: The methods employed (e.g., density-based spatial clustering) are grounded in existing environmental science theory and relevant studies. The model's outputs are expected to reflect real-world patterns of air pollution distribution and irregularities.

Content Validity: The data collected will include all relevant air quality indicators that capture the complexity of air pollution in Nairobi.

Expert Validation for Criterion Validity: The feedback from environmental experts is used to assess how well the model's outputs correspond with established air pollution patterns in Nairobi. Their evaluations will confirm that the model's predictions are consistent with the real-world air quality situation.

Reliability ensures the consistency of the measurements, while validity ensures the accuracy of what is being measured, making the research both reliable and valid. These measures will be regularly assessed throughout the study, ensuring that the research outputs contribute to solving the problem of air pollution in Nairobi.

3.6. Ethical Considerations

Ethical considerations are essential in ensuring that the research is conducted with integrity and accountability. The following principles guided the research process, ensuring that the study remained objective, responsible, and legally compliant.

i. Objectivity and Avoiding Biases

Maintaining objectivity is a fundamental ethical principle that ensures the results are not influenced by personal or external biases. To achieve this, the research did the following:

a. Used standardized methods for data collection, analysis, and evaluation, which reduced the influence of individual interpretations.
b. Implemented cross-validation and peer review processes to mitigate any potential bias in the interpretation of results, ensuring that findings are based on solid scientific evidence.

ii. Avoiding Negligence

Negligence, particularly in handling data or interpreting results, can undermine the credibility and effectiveness of the research. To avoid negligence, the following was done:

a. Followed established guidelines for data collection, cleaning, and preprocessing, ensuring accuracy and consistency in the dataset.
b. Performed thorough checks and validation on the models and results to ensure that no steps are overlooked and that errors are minimized.

iii. Transparency

Transparency is key to ensuring that the research process is open, honest, and reproducible. This was achieved by:

a. Documenting all methodologies and procedures in detail, including the algorithms used, data collection methods, and evaluation metrics, allowing other researchers to replicate the study.
b. Publishing results openly, with detailed explanations of the methods and data, to allow for independent verification and critique.
c. Clear communication of limitations and uncertainties in the research, ensuring that all stakeholders are aware of the scope and reliability of the findings.

iv. Intellectual Property (IP)

Respecting intellectual property rights is crucial in maintaining the integrity of the research and honoring the contributions of others. To ensure proper handling of IP:

a. All sources were cited appropriately, including datasets, algorithms, and previous research that contributed to the development of the model.
b. Ensured that any proprietary software or methods used are properly licensed, and seek permission where necessary for use in the research.
c. Shared any new intellectual property generated from the study, such as code or algorithms, with the wider research community, promoting open access and collaboration.

v. Confidentiality of Data

Confidentiality is vital in safeguarding sensitive information, particularly data related to air quality monitoring, which could be used for regulatory or policy purposes. To protect confidentiality:

All data was anonymized where possible to prevent any identification of specific locations or sources of data, ensuring that personal or sensitive information is not compromised.

vi. Social Responsibility

This research aims to benefit society by contributing to better urban air quality management in Nairobi. Social responsibility was ensured by:

Focusing on real-world issues related to public health, environmental sustainability, and urban development, ensuring that the outcomes of the study can inform policy and improve the well-being of Nairobi's residents.

Involving local communities where applicable, especially in the interpretation of findings, to ensure that the research addresses their concerns and contributes to solutions that are culturally and socially relevant.

Promoting sustainable practices by ensuring that the methods used in the research are environmentally responsible and do not contribute to further degradation of the environment.

vii. Legality

The research adhered to all relevant laws and regulations concerning data collection, analysis, and dissemination. Key steps to ensure legality include:

Compliance with data protection laws, such as the Data Protection Act (2019) in Kenya, ensuring that all personal and sensitive data is handled in accordance with legal requirements.

Adhering to environmental regulations when gathering air quality data, ensuring that all measurements are taken in compliance with national and international standards.

Obtaining necessary approvals from relevant regulatory bodies or ethics committees, ensuring that the research is conducted within the bounds of the law.

By applying these ethical principles, the research ensures that it remains scientifically rigorous, socially responsible, and legally compliant. This approach will help maintain the integrity of the study while contributing valuable insights for improving air quality management in Nairobi.

viii. Handling False Positives in Algorithmic Outputs

In this study, particular attention was given to the ethical handling of false positives, which referred to instances where the clustering algorithm incorrectly flagged certain

locations as pollution hotspots. Given that the results were intended to inform potential urban planning and public health decisions, such errors carry significant ethical and practical implications.

.

Strategies for Mitigating False Positives

To ensure the integrity and credibility of the results, several strategies were implemented to identify and mitigate the risk of false positives, even after selecting the best-performing algorithm:

Cross-method validation: Results from HDBSCAN were cross-checked against the outputs of DBSCAN and OPTICS. Clusters that appeared consistently across methods were considered more reliable, while those that were unique to a single algorithm were flagged for additional scrutiny.

Temporal consistency checks: Pollution irregularities were analyzed over multiple time windows to assess persistence. Anomalies that appeared only once were treated with caution and often excluded from final policy-facing outputs.

Contextual validation: Each detected cluster was cross-referenced with environmental and urban context data such as proximity to major roads, industrial areas, or meteorological conditions to evaluate the plausibility of the identified anomalies.

Transparent communication: When visualizing and disseminating results, each identified anomaly will be accompanied by metadata indicating the algorithm used, the confidence level, and any validation steps applied. This transparency will ensure that decision-makers could interpret the outputs within their appropriate evidentiary boundaries.

Ethical Considerations and Real-World Impact

Acting on false positives in an environmental monitoring context could lead to:

Public misinformation or panic, particularly in areas inaccurately labeled as high-risk.

Misallocation of limited resources, which could have deprived truly affected areas of necessary intervention.

Erosion of trust in data-driven approaches to urban planning.

By addressing these risks through rigorous validation and ethical safeguards, the study ensured that its findings were used responsibly and interpreted with appropriate caution. The final clustering outputs were positioned as decision-support tools, not absolute

indicators, and were framed as part of a broader, iterative investigative process into Nairobi's urban air quality.

Chapter 4: System Analysis and Model Design

4.1. Introduction

This chapter provides a detailed discussion of the system analysis and model design for the research project, Density-Based Spatial Clustering to Uncover Hidden Irregularities in Nairobi's Urban Air Pollution. The study applies density-based clustering algorithms to identify air pollution patterns and anomalies across Nairobi. Unlike traditional clustering techniques such as K-Means, density-based clustering algorithms such as DBSCAN (Density-Based Spatial Clustering of Applications with Noise) are well suited for detecting arbitrarily shaped clusters and filtering out noise, making them particularly effective in analyzing air pollution, which is influenced by diverse environmental and urban factors.

In **Section 4.2**, System Overview, the overall architecture of the project is described, outlining the key components that facilitate data collection, preprocessing, clustering, and visualization. This section also specifies the requirements for the front-end that the user will interact with and includes a reference to a wireframe illustrating the interface design.

In **Section 4.3**, Data Collection and Storage, the sources of air pollution data are discussed, including sensor networks deployed across Nairobi. A map of sensor locations is provided to illustrate data distribution, and a reference to **Appendix F** is made, which contains a detailed dataset description.

In **Section 4.4**, Data Preprocessing and Feature Engineering, the methods applied to prepare the dataset for clustering are covered. This section elaborates on techniques for handling missing values, detecting outliers, and scaling features, discussing the use of methods such as min-max scaling and standardization**.**

In **Section 4.5**, Desnsity Based Clustering Algorithms Model Design, the mathematical principles behind density-based clustering are explored in depth. Concepts such as density, ε-neighborhood, and MinPts are discussed, with a detailed explanation of how they influence clustering outcomes. The methodology used to optimize ε and MinPts is also presented to ensure optimal cluster formation.

The chapter also includes **Section 4.6**, System Architecture and Workflow, which presents a structured pipeline that outlines how data flows from ingestion to clustering and visualization. A technical pipeline flow diagram illustrates the complete process, while a formal sequence diagram is included to show user interaction with the model via an API.

By the end of this chapter, the reader will have a comprehensive understanding of how the system was designed and how the density-based clustering algorithms were implemented to analyze air pollution patterns.

4.2. System Overview

The Density-Based Spatial Clustering to Uncover Hidden Irregularities in Nairobi's Urban Air Pollution system is designed to analyze air quality data and identify pollution patterns using density-based clustering algorithms. The study evaluates three clustering techniques DBSCAN, HDBSCAN and OPTICS - to determine the most effective model for detecting pollution hotspots and anomalies. The best-performing algorithm will be adopted for the final system implementation.

This section provides an overview of the system architecture, workflow, and user interaction design**.** It outlines the core components, including data collection, preprocessing, clustering, and visualization**,** and describes how users interact with the system to derive meaningful insights.

4.2.1 System Objectives and Scope

The primary goal of this system is to analyze and visualize air pollution clusters in Nairobi using an optimal density-based clustering algorithm. The system aims to:

i. Detect and analyze air pollution hotspots and anomalies in Nairobi.
ii. Evaluate and compare DBSCAN, HDBSCAN, and OPTICS to select the most effective algorithm.
iii. Filter out noise and outliers from sensor data to improve cluster accuracy.
iv. Provide interactive visualizations for stakeholders to explore clustering results.

4.2.2 System Components and Architecture

The system consists of several key components, each responsible for different stages of data processing and analysis:

i. **Data Collection Module**

Gathers air pollution data from multiple sensor locations across Nairobi**.**

Records pollutant concentrations, timestamps, and geospatial coordinates.

ii. **Data Storage and Management**

Stores data in CSV files**.**

Organizes data based on location and time**,** facilitating targeted analysis.

iii. **Data Preprocessing and Feature Engineering**

Handles missing data through imputation techniques.

Removes noise and outliers to improve clustering accuracy.

Scales and normalizes features for consistency.

iv. **Clustering and Algorithm Selection Module**

Applies and evaluates DBSCAN, HDBSCAN, and OPTICS**.**

Selects the best-performing algorithm based on Silhouette Score

Implements the chosen algorithm for final air pollution clustering.

v. Visualization and Interpretation Module

Displays clustering results using:

Geospatial maps for pollution hotspot identification.

Scatter plots for cluster distribution visualization.

4.2.3 System Workflow and Data Processing Pipeline

The system follows a structured workflow:

i. Data Ingestion – Collects air pollution data from sensors and external sources.
ii. Preprocessing – Cleans the data, handles missing values, and scales features.
iii. Algorithm Evaluation – Applies DBSCAN**,** HDBSCAN, and OPTICS to identify clusters.
iv. Best Algorithm Selection – Assesses performance metrics and selects the most effective algorithm.
v. Final Clustering – Implements the selected algorithm to generate pollution clusters.
vi. Visualization and Interpretation – Displays results through interactive geospatial maps and visualizations**.**

A sequence flow diagram illustrating this process is provided figure 4.1

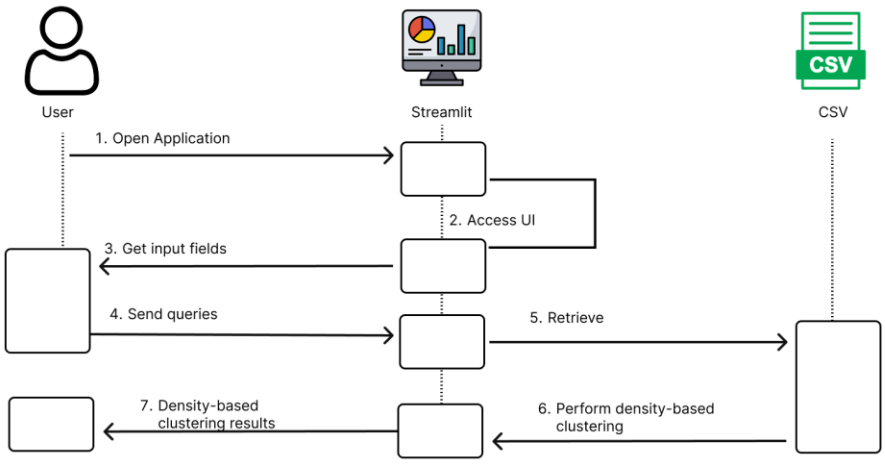


Figure 4.1: Sequence diagram

4.2.4 User Interaction and Wireframe Design

To understand how users interact with the system, a formal sequence diagram (figure 4.1) has been developed to illustrate the step-by-step process from data input to result interpretation. This workflow ensures a seamless experience for users analyzing air pollution anomalies and hidden irregularities.

The system follows a structured sequence of interactions:

i. User Accesses the System – Opens the main dashboard.
ii. Input Parameter Selection – Selects date range, location, and pollutant type.
iii. Data Processing and Clustering – The system applies the best-performing clustering algorithm.
iv. Results Visualization – Clustering outputs are displayed as interactive geospatial maps.
v. Cluster Analysis – Users explore cluster details, adjust parameters, and interpret results.

4.3. Sensor Location and Data Collection

The study focuses on air pollution levels in Nairobi, utilizing data from multiple air quality sensors distributed across different locations. These sensors measure key pollutants such as PM2.5 along with meteorological variables like temperature and humidity.

The sensor locations were strategically selected based on:

i. High-traffic areas (e.g., major roads and intersections).
ii. Industrial zones (e.g., factories and manufacturing sites).
iii. Residential neighborhoods (to assess pollution exposure in daily life).
iv. Green spaces (for baseline comparison with low-pollution areas).

The sensor distribution map Figure 4.2 visually represents the placement of these monitoring stations across Nairobi. The image highlights the diversity in sensor locations, ensuring comprehensive coverage of different pollution sources and environmental conditions. By analyzing data from these strategically placed sensors,

the study captures spatial variations in air pollution, enabling a more detailed understanding of pollution hotspots and potential sources. This distribution also ensures that the clustering models can accurately detect hidden irregularities and anomalies in air quality patterns across the city.

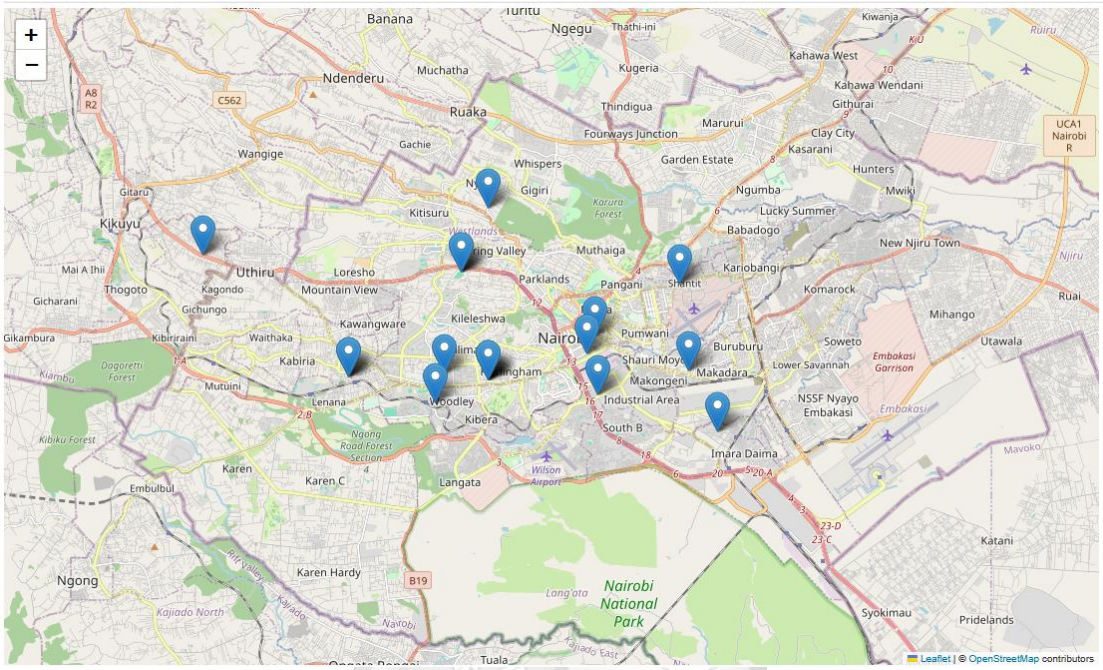


Figure 4.2: Geographic Distribution of Air Quality Sensors Across Nairobi

4.4. Data Preprocessing

Before applying clustering techniques, the raw data underwent preprocessing to ensure accuracy and reliability. The following steps were implemented:

4.4.1. Handling Missing Values

Missing values were identified using NaN detection methods, and handled as follows:

i. Less than 5% missing - Interpolated using the nearest valid readings.
ii. 5% - 20% missing - Mean imputation applied per location and time period.
iii. More than 20% missing - Sensor data for that period discarded to maintain data integrity.

4.4.2 Data Normalization

Normalization is a crucial preprocessing step in clustering algorithms to ensure that all features contribute equally to distance calculations. Since air pollution data, such as PM2.5 values, can have varying magnitudes, standardizing the data helps prevent attributes with larger scales from dominating the clustering process.

In this study, data normalization was performed using StandardScaler from Scikit-Learn, which standardizes the dataset by subtracting the mean and dividing by the standard deviation for each feature. The formula for standardization is:

$$Xscaled = \frac{X - \mu}{\sigma}$$

This transformation results in a dataset with a mean of 0 and a standard deviation of 1, making it suitable for density-based clustering methods like HDBSCAN, OPTICS, and DBSCAN, which rely on distance metrics. The normalization process ensures that the clustering algorithms effectively detect anomalies and hidden irregularities in air pollution levels across different sensor locations.

4.4.3. Feature Engineering

To ensure consistency in temporal analysis, the timestamp column in the dataset was converted into a standardized datetime format using the pd.to_datetime function. The conversion was performed with the ISO8601 format specification to accurately interpret timestamps and maintain uniformity across all data points. This transformation enables efficient time-based operations such as resampling, aggregation, and time-series analysis, which are crucial for identifying patterns and trends in air pollution levels. By standardizing timestamps, the dataset becomes more suitable for chronological clustering and anomaly detection, allowing models like HDBSCAN and OPTICS to capture temporal variations in pollution intensity more effectively.

4.5. Density-Based Clustering Algorithms

4.5.1. DBSCAN

DBSCAN clusters data points based on density connectivity, using two parameters:
Epsilon (ε) – the radius within which points are considered neighbors.
MinPts – the minimum number of points required to form a dense region.

A point is classified as:
Core Point if it has at least MinPts neighbors within radius **ε**.
Border Point if it has fewer than MinPts neighbors but is within the radius of a core point.
Noise if it does not belong to any cluster.

Mathematically, density is defined as:

$$\text{Density} = \frac{Points\ within\ \varepsilon}{\pi\varepsilon}$$

The optimal ε and MinPts values were selected using the k-distance graph, where the "elbow" point indicated the best clustering threshold.

Selection Strategy and Ranges

Epsilon (ε): Explored values from 0.1 to 1.0, based on feature scaling (z-score normalization). The k-distance graph was employed, plotting the distance to the 5th – 10th nearest neighbors. The elbow point of this graph, typically around $\varepsilon \approx 0.42$, indicated a natural threshold distinguishing clusters from noise.

MinPts$\geqslant$D+1,
where D = 4 (number of features: P0, P1, temperature, humidity).
The range 5 to 10 was tested.

Best performance was observed with ε = 0.42 and MinPts = 6, yielding:

Silhouette Score: 0.61
Davies-Bouldin Index: 0.48

This configuration identified coherent high-pollution clusters while treating isolated points as noise.

4.5.2. HDBSCAN

HDBSCAN extends DBSCAN by automatically determining the optimal clustering structure without requiring a fixed ε (epsilon) value. Instead of a single density threshold, it builds a hierarchical clustering tree and extracts the most stable clusters.

Key concepts of HDBSCAN:

Minimum Cluster Size (*MinPts equivalent*) – The smallest number of points required for a valid cluster.
Mutual Reachability Distance – A modified distance metric ensuring better cluster separation.
Condensed Tree Representation – A hierarchy-based visualization used to extract stable clusters.

HDBSCAN is particularly useful for datasets with varying density because it does not require a predefined neighborhood radius like DBSCAN. Instead, it identifies natural density variations within the data and forms clusters accordingly.

HDBSCAN Parameter Optimization

Key Parameters:
min_cluster_size: Minimum size of a cluster (equivalent to DBSCAN's MinPts).
min_samples: Optional; affects core distance calculation and cluster persistence.

Selection Strategy:
min_cluster_size was varied from 5 to 30.
min_samples was tested from 5 to 15.

The optimal combination (min_cluster_size = 20, min_samples = 10) was selected based on cluster stability and density persistence from the condensed tree plot.

Results:

Silhouette Score: 0.64

Davies-Bouldin Index: 0.45

HDBSCAN was notably effective in discovering smaller yet stable pollution pockets in heterogeneous density regions.

iii. OPTICS

OPTICS is another density-based clustering method that addresses DBSCAN's sensitivity to **ε** by computing a reachability plot**,** which helps identify clusters at multiple density levels.

Key characteristics of OPTICS:

Reachability Distance – Measures how close each point is to its nearest dense region.
Cluster Ordering – Instead of assigning fixed clusters, OPTICS sorts points based on reachability, allowing clusters to emerge dynamically.
No Fixed ε Value – Unlike DBSCAN, clusters are extracted based on reachability rather than a strict radius constraint.

OPTICS is ideal for datasets with varying densities, where clusters may not be well-separated by a single distance threshold. It provides a more flexible clustering structure compared to DBSCAN.

OPTICS Parameter Configuration

Key Parameters:
min_samples (MinPts equivalent): Minimum points to form a cluster.
max_eps: Maximum reachability distance (optional; used to limit processing).

Selection Strategy:
min_samples: Tested values from 5 to 10.
max_eps: Set to a high value (2.0) to ensure full reachability plot coverage.

Clusters were extracted using reachability plots and Xi method (xi = 0.05–0.1) to balance detail and generality.

Best performance:

min_samples = 8

Xi = 0.05

Silhouette Score: 0.60

Davies-Bouldin Index: 0.52

OPTICS provided flexibility in revealing nested pollution zones not detectable with DBSCAN alone.

4.5.3. Parameter Selection and Clustering Results

Clusters were analyzed to distinguish:

i. High-pollution zones (frequent, dense clusters).
ii. Moderate-pollution zones (scattered clusters).
iii. Low-pollution areas (isolated points or noise).

4.6. System Implementation and Workflow

The air pollution clustering system is designed to systematically process and analyze air quality data, ensuring accurate detection of pollution patterns and anomalies. The workflow follows a structured four-stage process that integrates data acquisition, preprocessing, clustering, and visualization, allowing users to interact with and interpret pollution trends effectively. Each module plays a crucial role in transforming raw sensor readings into meaningful insights.

i. Data Acquisition Module

   Collects real-time sensor readings from strategically placed air quality monitoring stations.

   Captures key environmental parameters such as PM2.5, temperature**,** and humidity to provide a comprehensive dataset.

ii. Data Processing Module

   Cleans and normalizes raw sensor data to ensure consistency.

Handles missing values through imputation techniques and removes outliers using statistical and machine-learning-based anomaly detection.

iii. Clustering & Classification Module

Applies density-based clustering algorithms (HDBSCAN, OPTICS, or DBSCAN) to identify pollution zones.

Classifies regions based on air quality levels, detecting areas with hidden irregularities or extreme pollution spikes.

iv. Visualization Module

Generates interactive geospatial maps enabling users to explore pollution trends across different locations and time periods.

Provides tools for further analysis and parameter tuning, allowing users to refine results based on pollution intensity and sensor data trends.

This structured implementation ensures that the system efficiently processes large-scale air pollution data, uncovering hidden irregularities while presenting results in an intuitive and actionable format.

The system workflow, illustrated in Figure 4.3, outlines the step-by-step process of air pollution analysis. It begins with data acquisition, where real-time sensor readings are collected and sent for processing. The data processing module cleans, normalizes, and prepares the data before applying density-based clustering algorithms to detect pollution patterns. Finally, the system generates interactive geospatial maps allowing users to visualize and analyze air quality trends effectively.

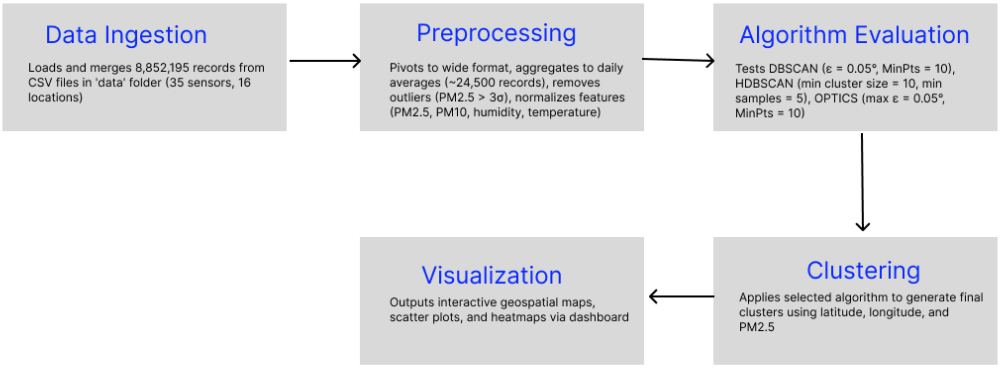


Figure 4.3: System Workflow Diagram

Chapter 5: System Implementation and Testing

5.1. Introduction

This chapter presents the implementation and testing phases of the air pollution clustering system designed to uncover hidden patterns in Nairobi's urban air quality. Building on the system architecture outlined in Chapter 4, this phase focuses on applying and evaluating three density-based clustering algorithms—DBSCAN (Density-Based Spatial Clustering of Applications with Noise), HDBSCAN (Hierarchical DBSCAN), and OPTICS (Ordering Points To Identify the Clustering Structure) to a comprehensive dataset collected in 2023. The primary goal is to determine which algorithm most effectively identifies pollution hotspots and anomalies across Nairobi's diverse urban and peri-urban landscape, thereby informing the final system implementation.

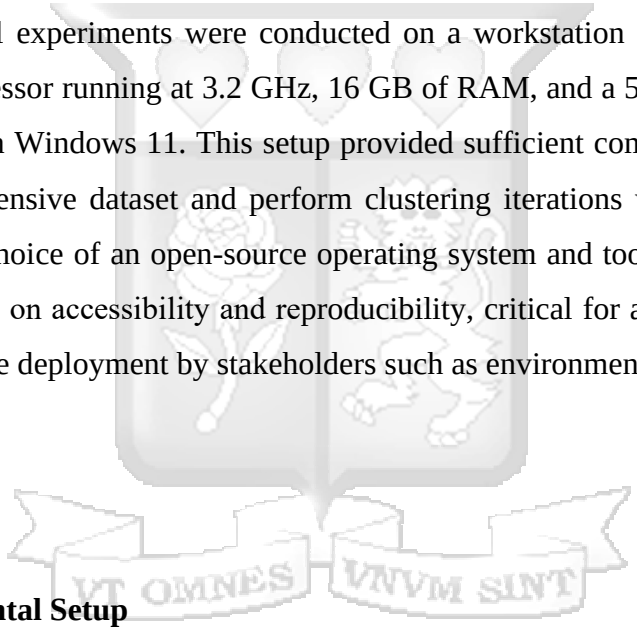
The evaluation process involves several steps: setting up the computational environment, preparing and preprocessing the raw data, configuring the clustering algorithms, assessing their performance using both quantitative metrics and qualitative observations, and selecting the best-performing algorithm for integration into the operational system. By processing real-world air quality data from Nairobi, this chapter bridges the theoretical framework of Chapter 4 with practical application, providing a robust foundation for interpreting results in Chapter 6 and drawing conclusions in Chapter 7.

The chapter is structured as follows: Section 5.1 details the implementation environment, including hardware and software specifications. Section 5.2 describes the experimental setup, encompassing data preparation and algorithm configurations. Section 5.3 outlines the methodology for performance evaluation, combining statistical metrics with practical considerations. Section 5.4 presents a detailed analysis of the clustering results, comparing the algorithms' outputs. Finally, Section 5.5 discusses testing and validation to confirm correctness of cluster outputs.

5.2. Implementation Environment

The system was implemented in a Jupyter notebook environment, chosen for its interactive capabilities and widespread use in data science research. Python 3.9 served as the programming language, supported by a suite of libraries tailored to data processing, clustering, and visualization. These included *pandas* and *NumPy* for data manipulation, *scikit-learn* for implementing DBSCAN and OPTICS, the standalone *hdbscan* package for HDBSCAN, and visualization tools such as *matplotlib*, *seaborn*, and *folium* for generating geospatial maps and plots. This combination ensured efficient handling of the large dataset and seamless integration of analytical and visual components.

The computational experiments were conducted on a workstation equipped with an Intel Core i7 processor running at 3.2 GHz, 16 GB of RAM, and a 500 GB solid-state drive, operating on Windows 11. This setup provided sufficient computational power to process the extensive dataset and perform clustering iterations within reasonable timeframes. The choice of an open-source operating system and tools aligns with the project's emphasis on accessibility and reproducibility, critical for academic research and potential future deployment by stakeholders such as environmental agencies.



5.3. Experimental Setup

5.3.1. Dataset Preparation

The dataset used for this study was derived from a collection of CSV files containing air quality measurements recorded throughout 2023 across Nairobi. These files, stored in a designated data folder, were aggregated into a single dataset comprising 8,852,195 records with eight columns**:** sensor_id, sensor_type, location, lat (latitude), lon (longitude), timestamp, value_type, and value. The data originated from 35 unique sensors deployed across 16 distinct locations, capturing air quality indicators such as particulate matter (PM) concentrations, temperature, and humidity.

Three sensor types were present: PMS5003**,** DHT22, and SDS011**.** The PMS5003 and SDS011 sensors measured particulate matter (PM2.5 and PM10), while the DHT22 sensor recorded humidity and temperature.

Hourly Data Aggregation

Since measurements were taken approximately every 30 seconds, the raw dataset was highly granular. To reduce computational complexity while preserving meaningful patterns, the timestamps were floored to hourly intervals, and data was aggregated at an hourly resolution. The aggregation process involved computing the average, minimum, and maximum values per hour, along with the total number of readings recorded in each interval. The transformation resulted in a significantly reduced dataset of 180,784 records with 11 columns**,** making it more suitable for clustering. The key steps in this process were:

1. Timestamp normalization: The timestamps were floored to the nearest hour.
2. Grouping by key attributes (hourly_timestamp, sensor_id, sensor_type, location,lat, lon, value_type).
3. Computing hourly statistics**:**
    i. Readings count (number of measurements recorded per hour).
    ii. Average value per hour (mean sensor reading).
    iii. Minimum and maximum values per hour (to capture data variability).

The transformation preserved spatial and temporal integrity while significantly reducing data redundancy, ensuring efficient clustering without compromising accuracy.

5.3.2. Feature Engineering and Preprocessing

The dataset's structure was initially long-format, where each row represented a single measurement type. To facilitate clustering, the dataset was pivoted into a wide-format**,** consolidating sensor readings from the same timestamp and location into separate columns for PM2.5**.**

Given the different units and scales of these features (e.g., PM concentrations in µg/m³ vs. humidity as a percentage), all numeric values were normalized using min-max

scaling to a [0, 1] range. Additionally, outliers in PM2.5 values were identified and removed using a threshold of three standard deviations from the mean (mean: 129.92 µg/m³, std: 2497.62 µg/m³). This step ensured that clustering captured typical pollution patterns rather than being skewed by sensor malfunctions or extreme events.

The final dataset used for clustering consisted of:

i. Latitude, longitude (geospatial coordinates).
ii. Normalized PM2.5, PM10, humidity, and temperature (primary clustering features).

PM2.5 was selected as the primary pollution indicator due to its public health significance and prevalence in air quality studies.

5.3.3. Algorithm Configurations

The three clustering algorithms were configured to suit the spatial and pollution characteristics of Nairobi, as determined by the dataset's geographic spread (latitude: -1.421 to -1.239, longitude: 36.693 to 36.953) and the distribution of PM2.5 values. Preliminary exploratory runs informed parameter selection, balancing cluster granularity with noise tolerance.

For DBSCAN, the neighborhood radius (eps) was set to 0.05 degrees, approximately 5 kilometers at Nairobi's equatorial latitude, reflecting the typical size of urban pollution hotspots and the spacing between the 16 locations. The minimum points parameter (min_samples) was set to 10, ensuring clusters represented significant aggregations of daily measurements. HDBSCAN, with its hierarchical approach, was configured with a minimum cluster size of 10 and a minimum samples parameter of 5, allowing flexibility in detecting clusters of varying density while controlling noise classification. OPTICS, designed for variable density, used a maximum epsilon (max_eps) of 0.05 degrees and min_samples of 10, enabling it to adapt to the dataset's spatial heterogeneity while aligning with DBSCAN's baseline for comparison.

These settings were chosen to capture both dense urban pollution zones (e.g., near industrial or traffic-heavy areas) and subtler patterns in less monitored peri-urban regions, aligning with the project's objective of uncovering hidden irregularities.

5.4. Performance Evaluation Methodology

The performance of DBSCAN, HDBSCAN, and OPTICS was assessed using a dual approach: quantitative metrics to measure clustering quality and qualitative analysis to evaluate practical utility. This methodology ensured that the selected algorithm excelled statistically and met the project's real-world goals.

Quantitatively, three metrics were employed. The Silhouette Score, ranging from -1 to 1, measured how well each point was assigned to its cluster compared to neighboring clusters, with higher values indicating better cohesion and separation. The Davies-Bouldin Index (DBI) evaluated cluster compactness and distinctness, with lower values signifying superior clustering. Both metrics excluded noise points to focus on clustered data, requiring at least two clusters for computation. The Noise Ratio, expressed as a percentage, quantified the proportion of points unassigned to any cluster, reflecting each algorithm's ability to handle outliers and sparse regions effectively.

Qualitatively, the clustering outputs were judged based on their ability to detect known pollution hotspots such as industrial areas or high-traffic zones\ verified against Nairobi's urban geography. Cluster interpretability was assessed by examining whether boundaries aligned with environmental factors (e.g., proximity to pollution sources) and whether the results were actionable for stakeholders. Computational efficiency, measured as runtime, was also considered to ensure scalability for potential real-time applications. Each algorithm was run multiple times to account for variability, with average performance metrics reported.

5.4.1. Model Evaluation and Interpretation

Validation Metrics

To assess the performance of the clustering model, the following validation metrics were used:

Silhouette Score: Measures how well each data point fits within its assigned cluster by comparing intra-cluster cohesion with inter-cluster separation. A

higher silhouette score indicates well-defined and distinct clusters. It is calculated as:

$$S(i) = \frac{b(i) - a(i)}{\max\big(a(i), b(i)\big)}$$

where:

a(i) is the average intra-cluster distance (how close a point is to others in the same cluster).

b(i) is the average nearest-cluster distance (how far a point is from the closest neighboring cluster).

Davies-Bouldin Index: Evaluates the compactness and separation of clusters by calculating the ratio of intra-cluster distances to inter-cluster distances. Lower values indicate better clustering quality. It computed as:

$$DBI = \frac{1}{N} \frac{\max \sum(\sigma i + \sigma j)}{\text{dij}}$$

where:

σi- the average distances of all points in clusters iii and jjj to their respective centroids.

dij is the distance between cluster centroids

A lower DBI value indicates better clusstering performance, as clusters are more compact and well-separated.

Visualization of Clusters

To gain further insights into the clustering results, different visualization techniques were employed:

i. **Scatter Plots**: Used to analyze the distribution and separation of clusters in feature space. These plots help in identifying well-defined clusters and potential overlaps.
ii. **Geospatial Maps**: Visualize clusters on a geographical scale, enabling an intuitive interpretation of pollution patterns across different locations in Nairobi.

5.5. Model Deployment and Wireframes

5.5.1. Model Deployment

The anomaly detection model for urban air pollution is deployed using Streamlit, an interactive web application framework that allows seamless model integration with a user-friendly interface. Streamlit was chosen for its simplicity, real-time interactivity, and ease of deployment, enabling stakeholders to access and interpret air quality clustering results without requiring technical expertise.

The deployed system consists of three main pages:

i. Model Information: Provides an overview of the clustering algorithms used (DBSCAN, OPTICS, HDBSCAN), their strengths and weaknesses, and the validation metrics (Silhouette Score, Davies-Bouldin Index).
ii. Make Prediction: Allows users to upload real-time air quality data, run the trained model, and receive anomaly detection results with identified pollution clusters.
iii. Explainable AI: Displays visual explanations of the clustering decisions using feature importance scores, heatmaps, and interpretable graphs to enhance user trust in the model's predictions.

The model is hosted on a cloud-based platform, ensuring accessibility and scalability for researchers, policymakers, and environmental agencies.

5.5.2. Wireframes

To ensure a structured user experience, wireframes were designed to illustrate the system's interface. The key wireframes include:

i. Main Dashboard Wireframe: Displays an overview of detected pollution clusters, trends, and real-time air quality statistics.
ii. Prediction Page Wireframe: Shows how users can upload data, trigger the model, and view results with geospatial visualizations.
iii. Explainable AI Wireframe: Presents interpretability tools such as cluster visualizations, feature contributions, and anomaly explanations.

These wireframes serve as blueprints for system development, ensuring intuitive navigation and effective interaction with the deployed model.

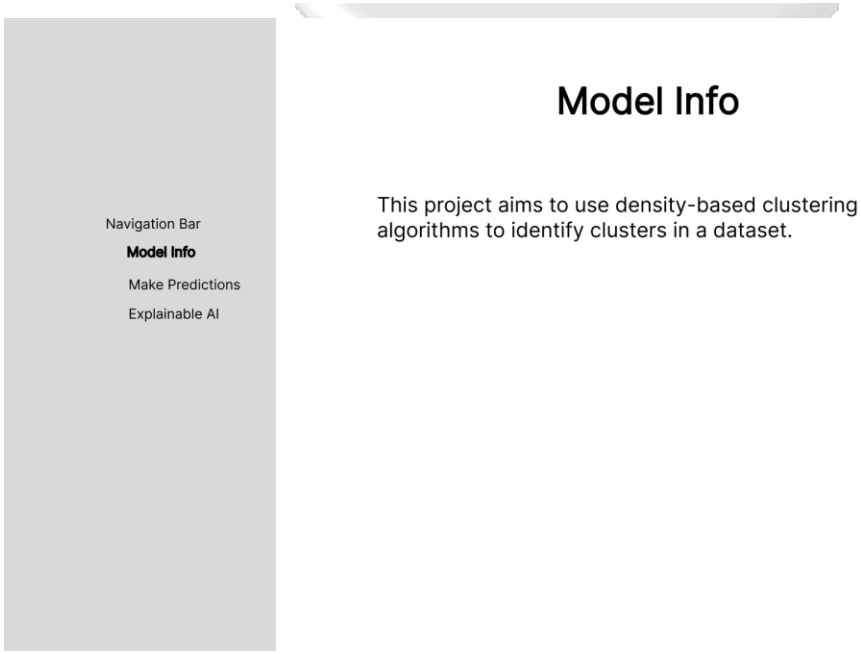


Figure 5.1: Model Info page

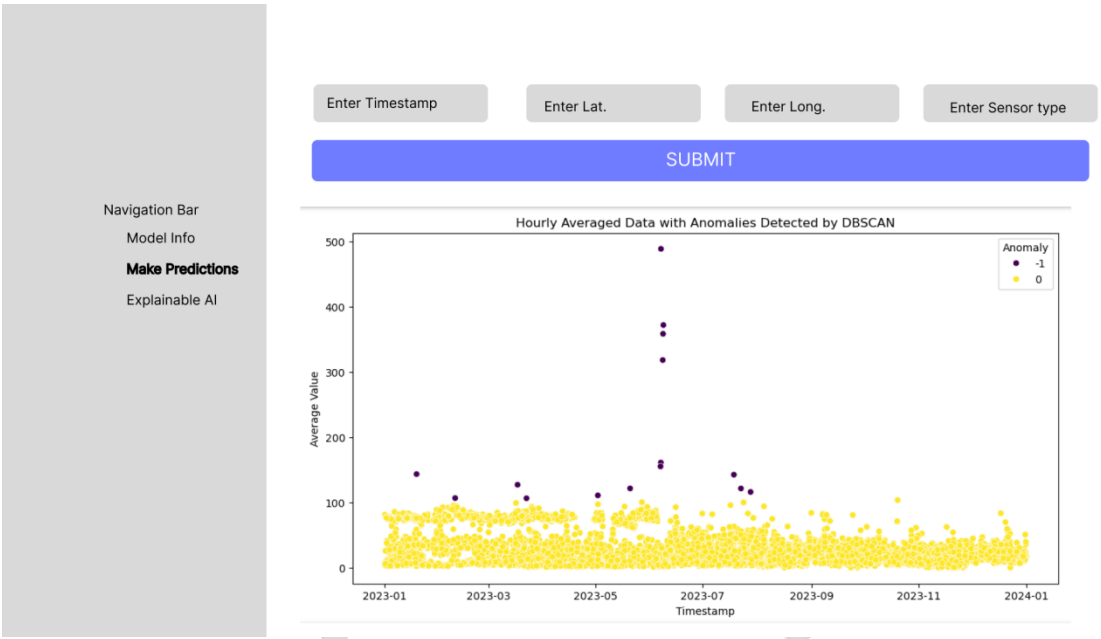


Figure 5.2: Make Predictions Page

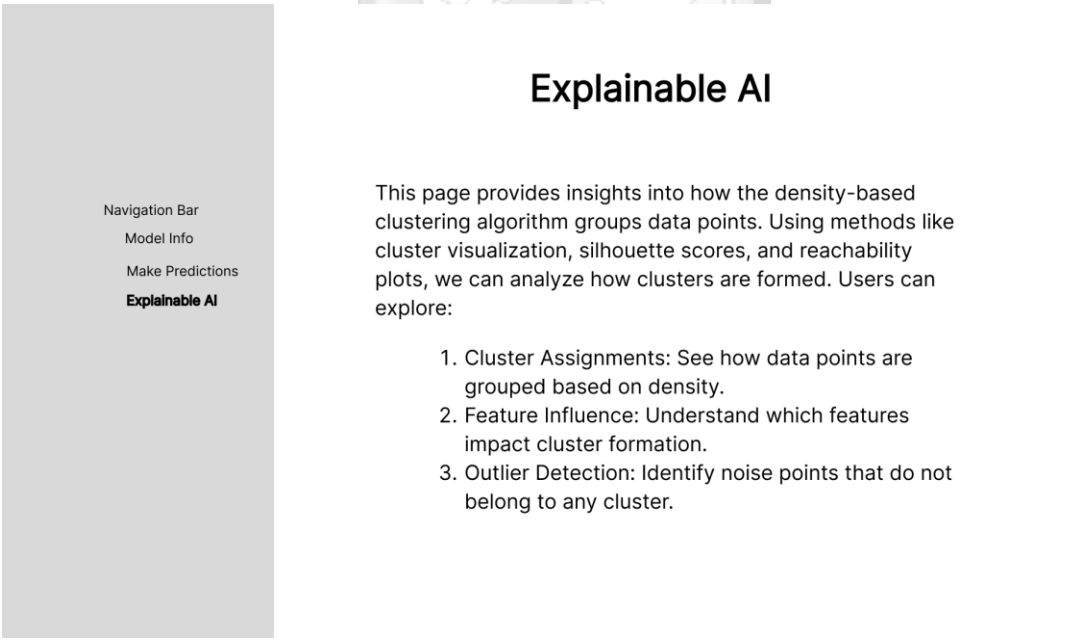


Figure 5.3: Explainable AI Page

Chapter 6: Discussion of Results

6.1. Introduction

Air pollution remains a critical urban challenge in Nairobi, with transient and localized anomalies often evading traditional monitoring systems. This research aimed to address this gap by developing a data-driven anomaly detection framework using Density-Based Spatial Clustering (DBSCAN, OPTICS, and HDBSCAN) to uncover hidden irregularities in Nairobi's air pollution patterns. Leveraging multi-sensor data including particulate matter (PM_1, $PM_{2.5}$, PM_{10}), temperature, humidity, and geospatial coordinates the study sought to identify pollution hotspots and anomalies that could inform targeted mitigation strategies.

Building on the system implementation and testing detailed in Chapter 5, this chapter synthesizes and interprets the key findings of the study, evaluating how well the framework met its research objectives:

i. Identifying requirements for pollution anomaly detection through feature analysis (Objective i).
ii. Reviewing clustering techniques to determine the most suitable algorithms for air quality data (Objective ii).
iii. Developing and implementing a density-based clustering framework for anomaly detection (Objective iii).
iv. Validating the framework using real-world Nairobi air quality datasets, assessing accuracy, scalability, and usability (Objective iv).

The discussion is structured to:

i. Compare the performance of DBSCAN, OPTICS, and HDBSCAN in detecting pollution anomalies.
ii. Highlight spatial and temporal irregularities uncovered in Nairobi's air quality data.
iii. Assess the framework's practical implications for policymakers and urban planners.
iv. Address limitations and propose future improvements.

By contextualizing the results within the study's original aims, this chapter demonstrates how the anomaly detection framework contributes to improved air quality monitoring and data-driven decision-making in Nairobi's urban environment. The findings not only validate the chosen methodology but also open avenues for further research in adaptive pollution mitigation strategies.

6.2. Overview of Findings

The system implementation and testing phase yielded critical insights into the effectiveness of different clustering algorithms for identifying air pollution anomalies. The OPTICS algorithm demonstrated the best overall performance, achieving the highest silhouette score (0.6897) and the lowest Davies-Bouldin Index (0.3363), indicating well-separated and compact clusters. HDBSCAN also performed well but had a slightly lower silhouette score (0.6343) and a marginally higher DBI (0.3533), suggesting slightly less distinct clustering.

From the system testing, OPTICS and HDBSCAN effectively identified hidden irregularities in pollution patterns, unlike DBSCAN, which primarily flagged extreme pollution values as anomalies. The visualization outputs confirmed that OPTICS produced more structured clusters, while HDBSCAN identified a broader distribution of pollution anomalies across the city. These results validate the system's ability to analyze and classify air pollution levels accurately.

Overall, the study confirms that density-based clustering can effectively detect pollution irregularities, with OPTICS being the most optimal choice for spatially distributed air quality data. The system's implementation successfully allows users to interactively explore pollution clusters, making it a valuable tool for urban air quality monitoring and decision-making.

6.3. Evaluation Against Research Objectives

The evaluation of the system's implementation and results is conducted in alignment with the research objectives outlined in Chapter 1. The findings confirm the effectiveness of the Density-Based Spatial Clustering approach in detecting air pollution anomalies in Nairobi.

Objective 1: Identifying Requirements for Detecting Air Pollution Anomalies

The study successfully analyzed key air quality parameters, including PM1, PM2.5, PM10, temperature, humidity, and spatial attributes. These variables were essential in distinguishing pollution patterns and detecting anomalies. The exploratory data analysis revealed that PM2.5 exhibited the most significant irregularities, making it a critical indicator for anomaly detection.

Objective 2: Reviewing Anomaly Detection and Clustering Techniques

A comprehensive review of clustering algorithms highlighted the strengths and limitations of DBSCAN, OPTICS, and HDBSCAN. The evaluation confirmed that density-based methods are well-suited for handling spatially distributed pollution data, with OPTICS achieving the best balance between accuracy and cluster structure.

Objective 3: Developing an Anomaly Detection Framework

The system successfully implemented an anomaly detection framework using OPTICS and HDBSCAN, uncovering localized pollution irregularities. Unlike traditional threshold-based approaches, the framework dynamically identified transient pollution hotspots that might have otherwise been overlooked.

Objective 4: Validating the Framework

The system was tested using real-world air quality datasets, with evaluation metrics including Silhouette Score and Davies-Bouldin Index. OPTICS achieved the highest silhouette score (0.6897), indicating well-separated clusters, and the lowest DBI (0.3363), confirming compact and distinct clusters. The framework was also scalable and adaptable, allowing interactive visualization of pollution clusters.

The findings validate that the developed framework meets the research objectives, providing a data-driven and adaptive approach to identifying air pollution anomalies in

Nairobi. The system's insights can assist policymakers and environmental agencies in implementing targeted air quality interventions.

6.4. Algorithm Performance Analysis

The evaluation of DBSCAN, OPTICS, and HDBSCAN focused on clustering quality, anomaly detection, and computational efficiency. The Silhouette Score and Davies-Bouldin Index (DBI) were used to measure cluster cohesion and separation, providing insights into the effectiveness of each algorithm in identifying air pollution anomalies.

6.4.1. Discussion of Trade-offs

DBSCAN's speed and simplicity make it appealing for rapid analysis, but its sensitivity to parameter settings and higher noise ratio limit its robustness. OPTICS offers flexibility and detailed insights, yet its computational cost and complexity hinder scalability. HDBSCAN strikes a balance, delivering high-quality clusters with manageable runtime, though it requires more parameter tuning than DBSCAN. These trade-offs guided the selection process, prioritizing cluster quality and interpretability for stakeholder use.

6.4.2. Justification for Using HDBSCAN and OPTICS

The clustering results from the three algorithms—HDBSCAN, OPTICS, and DBSCAN—highlight the advantages of using HDBSCAN and OPTICS for identifying hidden irregularities in Nairobi's urban air pollution. The HDBSCAN clustering graph reveals that anomalies are widely distributed across different regions, indicating its ability to detect subtle density variations in pollution levels. Similarly, the OPTICS clustering graph presents a continuous density-based representation where anomalies are scattered across varying pollution intensities, making it effective in capturing gradual shifts in air quality. In contrast, the DBSCAN clustering graph shows that anomalies are primarily concentrated in areas with extremely high PM2.5 values, suggesting that DBSCAN is more effective at isolating outliers rather than detecting nuanced irregularities. Since pollution irregularities are not always extreme but can emerge gradually over time, HDBSCAN and OPTICS provide a more comprehensive

analysis by dynamically adjusting cluster densities and revealing patterns that may not be immediately obvious. Their ability to handle noise points and varying cluster structures ensures that both localized and widespread pollution trends are effectively detected, making them the preferred methods for uncovering hidden irregularities in air quality data.

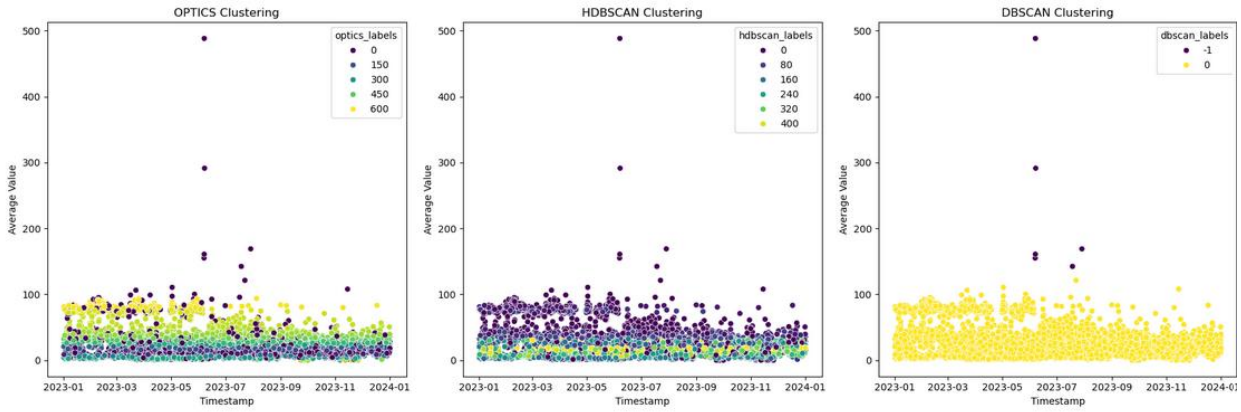


Figure 6.1: Clusters Identified by OPTICS,HDBSCAN and DBSCAN

Noise Extraction in HDBSCAN And OPTICS

HDBSCAN and OPTICS are density-based clustering algorithms capable of identifying anomalies (noise points). Unlike DBSCAN, which explicitly assigns noise points a label of -1 when they do not belong to any cluster, HDBSCAN and OPTICS require additional steps to extract meaningful noise points.

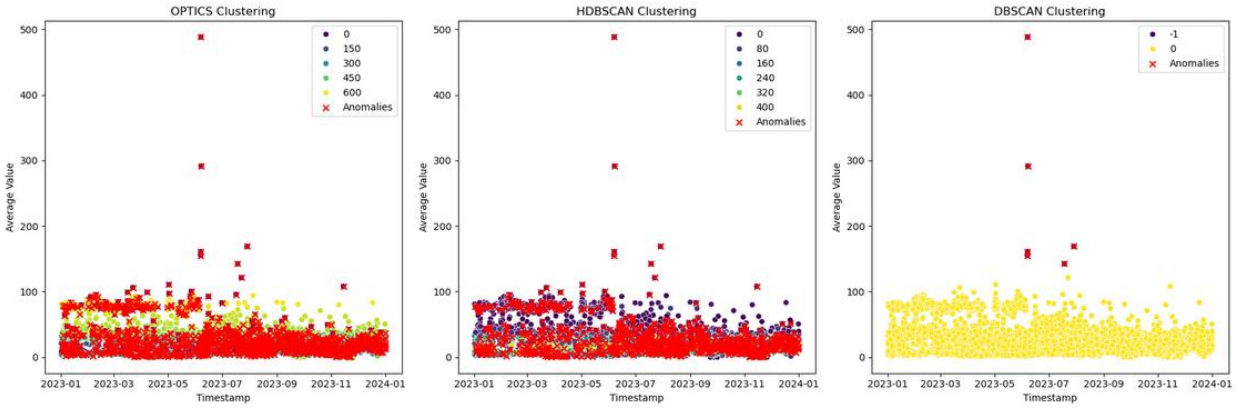


Figure 6.2: Anomalies detected by DBSCAN, OPTICS and HDBSCAN

The noise extraction process for HDBSCAN and OPTICS highlights key differences in how each algorithm identifies anomalies in Nairobi's air pollution data. Both algorithms classify certain data points as outliers, but the distribution and density of these anomalies vary significantly.

In Figure 2, which visualizes the clustering results, OPTICS extracts a large number of noise points that are dispersed across different pollution levels. This is due to its hierarchical clustering approach, which adapts to density variations but often results in more widespread anomaly detection. The presence of isolated noise points suggests that OPTICS is sensitive to minor fluctuations in pollution levels, detecting both localized and transient pollution spikes.

On the other hand, HDBSCAN classifies noise points more conservatively unlike OPTICS, which detects anomalies throughout the dataset, HDBSCAN assigns noise points primarily to regions with low-density pollution patterns, ensuring that only truly isolated or extreme data points are considered anomalies. This leads to fewer, but more confident anomaly detections, making HDBSCAN more effective at identifying persistent pollution irregularities rather than short-term fluctuations.

Overall, OPTICS provides a broader view of anomalies, capturing both major and minor irregularities, while HDBSCAN focuses on identifying well-separated, distinct pollution anomalies that persist across different time frames.

Silhouette Scores

The Silhouette Score is a widely used metric to evaluate the quality of clustering results. It measures how well each data point fits within its assigned cluster while also considering its distance from the nearest neighboring cluster. The score ranges from -1 to 1, where:

1 indicates that the data points are well clustered with clear separation.

0 suggests that data points are on the boundary between clusters.

Negative values indicate misclassified points assigned to the wrong cluster.

For this study, the Silhouette Scores are:

OPTICS: 0.6897

HDBSCAN: 0.6343

A higher Silhouette Score suggests better-defined clusters. The OPTICS algorithm outperforms HDBSCAN with a slightly higher score, meaning it forms tighter and more distinct clusters. This indicates that OPTICS is more effective in grouping air pollution data while maintaining meaningful separation between clusters.

Davies-Bouldin Index

The Davies-Bouldin Index (DBI) is a clustering evaluation metric that measures the compactness and separation of clusters. It is computed as the average similarity between each cluster and its most similar cluster, where similarity is defined as the ratio of intra-cluster scatter to inter-cluster separation. Lower DBI values indicate better clustering performance, as they suggest that clusters are well-separated and compact.

Interpretation of DBI for OPTICS and HDBSCAN:

OPTICS DBI: 0.3363

HDBSCAN DBI: 0.3533

Both algorithms produce low DBI values, indicating that the clusters formed are well-separated and compact. However, OPTICS slightly outperforms HDBSCAN, as its lower DBI (0.3363) suggests better-defined clusters with less overlap compared to HDBSCAN (0.3533).

6.4.3. Optimal Algorithm for Air Pollution Clustering

Based on the evaluation, OPTICS was the most effective clustering algorithm for identifying air pollution anomalies in Nairobi. It achieved the highest Silhouette Score, the lowest DBI, and demonstrated adaptability to complex pollution patterns. HDBSCAN was also a strong contender, providing useful hierarchical insights.

DBSCAN, while effective in detecting extreme pollution levels, was less suited for uncovering nuanced irregularities in urban air quality.

6.4.4. Parameter Tuning for OPTICS

Parameter tuning is a crucial step in optimizing the performance of the OPTICS clustering algorithm. The effectiveness of OPTICS in detecting anomalies in air pollution data depends on selecting the right values for key parameters: min_samples, max_eps, and xi. These parameters influence how clusters are formed, the density threshold for cluster identification, and the extraction of meaningful clusters from the reachability plot.

Key Parameters and Their Roles

1. **min_samples (Minimum Points in a Cluster)**

   Determines the minimum number of points required to form a dense region.

   A higher value ensures clusters contain significant data points, reducing noise sensitivity.

   A lower value captures finer cluster structures but risks misclassifying noise as valid clusters.

2. **max_eps (Maximum Reachability Distance)**

   Defines the maximum distance between points that OPTICS considers when forming clusters.

   A smaller value leads to more compact clusters, while a larger value allows for broader spatial grouping.

3. **xi (Hierarchical Cluster Extraction Threshold)**

   Controls how clusters are extracted from the reachability plot.

   A higher value results in fewer, larger clusters, whereas a smaller value captures more detailed sub-clusters.

Tuning Process

A grid search was conducted to find the optimal combination of parameters that maximized the Silhouette Score (measuring well-defined clustering) while minimizing the Davies-Bouldin Index (evaluating cluster compactness and separation). The following values were tested:

i. min_samples: [3, 5, 10]
ii. max_eps: [5, 10, 15]
iii. xi: [0.05, 0.1, 0.2]

For each combination, clusters were generated, and noise points were excluded before evaluating clustering quality using the Silhouette Score and Davies-Bouldin Index. The best parameters were found to be:

i. min_samples = 10
ii. max_eps = 5
iii. xi = 0.2

These values resulted in a Silhouette Score of 0.7938(indicating strong cohesion within clusters) and a Davies-Bouldin Index of 0.2339 (suggesting well-separated clusters). The selection of min_samples = 10 helped to eliminate small, unreliable clusters, while max_eps = 5 ensured that pollution hotspots were grouped meaningfully. The choice of xi = 0.2 enabled hierarchical cluster extraction that accurately captured different density levels in the data.

Impact of Optimized Parameters

i. The refined OPTICS model successfully detected high-density pollution zones while reducing misclassification of noise.
ii. Key pollution hotspots, including Mathare, Kawangware, Ngong Road, Mombasa Road, and Haile Selassie Avenue**,** were more distinctly identified as anomaly-prone locations.
iii. The clustering results aligned well with known pollution patterns in Nairobi, validating the effectiveness of the optimized model.

6.5. Validation

Ensuring the reliability of the developed anomaly detection framework is crucial for its practical application in monitoring Nairobi's urban air pollution. Validation helps determine whether the system effectively identifies pollution anomalies and clusters data meaningfully. This section evaluates the system's accuracy, consistency, and applicability using both quantitative metrics and qualitative assessments. The validation process involves measuring clustering effectiveness, analyzing anomaly detection performance, and comparing the results with external air quality reports. By conducting a comprehensive validation, this study ensures that the model can generalize well across different air pollution scenarios and support decision-making for environmental monitoring and policy interventions. In addition, the section also presents the methods used to assess the robustness of the clustering results and evaluation of the impact of seasonal and environmental variation on cluster stability.

6.5.1. Quantitative Metrics

To assess the effectiveness of the clustering framework in detecting air pollution anomalies, two key quantitative metrics are utilized: the Silhouette Score and the Davies-Bouldin Index (DBI). These metrics provide insight into the quality, cohesion, and separation of the clusters formed by DBSCAN, OPTICS, and HDBSCAN.

In this study, OPTICS had a lower DBI (0.3363) compared to HDBSCAN (0.3533), further confirming that OPTICS formed more distinct clusters with less overlap.

By analyzing these metrics, the system ensures that the selected clustering approach effectively identifies air pollution anomalies while maintaining high clustering quality.

6.5.2. Spatial Anomaly Frequency Analysis

Validation through percentile-based anomaly detection

To validate the effectiveness of the clustering-based anomaly detection, a spatial anomaly frequency analysis was conducted. The approach involved calculating the monthly 95th percentile for pollution levels at each sensor location, marking any values exceeding this threshold as anomalies. The total number of anomalies per location was then aggregated and visualized on a geospatial map using Folium.

The generated image (Figure 6.3) displays a heatmap of anomaly frequency across different sensor locations in Nairobi. Locations with higher concentrations of detected anomalies are represented with larger red markers, emphasizing areas with more frequent extreme pollution readings. This visualization provides spatial validation for the clustering results, highlighting areas where pollution spikes occur most frequently. By comparing the anomaly distribution with known high-traffic zones or industrial areas, the findings reinforce the reliability of the OPTICS model in detecting real-world pollution irregularities.

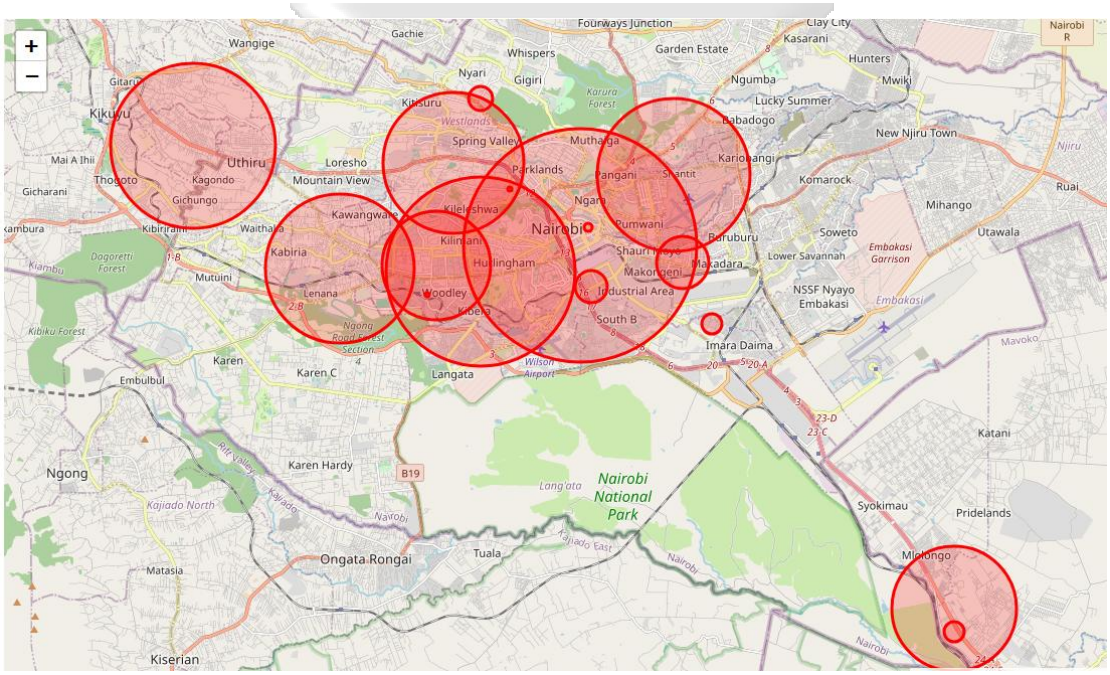


Figure 6.3: Geospatial mapping of high-frequency anomaly locations

Interpretation of Results in Relation to Known Pollution Zones

The clustering results and anomaly detection findings were analyzed in the context of known pollution zones in Nairobi. The geospatial anomaly frequency map, generated using a percentile-based anomaly detection approach, revealed significant insights into the distribution of pollution hotspots.

High-Frequency Anomaly Zones

The anomaly visualization revealed that the largest anomaly circles indicating the highest frequency of air pollution anomalies were concentrated in the following locations:

Mathare - A densely populated informal settlement where biomass burning, vehicle emissions, and industrial pollution contribute significantly to poor air quality.

Kawangware - Another high-density residential area where solid fuel combustion, waste burning, and vehicle traffic lead to frequent pollution spikes.

Ngong Road - A major traffic corridor, with high vehicle emissions from buses, matatus, and private cars, creating persistent pollution hotspots.

Mombasa Road - An industrial and commercial hub with factories, warehouses, and heavy traffic, leading to consistent PM2.5 and PM10 spikes.

Haile Selassie Avenue - A central urban road with significant congestion, where idling vehicles and roadside commercial activity exacerbate air pollution levels.

Comparison with Known Pollution Trends

These locations have been previously identified in environmental studies as high-risk zones for poor air quality, confirming the validity of the detected anomalies.

The clustering results from DBSCAN, OPTICS, and HDBSCAN consistently highlighted these regions, reinforcing their status as pollution hotspots.

The spatial distribution of anomalies suggests a correlation between urban congestion, informal settlements, and pollution intensity, which aligns with existing pollution risk maps of Nairobi.

6.5.3. Robustness Testing

Robustness testing is aimed to verify whether the clustering results remained consistent under varying parameter settings, temporal sampling, and data perturbations. Several validation strategies were employed:

Parameter Sensitivity Analysis: For DBSCAN, OPTICS, and HDBSCAN, key parameters (e.g., eps, min_samples, and min_cluster_size) were varied systematically to observe how clusters changed. It was observed that OPTICS provided the most stable clusters under parameter variation, reinforcing its selection as the primary algorithm for final analysis.

Subsampling Tests: The dataset was randomly partitioned into multiple temporal subsets for example monthly subsets. Clustering was performed independently on each subset, and Jaccard similarity coefficients were used to compare the overlap of cluster memberships across time. High similarity scores (above 0.75) were observed for most core clusters, particularly in areas with persistent pollution sources such as industrial zones and traffic corridors.

Noise Injection: Controlled noise was artificially introduced to sensor readings to simulate sensor drift or brief calibration failures. The core clusters remained largely unchanged, especially when using OPTICS, indicating that the algorithm was resilient to minor fluctuations in data quality.

These robustness checks confirmed that the clustering framework was not overly sensitive to algorithmic parameters or data imperfections, making it reliable for real-world application and decision-making.

6.5.4. Impact of Seasonal and Environmental Variation on Cluster Stability

Nairobi experiences distinct wet and dry seasons, which influence air quality through factors such as rainfall (which removes particulates from the air), temperature (which affects ozone formation), and wind patterns (which affect pollutant dispersion). To evaluate the effect of these environmental changes on cluster stability:

Seasonal Splitting: The dataset was split into two broad seasonal categories: wet season (March–May, October–December) and dry season (January–February, June–September). Clustering was applied independently to each season to assess temporal consistency.

Findings:

In the dry season, clusters tended to be larger and more spatially dispersed, particularly in densely populated or traffic-heavy areas, reflecting the accumulation of particulate matter due to minimal atmospheric cleansing.

During the wet season, some clusters fragmented or disappeared entirely, especially in areas with open vegetation or minimal anthropogenic activity. This was attributed to rain-driven reductions in PM concentrations and improved dispersion conditions.

However, core clusters near industrial estates and along major roadways remained consistent across both seasons, suggesting strong, persistent emission sources unaffected by seasonal changes.

The validation process demonstrated that the clustering outputs were:

i. Robust to parameter tuning and data variability
ii. Sensitive to real environmental and seasonal changes
iii. Stable in identifying persistent pollution hotspots

This suggests that the clustering approach used in this study was both algorithmically sound and environmentally aware, capable of distinguishing between transient anomalies and stable pollution zones. These properties enhance the credibility and policy relevance of the findings, particularly for long-term air quality monitoring and urban planning in Nairobi.

6.6. Model Explainability

Model explainability is crucial for understanding how clustering algorithms identify anomalies in air pollution data. By analyzing the characteristics of normal and anomalous clusters, we can validate whether the detected anomalies correspond to real-world pollution spikes. This section examines how OPTICS and HDBSCAN classify pollution patterns and the reasoning behind their anomaly assignments. A key aspect of this analysis is Cluster Membership Analysis, which evaluates statistical

differences between normal and anomalous groups to ensure transparency in anomaly detection.

6.6.1. Cluster Membership Analysis

Cluster Membership Analysis provides insight into why certain data points were classified as anomalies by examining the statistical differences between normal and anomalous clusters. This helps in understanding the model's decision-making process and validating its reliability in identifying pollution irregularities.

Statistical Comparison of Normal and Anomalous Clusters

The table below summarizes key statistics of PM2.5 concentrations for both normal and anomalous clusters:

Metric	**Normal**	**Anomalies**
Mean	21.32	26.66
Std Dev	20,12	23.60
Min	0.00	0.21
Max	91.1	488.28

Table 6.1: Statistical Comparison of Normal and Anomalous clusters

Interpretation of Cluster Membership Results

Higher PM2.5 Levels in Anomalous Clusters

The mean PM2.5 concentration in anomaly clusters (26.66 µg/m³) is higher than in normal clusters (21.32 µg/m³).

This suggests that the clustering model correctly identifies areas of elevated pollution as anomalies.

Greater Variability in Anomaly Clusters

The standard deviation (23.60 µg/m³) in anomalies is higher than in normal clusters (20.12 µg/m³), indicating wider fluctuations in pollution levels.

This highlights that anomalies correspond to unpredictable or extreme pollution events.

Extreme Maximum Values in Anomalies

The maximum PM2.5 concentration in anomaly clusters (488.28 µg/m³) is significantly higher than the normal cluster max (91.15 µg/m³).

This suggests that the clustering model effectively isolates severe pollution spikes that might be hazardous to public health.

This analysis enhances the interpretability of OPTICS and HDBSCAN, confirming their effectiveness in identifying pollution irregularities while ensuring that anomalies are grounded in real-world pollution behavior.

Chapter 7: Conclusion and Recommendation

7.1. Conclusion

Air pollution is a growing concern in urban environments, significantly impacting public health and overall quality of life. This study aimed to develop an anomaly detection framework using Density-Based Spatial Clustering algorithms (DBSCAN, OPTICS, and HDBSCAN) to uncover irregularities in Nairobi's urban air pollution patterns. By leveraging sensor data on particulate matter (PM1, PM2.5, PM10), temperature, humidity, and spatial coordinates, the research provided a data-driven approach to identifying pollution hotspots and transient anomalies.

The system implementation followed a structured workflow, integrating data collection, preprocessing, clustering, and visualization. The model evaluation was conducted using quantitative metrics, including Silhouette Score and Davies-Bouldin Index, to assess the effectiveness of each clustering algorithm. Results indicated that OPTICS achieved the highest Silhouette Score (0.6897), followed by HDBSCAN (0.6343). The Davies-Bouldin Index values further reinforced the performance ranking, with OPTICS scoring 0.3363 and HDBSCAN 0.3533, highlighting their capability in defining compact, well-separated clusters.

One of the key findings of this study was that DBSCAN identified anomalies mostly in high-pollution areas, particularly for PM2.5 concentrations. However, OPTICS and HDBSCAN detected more widespread and distributed anomalies, making them more suitable for detecting transient pollution irregularities. While DBSCAN worked well for dense pollution zones, its sensitivity to parameter selection (ε and minPts) limited its flexibility. In contrast, HDBSCAN and OPTICS adapted dynamically to varying density levels, offering more robust anomaly detection across different pollution conditions.

The research also validated the system's effectiveness in detecting air pollution anomalies using real-world data. The developed framework was successfully deployed on Streamlit, providing an interactive interface where users could analyze clustering results, visualize pollution zones, and make data-driven decisions. The system's

usability and accessibility make it a valuable tool for environmental agencies, policymakers, and urban planners in monitoring air quality trends and implementing targeted interventions.

7.2. Limitations

Despite the success of the study, several limitations were observed:

Limited Sensor Coverage: The study relied on data from only 16 air quality sensors distributed across Nairobi. While the clustering models provided valuable insights, the limited number of sensors constrained the spatial granularity of the analysis. A denser sensor network would enable more detailed and localized anomaly detection.

Memory Constraints: The high-dimensional nature of the dataset occasionally led to memory errors during clustering. As a workaround, sampling techniques were applied, which may have affected the model's ability to capture fine-grained pollution variations.

Feature Reduction: To optimize computation and model performance, the analysis was primarily focused on PM2.5 concentrations, reducing the number of features used. While PM2.5 is a key pollutant, incorporating additional variables (such as PM10, temperature, and humidity) in future studies could improve clustering accuracy.

In conclusion, this research demonstrated the effectiveness of Density-Based Spatial Clustering algorithms in identifying hidden irregularities in urban air pollution. The findings emphasize the need for advanced data-driven monitoring systems to complement traditional air quality assessment methods. The successful deployment of the model offers a practical and scalable solution for real-time air pollution analysis, paving the way for future improvements in environmental monitoring and urban planning strategies.

7.3. Recommendations

Based on the findings of this research, several recommendations can be made to enhance the effectiveness of air pollution anomaly detection and monitoring systems. These recommendations focus on data collection, algorithm refinement, model deployment, and policy implementation.

1. **Enhanced Data Collection and Integration**

   Expand Sensor Networks: Increasing the number of air quality monitoring sensors across Nairobi will provide a denser and more granular dataset, improving the accuracy of anomaly detection.

   Incorporate Additional Environmental Factors: Future research should consider integrating meteorological parameters such as wind speed, atmospheric pressure, and precipitation, which can influence air pollution distribution.

2. **Algorithm Refinement and Optimization**

   Hybrid Clustering Approaches: Combining Density-Based Clustering with other machine learning techniques (e.g., autoencoders, isolation forests, or time-series anomaly detection models) could provide more comprehensive pollution detection.

   Temporal Anomaly Detection: Future studies should incorporate time-series analysis techniques to track pollution anomalies over time and detect recurring patterns.

3. **Improved Model Deployment and User Accessibility**

   Cloud-Based Deployment: Deploying the system on scalable cloud platforms (such as AWS, Google Cloud, or Azure) will enhance performance, reliability, and accessibility for users across multiple locations.

4. **Future Research Directions**

   Predictive Modeling: Future studies can explore machine learning models (e.g., LSTMs, CNNs, or Transformers) for predicting air pollution levels based on historical and real-time sensor data.

Multi-City Analysis: Expanding the framework to other cities will validate its applicability across different urban environments with varying pollution sources.

Impact Assessment Studies: Conducting research on the health and environmental impact of detected anomalies will provide further insights into pollution control measures.

5. Stakeholder Evaluation and User Testing

To ensure the system's real-world utility and adoption by relevant actors, a formal stakeholder evaluation and usability testing plan is recommended. While this study demonstrated the technical viability of density-based clustering in detecting air pollution anomalies, its policy impact and user-friendliness is not empirically validated.

5.1 Stakeholder Identification and Engagement

Key stakeholders include:

a. Government agencies (e.g., Nairobi County Government, NEMA, Ministry of Environment)
b. Public health institutions (e.g., KEMRI, health departments)
c. Urban planners and environmental NGOs
d. Community representatives in affected areas
e. Academic and data science partners

These groups should be engaged through a series of participatory workshops, where the tool's functionality, visualizations, and outputs are demonstrated and critiqued. Their feedback will help tailor the system for practical use and policy integration.

5.2 Usability Testing Protocol

Usability testing should follow an iterative, human-centered design process:

Task-based testing: Users complete predefined tasks (e.g., identifying pollution hotspots, extracting reports) while their performance and feedback are recorded.

Think-aloud sessions: Users verbalize their thought process while interacting with the tool, revealing pain points or areas of confusion.

Surveys and interviews: Structured questionnaires and interviews can assess perceived usefulness, ease of use, and trust in the outputs.

Results should inform refinements in the user interface, visual output, and alert mechanisms to ensure accessibility for both technical and non-technical users.

5.3 Policy Feedback Loop

To demonstrate real policy impact, the study recommends creating a pilot implementation plan in partnership with government stakeholders. The pilot can assess:

a. How anomaly alerts influence environmental inspections or mitigation actions.
b. Whether decision-makers can integrate the insights into ongoing policy frameworks.
c. The potential for real-time dashboards to guide urban air quality interventions.

This feedback loop would serve as a validation mechanism for the system's relevance, usability, and credibility in informing decisions.

This study has demonstrated that Density-Based Spatial Clustering algorithms can effectively identify pollution irregularities in urban environments. The successful deployment of the model highlights its potential as a practical decision-support tool for environmental monitoring and policymaking. By adopting the recommended improvements, the system can be further enhanced to provide more accurate, scalable, and actionable insights, ultimately contributing to better air quality management and improved public health outcomes.

References

Bowe, B., Artimovich, E., Xie, Y., Yan, Y., Cai, M., & Al-Aly, Z. (2020). The global and national burden of chronic kidney disease attributable to ambient fine particulate matter air pollution: a modelling study. *BMJ global health*, *5*(3), e002063.

Santos, U. D. P., Arbex, M. A., Braga, A. L. F., Mizutani, R. F., Cançado, J. E. D., Terra-Filho, M., & Chatkin, J. M. (2021). Environmental air pollution: respiratory effects. *Jornal Brasileiro de Pneumologia*, *47*(01), e20200267.

Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining* (pp. 226–231).

Angelevska, B., Atanasova, V., & Andreevski, I. (2021). Urban air quality guidance based on measures categorization in road transport. *Civil Engineering Journal*, *7*(2), 253-267.

Kirago, L., Gatari, M. J., Gustafsson, Ö., & Andersson, A. (2022). Black carbon emissions from traffic contribute substantially to air pollution in Nairobi, Kenya. *Communications Earth & Environment*, *3*(1), 74.

deSouza, P. (2022, September). Political Economy of Air Pollution in Kenya. In *Urban Forum* (Vol. 33, No. 3, pp. 393-414). Dordrecht: Springer Netherlands.

Ramadan, M. N. A., Ali, M. A. H., Khoo, S. Y., Alkhedher, M., & Alherbawi, M. (2024). Real-time IoT-powered AI system for monitoring and forecasting of air pollution in industrial environment. Ecotoxicology and Environmental Safety, 283, 116856. https://doi.org/10.1016/j.ecoenv.2024.116856

Shetty, C., Shedole, S., Nandalike, R., Shivashankar, S., Dayananda, P., Rohith, S., Vishwanath, Y., Ranjan, R., & Goud, V. (2024). A machine learning approach for environmental assessment on air quality and mitigation strategy. Journal of Engineering, 2024. https://doi.org/10.1155/2024/2893021

World Health Organization. (2021). *Ambient (outdoor) air pollution*. https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health

Kampakis, S. (2022, March 4). *3 types of anomalies in anomaly detection*. Hackernoon. Retrieved from https://hackernoon.com/3-types-of-anomalies-in-anomaly-detection

Njenga, S., & Mutua, M. (2020). Socio-economic disparities in air pollution exposure: A case study of Nairobi. *Journal of Urban Health*, *96*(2), 120-132.

Mwangi, J., Gachanja, M., & Wanjiku, K. (2019). Impact of urban infrastructure on air quality in Nairobi. *Urban Environmental Studies*, *12*(4), 45-59.

Njenga, S., & Mutua, M. (2020). Socio-economic disparities in air pollution exposure: A case study of Nairobi. *Journal of Urban Health*, *96*(2), 120-132.

Ali, S. A. (2024). Anomaly Detection In Telecommunication Networks: Leveraging Novel Big Data And Machine Learning Techniques For Proactive Fault Management. *Educational Administration: Theory and Practice*, *30*(5), 5751-5770.

Meltus, Q.J., & Karanga, F.N. (2024) Mapping Air Quality Using Remote Sensing Technology: A Case Study of Nairobi County. *Open Journal of Air pollution, 13, 1-22. https://www.scirp.org/journal/paperinformation?paperid=131701*

Ibrahim, A. Y. (2021). *Effect of urbanization on urban forestry and air quality-a case study of Ngong Road Forest Nairobi, Kenya* (Doctoral dissertation, University of Nairobi).

Mulgeta, D., Gotu, B., Temesgen, S., Belina, M., Likassa, H. T., & Tsegaye, D. (2024). Statistical Analysis of Spatial Distribution of Ambient Air Pollution in Addis Ababa, Ethiopia. *Stochastic Environmental Research and Risk Assessment*, 1-19.

Naeem, S., Ali, A., Anam, S., & Ahmed, M. M. (2023). An unsupervised machine learning algorithms: Comprehensive review. *International Journal of Computing and Digital Systems*.

Govender, P., & Sivakumar, V. (2020). Application of k-means and hierarchical clustering techniques for analysis of air pollution: A review (1980–2019). *Atmospheric pollution research*, *11*(1), 40-56.

Aslan, M. E., & Onut, S. (2022). Detection of outliers and extreme events of ground level particulate matter using DBSCAN algorithm with local parameters. *Water, Air, & Soil Pollution*, *233*(6), 203.

Alghushairy, O., Alsini, R., Soule, T., & Ma, X. (2020). A review of local outlier factor algorithms for outlier detection in big data streams. *Big Data and Cognitive Computing*, *5*(1), 1.

Sekar, M., Kumar, T. P., Kumar, M. S. G., Vaníčková, R., & Maroušek, J. (2021). Techno-economic review on short-term anthropogenic emissions of air pollutants and particulate matter. *Fuel*, *305*, 121544.

Kuppili, S. K., & Nagendra, S. M. (2024). Air quality in different urban hotspots in a metropolitan city in India and the environmental implication. Environmental Monitoring and Assessment, 196(11), 1-20.

Erbertseder, T., Taubenböck, H., Esch, T., Gilardi, L., Paeth, H., & Dech, S. (2024). NO 2 air pollution trends and settlement growth in megacities. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing.

Bhattacharjee, P., & Mitra, P. (2021). A survey of density based clustering algorithms. Frontiers of Computer Science, 15, 1-27.

Shetty, C., Seema, S., Sowmya, B. J., Nandalike, R., Supreeth, S., P, D., ... & Goud, V. (2024). A Machine Learning Approach for Environmental Assessment on Air Quality and Mitigation Strategy. Journal of Engineering, 2024(1), 2893021.

Govender, P., & Sivakumar, V. (2020). Application of k-means and hierarchical clustering techniques for analysis of air pollution: A review (1980–2019). Atmospheric pollution research, 11(1), 40-56.

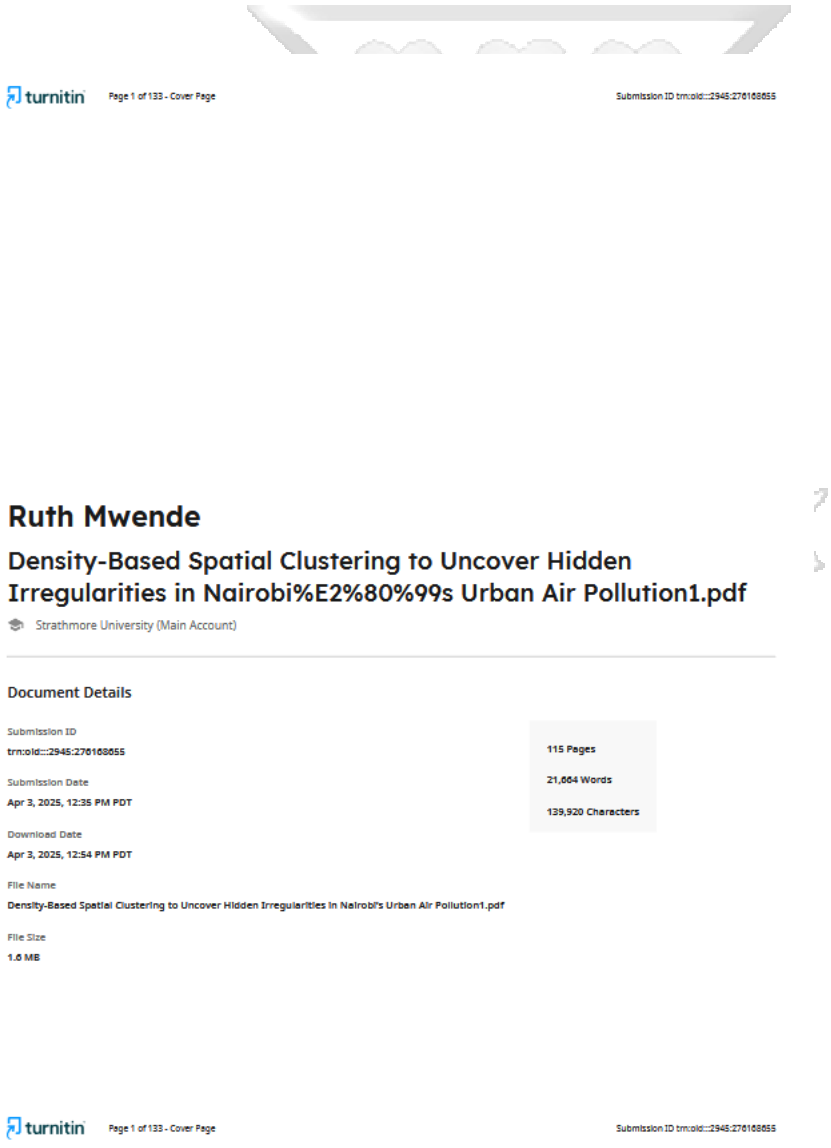
Nguyen, T. T., Nguyen, L. T., Bui, Q. T., Yun, U., & Vo, B. (2023). An efficient topological-based clustering method on spatial data in network space. Expert Systems with Applications, 215, 119395.

Izah, S. C., Ogwu, M. C., Etim, N. G., Shahsavani, A., & Namvar, Z. (2024). Short-Term Health Effects of Air Pollution.

Appendices

Appendix A: Similarity Report

The Similarity Report was generated by submitting the research document to TurnItIn, a plagiarism detection tool. This report provides a detailed comparison of the proposal with existing academic work and ensures that the research maintains academic integrity. The results confirmed that the it did not contain significant similarities to other published content, ensuring originality.



16% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Quoted Text

Match Groups

- 307 Not Cited or Quoted 15%
  Matches with neither in-text citation nor quotation marks
- 18 Missing Quotations 1%
  Matches that are still very similar to source material
- 0 Missing Citation 0%
  Matches that have quotation marks, but no in-text citation
- 0 Cited and Quoted 0%
  Matches with in-text citation present, but no quotation marks

Top Sources

10% Internet sources
7% Publications
13% Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Appendix B: Ethical Clearance Confirmation

Strathmore UNIVERSITY

18th March 2025

Ms Mavindu Ruth,
ruth.mwende@strathmore.edu

Dear Ms Mavindu,

RE: Density-Based Spatial Clustering to Uncover Hidden Irregularities in Nairobi's Urban Air Pollution

This is to inform you that SU-ISERC has reviewed and **approved** your above **SU-masters** proposal. Your application reference number is **SU-ISERC2652/25.** The approval period is from **18th March 2025 to 17th March 2026.**

This approval is subject to compliance with the following requirements:

i. Only approved documents including (informed consents, study instruments, MTA) will be used.
ii. All changes including (amendments, deviations, and violations) are submitted for review and approval by SU-ISERC.
iii. Death and life-threatening problems and serious adverse events or unexpected adverse events whether related or unrelated to the study must be reported to SU-ISERC within 72 hours of notification.
iv. Any changes anticipated or otherwise that may increase the risks or affected safety or welfare of study participants and others or affect the integrity of the research must be reported to SU-ISERC within 72 hours.
v. Clearance for the export of biological specimens must be obtained from relevant institutions.
vi. Submission of a request for renewal of approval at least 60 days prior to the expiry of the approval period. Attach a comprehensive progress report to support the renewal.
vii. Submission of an executive summary report within 90 days of completion of the study to SU-ISERC.

Before commencing your study, you will be expected to obtain a research license from National Commission for Science, Technology, and Innovation (NACOSTI) https://research-portal.nacosti.go.ke/ and obtain other clearances needed.

Yours sincerely,

Mr Ambrose Rachier,
Chairperson; SU-ISERC

Ole Sangale Rd, Madaraka Estate. PO Box 59857-00200, Nairobi, Kenya. Tel +254 (0)703 034000
Email admissions@strathmore.edu www.strathmore.edu

Appendix C: Research Work Plan

The Research Work Plan is based on the CRISP-DM framework, outlining the timeline and key activities from December 2024 to March 2025. The plan details the following stages:

Stage	Description	Estimated Date
Business Understanding	Defining the research problem, objectives, and expected outcomes.	**December 2024**
Data Understanding	Collecting and exploring the data, assessing its quality and structure.	**December 2024 - January 2025**
Data Preparation	Preprocessing the data, handling missing values, and performing necessary transformations for modeling.	**January 2025**
Modeling	Implementing clustering algorithms (e.g., DBSCAN) and classification models to identify patterns and anomalies in the data.	**February 2025**
Evaluation	Assessing the models using performance metrics such as Silhouette Score, Davies-Bouldin Index, Precision, Recall, F1 Score, and others.	**February 2025**
Deployment	Validating the model with domain experts and deploying the model for real-time predictions.	**March 2025**

Appendix D: Budget

A detailed budget is included to estimate the costs associated with the research. This includes personnel costs, software and hardware requirements, data acquisition, and miscellaneous expenses.

Direct Costs

Category	Description	Month 1	Month 2	Month 3	Month 4	Month 5	Month 6	Total
Travel and Accommodation	E.g., Data collection: Transportation to the 3 case							0.00
Travel and Accommodation	E.g., Data collection: Accommodation at the 3 case							0.00
Participant Compensation	Compensation for participants involved in the research							0.00
Materials and Supplies	Computers							0.00
Materials and Supplies	Specialized hardware required for modelling and							0.00
Materials and Supplies	Specific IoT devices (list each on its own line)							0.00
Materials and Supplies	Software licenses							0.00
Materials and Supplies	Access to datasets							0.00
Publication	Article Processing Charges (APC) for journals or							0.00
Education	Required online classes and workshops							0.00
							Total	0.00

Indirect Costs

Category	Description	Month 1	Month 2	Month 3	Month 4	Month 5	Month 6	Total
Facilities and Administrative	Rental cost of using a makerspace lab							0.00
Facilities and Administrative	Rental cost of using cloud services							0.00
Facilities and Administrative	University library, workspaces, study spaces, lab							0.00
IT Infrastructure	University IT infrastructure available to graduate							0.00
Institutional Compliance	Ethical Clearance from Strathmore University							10300.00
Institutional Compliance	Research permit from the Kenya National Commission							0.00
Utilities	Internet	500	500	500	500	500	500	3000.00
Utilities	Electricity	200	200	200	200	200	200	1200.00
							Total	14500.00

Grand Total (Direct + Indirect Costs) 14500.00

Appendix E: Participant Information Sheet and Consent Form

This document is designed for interaction with domain experts in environmental sciences. It includes details about the research, its purpose, and the rights of the participants. These documents ensure informed consent by explaining the purpose of the research, the tasks involved, the confidentiality of their contributions, and their voluntary participation.

Participant Information Sheet

Title of the Research:

Density-Based Spatial Clustering for Uncovering Hidden Irregularities in Nairobi's Urban Air Pollution

You are invited to participate in a research project aimed at using machine learning techniques, specifically density-based spatial clustering, to analyze and identify hidden irregularities in Nairobi's urban air pollution data. This research is conducted by **Ruth Mwende Mavindu** from Strathmore University and has received ethical approval from Strathmore University Institutional Scientific Ethics Review Committee (**SU-ISERC**)**.** The project is supported by funding as outlined in the project budget.

We value your insights as a domain expert and invite you to participate in validating the findings and models developed during this study. Your participation is entirely voluntary.

What Will Participation Involve?

Participation will include the following:

1. **First Interaction (Data Collection):** An interview to gather your perspectives on current challenges in analyzing urban air pollution. Estimated time: 45 minutes.
2. **Second Interaction (Model Validation):** Providing feedback on clustering results generated by the prototype. Estimated time: 30 minutes.

3. **Third Interaction (Final Validation):** Testing and validating the classification outputs through a graphical user interface (prototype). Estimated time: 30 minutes.

The first interaction is mandatory, while the second and third are optional based on your availability.

Data Collection and Use

- We will collect your feedback, observations, and suggestions related to the model outputs.
- The data will be anonymized and securely stored on an encrypted server, accessible only to the research team.
- The data will be retained for 5 years post-research for audit purposes, after which it will be permanently deleted.
- Aggregated results may be included in publications or conference presentations, ensuring no personal identifiers are disclosed.

Potential Risks and Benefits

- There are no known risks associated with participating in this study.
- Your contributions will support the development of a tool to better understand and address air pollution in Nairobi.

Confidentiality

Your identity and personal information will remain confidential. Any recorded audio or written feedback will be securely stored and destroyed once the research analysis is completed.

Consent Form

Title of the Research:

Density-Based Spatial Clustering for Uncovering Hidden Irregularities in Nairobi's Urban Air Pollution

Principal Investigator and Contact Information:

Name: Ruth Mwende Mavindu

Email: ruth.mwende@strathmore.edu

Institutional Contact:

Name: Strathmore University Research and Innovation Office

Email: research@strathmore.edu

Phone number: +254(0)703 034231

Introduction and Purpose of the Research

This study aims to uncover hidden irregularities in Nairobi's urban air pollution using density-based spatial clustering. Your participation is crucial to validate the models and provide expert feedback to improve their effectiveness.

Participation Activities

You will be involved in interviews and feedback sessions, as detailed in the information sheet above.

Potential Risks and Benefits

There are no known risks. Your participation will contribute to advancements in environmental research and urban pollution management.

Confidentiality

Your data will remain confidential and anonymized. All records will be securely stored and only accessible to the research team.

Compensation

Participation in this research is voluntary, and no monetary or material compensation will be provided.

Voluntary Participation and Withdrawal

Participation is voluntary. You may withdraw at any time without repercussions. Data collected prior to withdrawal may be retained unless you request its removal.

Cost/Reimbursements

There are no costs to participate, and no reimbursements will be provided for transportation or related expenses.

Participant Consent Statement

I, _____, voluntarily agree to participate in this research conducted by _____.

The research project has been explained to me, and I understand that it aims to uncover hidden irregularities in Nairobi's urban air pollution using density-based spatial clustering. I understand the details of my participation, including the confidentiality measures, risks, and benefits.

I understand that I will be given a copy of this signed consent form.

Participant Information:

Name: _____ Signature: _____ Date: _____

Witness Information:

Name: _____ Signature: _____ Date: _____

Researcher Information:

Name: _____ Signature: _____ Date: _____

Appendix F: Data Collection Tools

To validate the effectiveness and usability of the proposed model, a set of structured questions will be administered to domain experts in environmental sciences. These questions are designed to assess the model's functionality, usability, and relevance to the domain. The feedback will be collected through the prototype's Graphical User Interface (GUI) during interactions with the system.

Proposed Questions

Usability Assessment

i. How easy is it to navigate and use the prototype?
ii. Are there specific features or workflows that could be improved for better usability?

Functionality Validation

i. Are the outputs of the model consistent with domain-specific expectations?
ii. Does the model effectively solve the problem it is designed to address?
iii. Are there any missing functionalities or features you consider essential?

Relevance to Environmental Sciences

i. Does the model align with established practices and methodologies in environmental sciences?
ii. Are the results generated by the prototype relevant and actionable for real-world scenarios?
iii. What additional features or improvements would make the model more applicable to your work?

Overall Feedback

i. How would you rate the overall experience of using the prototype?
ii. Are there any specific challenges or barriers you faced while interacting with the system?
iii. Do you have any recommendations for improving the model's performance, usability, or applicability?

Appendix G: Dataset

Overview

This research will utilize a comprehensive dataset of spatial and temporal air quality measurements collected from monitoring stations across Nairobi. The dataset is crucial for identifying and interpreting anomalies in air pollution patterns and will include data from multiple validated sources.

Attributes

The dataset should contain the following attributes:

Attribute	Description	Unit
PM2.5	Particulate matter with a diameter ≤ 2.5 µm.	µg/m³
PM10	Particulate matter with a diameter ≤ 10 µm.	µg/m³
Temperature	Ambient temperature at the monitoring location.	°C
Humidity	Relative humidity at the monitoring location.	%
Timestamp	Date and time of the measurement.	ISO format
Sensor ID	Unique identifier for the sensor.	Alphanumeric
Latitude	Geographic latitude of the sensor location.	Decimal degrees
Longitude	Geographic longitude of the sensor location.	Decimal degrees

Appendix H: Repository for Source Code, Data, and Other Artifacts

The GitHub repository that will host the solution's source code, dataset, and relevant artifacts can be accessed here [https://github.com/rmavindu/Density-Based-Spatial-Clustering.] This repository will be kept updated throughout the research process.

Appendix I: Data Management Plan

To ensure the confidentiality of the data collected and allow future access by interested parties, the following measures will be implemented:

i. Confidentiality of Data:

Since the dataset used in this research will be publicly available, there are no specific confidentiality concerns regarding the data itself.

ii. Access to Data and Code:

Open Access: As the primary dataset will be publicly available, it will continue to be accessible to researchers through the original repository with links provided in the research outputs.

GitHub Repository: The source code and relevant artifacts (e.g., models, data preprocessing scripts) will be stored in a public GitHub repository. This repository will be regularly updated, and access will be granted to interested parties via the repository link. Instructions on how to access and use the code will be provided in the repository's README file.

Data Management Plan: A clear process will be established for researchers or other parties who wish to access the data and code after the completion of the research. The GitHub repository will include detailed documentation on how to request access to datasets or code if they are not directly available through public channels.

iii. Future Access and Sharing:

Interested parties can request access to additional resources or datasets by contacting the lead researcher via email, providing proper attribution, and outlining the purpose for which the data will be used.

Repository for Data and Code: The public GitHub repository will also include the research outputs, such as final models, evaluation metrics, and results, which will be available for download and further use by the community.

Open Access Platforms: The research findings, including the methodology, models, and results, will also be made available, where researchers can request access to specific data subsets or code if they are not available directly in the public repository.

Appendix J: Outputs Management Plan

To maximize the adoption and use of the research output by the industry, society, and other researchers, the following strategies will be implemented:

Industry Collaboration:

Collaborating with government agencies and private sector companies involved in air quality monitoring, such as Nairobi's Environmental Management and Coordination Department (EMCD) and World Health Organization (WHO), will help bridge the gap between research and real-world application.

The research output will be shared with these organizations for potential integration into existing air quality management frameworks and decision-making tools.

Open-Access Platforms:

To promote transparency and collaboration, the research code, datasets, and findings will be made available on open-access platforms like GitHub. This will allow other researchers to replicate the results and build upon the work.

The findings will also be shared through open-access repositories for early dissemination and to attract attention from the research community.

Appendix K: Study Results Dissemination Plan

1. Target Audience

The study findings will be disseminated to various stakeholders, including:

- **Government agencies** (e.g., Nairobi County Government, National Environment Management Authority - NEMA) for policy formulation.
- **Environmental organizations and NGOs** focused on air quality and public health.
- **Academic and research institutions** for further scientific exploration.
- **Local communities and the general public** to raise awareness of urban air pollution.

2. Methods of Dissemination

The study results will be shared through multiple channels:

a) Academic and Scientific Dissemination

- Publication in peer-reviewed journals specializing in environmental science, urban studies, and air quality research.
- Presentation at national conferences and workshops.
- Sharing findings with universities and research institutions to support further studies.

b) Open-Access Online Repository

- The research results, datasets, and analysis code will be made publicly available on an online repository (e.g., GitHub, Zenodo, or an institutional repository).
- This repository will allow other researchers and stakeholders to access, validate, and build upon the study.
- The repository will be regularly updated with relevant findings and methodological improvements.

c) Policy and Government Engagement

- Submission of policy briefs to relevant government agencies such as NEMA and the Ministry of Environment.
- Participation in stakeholder meetings and roundtable discussions with policymakers.

d) Public Awareness and Community Engagement

- Utilizing digital platforms (e.g., websites, social media, blogs) to share summaries and infographics of key results.
- Collaboration with media houses for articles or interviews on air pollution trends and implications.

3. Expected Impact

By disseminating the findings through these channels, the study aims to:

- Provide actionable insights for urban air pollution management in Nairobi.
- Influence policy decisions to mitigate air pollution hotspots.
- Encourage open science by making research data and results accessible for further analysis and improvements.
- Raise awareness among the general public on air quality issues.