# III. METHODOLOGY

A methodical integration of diverse datasets, including satellite observations, ground sensor data, and forecasted models, is necessary to investigate urban air quality and mobility. The COVID-19 pandemic provided a rare chance to investigate the effects of anthropogenic mobility decrease on pollution, highlighting the urgent need for interdisciplinary methods [1], [2]. This study employs a two-phase methodology: Phase One conducts a historical analysis based on observed data (2020–2022), and Phase Two projects future risk scenarios using forecasted data (2023–2025). Figure 1 illustrates the overall methodology flow adopted in this study.

Figure 1 Methodology Overview

## Phase One: Historical Analysis Using Observed Data (2020–2022)

### A. Data Acquisition and Preprocessing

We obtained four main datasets: Google's UAE Mobility Reports (2020–2022), OpenAQ ground sensor air quality measurements, satellite pollution rasters (NO₂, SO₂, aerosol indices), and preliminary regional mobility forecasts. Air quality and mobility data were cleaned, parsed, and time-standardized to monthly intervals. To ensure consistent EPSG 4326 projection, raster datasets were processed using geospatial libraries. Following spatial pollution monitoring frameworks [3], raster data were clipped to the UAE boundary and preprocessed to mask invalid pixel values. Table I summarizes the datasets utilized in Phase 1, including their key features and corresponding data sources:

TABLE I: Datasets Used for Phase One (Historical Analysis, 2020–2022)

|  |  |  |
| --- | --- | --- |
| Dataset | Features / Variables | Source |
| Google Mobility Reports | Workplace mobility, Retail mobility, Transit station mobility, Residential mobility trends | [10] |
| OpenAQ Air Quality Data | NO₂ concentration, PM2.5 concentration, Timestamp, Location (Latitude, Longitude) | [11] |
| Sentinel-5P Pollution Raster Data | NO₂ Vertical Column Density, SO₂ Density, Aerosol Index, Multiband raster images (2020–2022) | [12] |

### B. Feature Engineering and Geohashing

Normalized pollution indices, mobility change rates, and spatial interaction terms were produced by feature engineering. To facilitate effective spatial aggregation and analysis beyond localized urban zones, latitude and longitude coordinates were geohashed at a precision of 5–6 characters (about 1–5 km²) [4]. This geospatial discretization enhanced computational tractability and enabled scaled spatial joins to be performed throughout the modeling and simulation stages.

### C. Raster Difference Analysis and Validation

Validation of satellite pollution rasters across years employed pixel-level statistical analysis. The Mean Absolute Difference (MAD) computed quantifies the average magnitude of pixel-level differences between two raster datasets, disregarding directionality. It provides a robust measure of the overall discrepancy between observed and forecasted pollutant concentrations. The MAD is computed as:

Where and The corresponding pixel intensities in different years. Maximum Absolute Difference (MAXD) was also computed to capture extreme spatial changes [5]. In addition to reducing satellite noise, this ensured that the raster layers utilized for analysis caught actual atmospheric fluctuations.

### D. Clustering Analysis and Optimal Cluster Selection

After normalizing all variables, the mobility-pollution geohash features were subjected to K-Means clustering to reveal underlying urban trends. The optimal number of clusters was selected using the Elbow Method, minimizing within-cluster sum of squares and balancing model complexity and interpretability [6]. The formal expression for K-Means clustering is to minimize the variance inside each cluster, which is as follows:

### where represents the centroid of the cluster ​.

### E. Predictive Modeling with Linear Regression

OLS linear regression was used to estimate the annualized workplace mobility and NO₂ averages for 2020–2022. The fit was assessed using the coefficient of determination (R² score). OLS linear regression was used to estimate the annualized workplace mobility and NO₂ averages for 2020–2022. For Phase Two validation, baseline estimates were generated by extrapolating to 2023–2025.

### F. Simulation Framework and Boosting Techniques

We created a simulation framework with three synthetic scenarios to evaluate the model's resilience:

* **Synthetic Mobility Scenarios**  
  We introduced controlled variations in observed mobility data to simulate fluctuations in urban movement. The synthetic mobility Was generated by applying proportional adjustments:

Here, *δ* represents synthetic changes of ±10%, ±20%, and ±30%.

* **Spatial Perturbations**  
  To simulate real-world disturbances like zoning regulations or new infrastructure, we randomly reassigned the mobility values between nearby geohashes [7]. This examined how sensitive the model was to regional spatial changes.
* **Population Weight Perturbations**  
  To replicate demographic fluctuation, we used normally distributed noise to alter population weights (P).

Where *γ* follows a normal distribution, introducing realistic randomness.

We employed boosting approaches that approximated second-order interactions between geohash clusters, mobility, and pollution to improve the simulations. [8]. This allowed the framework to capture non-linear urban dynamics, improving its adaptability to complex real-world scenarios. Among the simulated models, the model that performed the best based on evaluation parameters such as mean absolute difference (MAD), standard deviation (σ), and geographic consistency with historical data was chosen for Phase Two forecasting and scenario risk evaluation.

## Phase Two: Forecast Scenario Analysis Using Projected Data (2023–2025)

### A. Forecast Data Integration and Preprocessing

As stand-ins for human activity, projected Ericsson Mobility statistics (GB/user/month) for 2023–2025 were used. Forecast rasters for NO₂, SO₂, and aerosol indices that were generated from satellites were loaded concurrently. Forecasted pollution rasters were matched with historical geohash zones created in Phase One, per best standards for spatial forecasting integration [1]. Table II summarizes the datasets utilized in Phase Two, including their key features and corresponding data sources

TABLE II: Datasets Used for Phase Two (Forecast Scenario Analysis, 2023–2025)

|  |  |  |
| --- | --- | --- |
| Dataset | Features / Variables | Source |
| Ericsson Mobility Forecast | Forecasted GB/user/month data for the Middle East and Africa, including the UAE | [13] |
| Sentinel-5P Forecasted Pollution Raster Data | Predicted NO₂ Vertical Column Density, SO₂ Density, Aerosol Index (2023–2025) | [12] |

### B. Raster Comparative Evaluation and Change Quantification

Forecasted raster values were compared against historical baselines using MAD and MAXD metrics. Additionally, the standard deviation (σ) was computed to measure the dispersion of forecasted pollution values across geohash zones relative to the mean. It provides insight into the spatial variability and uncertainty associated with future pollution patterns. A higher standard deviation indicates greater variability in predicted pollution intensities. The standard deviation is calculated as:  
This enabled magnitude and uncertainty evaluation of projected pollution changes [5].

### C. Visual Analytics and Interpretation

Visual analytics included time series plots, raster heatmaps, cluster overlays, and spatial standard deviation maps. These visualizations emphasized trends and variation among UAE cities [9]. Visual analytics included:

* **Time Series Plots**:  
  Mobility and composite pollution indicators are plotted annually (2020–2025) in comparison, with the observed and forecasted years separated.
* **Raster Heatmaps**:  
  Visualizations of pixel-level differences between historical and forecasted pollution levels [9].
* **Cluster-Enhanced Maps**:  
  K-Means clusters derived in Phase One were overlaid on forecast maps to assess differential urban risk evolution.
* **Standard Deviation Spatial Maps**:  
  Zones with the highest forecast variability were identified as potentially unstable environmental risk zones.

### D. Scenario Risk Evaluation

A methodology for evaluating risk was developed by integrating mobility growth rates, forecasted pollution deterioration, and variability measures. High-risk geohashes were flagged as priorities for urban mitigation strategies, aligning with recommendations from recent urban pollution scenario studies [2], [5].

### REFERENCES

[1] A. Al Jawarneh, et al., "AI-driven Urban Analytics: Data-Driven Models for Mobility and Pollution," *Future Internet*, 2023.

[2] O. Ghaffarpasand, et al., "The impact of urban mobility on air pollution in Kampala," *Atmospheric Pollution Research*, 2024.

[3] G. Gianquintieri, et al., "Atmospheric satellite validation strategies," *Atmosphere*, 2024.

[4] M. Morley, et al., "Improving spatial exposure models via geocoding," *Environmental Pollution*, 2016.

[5] S. Oh, et al., "Enhancing air pollution forecasts using data-driven deweathering," *Atmospheric Pollution Research*, 2024.

[6] R. Mulyana, et al., "K-Means Clustering for Source Attribution," *Environmental Monitoring and Assessment*, 2025.

[7] A. Cummings, et al., "Urban mobility and mobile source emissions," *Frontiers in Built Environment*, 2021.

[8] R. Dong, et al., "Air Pollution Risk Simulations Using Demographic Weighting," *Scientific Reports*, 2020.

[9] O. Obiefuna, et al., "GIS-based pollution mapping for urban management," *Environment and Ecology Research*, 2021.

[10] Google, "COVID-19 Community Mobility Reports," Google, 2022. [Online]. Available: https://www.google.com/covid19/mobility/. [Accessed: 26-Apr-2025].

[11] OpenAQ, "Open Air Quality Data Platform," OpenAQ, 2022. [Online]. Available: <https://openaq.org/>. [Accessed: 26-Apr-2025].

[12] Copernicus Open Access Hub, "Sentinel-5P Satellite Data," European Space Agency, 2022. [Online]. Available: https://scihub.copernicus.eu/. [Accessed: 26-Apr-2025].

[13] Ericsson, "Ericsson Mobility Report – Forecasts for Mobile Data Traffic," Ericsson, 2023. [Online]. Available: https://www.ericsson.com/en/reports-and-papers/mobility-report. [Accessed: 26-Apr-2025].