Abstract— High-resolution human mobility and air-quality datasets offer unprecedented opportunities to analyze anthropogenic pollution drivers at fine spatial scales, yet fusing heterogeneous streams—smartphone mobility traces, ground sensors, and satellite imagery—remains computationally demanding. In this paper, we develop a Python-based, two-phase framework that integrates Google Community Mobility Reports, OpenAQ ground NO₂ measurements, and Sentinel-5P satellite rasters on a unified 5-character geohash grid, enabling scalable spatial aggregation. In Phase I (2020–2022), feature engineering (mobility change rates, log-population density), K-Means clustering, and ordinary least squares regression uncover mobility–pollution regimes and quantify their linear interdependencies. In Phase II (2023–2025), three synthetic mobility scenarios—proportional shifts, spatial perturbations, and population-weight variations—are generated and fed into a GradientBoostingRegressor (100 estimators, max\_depth = 3, learning\_rate = 0.1) to capture non-linear mobility impacts on NO₂ concentrations. On held-out data, the Synthetic Mobility scenario achieves CV R² = 0.966 and Test R² = 0.959, reducing MAE to 1.0 × 10⁻⁵ and RMSE to 1.4 × 10⁻⁵—outperforming alternative strategies. This end-to-end pipeline produces fine-grained risk maps that empower urban planners and public-health officials to target air-quality interventions with minimal computational overhead.

Keywords— geohash; human mobility; air-quality forecasting; K-Means clustering; gradient boosting; scenario simulation.

INTRODUCTION

The rapid proliferation of smartphone-based mobility tracking and dense air-quality sensing has generated unprecedented volumes of georeferenced data, promising new insights into how human movement drives urban pollutant concentrations [1]. Yet these heterogeneous streams—from Google Community Mobility Reports through OpenAQ ground-station NO₂ measurements to Sentinel-5P satellite rasters—arrive in disparate formats and resolutions, complicating efforts to integrate them at fine spatial scales.

Natural experiments such as the COVID-19 lockdowns demonstrated the sensitivity of NO₂ levels to abrupt drops in workplace and transit activity [2], [3], revealing clear mobility–pollution couplings. However, most analyses to date have focused on historical correlation within coarse spatial aggregates or have employed linear models that fail to capture localized heterogeneity and non-linear interactions. Moreover, few end-to-end frameworks exist that both explain past pollutant dynamics and forecast future air-quality risk under varying mobility scenarios.

To fill these gaps, we develop a fully reproducible, two-phase Python pipeline that harmonizes multi-source mobility and pollution datasets on a common 5-character geohash grid, then applies clustering, regression, and boosting-based simulation to interpret past trends and project future hotspots. In Phase I (2020–2022), workplace mobility data, ground-station NO₂ readings, and satellite rasters are discretized into approximately 200 geohash zones; feature engineering (e.g., mobility change rates, log-population density) is followed by K-Means clustering to uncover mobility–pollution regimes and ordinary least squares regression to quantify their linear dependencies. In Phase II (2023–2025), three “what-if” scenarios—proportional mobility shifts, spatial perturbations, and population-weighted noise—are generated and fed into a GradientBoostingRegressor (100 trees, depth = 3, learning\_rate = 0.1) to capture non-linear mobility impacts on NO₂ concentrations.

The principal contributions of this work are:

1. A unified geohash-based data-fusion framework for high-resolution mobility and air-quality datasets;
2. A two-phase analytical suite combining clustering, OLS regression, and gradient boosting to both explain past pollutant dynamics and forecast future risk; and
3. Fine-grained risk maps under alternative mobility scenarios, enabling urban planners and public-health officials to target interventions where they are most needed.

The remainder of this paper is organized as follows. Section II surveys related work on mobility-driven pollution modeling and spatial forecasting. Section III details our data sources, preprocessing steps, and methodological pipeline. Section IV presents experimental results and scenario analyses. Finally, Section V concludes and outlines avenues for future research.

Reference:

[1] H.-Y. Liu, E. Skjetne, and M. Kobernus, “Mobile phone tracking: in support of modelling traffic-related air pollution contribution to individual exposure and its implications for public health impact assessment,” *Environmental Health*, vol. 12, no. 1, Nov. 2013, doi: <https://doi.org/10.1186/1476-069x-12-93>.

[2] M. Bauwens *et al.*, “Impact of Coronavirus Outbreak on NO 2 Pollution Assessed Using TROPOMI and OMI Observations,” *Geophysical Research Letters*, vol. 47, no. 11, Jun. 2020, doi: <https://doi.org/10.1029/2020gl087978>.

[3] F. Liu *et al.*, “Abrupt decline in tropospheric nitrogen dioxide over China after the outbreak of COVID-19,” *Science Advances*, vol. 6, no. 28, p. eabc2992, Jul. 2020, doi: https://doi.org/10.1126/sciadv.abc2992.

‌

‌

‌