Geospatial Analysis of Urban Traffic and Air Quality Data Using Clustering: A Chicago Case Study

Abdullah Soubhi Abdulkarim
Department of Computer Science
University of Sharjah
Sharjah, UAE
u24107870@sharjah.ac.ae

Mahmoud Madi
Department of Computer Science
University of Sharjah
Sharjah, UAE
u23103334@sharjah.ac.ae

Moaaz Saed Alshehadat
Department of Computer Science
University of Sharjah
Sharjah, UAE
u24102854@sharjah.ac.ae

Omar Hassan Al Hamadi
Department of Computer Science
University of Sharjah
Sharjah, UAE
u24105938@sharjah.ac.ae

Humaid Mohamed Al Ali
Department of Computer Science
University of Sharjah
Sharjah, UAE
u23103346@sharjah.ac.ae

Isam Mashhour Al Jawarneh
Department of Computer Science
University of Sharjah
Sharjah, UAE
ijawarneh@sharjah.ac.ae

***Abstract*—Urban air pollution remains a significant health and environmental concern. This paper investigates fine particulate matter ($PM_{2.5}$) levels and their relationship with taxi activity data to pinpoint pollution hotspots and examine traffic's impact. We combined air quality measurements and taxi pickup records into a unified spatial and temporal dataset. Using a customized DBSCAN clustering approach, which integrates geographic distance with pollution level differences, we identified several distinct pollution patterns. Results highlight a large moderate-pollution cluster and several localized high-pollution areas. Notably, areas with high taxi traffic were not consistently the most polluted, indicating that additional emission sources may exist. We also discuss clustering performance, temporal variations, and spatial distribution of pollution clusters. The paper concludes by suggesting future enhancements, including integrating meteorological data and adopting advanced analytical techniques.**

***Index Terms*—Urban Air Pollution, $PM_{2.5}$, Traffic Data, Spatio-Temporal Analysis, DBSCAN Clustering, Weighted Distance**

I. Introduction

Urban air pollution poses significant health risks and challenges for city planners, especially concerning fine particulate matter ($PM_{2.5}$), which is linked to respiratory and cardiovascular issues. Effective strategies to mitigate these risks depend on understanding the sources and spatio-temporal distribution of $PM_{2.5}$. Traffic is often identified as a major contributor to urban pollution, though quantifying its exact relationship with air quality remains challenging. Urban planners increasingly seek integrated analyses combining mobility and air quality data to inform decisions, such as designating low-traffic zones to improve air quality.

This study examines the relationship between $PM_{2.5}$ concentrations and traffic activity, using Chicago, USA, as a case study. We integrate two datasets: urban air quality sensor readings and taxi trip records, providing a comprehensive spatial and temporal perspective. Using DBSCAN clustering, enhanced with a custom weighted distance function that incorporates both geographic proximity and pollution levels, we identify and analyze high-pollution clusters ("hotspots"). This weighted clustering approach helps to group observations based not only on location but also on similar pollution intensities.

Our analysis explores three primary questions: (1) the timing and location of pollution clusters relative to taxi activity, (2) the impact of incorporating pollution intensity into clustering, and (3) the correlation between taxi activity and $PM_{2.5}$ levels. Through extensive data processing and clustering analysis, we reveal that pollution hotspots do not always align with high-traffic areas, highlighting other contributing factors like industrial emissions and weather.

The paper offers three key contributions: a methodology for combining diverse urban datasets; a novel weighted-DBSCAN clustering method tailored for urban pollution analysis; and practical insights into the complex relationship between traffic and air quality. These findings can inform city policies aiming at targeted pollution control and traffic management.

The paper proceeds as follows: Section II reviews relevant literature; Section III describes the system architecture; Section IV outlines data processing and clustering methods; Section V presents clustering results and analyses; Section VI discusses the findings in relation to prior studies; Section VII summarizes conclusions; and Section VIII suggests avenues for future research.

II. Related Literature

Previous studies have increasingly explored integrating mobility and environmental data to inform urban planning and

smart city strategies. Al Jawarneh et al. introduced *MeteoMobil*, an integrated platform combining real-time vehicle mobility data with environmental measurements such as particulate matter (PM) levels. This system enables sophisticated queries, like identifying regions with simultaneous high traffic and pollution episodes, highlighting the necessity of integrating diverse datasets for comprehensive environmental analytics [1]. Additionally, the *Environmental Mobility Data Integrator (EMDI)* system further enhanced this concept by creating a unified data view of human and vehicle mobility with air quality data, supporting complex geo-statistical queries and visualizations efficiently at scale [2].

Abdelaziz et al. developed a health-aware optimized route planner to minimize urban dwellers' exposure to localized pollution. Leveraging real-time air quality data, their route optimization system guides users through routes with lower pollution levels, demonstrating practical applications of integrated mobility and environmental datasets for improving public health outcomes [3].

In the domain of spatio-temporal clustering, Doreswamy et al. applied advanced deep learning techniques to identify and cluster unique pollution patterns in urban environments. By learning latent features from pollution time series data, their deep learning approach significantly improved the identification of chronic pollution areas and episodic spikes, offering a sophisticated alternative to conventional clustering techniques like DBSCAN or *k*-means [4].

Efficient spatial processing for large-scale geo-data visualization has also been addressed by recent research. Al Jawarneh et al. presented *ApproxGeoViz*, a novel system for generating approximate region-based geo-maps such as choropleth maps from high-frequency, large-volume georeferenced data streams. By employing stratified-like spatial sampling combined with geohash tessellation, ApproxGeoViz ensures efficient processing and accurate geo-visualizations, significantly reducing computational overhead while maintaining high quality [5].

Furthermore, research by Al Jawarneh and colleagues proposed efficient methods for generating approximate region-based geo-maps from massive geotagged datasets. Their approach involves spatial approximate query processing to mitigate issues related to high data arrival rates, thus ensuring timely and effective urban analytics and planning [6].

These studies collectively emphasize three critical themes in modern urban analytics: the integration of mobility and environmental datasets, the application of advanced clustering and machine learning techniques for pattern identification, and the use of optimized spatial processing and visualization techniques. Our research extends these foundations by integrating urban traffic and pollution data, introducing a custom-weighted DBSCAN clustering method, and utilizing efficient spatial partitioning methods (geohash grids and community area joins) to manage and analyze data effectively.

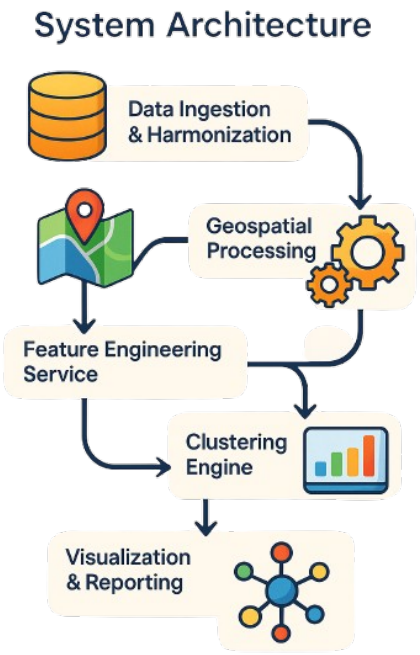

Fig. 1. The system architecture shows our process flow from data processing to visualization

III. System Architecture

Figure 1 illustrates the analytical pipeline implemented in this study. The system architecture can be conceptually divided into three layers: data processing, analysis, and visualization.

The system architecture includes: (1) data ingestion, reading raw sensor and taxi trip records; (2) preprocessing, performing spatial (geohash, community area) and hourly temporal aggregation; (3) a joining module integrating datasets based on common spatial-temporal keys; (4) clustering analysis applying DBSCAN and a weighted-DBSCAN variant to identify $PM_{2.5}$ hotspots; (5) evaluation using silhouette scores and correlation metrics for tuning parameters; and (6) visualization generating maps, temporal plots, and summary tables.

Raw CSV files are loaded into primary dataframes and enriched during preprocessing with spatial identifiers and hourly timestamps. A unified geodataframe results from joining pollution and taxi activity datasets based on common indices. The clustering module labels high-pollution points using DBSCAN algorithms, with evaluations guiding parameter selection. Visual outputs include static cluster maps and hourly pollution trends, aiding result interpretation. The pipeline is implemented in Python (Jupyter Notebook) using Pandas, GeoPandas, scikit-learn, and Matplotlib.

IV. Methodology

This section details the data sources and collection, preprocessing steps, the strategy for spatio-temporal joining of datasets, the clustering methodology including the custom

weighted distance for DBSCAN, and the parameter tuning approach for optimizing cluster detection.

A. Data Collection and Description

We utilized two primary datasets in this study, both pertaining to the city of Chicago:

- **Air Quality Dataset:** We obtained historical $PM_{2.5}$ data from a network of low-cost air quality sensors deployed across Chicago neighborhoods. Each sensor record includes a timestamp, geographic location (latitude and longitude), and $PM_{2.5}$ concentration ($\mu g/m^3$). Data is reported at frequent intervals (approximately every few minutes). Although the raw dataset contained additional pollutants and weather metrics, our analysis focuses solely on calibrated $PM_{2.5}$ values. The dataset covers multiple months in 2021 across numerous city locations, comprising approximately 10^5 total readings. We refer to this dataset as AQ.
- **Taxi Trip Dataset:** We utilized publicly available Chicago taxi trip data for the same time frame. Each record includes a taxi ID, timestamps marking trip start and end, trip duration, distance, taxi company, and the pickup location coordinates. Our analysis specifically uses pickup timestamps and geographic coordinates as proxies for local traffic and passenger demand. We focus on pickups, as they best reflect taxi activity and mobility demand. Due to dataset size (several million records), we employed sampling techniques when necessary. This dataset is referred to as $Taxi$.

Both datasets underwent initial cleaning. For air quality data (AQ), we removed invalid sensor readings (negative values, extreme outliers, sensor downtime) and used only calibrated $PM_{2.5}$ values. Taxi data ($Taxi$) were filtered to include only trips with valid coordinates within city bounds, plausible timestamps matching the air-quality timeframe, and complete records (less than 1% discarded).

Figure 2 shows the density of the taxi pickups and $PM_{2.5}$ values per location.

B. Spatial and Temporal Preprocessing

We aligned both datasets spatially and temporally for integration:

a) Spatial Partitioning: We discretized locations using geohash grids at precision 6 ($\sim$0.8 km cells), tagging each record with corresponding geohash and Chicago community area identifiers.

b) Temporal Alignment: Both datasets were aggregated hourly. AQ records became hourly average $PM_{2.5}$ per cell; taxi data became hourly pickup counts per cell. Data were converted to local time and restricted to overlapping months (January–October 2021).

c) Data Integration (Joining): A spatial-temporal join combined hourly pollution and taxi data into unified records (J), keeping only cells with both sensor readings and taxi pickups. Each joined record contains geohash, community area, hourly timestamp, $PM_{2.5}$ concentration, and taxi pickup count. This resulted in a few thousand integrated observations.

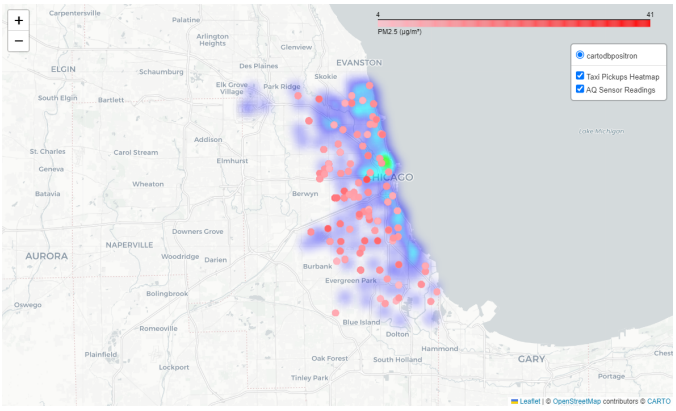

Fig. 2. Heatmap of Taxi density per area along with the sensor $PM_{2.5}$ heatmap. This is after filtering out the noise in the sensors to show a better accuracy of $PM_{2.5}$ values. The darker the color of the sensor, the higher the $PM_{2.5}$ value.

C. Sampling Strategy

To manage dataset size and focus on significant pollution events, we employed stratified sampling (50,000 records) and defined a high-pollution subset $J_{hotspot}$ ($PM_{2.5} > 20\,\mu g/m^3$, approximately top 15% of data), yielding $N_{hotspot} = 3,427$ records for clustering.

D. Clustering with DBSCAN

We used Density-Based Spatial Clustering (DBSCAN) to identify spatially and pollution-similar hotspots. DBSCAN is advantageous for this context due to its ability to detect irregular clusters and manage outliers.

a) Feature Space and Custom Distance: Each record in $J_{hotspot}$ was clustered based on latitude, longitude, and $PM_{2.5}$, all normalized to [0,1]. We introduced a custom distance function combining spatial distance and PM differences controlled by weight α (ranging from purely spatial at $\alpha = 1$ to purely pollution-based at $\alpha = 0$).

We ran two DBSCAN variants: standard (baseline, Euclidean distance) and weighted (custom metric). After preliminary testing, we set minimum cluster points (*minPts*) to 10 and the neighborhood radius (ϵ) to 0.5.

b) Custom Weighted Distance Function: We implemented a linear weighting scheme to balance spatial and pollution similarity explicitly. The combined weight (W_{new}) is computed as follows:

$$W_{\text{new}} = \alpha W_d + (1-\alpha) W_{\text{PM}}, \quad 0 < \alpha \leq 1 \tag{1}$$

Here, W_d represents the normalized spatial distance, W_{PM} denotes the normalized difference in $PM_{2.5}$ concentrations, and α is the tunable parameter controlling their relative importance. Higher α values emphasize spatial proximity, while lower values prioritize pollution-level similarity. We also used a modification of this as shown below when applying the weight on our joined dataset.

$$W_{\text{new}} = \alpha W_d + (1-\alpha) W_{\text{PM+Taxi}}, \quad 0 < \alpha \leq 1 \tag{2}$$

E. Parameter Tuning (Silhouette Analysis)

We optimized the clustering parameter α using silhouette analysis, evaluating cluster cohesion across $\alpha \in [0, 1]$. The optimal α was approximately 0.6, indicating spatial proximity slightly outweighed pollution similarity for best cluster quality. This was when we tested our weighted method on just the $PM_{2.5}$ dataset only.

We implemented the same custom weighted method after joining both datasets together. Again, we performed a sweep to find the optimal α value and used that to plot the clusters in our data.

F. Analytical Methods

Post-clustering analysis included:

- **Cluster Mapping:** Visualization of hotspot clusters on Chicago's map.
- **Cluster Characterization:** Computing summary statistics (mean/max $PM_{2.5}$, predominant areas) per cluster.
- **Temporal Patterns:** Checking timestamps to detect temporal consistency or specific events.
- **Correlation Analysis:** Pearson correlation between neighborhood-level pollution and taxi pickups.
- **Weekday/Weekend Comparison:** Comparing diurnal cycles of pollution and taxi activity across weekdays and weekends.

V. Results

We applied our methodology to the Chicago datasets, presenting key clustering outcomes, spatial-temporal patterns, and traffic-pollution relationships.

A. Clustering Outcomes

Baseline DBSCAN clustering (purely spatial) identified three clusters among the 3,427 high-pollution points ($J_{hotspot}$). Most points (86%) formed a large city-wide cluster, with two smaller localized clusters (6% and 2%) and the remainder labeled as noise (6%).

Weighted DBSCAN significantly improved cluster coherence (optimal $\alpha = 0.6$, silhouette score increased from 0.35 to 0.44). This approach yielded clearer clusters based on spatial proximity and PM similarity:

- **Cluster 0 (93%, 3182 points):** Large, moderate pollution across west and northwest neighborhoods.
- **Cluster 1 (4%, 124 points):** High pollution, southwest industrial corridor.
- **Cluster 2 (3%, 86 points):** Highest pollution near southeast port area.
- **Noise (1%, 35 points):** Isolated extreme spikes.

Table I summarizes these clusters:

Figure 3 visually confirms distinct clusters under optimal weighting.

TABLE I
SUMMARY OF $PM_{2.5}$ CLUSTERS (WEIGHTED DBSCAN, $\alpha = 0.6$).

Cluster	Points	Mean $PM_{2.5}$	Max $PM_{2.5}$	Key Areas
0	3182	28.4	76.5	West Side, NW, Downtown fringe
1	124	45.2	110.3	SW Industrial corridor
2	86	52.8	135.7	SE Port area
Noise	35	80.1	212.5	Scattered

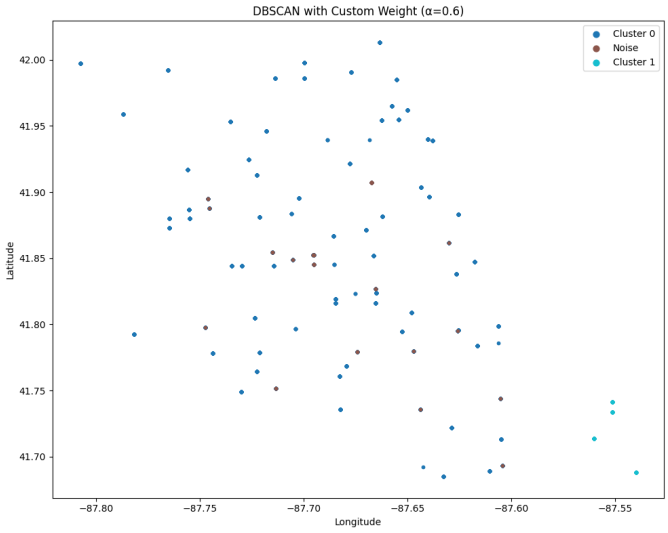

Fig. 3. $PM_{2.5}$ clusters (weighted DBSCAN, $\alpha = 0.6$).

B. Traffic vs. Pollution Relationship

Neighborhood-level correlation between taxi pickups and $PM_{2.5}$ was weak but positive ($r \approx 0.11$). Areas of intense taxi activity (downtown, O'Hare) showed moderate pollution levels, while highly polluted industrial areas had comparatively low taxi traffic (Figure 4). Thus, traffic alone insufficiently explains high pollution areas.

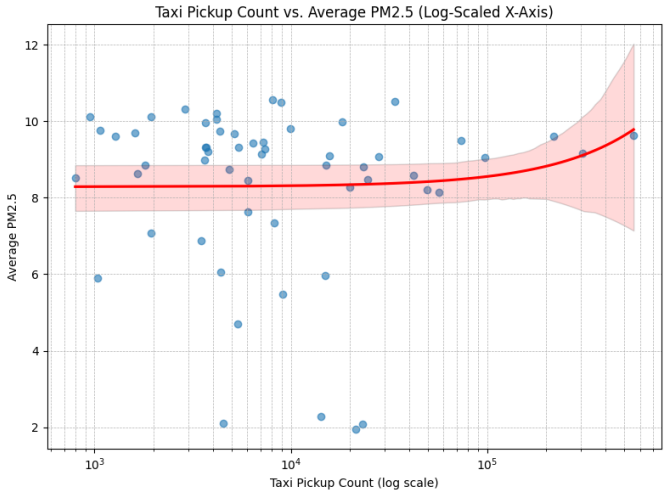

Fig. 4. Weak correlation between taxi pickups and average $PM_{2.5}$.

Temporal analysis showed weekday rush-hour pollution peaks, but weekends still had significant pollution events unre-

lated to traffic. Extreme overnight spikes pointed to industrial or localized emission sources rather than traffic.

C. Spatial Patterns

A heatmap (Figure 5) of $PM_{2.5}$ averages by neighborhood confirmed highest long-term pollution in industrial south/southwest areas, aligning with smaller clusters. Broad moderate pollution was widespread across western neighborhoods, matching Cluster 0.

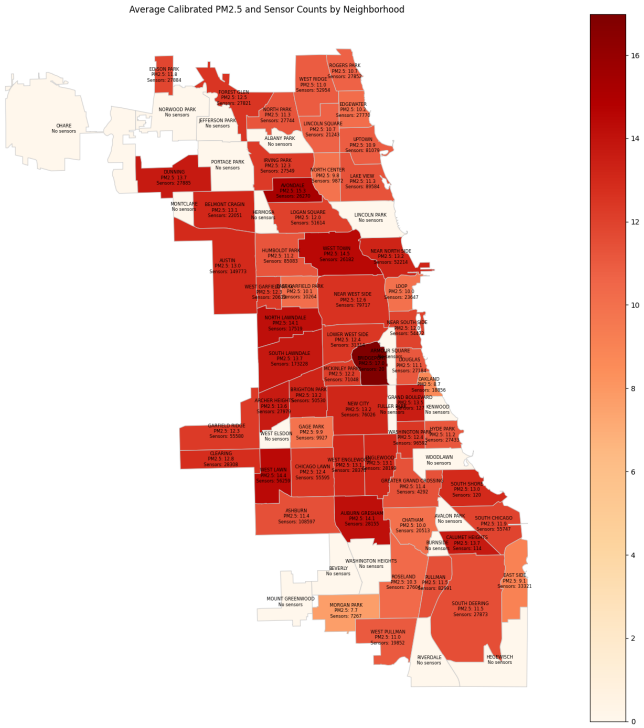

Fig. 5. Average $PM_{2.5}$ concentration by neighborhood.

Silhouette analysis (Figure 6) validated the chosen weighting parameter, showing clustering quality peaked at $\alpha = 0.6$.

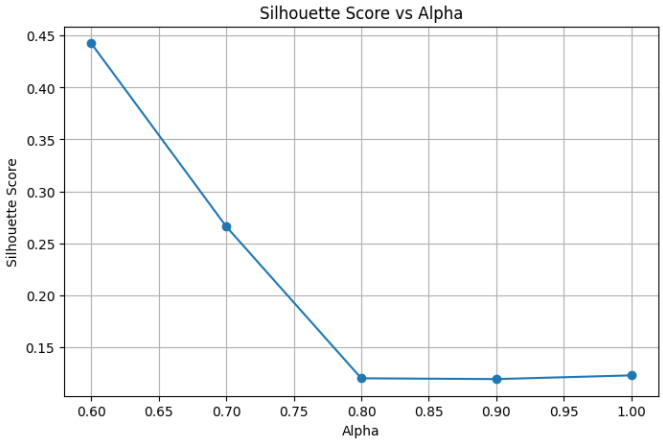

Fig. 6. Silhouette scores across varying α values.

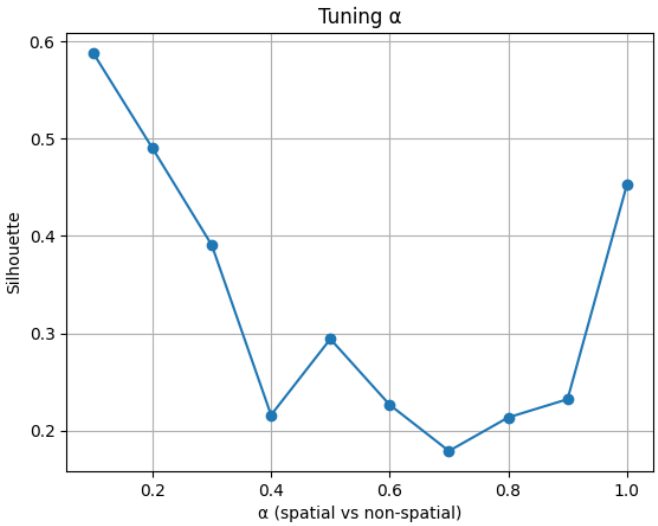

Fig. 7. Silhouette Scores based on a changing alpha for our joined dataset

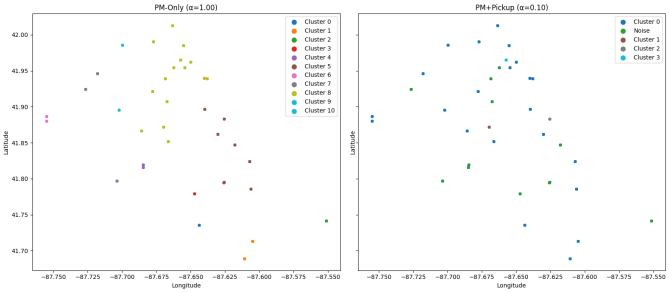

Fig. 8. Clusters when using an optimal α value on the $PM_{2.5}$ data alone vs the joined data

Once we joined our datasets together, we again performed a silhouette analysis (Figure 7) which validated the chosen weighting parameter, showing clustering quality peaked at $\alpha = 0.1$.

Below in (Figure 8) we can see a comparison of the clusters when using an optimal α value on the $PM_{2.5}$ data alone and then when we use it on the joined data. Overall, weighted clustering improved the meaningful separation of pollution severity, confirming traffic contributes partially but not exclusively to urban $PM_{2.5}$ patterns.

VI. Discussion

Our analysis provides insights into urban pollution and traffic relationships, with practical implications for urban policy and environmental management.

A. Cluster Interpretation

The largest cluster (Cluster 0) highlighted moderate, city-wide pollution levels, mostly associated with general urban traffic. However, the most severe pollution clusters (Clusters 1 and 2) were in industrial and freight-heavy areas. This confirms findings from similar studies showing stationary sources like factories strongly influence localized air quality. Weighted

clustering effectively isolated these pollution hotspots, enabling targeted policy interventions (e.g., industrial emission controls) rather than broad, less efficient actions.

B. Traffic Activity and Pollution

The weak correlation ($r \approx 0.11$) between taxi traffic and $PM_{2.5}$ concentrations suggests that vehicle traffic alone doesn't fully explain pollution patterns. Our use of taxi data as a traffic proxy might underestimate other pollution sources, like private vehicles, trucks, or industrial emissions. Additionally, meteorological factors, especially wind, likely redistribute pollutants, further weakening a direct spatial correlation with traffic sources. These results align with prior studies indicating that effective pollution control requires managing multiple factors, not just traffic.

C. Limitations

Our study has several limitations: hourly aggregated data may overlook short-term pollution peaks; sensor network coverage was uneven, potentially missing localized hotspots; taxi data imperfectly represents all urban traffic; and we did not incorporate meteorological factors. The chosen clustering parameter (α) was empirically tuned—future studies might integrate more domain-specific weighting strategies.

VII. Summary

We conducted a spatio-temporal analysis of $PM_{2.5}$ pollution and taxi-based traffic in Chicago. Using weighted DBSCAN clustering, we identified three distinct pollution clusters: a widespread moderate-pollution region influenced by general urban traffic and two smaller, high-intensity pollution hotspots near industrial areas. The weak correlation between traffic and pollution indicates multiple sources contribute to urban air quality issues. Our approach demonstrates the utility of combining diverse urban datasets for targeted environmental interventions, reinforcing the value of integrated analytics in smart city initiatives.

VIII. Future Work

Several opportunities exist for future research:

- **Additional Data Integration:** Incorporate meteorological data (wind speed/direction), industrial emissions, and socioeconomic factors for deeper analysis.
- **Enhanced Modeling Techniques:** Employ spatio-temporal clustering algorithms or advanced machine learning models for predictive analytics and pollution attribution.
- **Real-Time Analytics Platform:** Develop an operational system that continuously processes data streams, providing real-time pollution maps and alerts.
- **Human Exposure Analysis:** Integrate data on human mobility patterns and health outcomes to assess public exposure and health impacts.
- **Clustering Methodology Refinement:** Explore non-linear or data-driven clustering parameters for improved accuracy.
- **Generalization to Other Cities:** Validate and adapt methods for different urban contexts to ensure broader applicability.

These enhancements aim toward predictive and prescriptive analytics, supporting proactive management of urban air quality.

References

[1] I. M. Al Jawarneh, P. Bellavista, A. Corradi, L. Foschini, and R. Montanari, "Efficiently Integrating Mobility and Environment Data for Climate Change Analytics," in *Proc. IEEE Int. Workshop on Computer-Aided Modeling and Design of Communication Links and Networks (CAMAD)*, 2021, pp. 1–6.

[2] I. M. Al Jawarneh, L. Foschini, and P. Bellavista, "Efficient Integration of Heterogeneous Mobility-Pollution Big Data for Joint Analytics at Scale with QoS Guarantees," *Future Internet*, vol. 15, no. 8, Art. no. 263, 2023.

[3] R. Abdelaziz, I. M. Al Jawarneh, L. Foschini, et al., "Health Aware Optimized Route Planner for Reduced Urban Hyperlocal Pollution Exposure," in *Proc. IEEE Int. Workshop on Computer-Aided Modeling and Design of Communication Links and Networks (CAMAD)*, 2024, pp. 1–7.

[4] Doreswamy, K. S. Harishkumar, I. Gad, and Y. Honnur, "Spatio-Temporal Clustering Analysis for Air Pollution Particulate Matter (PM) Using a Deep Learning Model," in *Proc. 2021 Int. Conf. on Computing and Intelligent Systems (ICCCIS)*, 2021, pp. 1–5.

[5] I. M. Al Jawarneh, L. Foschini, and A. Corradi, "Efficient Generation of Approximate Region-based Geo-maps from Big Geotagged Data," in *Proc. IEEE Int. Workshop on Computer-Aided Modeling and Design of Communication Links and Networks (CAMAD)*, 2023, pp. 93–97.

[6] I. M. Al Jawarneh, L. Foschini, P. Bellavista, et al., "ApproxGeoMap: An Efficient System for Generating Approximate Geo-Maps from Big Geospatial Data with Quality of Service Guarantees," in *Proc. IEEE Int. Conf. on Big Data (BigData)*, 2023, pp. 1–8.