Optimizing Sensor Locations in Urban Environments: A Recommender System Based on Clustering and Multi-Source Data

Amna Alsuwaidi*, Fatima Alketbi*, Ayesha Aldoobi*, Alanood Almansoori*, Isam Aljawarneh*
*dept. of computer science, University os Sharjah, Sharjah, UAE
{U23200512, U24102845, U23107181, U24103123, ijawarneh}@sharjah.ac.ae

***Abstract*— *This paper explores the use of air quality and taxi mobility data to build a data-driven recommender system for optimizing sensor locations in metropolitan settings. By use of spatial-temporal clustering, this approach identifies locations of great environmental and population exposure risk by combining PM2.5 pollution levels with taxi trip densities. Using 5-hour intervals and geographically aggregated via geohashes, data from air quality sensors and taxi journeys was temporally synced. The paper is designed to depict the interaction between pollution intensity and human activity, a composite feature giving pollution ($\lambda = 0.6$) top priority. Using the Silhouette Score, clustering algorithms DBSCAN and K-Means were applied and assessed; DBSCAN outperformed K-Means (0.7900 vs. 0.4321) because it could represent irregular spatial patterns and filter noise. Extracted as ideal sensor sites, cluster centroids were tested against real air quality data to provide a mean absolute error of 4.98 µg/m³. Spatial trends were shown using visualization tools, including Folium and Seaborn, and stratified sampling guaranteed fair geographic representation in training and testing datasets. The paper results show that combining unsupervised learning with multi-source data produces a strong framework for evidence-based urban sensor deployment. The strategy is useful in promoting public health and planning decisions and improves urban environmental monitoring. Real-time integration and further data layers will be features of the next projects.*

Keywords*— *Urban sensing, sensor placement, recommender systems, clustering algorithms, IoT.

I. INTRODUCTION

Urban environments are increasingly affected by air pollution and high population density, which pose a significant risk to public health. Reducing these dangers depends on effective air quality monitoring [1]. This study presents a data-driven recommender system that integrates multi-source urban data, namely, PM2.5 concentration levels and taxi mobility data, to identify areas of high pollution and human activity, optimizing the location of air quality sensors [2]. The data obtained is compiled using geohash encoding and arranged into 5-hour time blocks using spatial-temporal processing methods to guarantee temporal consistency and meaningful daily segmentation [8]. A composite weighted feature gives air quality more weight as it directly affects public health through environmental and population exposure risk analysis. This feature set is examined using DBSCAN and K-Means, among other clustering techniques; DBSCAN shows better performance in spotting noise and irregular urban patterns. Low mean absolute error results from validation against ground-truth data and recommendations for ideal sensor sites based on centroids of resultant clusters [6]. Geohash-based stratified sampling and visualizing tools help to guarantee fair geographic representation and improve interpretability [10]. Combining real-world data, spatial analytics, and machine learning helps this system offer more exact, evidence-based sensor deployment techniques for urban air quality monitoring [9]. The approach has wider ramifications for the design of smart cities and might be developed going forward to include real-time data and other urban metrics.

II. RELATED LITERATURE

Many past studies focusing on sensor placement optimization have mostly concentrated on urban mobility or environmental monitoring. Often ignoring changing urban activity patterns, several studies use spatial interpolation or stationary models to find sensor locations [4]. Other researchers have used clustering methods for sensor deployment without including real-time mobility metrics such as taxi density. This study provides a more complete and context-aware method of sensor suggestion by aggregating air quality data with mobility patterns to quantify spatio-temporal exposure risk.

Clustering algorithms are critical to sensor network optimization, based on recent urban sensing study. The corresponding benefits of the DBSCAN and k-means algorithms for spatial data analysis in smart city applications are demonstrated by Al Jawarneh's work. According to the authors of a thorough review of crowd-sensing systems [1], k-means clustering provides better computational efficiency for large urban datasets, particularly when domain knowledge offers pre-defined numbers of clusters. DBSCAN is a density-based clustering algorithm that groups data points as a function of their spatial density. Key concepts: [11]. Core Point is A point with at least minPts neighbors within distance ε (eps). While Border Point is a point within ε of a core point but lacks sufficient neighbors and a Noise Point: A point not reachable from any core point. [6].

K-Means surpasses DBSCAN for datasets of well-separated, spherical clusters (e.g., Gaussian mixtures) [10]. DBSCAN works better with real-world data with noise and non-convex shape (e.g., geospatial data) [7]. Hybrid approaches (e.g., combination of K-Means and DBSCAN) have been considered to leverage both the strengths [8]. Therefore, future research should focus on improving the present system by including extra urban information such as weather conditions, transportation congestion levels, and

land use classifications to offer a richer analytical environment [5]. Additionally, real-time data should be included to provide adaptive sensor placement recommendations since urban dynamics are prone to change [7]. It is also crucial to note that extending the technology to assist other towns will test its scalability and generalizability [3]. Finally, working with local planning departments could allow field validation and actual implementation of the suggested sensor sites, bridging the gap between research and useful urban policy execution.

III. Methodology and System Architecture

This section describes the end-to-end methodology and architectural design employed to process, integrate, and analyze urban air quality and taxi trip data for clustering and sensor placement planning.

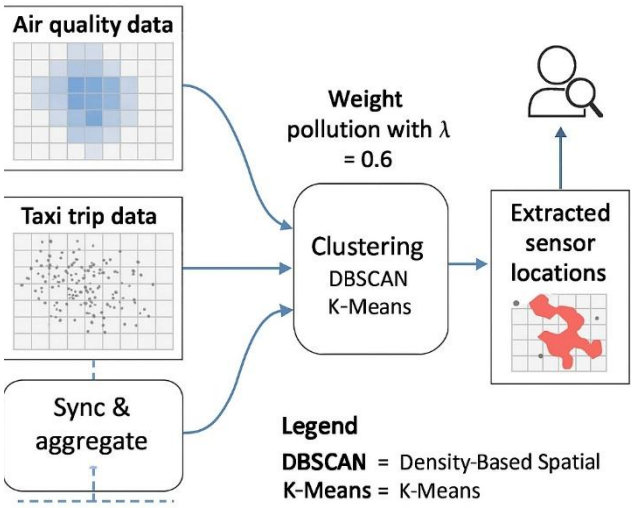

Figure 1. Methodology overview

A. System Setup and Data Loading

The project utilizes the PyGeohash library for spatial indexing, enabling efficient encoding and decoding of latitude and longitude coordinates into geohash strings. This facilitates spatial grouping and aggregation. Core Python libraries such as pandas, numpy, matplotlib, and scikit-learn were also imported for data manipulation, visualization, and machine learning.

The system loads:

- Air Quality (AQ) data from 19 CSV files, each containing readings of PM2.5 levels and corresponding locations.
- Taxi Trip data from a separate CSV file, representing pickup locations and trip start times.

Data from both sources are standardized for consistent column naming and validated using previews and record counts. This ensures clean integration for downstream processing.

B. Temporal Aggregation and Preprocessings

To align the datasets temporally, string-formatted timestamps are converted into Python datetime objects using pd.to_datetime(). Each timestamp is then rounded down to the nearest 5-hour window using .dt.floor('5H'). This temporal binning facilitates consistent time-based grouping and aggregation between AQ and taxi data.

The 5-hour window is selected based on three key factors:

1- Granularity vs. Sparsity Trade-off:

- Shorter windows, for example, 1 hour can result in excessive sparsity.
- Longer windows, for example, 24 hours obscure temporal variations.
- 5-hour blocks capture meaningful daily patterns without compromising data density.

2- Alignment with Urban Activity Cycles:

The 5-hour segments naturally divide the day into interpretable activity windows such as morning rush hour (5–10 AM), daytime (10 AM–3 PM), and evening peak (3–8 PM).

3- Merging Compatibility:

Ensures both AQ and Taxi datasets are aligned temporally for seamless spatial-temporal feature merging.

C. Spatio-Temporal Feature Engineering

Spatial aggregation is achieved by converting geographic coordinates to geohash strings (precision=5). This encodes latitude and longitude into a compact alphanumeric string representing spatial grids, suitable for grouping nearby locations.

For each geohash and 5-hour time block:

- Taxi Density is computed as the count of trips within the grid-time unit.
- Average PM2.5 values are calculated to capture local air pollution levels.

Both features are then merged into a unified dataset on geohash and time block, producing a spatio-temporal feature set suitable for clustering.

D. Weighted Feature Construction

A composite feature called weighted_feature is created to combine pollution (PM2.5) and mobility (taxi density) data. This is calculated as shown in equation (1):

$$\text{Weighted Feature} = \lambda * \text{PM2.5} + (1-\lambda) * \text{Taxi Density} \quad (1)$$

Where $\boldsymbol{\lambda = 0.6}$, prioritizing air quality for the following reasons:

1- Public Health Impact: Air quality has a stronger influence on human well-being than mobility.

2- Scale Adjustment: PM2.5 typically spans a narrower range than taxi counts; higher weighting corrects for this imbalance.

3- Analytical Focus: The project aims to explore how pollution patterns interact with human activity, hence prioritizing pollution.

Finally, geohashes are decoded back into latitude and longitude to retain geographic interpretability for visualization and mapping.

E. Clustering and Evaluation

To identify spatial clusters for analysis and potential sensor deployment, the following process is executed:

1- Feature Normalization:

Longitude, latitude, and the weighted feature are scaled to a [0, 1] range using Min-Max normalization to ensure fair distance comparisons during clustering.

2- Clustering Algorithms:

- DBSCAN (Density-Based Spatial Clustering of Applications with Noise): Detects arbitrary-shaped clusters and filters out noise using:
  - eps = 0.05: Radius for neighborhood definition in normalized space.
  - min_samples = 5: Minimum points to form a cluster.
- KMeans: Applied with n_clusters = 8, selected via experimentation. The parameter random_state = 42 ensures reproducibility.

3- Evaluation:

Clustering results are evaluated using the Silhouette Score, which measures how well-separated clusters are. Comparisons are made between DBSCAN and KMeans to validate quality.

4- Train-Test Split:

A stratified sampling strategy is used based on geohash values, ensuring that both training and testing datasets maintain a balanced spatial distribution.

- test_size = 0.3: 30% of data used for testing.
- random_state = 42: Ensures reproducibility.

F. Sensor Placement Strategy

To identify optimal locations for air quality sensors:

- Data is grouped by DBSCAN cluster labels, excluding noise (label = -1).
- For each cluster, the mean longitude, latitude, and PM2.5 values are computed to represent cluster centroids.
- These centroids are treated as candidate sensor placement sites and are exported to a CSV file (sensor_candidates.csv) for downstream deployment planning.

This strategy supports evidence-based urban planning, guiding sensor deployment in regions with distinct spatio-temporal pollution and mobility patterns.

IV. Results and Discussion

The paper outcome analysis examines a number of aspects. Initially, a composite feature integrating pollution intensity (PM2.5) and people mobility (taxi density) is constructed to represent regions with increased environmental and population exposure risk. Clustering methods, such as DBSCAN and K-Means, are then used and assessed using the Silhouette Score to identify spatial patterns and optimize sensor location. The chosen sensor sites are verified against actual air quality measurements to evaluate the accuracy of the suggestions. Subsequently, other geospatial datasets, such as public transportation, urban agriculture, and socioeconomic vulnerability, are included to enhance the contextual significance of each candidate site. Finally, interactive maps and stratified sampling techniques are used to illustrate and facilitate equitable geographical representation throughout the analysis.

A. Composite Feature Construction

A weighted feature was developed by integrating PM2.5 concentration with taxi density to include both pollution intensity and human mobility. A weight of $\lambda = 0.6$ was assigned, prioritizing air quality. This composite measure underpinned the clustering, guaranteeing that suggested sensor sites account for both environmental risk and population exposure.

B. Clustering Analysis and Evaluation

Clustering was performed using DBSCAN and K-Means on normalized geographic and weighted feature data. DBSCAN obtained a Silhouette Score of 0.7900, significantly outperforming K-Means, which achieved a score of 0.4321. DBSCAN effectively identified noise points and irregular cluster shapes, making it more appropriate for complex urban spatial patterns. To enhance clustering performance, both methods were tuned and visualized using Silhouette Score graphs as shown in Figure 2 and Figure 3. The plots validated DBSCAN's superior performance, with a maximum score of 0.7905 at eps = 0.04, in contrast to K-Means' highest score of 0.4578 at k = 10. A train-test split was applied using geohash-based stratification, resulting in 9,933 training samples and 4,258 testing samples, hence assuring geographic consistency for assessment.

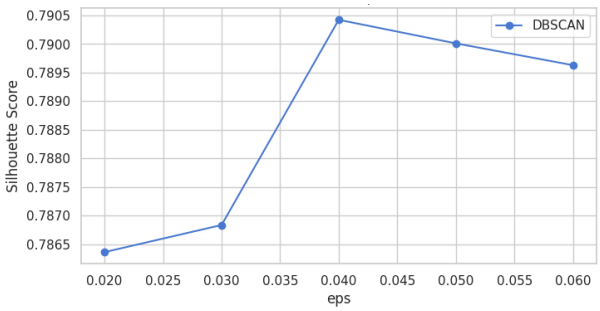

Figure 2. Silhouette Score vs. eps for DBSCAN

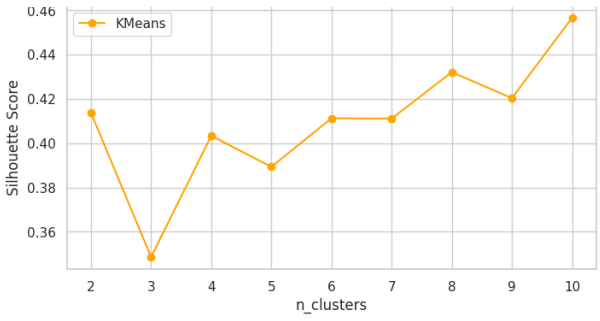

Figure 3. Silhouette Score vs. Number of Clusters (k) for K-Means

C. Sensor Recommendation and Validation

Sensor locations were selected by calculating the centroids of DBSCAN clusters, which signify regions with elevated pollution and mobility density. Outliers were eliminated to guarantee reliable suggestions. To evaluate precision, the PM2.5 value of each centroid was compared with the closest ground-truth station. The system attained a Mean Absolute Error (MAE) of 4.98 µg/m³, demonstrating a robust correlation between projected and actual air quality values.

D. Visualization of Sensor Distribution

The recommended sensor locations were interactively plotted using Folium, as seen in Figure 4. Each blue marking represents a cluster centroid obtained from DBSCAN, indicating regions of elevated pollution and mobility density. This map enables stakeholders to visually evaluate spatial coverage across the city. In addition, cluster comparison plots using Seaborn have been generated to demonstrate the grouping of geographic data by each method as seen in Figure 5. The DBSCAN plot seen in Figure 5(a) shows clusters of varying shapes and densities, along with identified outliers (labeled as -1), showcasing their adaptability in representing urban complexities. In contrast, the K-Means plot in Figure 5(b) displays uniformly distributed, circular clusters that overlook outlier behavior and spatial irregularities. This visual data supports the conclusion that DBSCAN is better suited for identifying meaningful zones in heterogeneous city environments.

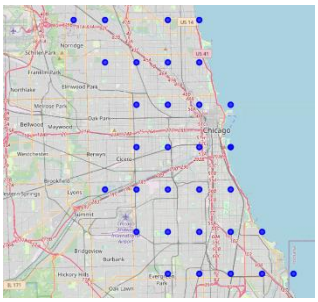

Figure 4. Recommended sensor locations across Chicago, mapped using Folium.

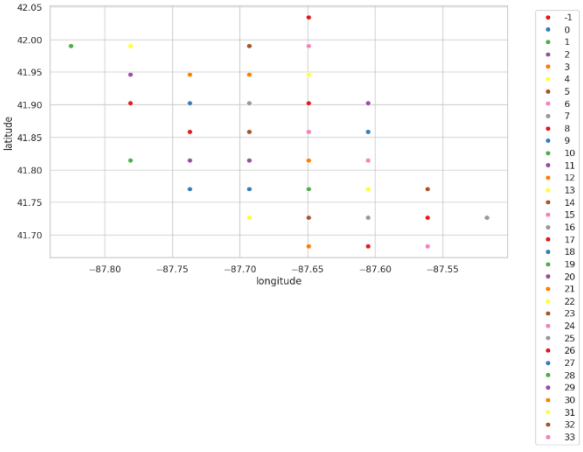

Figure 5(a). DBSCAN Clustering Results

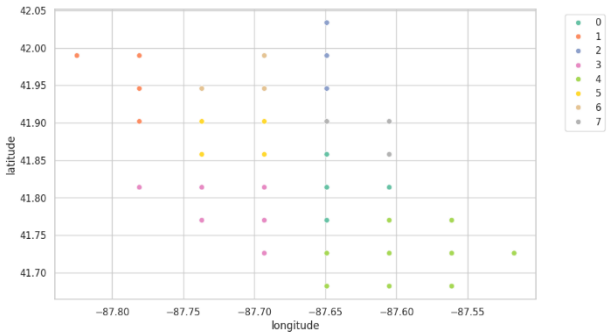

Figure 5(b). K-Means Clustering Results

E. Spatial Stratified Sampling

A stratified sampling approach based on geohash was implemented to ensure spatial fairness and mitigate geographic bias in model evaluation. Each geohash, denoting a particular geographical cell, contributed 30% of its data to the final sample, so assuring proportionate representation of both high-density metropolitan areas and outlying regions. Figure 6(a) demonstrates the entire dataset before sampling. The data points show wide distribution across Chicago, with clear clustering in certain regions for example, downtown and near the lakefront). Figure 6(b), illustrating the sampled dataset, indicates that the general geographical pattern is maintained. Despite a reduced data amount, each geographic location persists in contributing data points, therefore preserving the structural variety of the original dataset. The preservation of spatial heterogeneity in training and testing datasets is essential for enhancing generalizability, preventing model overfitting to overrepresented or centralized regions, and facilitating fair evaluation of clustering, validation, and enrichment processes across varied urban settings. The final dataset split included 9,933 training samples and 4,258 testing samples, providing a balanced foundation for further analysis without compromising geographic coverage.

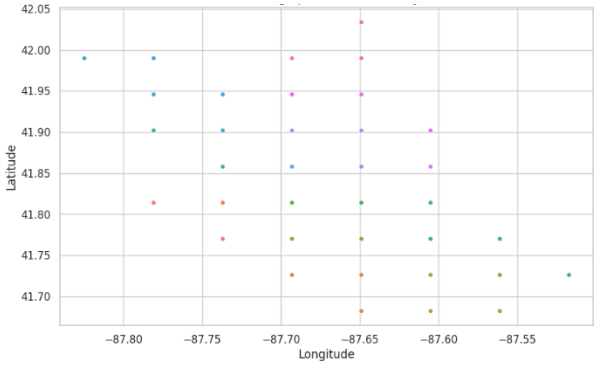

Figure 6(a). Full dataset - Geographical distribution by geohash

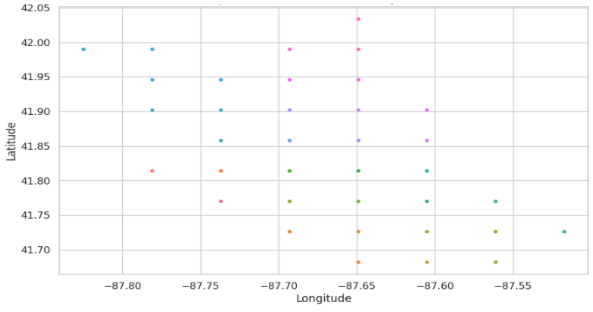

Figure 6(b). Sampled dataset - 30% stratified by geohash

V. Summary and Future work

To summarize our proposed paper approach that recommend combining PM2.5 pollution and taxi mobility data in urban areas, it'll indicate a data-driven method for suggesting the locations of air quality sensors. DBSCAN and K-Means were used for spatial clustering, and a hybrid weighted feature was built with pollution severity as a priority ($\lambda = 0.6$). DBSCAN did well with a Silhouette Score of 0.7900, cutting outliers and spotting tricky space groups. At check-up, sensor spots, group centers, had a mean absolute error of 4.98 $\mu g/m^3$ against true data. The model results and area spread were shown with tools like Seaborn and Folium. Striped sampling made sure every place was shown well during learning and checking.

Future improvements to our work may include:

- Putting sensors where needed using live city data for adaptive sensor placement.
- Adding extra city facts like weather, traffic jams, and land use for deeper info.
- Growing the system to new cities to check if it can work well and be used widely to test scalability.
- Working with city plan groups to roll out and test sensor spots in real-life to support real-world deployment and validation.
- Development of advanced sensing technologies to improve data accuracy and resolution.

References

[1] Reem Abdelaziz Alshamsi, A. Jawarneh, L. Foschini, and A. Corradi, “ApproxGeoMap: An Efficient System for Generating Approximate Geo-Maps from Big Geospatial Data with Quality of Service Guarantees,” Computers, vol. 14, no. 2, pp. 35–35, Jan. 2025, doi: https://doi.org/10.3390/computers14020035.

[2] A. Jawarneh, L. Foschini, and Paolo Bellavista, “Efficient Integration of Heterogeneous Mobility-Pollution Big Data for Joint Analytics at Scale with QoS Guarantees,” Future internet, vol. 15, no. 8, pp. 263–263, Aug. 2023, doi: https://doi.org/10.3390/fi15080263.

[3] A. Jawarneh, Paolo Bellavista, A. Corradi, L. Foschini, and R. Montanari, “Efficiently Integrating Mobility and Environment Data for Climate Change Analytics,” pp. 1–5, Oct. 2021, doi: https://doi.org/10.1109/camad52502.2021.9617784.

[4] I. M. Al Jawarneh, P. Bellavista, A. Corradi, L. Foschini, and R. Montanari, “QoS-Aware Approximate Query Processing for Smart Cities Spatial Data Streams,” Sensors, vol. 21, no. 12, p. 4160, Jun. 2021, doi: https://doi.org/10.3390/s21124160.

[5] G. Wang, Y. Wang, Y. Li, and T. Chen, “Identification of Urban Clusters Based on Multisource Data—An Example of Three Major Urban Agglomerations in China,” Land, vol. 12, no. 5, pp. 1058–1058, May 2023, doi: https://doi.org/10.3390/land12051058.

[6] H. Fu, W. H. K. Lam, H. Shao, W. Ma, B. Y. Chen, and H. W. Ho, “Optimization of multi-type sensor locations for simultaneous estimation of origin-destination demands and link travel times with covariance effects,” Transportation Research Part B Methodological, vol. 166, pp. 19–47, Oct. 2022, doi: https://doi.org/10.1016/j.trb.2022.10.006.

[7] J. Xing, W. Wu, Q. Cheng, and R. Liu, “Traffic state estimation of urban road networks by multi-source data fusion: Review and new insights,” vol. 595, pp. 127079–127079, Feb. 2022, doi: https://doi.org/10.1016/j.physa.2022.127079.

[8] G. Zhao et al., “Location Recommendation for Enterprises by Multi-Source Urban Big Data Analysis,” IEEE Transactions on Services Computing, pp. 1–1, 2017, doi: https://doi.org/10.1109/tsc.2017.2747538.

[9] F. Yang, Y. Hua, X. Li, Z. Yang, X. Yu, and T. Fei, “A survey on multisource heterogeneous urban sensor access and data management technologies,” Measurement: Sensors, vol. 19, p. 100061, Feb. 2022, doi: https://doi.org/10.1016/j.measen.2021.100061.

[10] D. Roy and M. Dutta, “A systematic review and research perspective on recommender systems,” Journal of Big Data, vol. 9, no. 1, May 2022, doi: https://doi.org/10.1186/s40537-022-00592-5.

[11] M. ESTER ET AL., "A DENSITY-BASED ALGORITHM FOR DISCOVERING CLUSTERS IN LARGE SPATIAL DATABASES WITH NOISE," PROC. 2ND INT. CONF. KNOWL. DISCOV. DATA MIN., PP. 226–231, 1996.

[12] R. J. G. B. Campello et al., "Density-based clustering based on hierarchical density estimates," Proc. Pacific-Asia Conf. Knowl. Discov. Data Min., pp. 160–172, 2013.

[13] L. Xu and M. Wunsch, "Survey of clustering algorithms," IEEE Trans. Neural Netw., vol. 16, no. 3, pp. 645–678, 2005.