Pakistan
Air Quality
Initiative

Population-Centric Optimization of Air Quality Monitoring Networks in Data-Sparse Urban Regions:

A Weighted K-Means Approach

29 April 2025

EGU 2025 Session AS5.9: Low-cost air quality sensors:
Challenges, opportunities, and collaborative strategies across the world

Abid Omar and Mahad Naveed
Pakistan Air Quality Initiative (PAQI پاکی) provides community-driven air quality research, data and resources to increase social awareness.
abidomar@pakairquality.com

EGU General Assembly 2025

Population-Centric Optimization of Air Quality Monitoring Networks

Data scarcity in resource limited regions.	Air quality monitoring networks are often sparse, unevenly distributed, or entirely absent. **Limited resources** and **data scarcity** in the Global South are challenges for evidence-based air quality management. where high population densities coincide with significant air pollution challenges.
Monitoring is critical infrastructure.	**Monitoring is critical infrastructure** for assessing population exposure to harmful pollutants, evaluating regulatory compliance, and informing policy decisions. This monitoring gap is particularly evident in rapidly growing urban centers like Lahore (Pakistan), Lagos (Nigeria), and Dhaka (Bangladesh), where high population densities coincide with significant air pollution challenges.
LCS is a solution for monitoring gaps.	The emergence of low-cost sensors (LCS) offers potential solutions to these monitoring gaps, particularly in resource-constrained settings, **but optimal spatial distribution** remains a critical research question.

Population-Centric Optimization of Air Quality Monitoring Networks

Optimal spatial distribution remains a question in data scarce regions.

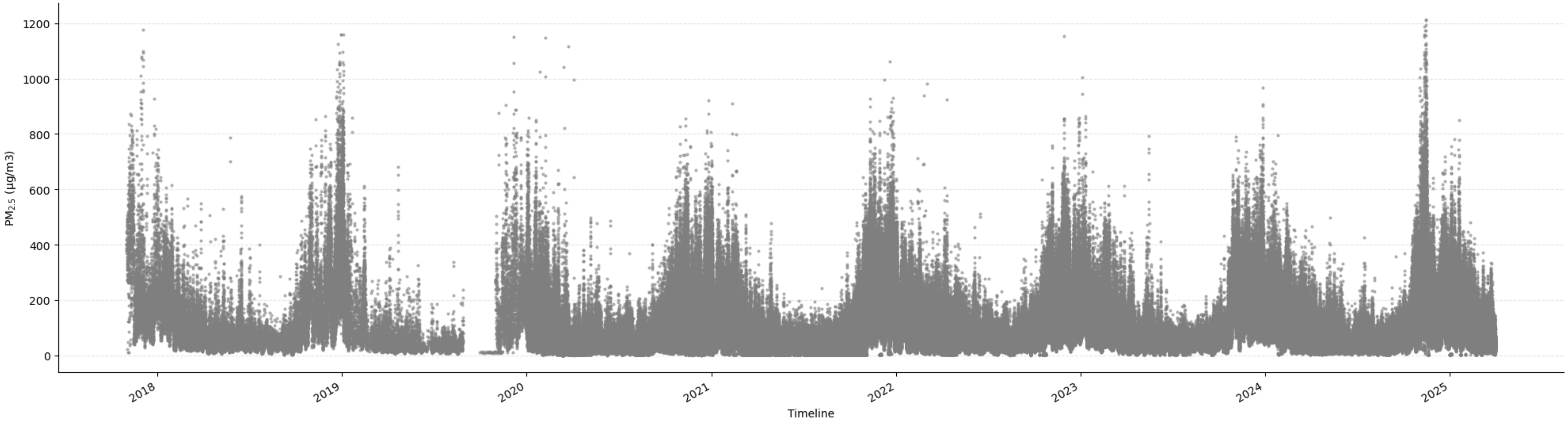
Develop a methodology to optimize monitor placement, improving data representativeness for population exposure:

1. Effectively **integrating population data** into air quality monitoring network design
2. Optimal **spatial distribution** of monitors in high- versus low-population density urban areas
3. ‘Validate’ by comparing with existing networks in ‘data rich’ regions.

Key Research Objectives

1. How can population data be effectively integrated into air quality monitoring network design?
2. What is the optimal spatial distribution of monitors in high versus low population density urban areas?
3. How do population-optimized monitoring networks compare to current configurations?

224.3 µg/m3 average PM2.5 in Lahore during the 2024 ‘smog season’.



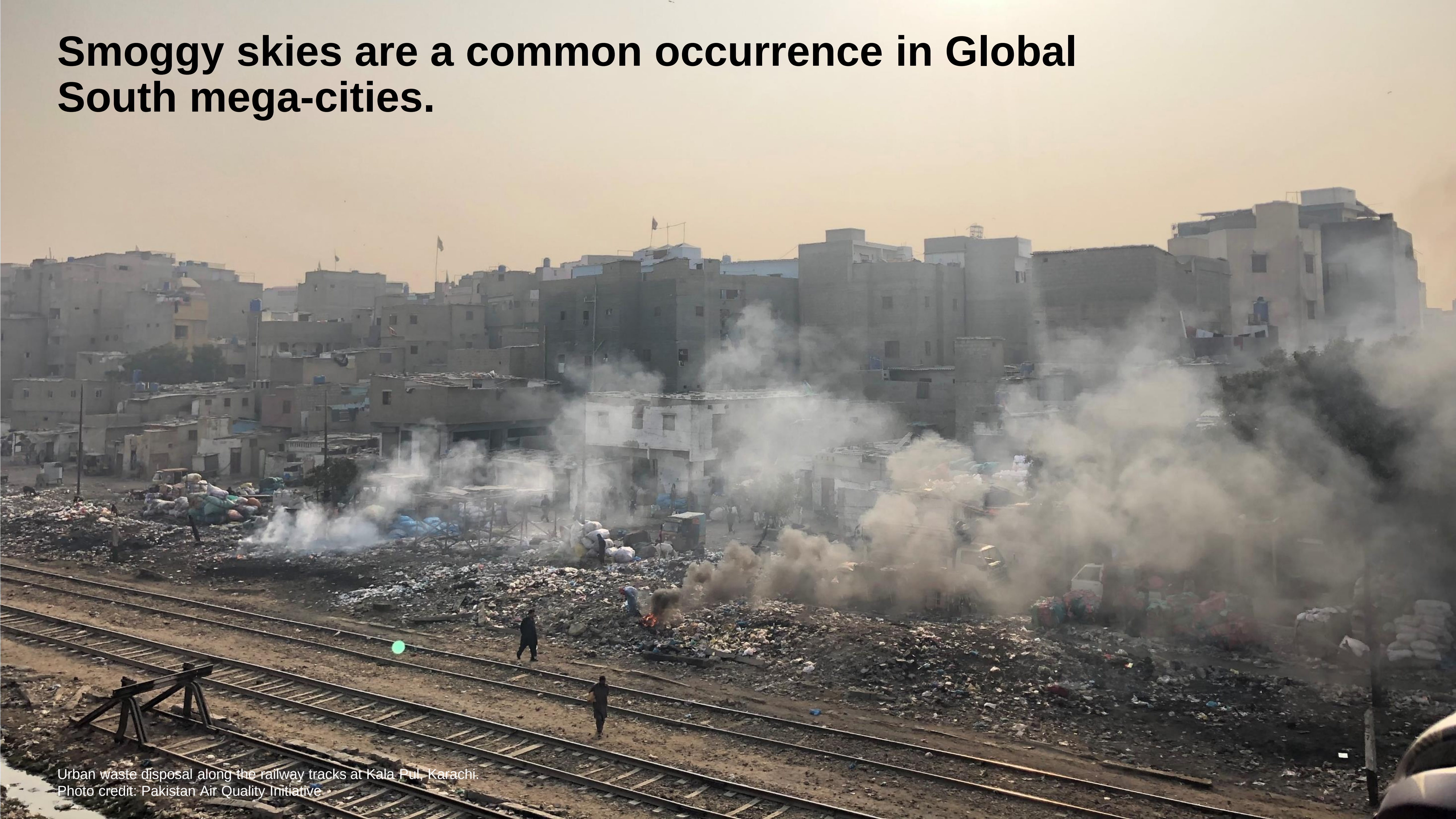
Lahore, hourly average time series of PM2.5 µg/m^3 conc. from 2018-2025.

Seasonal peaks in Lahore seen consistently since 2016 when the first air quality monitors were deployed. Fine particulate matter (PM2.5) concentrations frequently rise over 800 µg/m^3, well beyond WHO guidelines.

The WHO air quality guideline (AQG) states that annual average concentrations of PM2.5 should not exceed 5 µg/m3, while 24-hour average exposures should not exceed 15 µg/m3 more than 3-4 days per year.

Data source: Pakistan Air Quality Initiative

Smoggy skies are a common occurrence in Global South mega-cities.



Urban waste disposal along the railway tracks at Kala Pul, Karachi.
Photo credit: Pakistan Air Quality Initiative

Existing methods for air quality monitoring network design have limited applicability in data-scare regions.

Existing methods	Often focus on **pollution variability** (Kanaroglou et al., 2005), **spatial coverage**, or **cost** (Romero et al., 2020). Challenges exist in data-limited regions (Gupta et al., 2018).
Gap	Population exposure is often underrepresented. Population-weighted assessments differ significantly from spatial averaging (Weichenthal et al., 2015).
Low-Cost Sensors (LCS)	Offer potential solutions for expanding networks in resource-constrained settings, but optimal placement is key.
Our Approach	Building upon previous weighted clustering approaches, we adapt K-means clustering algorithm for data-sparse urban regions by integrating high-resolution population density data and using geospatial considerations.

- Kanaroglou et al. (2005) proposed a **location-allocation** model for optimizing a network of monitors, **using pollution variability** as the primary criterion.
- Romero et al. (2020) applied **multi-objective optimization** to balance spatial coverage, temporal representativeness, and cost considerations in network design.
- Gupta et al. (2018) highlighted the challenges of limited existing data for network optimization and proposed alternative approaches utilizing **land use and emission inventory** data.
- Weichenthal et al. (2015) demonstrated that **population-weighted exposure** assessments can differ substantially from assessments based on spatial averaging alone.

Existing research shows that population density is associated with pollution exposure.

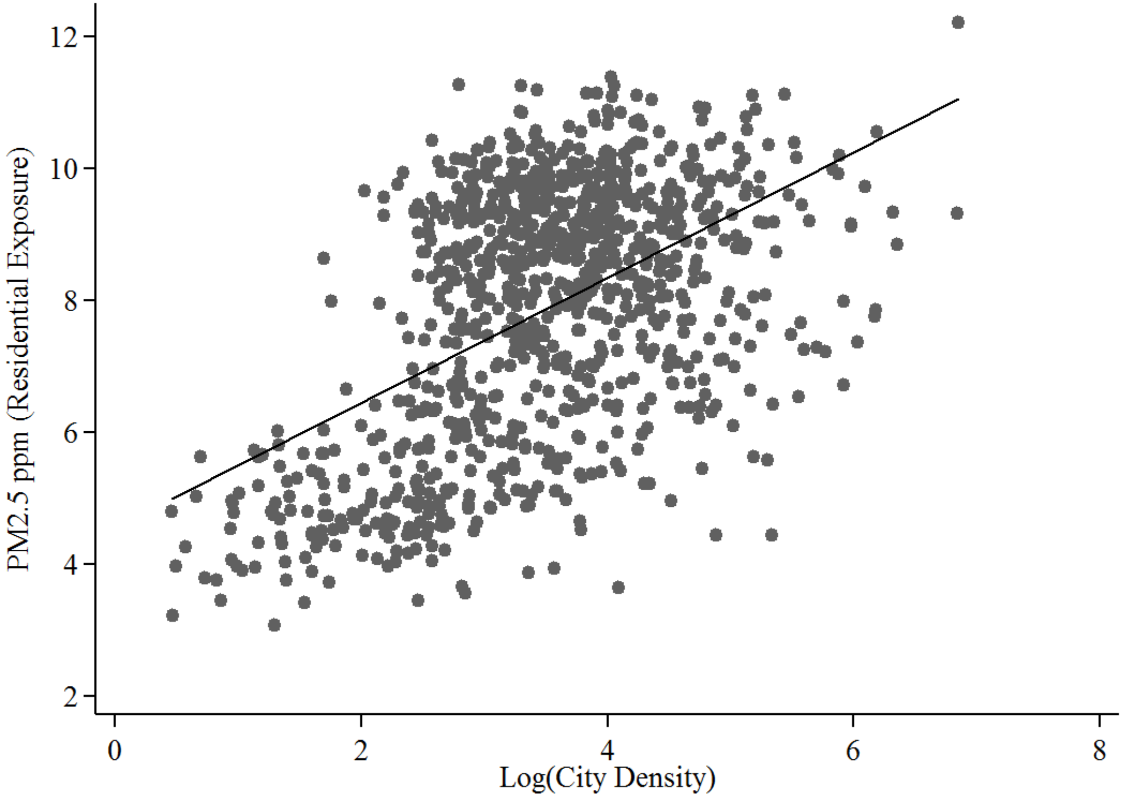
Pakistan Air Quality Initiative

Dirty Density: Air Quality and the Density of American Cities

Denser cities are associated with lower emissions, however the pollution exposure is higher.

LSE THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

CENTRE *for* ECONOMIC PERFORMANCE



The vertical axis represents PM2.5 average residential exposure (in µg /m3), as obtained from the satellite-derived measures. The horizontal axis represents the natural logarithm of population density. The points represent 933 CBSAs (metro and micropolitan areas). The black line is estimated by OLS using the underlying data.

Carozzi, F.: Dirty Density: Air Quality and the Density of American Cities. CEP Urban and Spatial Programme Blog (formerly SERC), 29 August 2019. https://spatial-economics.blogspot.com/2019/08/dirty-density-air-quality-and-density.html

Methodology: A Weighted K-Means algorithm to bias towards the ‘heavier’ point.

Pakistan Air Quality Initiative پاکی

Standard K-Means

The standard K-means algorithm computes the mean distance of x and y coordinates.

Partitions data into K clusters, minimizing within-cluster variance. Centroids are calculated as the simple mean of points in a cluster.

Limitation:

Treats all locations equally, regardless of population, can misplace monitors away from high-exposure areas.

The resulting centroid (monitor location) is influenced by the geometry of the cluster rather than where people actually live.

In areas with large rural or low-density zones, this can lead to misplaced monitors, far from where exposure is highest.

Weighted K-Means:

Utilizes a center-of-gravity integration to be more inclined towards the ‘heavier’ or more ‘dense’ point.

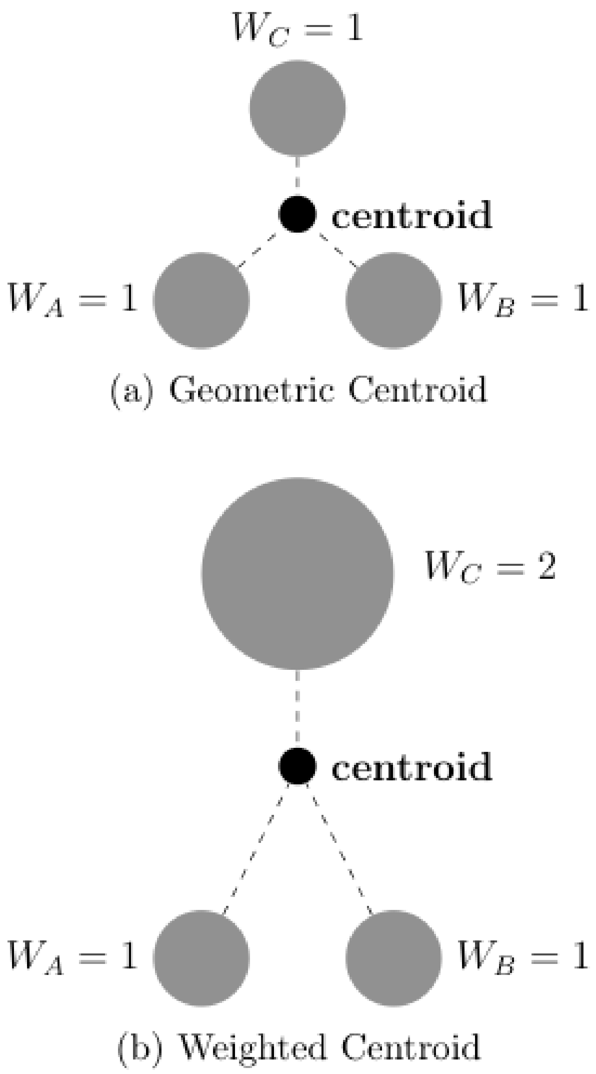
Assigns a weight (population count) to each grid cell, centroids are calculated as the weighted mean (center of gravity).

Prioritizes areas with higher population density.

This shifts the centroid closer to more populated areas within a cluster.

Reflects where people are most affected by air pollution, prioritizing human exposure.

Monitors placed using this method are more representative of real-world exposure patterns.



(a) Geometric Centroid

(b) Weighted Centroid

Methodology: Consider geospatial distances using the Haversine formula.

Standard K-Means

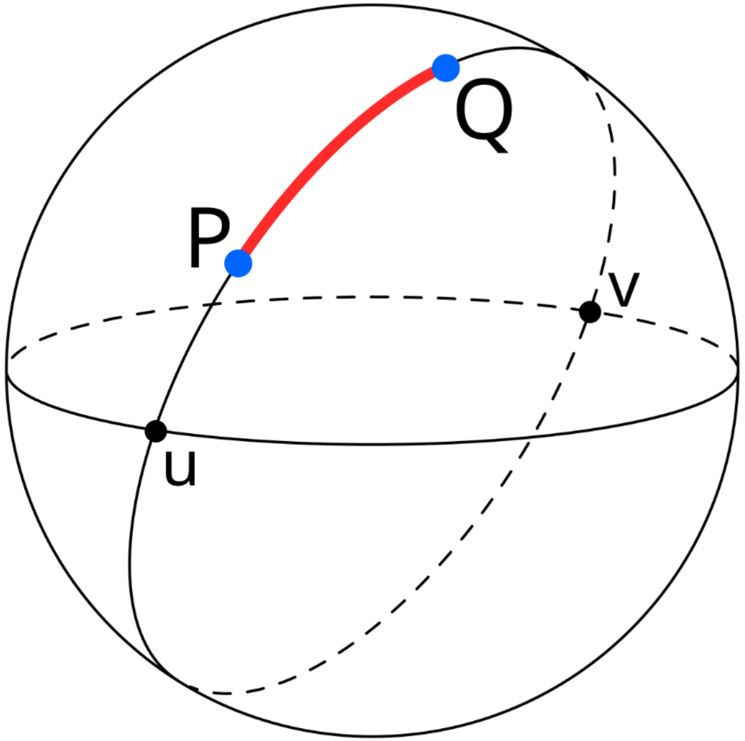
Standard K-means uses Euclidean distance, unsuitable for geographic coordinates (latitude/longitude) on a spherical surface.

Partitions data into K clusters, minimizing within-cluster variance. Centroids are calculated as the simple mean of points in a cluster.

Weighted K-Means:

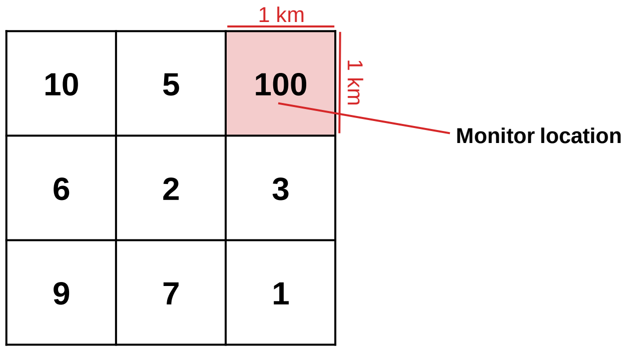
Use the Haversine formula to accurately calculate great-circle distances between points on Earth.

Integrate Haversine distance into the K-means algorithm's distance calculation step.



Methodology: Data & Workflow

Data Acquisition	**High-resolution population data** (100m): WorldPop (Constrained individual countries 2020 UN adjusted). **Administrative Boundaries**: Humanitarian Data Exchange (Shapefiles).
GIS Processing (QGIS)	**Define airshed** using a 0.01^{circ} x 0.01^{circ} grid. **Calculate zonal statistics**: Aggregate population counts per grid cell. Assign longitude/latitude to each grid cell.
Population Categorization	**Run standard K-means** (K=2) on population density per grid cell to define 'Low' and 'High' density zones. **Visualize low/high density** areas.
Weighted K-Means Application	Apply custom Weighted K-Means (using Haversine distance) separately to Low and High density zones. Determine number of monitors (K) based on target average population per cluster (e.g., 250k-350k for low-density, 500k-600k for high-density). Monitors located within X km of each other are merged by averaging their coordinates to form a single location. Algorithm iterates (max 300 times) until convergence.



10	5	100
6	2	3
9	7	1

Figure: Each grid cell represents population counts. Monitor location is closer to the grid cell with the highest sum of population count.

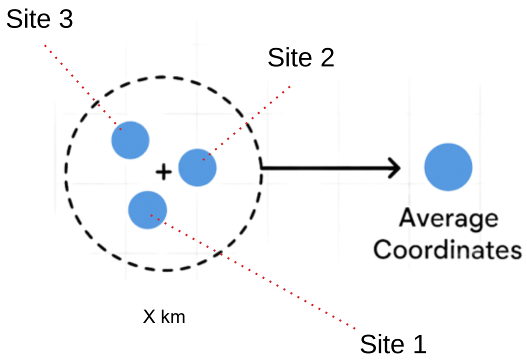


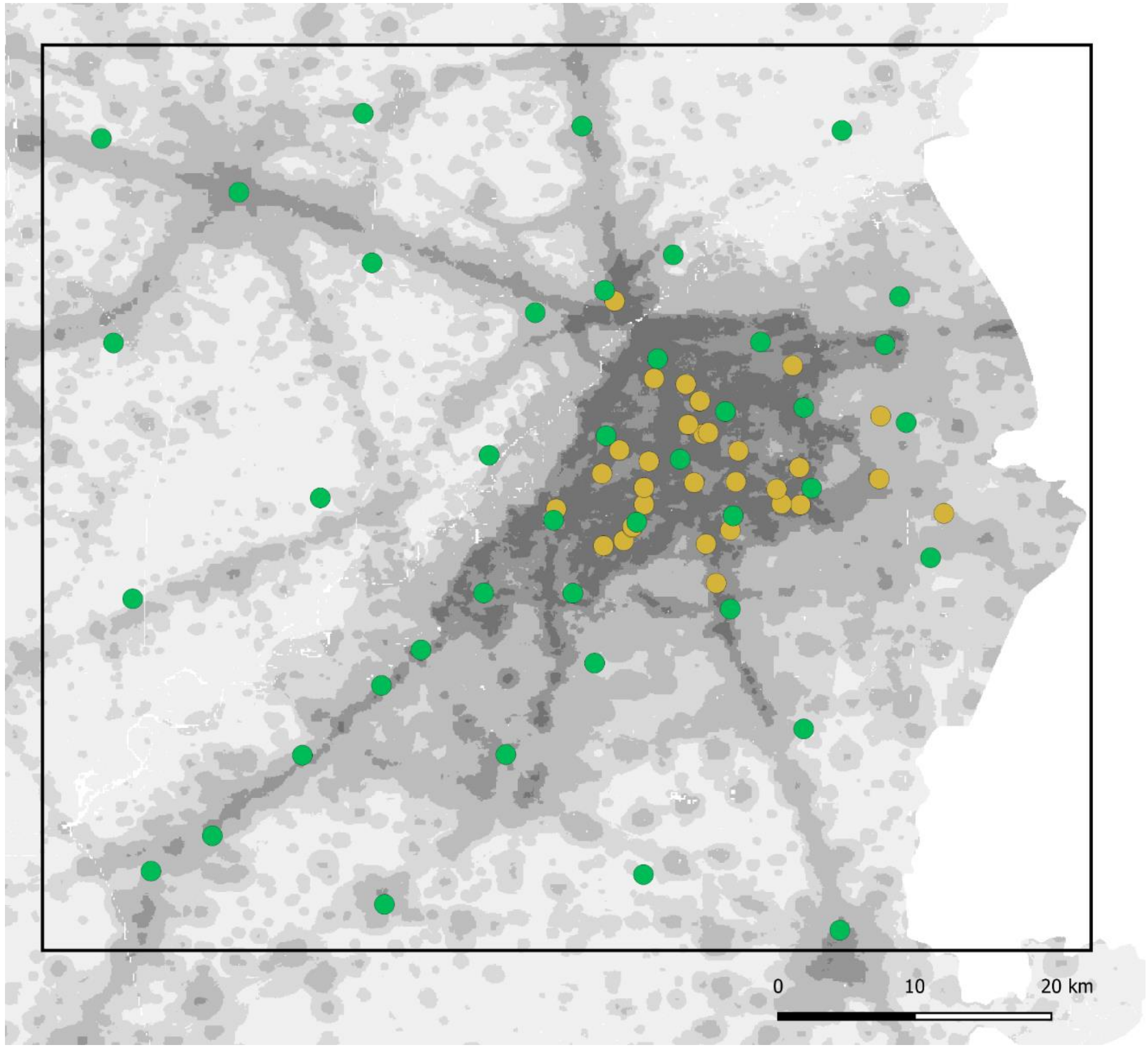
Figure: Monitors located within X km of each other are merged by averaging their coordinates to form a single location.

Results: Lahore shows an improved spatial spread, reflecting population weighting.

Pakistan Air Quality Initiative پاکی

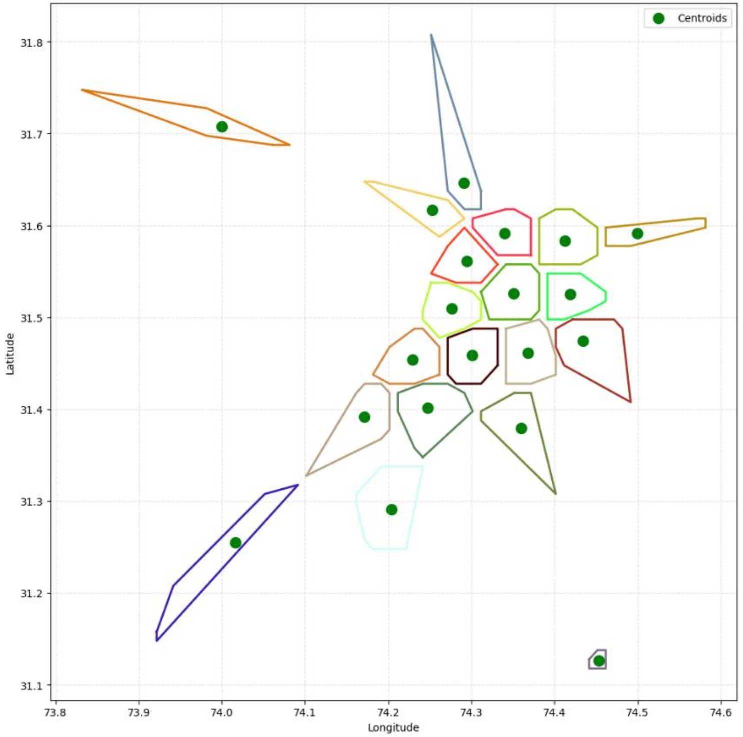
Population	**15.9 million**
Airshed	**70 x 81 km**
Number of monitors	**45** 25 low-density, 20 high-density
Population per monitor	**353,333**

- **Improved spatial spread** (green), as compared to existing ad-hoc placements (yellow) typical in data-scarce regions.
- Monitors are placed more sparsely in low-density regions and concentrated in densely populated zones, reflecting population weighting.

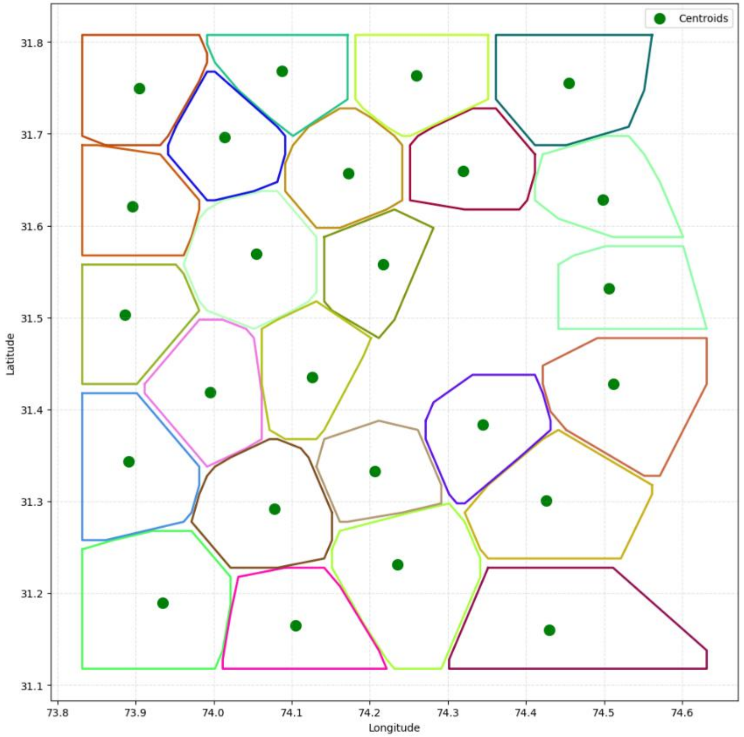


Green are suggested monitor locations overlaid on population density. Yellow is the existing ad-hoc low-cost monitoring network.

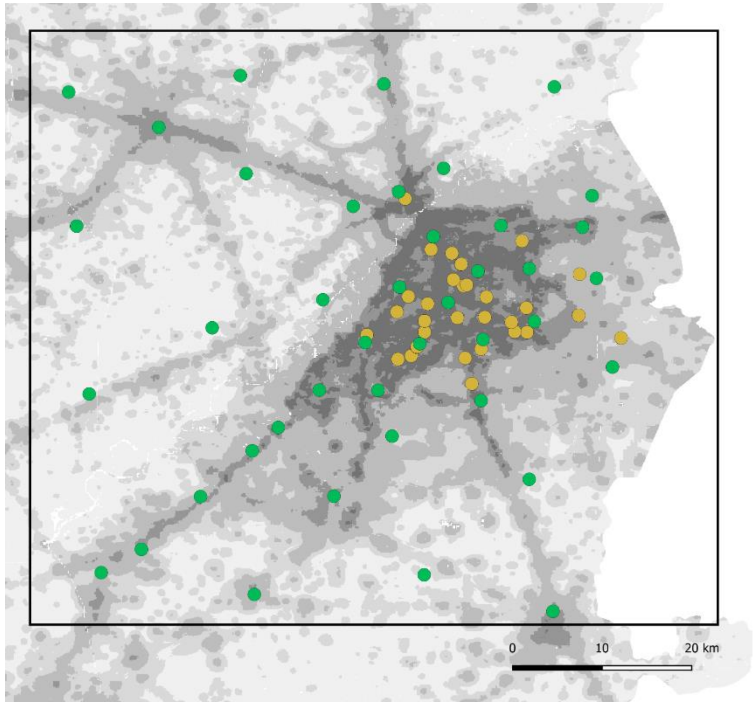
Results: Lahore shows an improved spatial spread, reflecting population weighting.



High-density clusters in Lahore



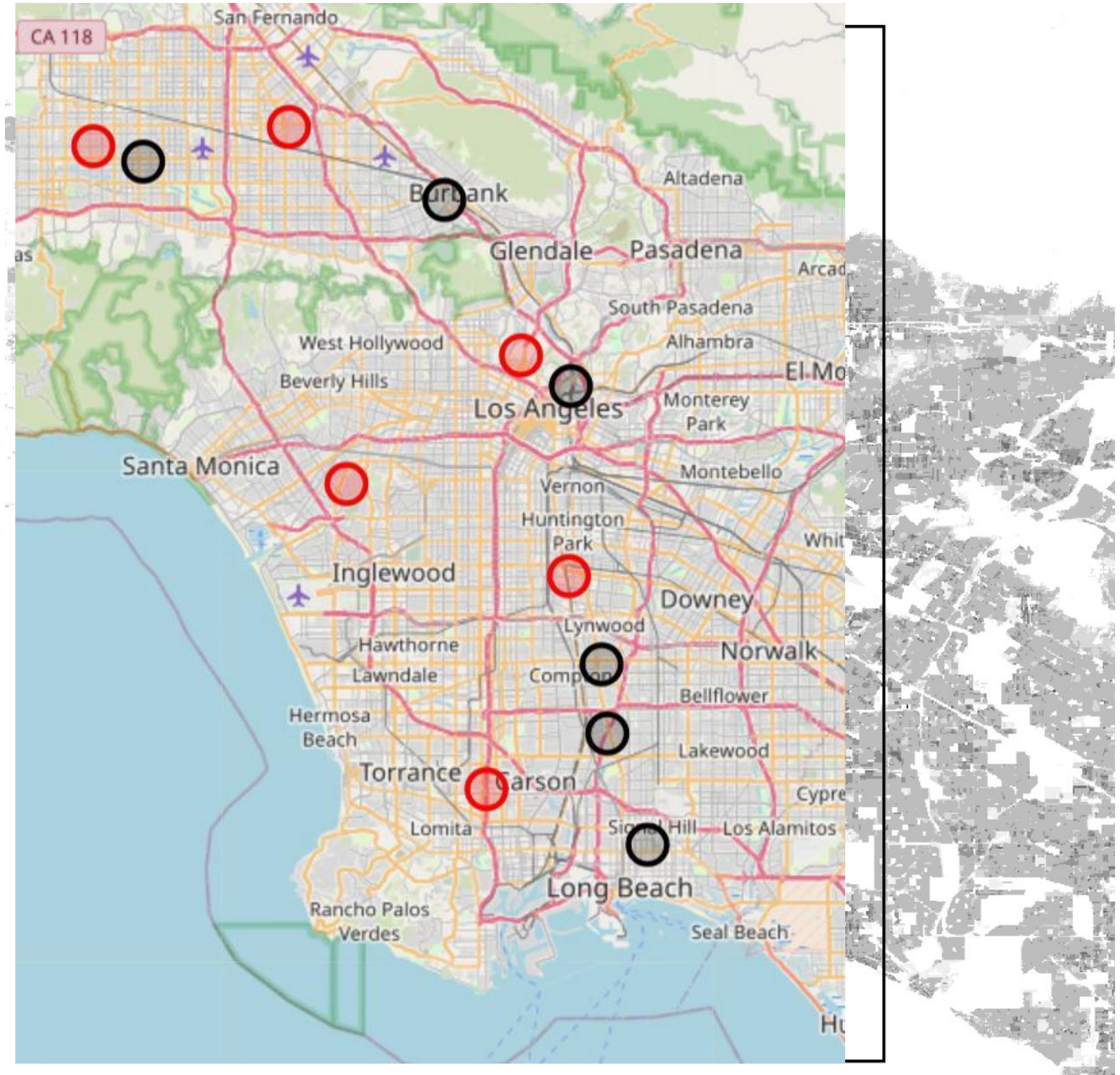
Low-density clusters in Lahore



Green are suggested monitor locations overlaid on population density. Yellow is the existing ad-hoc low-cost monitoring network.

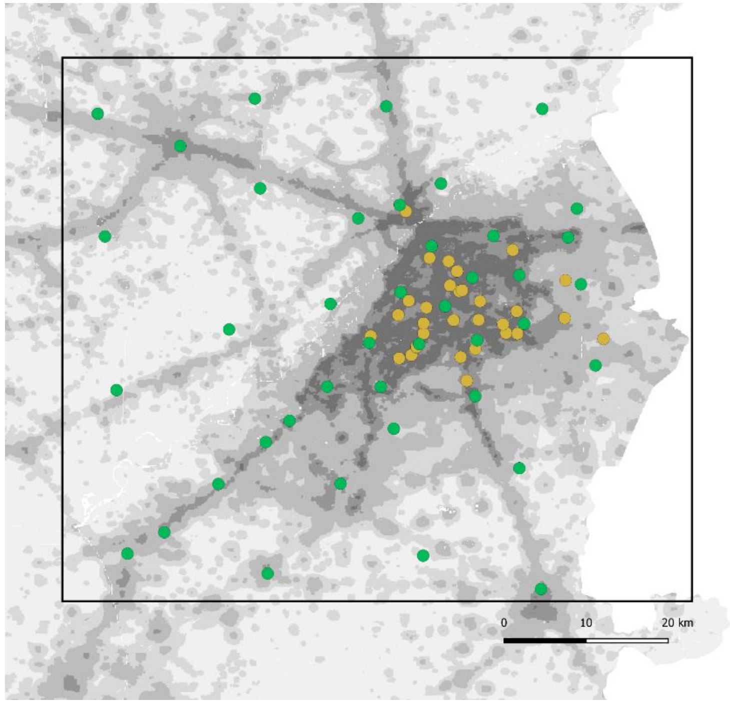
Results: Los Angeles monitor placement with consideration of population density.

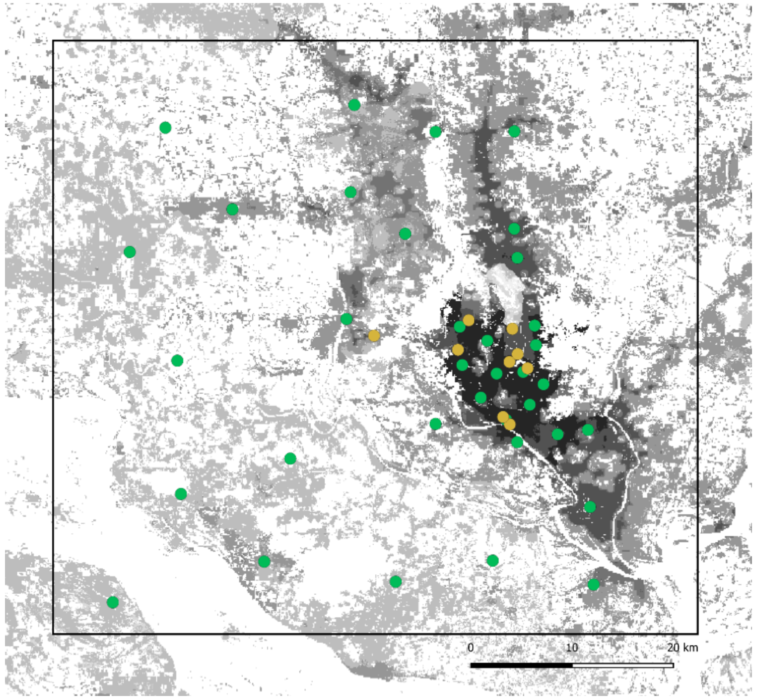
- **Improved spatial spread** (red), as compared to existing US-EPA (black) placement.

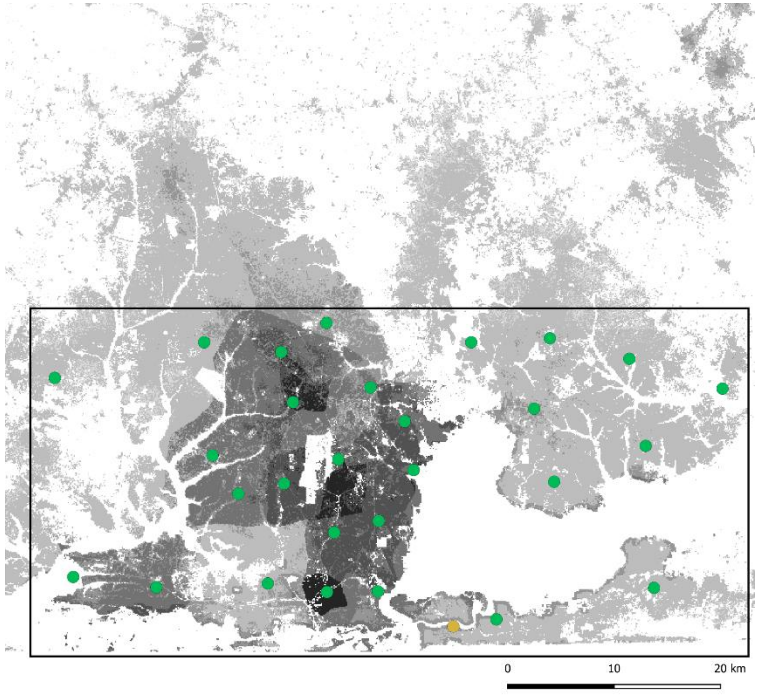


Black is the existing US-EPA monitoring network. Red is the optimized locations based on population density clusters.

Results: Consistent outcomes observed in different urban morphologies and population distributions.







	Lahore	Dhaka	Lagos
Population	15.9 million	21.9 million	13.6 million
Airshed	70 x 81 km	53 x 51 km	34 x 67 km
Proposed locations	45 (25 low-density, 20 high-density)	35 (18 low-density, 17 high-density)	27 (13 low-density, 14 high-density)
Population per monitor	353,333	503,704	625,714

Adaptable to different urban morphologies and population distributions, with consistent outcomes.

Data-Agnostic Flexibility	Requires only basic inputs like population data and can be adapted to different cities without heavy customization.
City Size Independence	Scales efficiently across cities of varying sizes and population distributions, from dense megacities to smaller urban clusters.
Transferable Across Geographies	Geography-neutral — the same framework works in cities with different urban forms (grid-like, radial, organic growth).

Limitations and future work

1. Comparison of population-weighted exposure capture
2. Statistical measures of spatial distribution quality
3. Incorporate emission sources, meteorology, land use, seasonal variations
4. Future direction: Multi-criteria clustering
5. Comparing results with other clustering algorithms

References

1. **Carozzi**, F.: Dirty Density: Air Quality and the Density of American Cities. CEP Urban and Spatial Programme Blog (formerly SERC), 29 August 2019. https://spatial-economics.blogspot.com/2019/08/dirty-density-air-quality-and-density.html
2. **Gupta**, S., Pebesma, E., Mateu, J., and Degbelo, A.: Air Quality Monitoring Network Design Optimisation for Robust Land Use Regression Models, Sustainability, 10, 1442, https://doi.org/10.3390/su10051442, 2018.
3. **Kanaroglou**, P. S., Jerrett, M., Morrison, J., Beckerman, B., Arain, M. A., Gilbert, N. L., and Brook, J. R.: Establishing an air pollution monitoring network for intra-urban population exposure assessment: A location-allocation approach, Atmos. Environ., 39, 2399–2409, https://doi.org/10.1016/j.atmosenv.2004.06.049, 2005.
4. **World Health Organization**: WHO Global Air Quality Guidelines: Particulate Matter (PM2.5 and PM10), Ozone, Nitrogen Dioxide, Sulfur Dioxide and Carbon Monoxide. Geneva, World Health Organization, 2021. Licence: https://creativecommons.org/licenses/by-nc-sa/3.0/igo/.
5. **U.S. Environmental Protection Agency**: Interactive Map of Air Quality Monitors. AirData, available at: https://www.epa.gov/outdoor-air-quality-data/interactive-map-air-quality-monitors, accessed 27 April 2025.

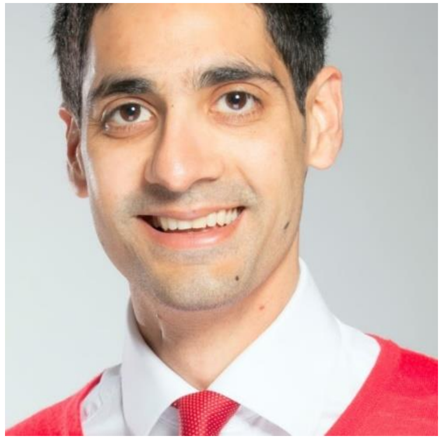
Thank you.

Pakistan Air Quality Initiative پاکی

EGU General Assembly 2025

Population-Centric Optimization of Air Quality Monitoring Networks in Data-Sparse Urban Regions: A Weighted K-Means Approach.
This presentation was shared at the EGU 2025 Session AS5.9: Low-cost air quality sensors: Challenges, opportunities, and collaborative strategies across the world, on 29 April 2025. Please contact Abid Omar or Mahad Naveed of the Pakistan Air Quality Initiative for any questions or comments at abidomar@pakairquality.com.

Pakistan Air Quality Initiative
About us



About our founder

Abid Omar has a social mission: how can I help improve the environment of Pakistan? He sees the air pollution emergency in Pakistan as a silver lining — an opportunity to drive environmental awareness and change that will clean up Pakistan for good.

A key indicator of environmental issues is pollution that affects our daily lives, specifically air pollution. He found but there is no data for this, so he founded the Pakistan Air Quality Initiative to monitor air quality in across major urban areas of Pakistan and to provide awareness for air quality and air pollution issues, and therefore provide impetus for change

About the Pakistan Air Quality Initiative

The Pakistan Air Quality Initiative (PAQI پاکی) is a researching organization, provides crowdsourced air quality data for Pakistan, PAQI is a community-driven initiative to set up low-cost, real-time monitors to capture air quality data and thereby increase social awareness. PAQI provides tools and information people need to thrive in polluted environments.

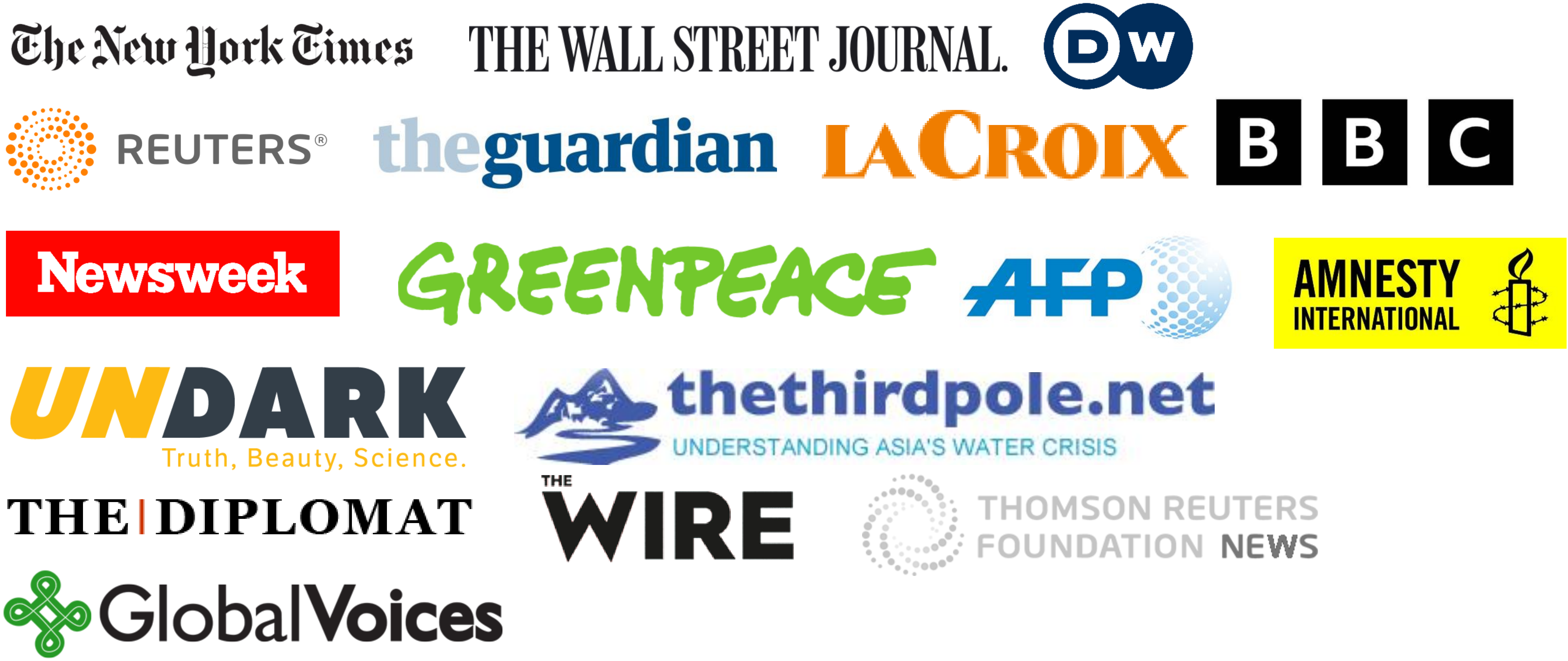
The Pakistan Air Quality Initiative was founded in 2016, and formally established as an atmospheric science research organization in 2024.

hello@pakairquality.com
+92 300 201 8617

5-3-1 Sector 15, Korangi Industrial Area
Karachi 74900, Pakistan

Media Coverage

International



Selected media coverage for work done by or data provided by the Pakistan Air Quality Initiative.