Method Article

Methods to Estimate Personal Exposure Levels to Air Pollution from Extensive Stationary Air Quality Dataset and Human Mobility Dataset

ABSTRACT

Accurately assessing personal exposure to air pollution has long posed a challenge due to limitations in conventional monitoring approaches. Most studies still rely on sparse, stationary regulatory monitors, assigning identical exposure values to individuals regardless of their movements. This approach neglects the dynamic nature of human mobility patterns and activity locations, leading to inferential errors. This method developed an approach by integrating high-resolution global positioning systems (GPS) trajectory data from 100 participants with air quality data from 213 PurpleAir low-cost stationary monitors across Eastern North Carolina. Using geostatistical modelling, which is an automated kriging (ordinary kriging) algorithm developed in Python, the method estimates individualised $PM_{2.5}$ exposure every minute over a 3-day window (two weekdays and one weekend day), which encompasses 129,600-minute points. The study offers an innovative fusion of spatial and temporal data that bridges the gap between environmental sensing and actual human experience, and the result is a transformative methodology that significantly enhances the precision of personal air pollution exposure assessments from stationary air quality sensors.

***Keywords**:* Personal exposure assessment; Human Mobility; Low-cost sensors; Geostatistical modelling; Uncertain Geographic Context Problem (UGCoP)

INTRODUCTION

Air pollution poses a threat to public and environmental health, and effective monitoring strategies are required to understand and mitigate its associated health impacts (Ki-Kwang et al., 2020; Chen et al., 2024; Wang et al., 2024). Conventionally, air quality exposure assessment studies rely on the use of outdoor air quality data from reference-based or regulatory monitors, which their installations are limited across space, often expensive, sparsely deployed, low temporal resolution, and too dispersed to capture hyper-local variation in air quality, especially in densely populated urban areas (Tang et al., 2025; Opejin & Park, 2024). Studies have utilized regulatory monitors data to assess people exposure to air quality at their residential home, assigning single exposure value to all residents and the findings of these studies revealed that using only regulatory monitors may generate bias and uncertain geographical context problems – a type of problem that causes inferential error when exposure assessment studies does not consider human mobility pattern– which could affect the finding of study (Bell, 2006; Son et al., 2010; Kwan, 2012; Park, 2020).

The limitations of the previous methods have led to the emergence of low-cost air quality sensors from distributed mobile low-cost sensor (e.g. GeoAir2, OPC-N3) and low-cost networks (e.g., PurpleAir), which are becoming increasingly viable for assessing exposure to air quality as it provide real-time air quality measurements through mobile sensors or a public web map, having the ability to get data from multiple locations and improve the spatiotemporal resolution of air quality data (Li et al., 2019; Streuber et al, 2022; Park et al., 2023). The accuracy of these low-cost sensors can vary significantly from that of regulatory monitors, and it is often essential to calibrate low-cost air sensors to ensure data reliability (de Souza et al., 2022).

Low-cost monitors are primarily classified into two types: mobile and stationary. Mobile low-cost monitors (e.g., GeoAir, OPC-N3) can be worn with a carabiner or belt clip or strapped to the shoulders and simultaneously record $PM_{2.5}$ readings (µg/m3), GPS positions, temperature, humidity, date, time in 1-minute interval, and they can collect a list of nearby

Wi-Fi media access control (MAC) addresses to ensure indoor geolocation accuracy (Park et al., 2021; Opejin & Park, 2024; Park et al., 2023). The flexibility of carrying a mobile, low-cost monitor facilitates the collection of high-resolution air quality data in real-time, capturing people's exposure to air quality as monitors record their exposure every time they visit (Relvas et al., 2025; Park, 2021). On the other hand, stationary low-cost monitors capture air quality data in real-time, including temperature and humidity, and record the geographical location of the monitor every minute; however, they do not directly assess people's actual exposure to air quality.

Therefore, an air pollution estimation model method was developed and used to assess people's exposure to air quality from a reasonable number of aggregated stationary low-cost monitors, based on their global positioning system (GPS) trajectory data, in contrast to the traditional method relying on using limited regulatory monitors.

METHOD DETAILS

Data preparation

This method utilises two different types of datasets: (1) Participants' GPS trajectory data extracted from GeoAir2.0, and (2) Air quality dataset from a network of PurpleAir low-cost sensors.

The Participants' GPS trajectory data were extracted from GeoAir2.0 – a low-cost mobile air sensor used in a pilot study conducted by Park's research team between October 2021 and March 2022 (Park et al., 2023). The pilot study included GPS trajectory data of 44 participants. The second round of data collection involved 56 participants, resulting in a total of 100 participants' GPS trajectory data across various daily activities (work, in-transit, grocery store, home, and restaurant) in Eastern North Carolina (Park et al., 2023; Opejin & Park, 2024). The GPS data captures each participant's indoor and outdoor trajectories over three days (two weekdays and one weekend), with timestamps at one-minute intervals. The

participants' GPS data spanned 90 unique calendar dates, and Figure 1 illustrates the data format of the GPS trajectory data.

Datetime	**latitude**	**longitude**	**P.ID**
10/12/2021 10:29	35.31244	-77.54323	P37
10/12/2021 10:29	35.50874	-77.35123	P38
10/12/2021 10:29	35.64404	-77.31435	P39
10/12/2021 10:29	35.31244	-77.31253	P42
10/12/2021 10:29	35.31244	-77.56332	P45
10/12/2021 10:29	35.31244	-77.51323	P46

table 1. Data structure of Participants' GPS trajectory data

Stationary outdoor air quality data used in this method was downloaded from the PurpleAir sensor network—a distributed network of low-cost sensors that uses the PMS5003 sensor (Plantower, Bejing, China) to detect PM ranging from 0.3 to 10 μm and the control board estimates the total mass for PM1, $PM_{2.5}$, and PM_{10} using particle counts calculated every second. The PurpleAir low-cost sensor network was established through a crowd-sourced community effort to monitor air pollution concentrations at an acceptable spatiotemporal resolution, filling data gaps in locations where regulatory monitoring stations are scarce. The data from PurpleAir sensors is transferred to the cloud every 80 seconds and logged, including outdoor $PM_{2.5}$ concentration measurements taken every minute, as well as the dates, times, and geographic coordinates of the monitor locations.

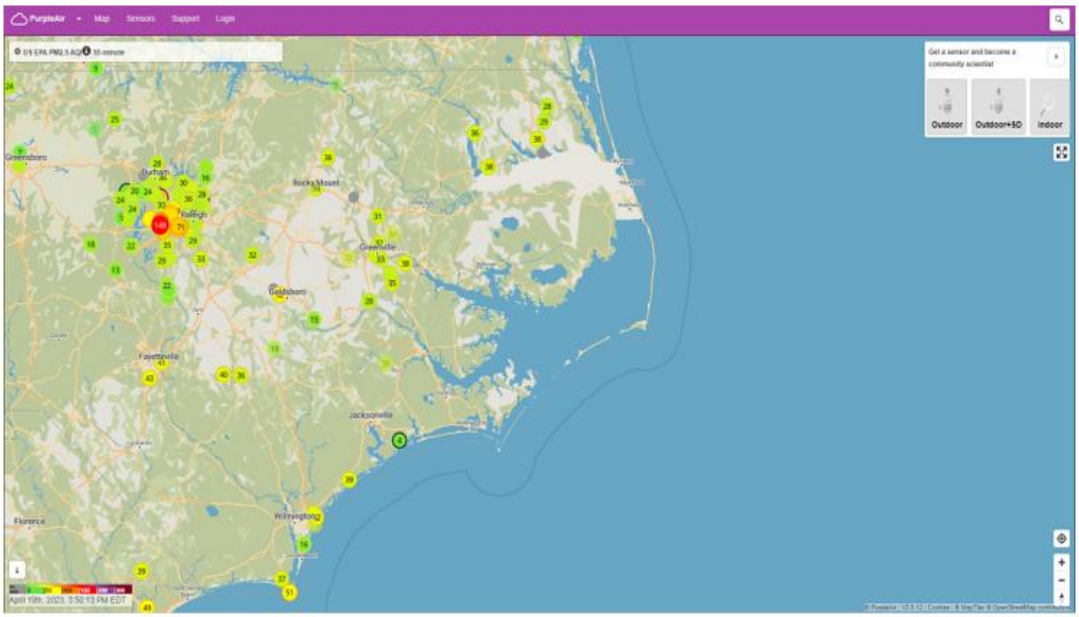


Figure 1. Distribution of PurpleAir monitors across Eastern North Carolina (PurpleAir web map)

The data accuracy of PurpleAir low-cost air sensors is lower than that of regulatory monitoring stations in general, even though the low-cost sensors allow for data acquisition from numerous locations at a community level and produce air quality data with higher spatiotemporal resolutions than those from a regulatory monitoring network.

	Time	PM2.5	Latitude	Longitude	Sensor Index
1	9/24/2021 00:12	10.685	34.216766	-77.80405	100641
2	9/24/2021 00:12	10.305	34.216766	-77.80405	100641
3	9/24/2021 00:12	9.67	34.216766	-77.80405	100641
4	9/24/2021 00:12	5.625	34.216766	-77.80405	100641
5	9/24/2021 00:12	4.835	34.216766	-77.80405	100641
6	9/24/2021 00:12	6.01	34.216766	-77.80405	100641
7	9/24/2021 00:12	18.915	34.216766	-77.80405	100641
8	9/24/2021 00:12	24.065	34.216766	-77.80405	100641
9	9/24/2021 00:12	22.465	34.216766	-77.80405	100641

table 2. The data format of the PurpleAir sensor dataset

Data validation

Field and laboratory evaluations were conducted to assess the accuracy of various low-cost monitors, revealing that GeoAir2.0 correlates significantly with the reference instrument in both laboratory and field settings (Streuber et al., 2022; Sousan et al., 2021). Sousan et al. (2021) revealed that GeoAir2.0 has extremely strong correlations with the reference instrument in both environmental ($r = 0.99$) and occupational situations ($r > 0.96$) while Streuber et al. (2021) proved that GeoAir2.0 strongly correlated with the reference instrument while detecting salt and Arizona road dust in regulated laboratory and field settings. PurpleAir monitors correlate effectively with a reference-grade beta attenuation monitor ($r^2 = 0.87$) and an optical particle counter ($r^2 = 0.98$) in two metropolitan settings in Greece over three seasons (Stavroulas et al., 2020). It also corresponded well with gravimetric concentrations ($r > 0.91$) when tested in indoor settings (Koehler et al., 2023).

Geostatistical operation

The overall idea of this method is that the algorithm was developed to systematically filter the air quality datasets (PurpleAir dataset) by utilizing the timestamp field in the GPS trajectory dataset to select all instances of PurpleAir monitors available for a particular minute. At that minute, the $PM_{2.5}$ values of the selected PurpleAir sensors were used to perform an ordinary kriging operation and generate an air pollution surface at that specific minute (Figure 2). The air pollution surface for that minute was generated, and people's GPS trajectory data (i.e., latitude and longitude) were dynamically overlaid on the ordinary kriging outputs (air pollution surface) to extract each participant's air quality exposure values for that corresponding minute. In this method, since the GeoAir2.0/GPS dataset (from timestamp) contained 129,600 unique minutes, which corresponds to the number of unique minutes in the PurpleAir data, the algorithm executed the process iteratively 129,600 times. Below are the detailed steps involved in algorithm development

***Step 1**: Setting up the geospatial environment and workspace definition*

The algorithm begins by setting up the geospatial environment using the arcpy Python package to perform geographic data analysis, as well as geopandas, pandas, csv, and shapely. The workspace was defined, and overwritten permission was enabled to facilitate seamless file operations. Also, three dedicated folders are created within the workspace: (1) one for storing interpolated air pollution surfaces, (2) the second for the exposure values extracted from those kriged surfaces, and (3) the third one for final output estimates of participant-level exposures.

***Step 2**: Importing Data and Preprocessing*

The two primary datasets are imported in comma-separated format. The PurpleAir sensor data contains real-time $PM_{2.5}$ concentrations with geolocated coordinates, and the participants' GPS trajectory data includes timestamped locations of individuals as they move from one place to another. Hence, each dataset is filtered to retain only the columns that are relevant for spatial and temporal analysis. Additionally, the algorithm was developed to extract unique timestamps from the GPS data, which form the temporal basis for minute-by-minute processing, ensuring that each computation synchronizes pollutant readings with participant locations in real time.

***Step 3**: Time-Synchronised GeoDataFrame and Shapefile Creation*

In this step, the two datasets are filtered to isolate data for each specific minute corresponding to every unique timestamp. Thereafter, these subsets are transformed into GeoDataFrames and saved as shapefiles, which allow the datasets to be compatible with ArcGIS spatial operations. The PurpleAir points represent the spatial distribution of pollution concentrations, while the GPS data points represent where individuals were located at that time (minute).

***Step 4**: Spatial Filtering and Geostatistical operation (*Ordinary *Kriging Interpolation)*

This step introduces the ordinary Kriging method, which uses the aggregation of PurpleAir sensor data for a minute to interpolate pollution levels across the study region based on available sensor readings, utilising the algorithm (Figure 2). Therefore, to maintain spatial relevance and avoid extrapolation errors, only PurpleAir sensor points within 60 miles of the study area boundary are selected for analysis. However, the algorithm was customised so that when no points fall within this threshold, the algorithm should intelligently skip that minute to ensure computational robustness.

***Step 5**: Extraction of Raster Value from the kriged surface and its mapping to Participant Locations*

The ordinary kriging surface is generated and clipped to the boundary of the study area using the clip tool. Additionally, to address potential data gaps, a focal statistics filter is applied, utilizing the mean of surrounding raster cells to fill in NoData regions, ensuring the resulting pollution surface is continuous and more reliable for exposure extraction.

The kriged surface (Figure 2) is then used to extract $PM_{2.5}$ values at each participant's GPS location using a spatial analysis tool that overlays points onto the raster and dynamically extracts raster values of each location using the ArcGIS extract tool. And these values are stored in individual data tables for each minute, which represent the pollution exposure level by individual at that minute.

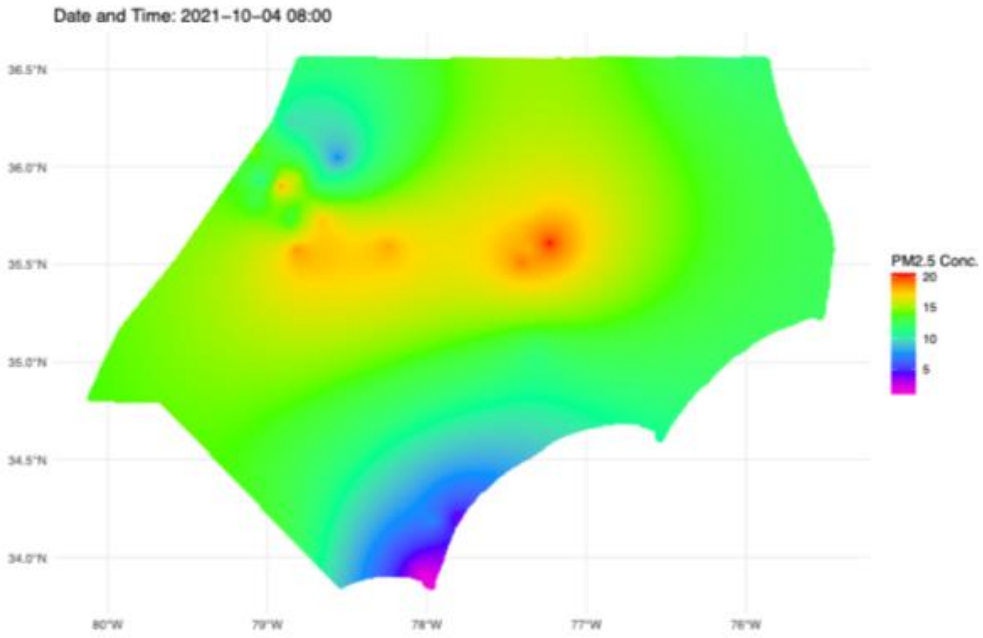
***Step 6**: Intermediate File Cleanup*

After the extraction operation, the kriged surfaces after each iteration and intermediate shapefiles (GPS and sensor shapefiles) used in the process are deleted to maintain the workspace, manage disk space, improve performance, and preserve only relevant data.

***Step 7**: Spatial Filtering and Geostatistical operation (*Ordinary *Kriging Interpolation)*

At the end of the minute-by-minute processing, all individual data tables saved into the initially created folder “One minute exposure values” are merged into a single shapefile. And the dataset consolidates the exposure values for all participants across all observation minutes. Lastly, the merged shapefile is converted into a CSV file, providing a tabular format suitable for statistical analysis, visualization, or further analysis and modelling.

For the algorithm source code, visit https://doi.org/10.5281/zenodo.15499114 or https://github.com/OpejinAbdulahi/Air-Pollution-Estimation-Model.



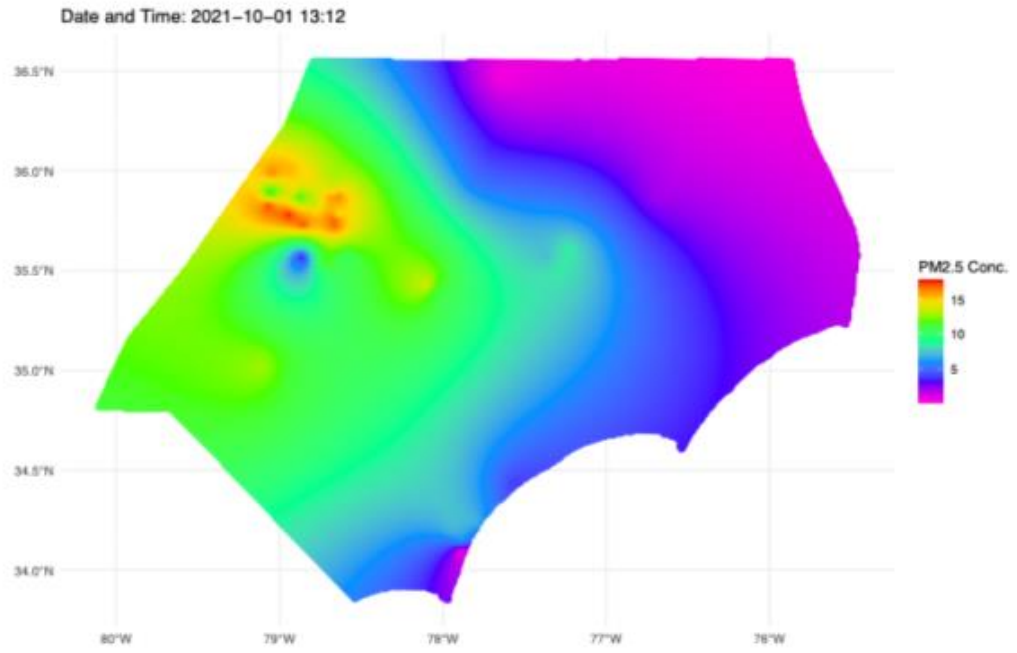


Figure 2. One-minute kriged outcomes for minute (October 01, 2021, 08:00 am) and (October 04, 2021, 08:00 am) based on the PurpleAir low-cost sensor network data. Reproduced from Opejin & Park (2024), Science of the Total Environment, © 2024 Elsevier. Reused with permission under Elsevier License Number [6180820514617].

DISCUSSION

This study introduces an effective and scalable approach for estimating personalized $PM_{2.5}$ exposure by combining dense, low-cost stationary sensor networks with high-resolution human mobility data. This method significantly enhances traditional exposure assessment techniques that depend on fixed home locations or limited regulatory monitors by automating minute-by-minute ordinary kriging and dynamically syncing interpolated pollution surfaces with individual GPS trajectories. The resulting exposure estimates more accurately represent the spatiotemporal dynamics of human mobility, thereby directly addressing the Uncertain Geographic Context Problem and minimizing exposure misclassification.

The reliability of the exposure estimates is fundamentally linked to the geostatistical assumptions underlying ordinary kriging, particularly second-order stationarity and the choice of variogram model. This study employed kriging at short temporal intervals (one minute) and over confined spatial extents, settings that support the notion of local stationarity. Although various variogram structures and parameterizations can affect smoothness and long-range dependence. The technique improves robustness by omitting time steps with inadequate sensor coverage, thus preventing unstable variogram fitting. While a formal sensitivity analysis of variogram parameter selection was not within the purview of this methodological paper, subsequent applications could systematically assess alternative variogram models or integrate adaptive or spatiotemporal models based on sensor density, spatial extent, and the spatial configuration of air pollution fluctuation estimates.

The applicability of this method is primarily contingent on sensor density, spatial extent, and the spatial configuration of air pollution fluctuations within the study area. Eastern North Carolina has a comparatively dense PurpleAir network; nonetheless, the method is designed to be scalable across diverse geographic contexts by adjusting spatial filtering thresholds, minimum sensor requirements require longer temporal aggregation periods, larger spatial neighborhoods, or hybrid modeling techniques that incorporate remains applicable across urban, suburban, and semi-rural contexts, making it a versatile tool for exposure assessment in data-constrained settings models based on sensor density, spatial extent, and the spatial configuration of air pollution fluctuations, the fundamental framework—time-synchronized geostatistical interpolation integrated with GPS trajectories—retains its applicability in urban, suburban, and semi-rural contexts, rendering it a versatile instrument for exposure assessment in data-constrained scenarios.

The method's computing efficiency is attained by a systematic, automated procedure that iteratively handles extensive spatiotemporal datasets, concurrently reducing disk utilization and memory overhead. The algorithm performs ordinary kriging up to 129,600 times,

discarding intermediate files after each iteration, thereby significantly reducing storage requirements and improving runtime efficiency. The processing duration is affected by variables such as the number of sensors per minute, the spatial resolution of interpolation surfaces, and hardware specifications. In a typical desktop computing environment, minute-level processing is achievable without high-performance computing resources, making the method accessible to a broad spectrum of researchers. The complete source code has been released to promote adoption and repeatability, and enable users to evaluate performance, modify parameters, and customize the workflow to their datasets and computational constraints .

CONCLUSION

This method presents an approach to measuring and understanding human exposure to air pollution using air quality datasets from distributed low-cost sensors. By coupling the spatial mobility of individuals with the temporal depth of a large number of air-quality sensors from stationary sensor networks, the method provides a scalable and effective means of estimating real-time personal exposure to air pollution. Therefore, the Python-based geostatistical model automates a time-intensive process and extensive dataset, enabling minute-by-minute assessments that reflect the nuanced realities of human behavior and movement.

LIMITATION

The method relies on geostatistical operations (ordinary kriging), which assume stationarity and isotropy and smooth extreme values, thereby attenuating them and reducing local variability. It primarily depends on variogram fitting, which may lead to a slightly uncertain exposure estimate that differs from the actual real-world estimate. Despite this limitation, this

method contributes to the approach used to estimate people's actual exposure to air quality, particularly in areas where regulatory monitors are limited.

Consent

Not Applicable

Ethics statements

This study did not involve human participants or animal subjects. The data used were obtained from publicly available sources and analyzed in accordance with applicable ethical guidelines.

Data availability

PurpleAir data is publicly available, while GeoAir2.0 data is confidential.

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Disclaimer (Artificial intelligence)
Author(s) hereby declare that NO generative AI technologies such as Large Language Models (ChatGPT, COPILOT, etc.) and text-to-image generators have been used during the writing or editing of this manuscript.

REFERENCES

1. Bell, M. L. (2006). The use of ambient air quality modeling to estimate individual and population exposure for human health research: A case study of ozone in the Northern Georgia Region of the United States. Environment International, 32(5), 586–593. https://doi.org/10.1016/j.envint.2006.01.005
2. Chen, F., Zhang, W., Mfarrej, M. F. B., Saleem, M. H., Khan, K. A., Ma, J., Raposo, A., & Han, H. (2024). Breathing in danger: Understanding the multifaceted impact of air pollution on health impacts. Ecotoxicology and Environmental Safety. https://doi.org/10.1016/j.ecoenv.2024.116532
3. Lee, K.-K., Park, Y., Han, S.-P., & Kim, H. C. (2020). The Alerting Effect from Rising Public Awareness of Air Quality on the Outdoor Activities of Megacity Residents. Sustainability, 12(3), 820. https://doi.org/10.3390/su12030820
4. Kwan, M.-P. (2012). The uncertain geographic context problem. Annals of the Association of American Geographers, 102(5), 958–968. https://doi.org/10.1080/00045608.2012.687349
5. Li, J., Mattewal, S. K., Patel, S., & Biswas, P. (2020). Evaluation of nine low-cost-sensor-based particulate matter monitors. Aerosol and Air Quality Research, 20(2), 254–270. https://doi.org/10.4209/aaqr.2018.12.0485

6. Opejin Abdulahi. (2024). Air-Pollution-Estimation-Model: v1.0.0. Zenodo.
7. Opejin, A., & Park, Y. M. (2024). Assessing bias in personal exposure estimates when indoor air quality is ignored: A comparison between GPS-enabled mobile air sensor data and stationary sensor network data. The Science of the Total Environment. https://doi.org/10.1016/j.scitotenv.2024.175249
8. Park, Y. M. (2020). Assessing personal exposure to traffic-related air pollution using individual travel-activity diary data and an on-road source air dispersion model. Health & Place, 63, 102351. https://doi.org/10.1016/j.healthplace.2020.102351
9. Park, Y. M. (2021). A GPS-enabled portable air pollution sensor and web-mapping technologies for field-based learning in health geography. Journal of Geography in Higher Education, 46(2), 241–261. https://doi.org/10.1080/03098265.2021.1900083
10. Park, Y. M., & Kwan, M. P. (2017). Individual exposure estimates may be erroneous when spatiotemporal variability of air pollution and human mobility are ignored. Health & Place, 43, 85–94. https://doi.org/10.1016/j.healthplace.2016.10.002
11. Park, Y. M., Chavez, D., Sousan, S., Figueroa-Bernal, N., Alvarez, J. R., & Rocha-Peralta, J. (2023). Personal exposure monitoring using GPS-enabled portable air pollution sensors: A strategy to

promote citizen awareness and behavioral changes regarding indoor and outdoor air pollution. Journal of Exposure Science and Environmental Epidemiology https://doi.org/10.1038/s41370-022-00515-9

12. Relvas, H., Lopes, D., & Armengol, J. M. (2025). Empowering communities: Advancements in air quality monitoring and citizen engagement. *Urban Climate*, *60*, 102344. https://doi.org/10.1016/j.uclim.2025.102344
13. Son, J., Bell, M. L., & Lee, J. (2010). Individual exposure to air pollution and lung function in Korea: Spatial analysis using multiple exposure approaches. Environmental Research, 110(8), 739–749. https://doi.org/10.1016/j.envres.2010.08.003
14. Sousan, S., Regmi, S., & Park, Y. M. (2021). Laboratory Evaluation of Low-Cost Optical Particle Counters for Environmental and Occupational Exposures. Sensors, 21(12), 4146. https://doi.org/10.3390/s21124146
15. Streuber, D., Park, Y. M., & Sousan, S. (2022). Laboratory and field evaluations of the GeoAir2 air quality Monitor for use in indoor environments. Aerosol and Air Quality Research, 22(8), 220119. https://doi.org/10.4209/aaqr.220119
16. Tang, D., Mi, T., Zheng, X., Yang, M., Grieneisen, M. L., Zhan, Y., & Yang, F. (2025). Harmonizing low-cost and regulatory air quality

monitoring networks with interpretable semi-supervised learning: Reducing exposure misclassification in underrepresented communities. Journal of Hazardous Materials, 491, 137893. https://doi.org/10.1016/j.jhazmat.2025.137893

17. Wang, S., Song, R., Xu, Z., Chen, M., Di Tanna, G. L., Downey, L., Jan, S., & Si, L. (2024). The costs, health and economic impact of air pollution control strategies: a systematic review. Global Health Research and Policy, 9(1). https://doi.org/10.1186/s41256-024-00373-y