Exploration of Bias and Variability between Low-Cost Air Quality Sensors in Urban Environments

Nail F. Bashan, Qi R. Wang

1. RESEARCH PROBLEM STATEMENT

Air pollution remains a significant global concern, with far-reaching impacts on public health. Particulate matter with an aerodynamic diameter of 2.5 μm or less ($PM_{2.5}$) is of particular concern due to its significant impact on respiratory health, its association with trillions of US dollars in economic costs, and its contribution to social inequalities (Burnett et al., 2018; Yin et al., 2021). In the United States, over 30% of the population resides in areas with unhealthy levels of air pollution, which has been associated with an annual toll of 85,000 to 200,000 deaths (American Lung Association, 2023). The dispersion of $PM_{2.5}$ is notably influenced by atmospheric factors and the urban built environment. Consequently, even communities located in close proximity to each other may experience varying levels of particulate matter concentrations. Therefore, to effectively mitigate the health burden of air pollution, understanding the hyperlocal spatial distribution of these pollutants, identifying potential sources, and implementing effective air quality monitoring mechanisms is essential (Childs et al., 2022).

The Environmental Protection Agency (EPA) in the United States is tasked with developing and enforcing regulations related to environmental protection, including air quality standards. Its foremost objective is to protect public health and the environment from significant risks. To assess compliance with the National Ambient Air Quality Standards (NAAQS), the EPA uses established standard techniques such as the Federal Reference Method (FRM) and the Federal Equivalent Method (FEM) for conducting $PM_{2.5}$ measurements. These methods are critical in determining whether regions meet the NAAQS or are classified as non-attainment areas (Noble et al., 2001). However, due to the substantial costs and maintenance demands associated with FRM and FEM monitoring stations, their deployment is sparse and uneven across the nation. Among the 3,100 counties in the US, only about 21 percent are equipped with $PM_{2.5}$ monitors. Moreover, the temporal resolution of the data from these monitors is typically limited to 1-hour average intervals, which may not adequately capture short-term fluctuations in $PM_{2.5}$ levels (Sullivan & Krupnick, 2018).

Recent advancements in low-cost air pollution sensors enable studies that offer more granular insights into air quality, enhancing our understanding at a hyperlocal scale (Gao et al., 2015). Despite this progress, data quality assurance remains a significant hurdle. These low-cost monitors, which typically utilize optical sensors, can deliver readings affected by ambient factors like temperature and humidity (Han et al., 2020). Although useful in normal conditions, their reliability is questionable during extreme pollution events, when accurate readings are most critical (Barkjohn et al., 2022). Issues such as sensor maintenance, lifespan, and signal integrity further complicate data quality. Variability in production standards can lead to disparities in measurements even among sensors of the same model. The positioning of these sensors can also affect precision, with seemingly identical sensors yielding different data (Feenstra et al., 2019). Furthermore, community-operated sensors may not adhere to essential setup protocols, such as correct installation height and orientation, compromising the collection of high-quality data.

Researchers have enhanced the accuracy of low-cost monitors and reduced bias by developing regression models incorporating land-use features or meteorological conditions (Barkjohn et al., 2021). Evidence from numerous studies suggests that with proper maintenance, the use of sophisticated data-cleaning algorithms, and suitable calibration, these monitors can achieve notably high-performance levels (Connolly et al., 2022).

In this study, we aim to examine the urban elements that contribute to the bias and variability of low-cost air quality monitors within city environments. Urban areas possess known pollution hotspots, such as industrial sites and major roadways, yet the unique characteristics of each neighborhood complicate the generalization of air pollution patterns. The design and structure of urban spaces are key in shaping hyperlocal air pollution, underlining the importance of research into these distinct variations. Our methodology involves correlating points of interest (POI) data from OpenStreetMap (OSM) with air quality measurements from PurpleAir sensors across the San Francisco Bay Area. This strategy explores the relationship between urban configurations and their effects on air quality.

2. BRIEF RESEARCH METHODOLOGY AND APPROACH

2.1 Air Quality Data Collection and Processing

PurpleAir is a community-operated, low-cost air quality monitoring network, primarily for particulate matter assessment, both indoors and outdoors. With its network of over 20,000 sensors in the United States, PurpleAir offers real-time air quality data to the public via an online map. For our study, we accessed historical data from 1,842 PurpleAir monitors located within the San Francisco Bay Area, utilizing the network's historical data access API. This region was chosen for its dense concentration of PurpleAir stations, ensuring a rich dataset. We collected data at 2-minute intervals from January 1 to January 31, 2020, a timeframe chosen to reflect pre-COVID travel patterns without the confounding effects of other pollution events. This selection is particularly pertinent for Northern California, where wildfire-induced pollution could skew results. By excluding periods of extreme pollution, our analysis aims to isolate the influence of urban environmental factors on air quality.

In our research, we applied a data cleaning and humidity correction method devised by Barkjohn et al. (2021) to improve data integrity. To filter out potential anomalies, we disregarded temperature readings above 1000°F or below -200°F, which likely indicate sensor malfunctions or incorrect air quality information. Our analysis utilized data from PurpleAir monitors, each equipped with two Plantower PMS5003 sensors (channels A&B) to allow cross-verification of results. We initially omitted any sensor data reliant on a single channel. Additionally, we excluded readings lacking temperature (T) or humidity (RH) values, as missing data may suggest signal errors. For monitors with dual channels, we flagged for exclusion any particulate matter measurements with an absolute difference exceeding 5 $\mu g/m^3$ or those with percent differences beyond 2 standard deviations (61%). Monitors failing to reach at least 75% data completeness for January 2020 were also removed from our analysis.

After refining the data, we juxtaposed the readings from PurpleAir monitors with those from FRM/FEM stations in the vicinity, finding a strong correlation in the daily and hourly $PM_{2.5}$ averages (p-values of 0.96 and 0.77, respectively). This correlation reinforces the notion that, given rigorous data quality management, low-cost sensors can provide reliable data for scientific

inquiries, including the objectives of this study. To examine the spatial variability of $PM_{2.5}$, we applied the Ordinary Kriging technique to interpolate the average $PM_{2.5}$ concentrations during the weekdays of January 2020, using data from 534 monitors. The interpolated map revealed distinct spatial patterns, with certain urban areas displaying higher $PM_{2.5}$ concentrations. Subsequent sections of this study will delve into the underlying factors contributing to these heightened concentrations.

2.2 Point of Interest Data

OpenStreetMap (OSM) is a participative project that allows global users to create, modify, and share extensive mapping details. It thrives on an open-source framework, relying on a broad community to maintain up-to-date and exhaustive geographical data. For our analysis, we utilized the "overpass-api" to methodically extract POI data within the San Francisco Bay Area, aligning with the range of our gathered PurpleAir sensor data. We categorized the POIs into seven key groups—food, health, education, industry, transit, green spaces, and shopping—to reflect varied urban characteristics that may influence air quality. The central coordinates of each POI were pinpointed for subsequent spatial correlation with the air quality monitoring data.

2.3 Spatial analysis

To evaluate the impact of POIs on air quality, we implemented a method to correlate each air quality sensor with its immediate urban environment. Following the EPA's guidelines for monitoring network design, which advise a middle-scale representation range from 100 to 500 meters (U.S. Environmental Protection Agency, 2017), we established a 300-meter buffer around each sensor. We chose this specific radius based on the premise that POIs within it have a substantial influence on sensor readings. Within these zones, POIs were cataloged and binary encoded—zero indicating absence and one indicating presence—per their main category. We computed the average air quality values for each sensor in January 2020, focusing on weekdays when the effects of industrial, educational, and healthcare activities are most pronounced. We then applied ordinary least squares regression (OLS) analysis, with $PM_{2.5}$ averages as the dependent variable and the binary encoded POI categories serving as independent variables.

3. KEY FINDINGS

After conducting the Ordinary Least Squares (OLS) regression, the model presented an $R^2 = 0.07$, signifying considerable variability. Atmospheric modeling, particularly concerning air pollution, demands complex, large-scale data sets and advanced computational techniques. Our initial model, based predominantly on nearby POIs to predict air pollution patterns, inherently exhibited high variability. Additional geographical, pollution transport, and atmospheric data could enhance the model's robustness and reduce its variance. Of the seven categories analyzed, five (health, food, shopping, green spaces, and transit) emerged as statistically significant ($p<0.05$), indicating that, despite the model's variance, the bias was relatively limited. Specific POIs within these significant categories appeared to influence air quality in either positive or negative ways. Notably, areas dense with food and shopping POIs, likely to experience increased vehicle traffic and idling, correlated positively with higher levels of air pollution. In contrast, regions with health facilities, green spaces, and transit hubs correlated negatively with $PM_{2.5}$ values, likely due to reduced emission-contributing activities. The categories of industry and education did not show statistically meaningful patterns ($p>0.05$), possibly due to the uniform presence of educational POIs in residential zones not significantly affecting pollution, and a

limited presence of heavy industries within the study area not sufficiently impacting the pollution levels to be detected by our model.

4. IMPLICATIONS

Through our analysis, we observed that areas with high-impact POIs, specifically food and shopping, displayed significantly increased $PM_{2.5}$ levels compared to areas associated with lower-impact POIs, such as health facilities, green spaces, and transit hubs. Conversely, sites featuring industrial and educational POIs, which were statistically insignificant, showed $PM_{2.5}$ readings akin to the broader monitor dataset. This investigation operates under certain presumptions and is subject to constraints that warrant additional scrutiny. OpenStreetMap, being a community-driven platform, might suffer from data accuracy issues, with the POIs potentially being outdated or misplaced. For greater accuracy, alternative datasets with more precise mobility records, like SafeGraph, may be superior. Moreover, despite thorough data cleaning, the low-cost monitors utilized may still harbor significant inaccuracies, and professional monitoring sources are likely to provide more reliable data. It is important to note that this study's scope was limited to January 2020 and confined to the Bay Area; hence, variations due to different months, seasonal effects, and other locations might lead to divergent findings.

REFERENCES

American Lung Association. (2023). *State of the Air by American Lung Association*. https://www.lung.org/research/sota/key-findings

Barkjohn, K. K., Gantt, B., & Clements, A. L. (2021). Development and application of a United States-wide correction for $PM_{2.5}$ data collected with the PurpleAir sensor. *Atmospheric Measurement Techniques*, *14*(6), 4617–4637. https://doi.org/10.5194/amt-14-4617-2021

Barkjohn, K. K., Holder, A. L., Frederick, S. G., & Clements, A. L. (2022). Correction and Accuracy of PurpleAir PM2.5 Measurements for Extreme Wildfire Smoke. *Sensors*, *22*(24), Article 24. https://doi.org/10.3390/s22249669

Burnett, R., Chen, H., Szyszkowicz, M., Fann, N., Hubbell, B., Pope, C. A., Apte, J. S., Brauer, M., Cohen, A., Weichenthal, S., Coggins, J., Di, Q., Brunekreef, B., Frostad, J., Lim, S. S., Kan, H., Walker, K. D., Thurston, G. D., Hayes, R. B., … Spadaro, J. V. (2018). Global estimates of mortality associated with long-term exposure to outdoor fine particulate matter. *Proceedings of the National Academy of Sciences*, *115*(38), 9592–9597. https://doi.org/10.1073/pnas.1803222115

Childs, M. L., Li, J., Wen, J., Heft-Neal, S., Driscoll, A., Wang, S., Gould, C. F., Qiu, M., Burney, J., & Burke, M. (2022). Daily Local-Level Estimates of Ambient Wildfire Smoke PM2.5 for the Contiguous US. *Environmental Science & Technology*, *56*(19), 13607–13621. https://doi.org/10.1021/acs.est.2c02934

Connolly, R. E., Yu, Q., Wang, Z., Chen, Y.-H., Liu, J. Z., Collier-Oxandale, A., Papapostolou, V., Polidori, A., & Zhu, Y. (2022). Long-term evaluation of a low-cost air sensor network for monitoring indoor and outdoor air quality at the community scale. *Science of The Total Environment*, *807*, 150797. https://doi.org/10.1016/j.scitotenv.2021.150797

Feenstra, B., Papapostolou, V., Hasheminassab, S., Zhang, H., Boghossian, B. D., Cocker, D., & Polidori, A. (2019). Performance evaluation of twelve low-cost PM2.5 sensors at an ambient air monitoring site. *Atmospheric Environment*, *216*, 116946. https://doi.org/10.1016/j.atmosenv.2019.116946

Gao, M., Cao, J., & Seto, E. (2015). A distributed network of low-cost continuous reading sensors to measure spatiotemporal variations of PM2.5 in Xi'an, China. *Environmental Pollution*, *199*, 56–65. https://doi.org/10.1016/j.envpol.2015.01.013

Han, J., Liu, X., Chen, D., & Jiang, M. (2020). Influence of relative humidity on real-time measurements of particulate matter concentration via light scattering. *Journal of Aerosol Science*, *139*, 105462. https://doi.org/10.1016/j.jaerosci.2019.105462

Noble, C. A., Vanderpool, R. W., Peters, T. M., McElroy, F. F., Gemmill, D. B., & Wiener, R. W. (2001). Federal Reference and Equivalent Methods for Measuring Fine Particulate Matter. *Aerosol Science and Technology*, *34*(5), 457–464. https://doi.org/10.1080/02786820121582

Sullivan, D. M., & Krupnick, A. (2018). *Using Satellite Data to Fill the Gaps in the US Air Pollution Monitoring Network*. https://www.rff.org/publications/working-papers/using-satellite-data-to-fill-the-gaps-in-the-us-air-pollution-monitoring-network/

U.S. Environmental Protection Agency. (2017). *Quality Assurance Handbook for Air Pollution Measurement Systems, Volume 2*. Office of Air Quality Planning and Standards Air Quality Assessment Division. https://www.epa.gov/sites/default/files/2020-10/documents/final_handbook_document_1_17.pdf

Yin, H., Brauer, M., Zhang, J. (Jim), Cai, W., Navrud, S., Burnett, R., Howard, C., Deng, Z., Kammen, D. M., Schellnhuber, H. J., Chen, K., Kan, H., Chen, Z.-M., Chen, B., Zhang, N., Mi, Z., Coffman, D., Cohen, A. J., Guan, D., … Liu, Z. (2021). Population ageing and deaths attributable to ambient PM2·5 pollution: A global analysis of economic cost. *The Lancet Planetary Health*, *5*(6), e356–e367. https://doi.org/10.1016/S2542-5196(21)00131-5