Enriching fixed stations air pollution monitoring with opportunistic mobile monitoring

Karine Zeitouni, Mohammad Abboud, Yehia Taher

Enriching fixed stations air pollution monitoring with opportunistic mobile monitoring

Mohammad Abboud, Karine Zeitouni and Yehia Taher

DAVID Lab, UVSQ - Université Paris-Saclay, Versailles, France

Abstract

The deteriorating air quality in urban areas, particularly in developing countries, has led to increased attention being paid to the issue. Daily reports of air pollution are essential to effectively manage public health risks. Pollution estimation has become crucial to expanding spatial and temporal coverage and estimating pollution levels at different locations. The emergence of low-cost sensors has enabled high-resolution data collection, either in fixed or mobile settings, and various approaches have been proposed to estimate air pollution using this technology. This study aims to enhance the data from fixed stations by incorporating opportunistic mobile participatory monitoring (MPM) data. The research question is: "How can we enrich fixed station data using MPM?" To overcome the limited availability of MPM data, we reuse existing data for periods with similar pollution maps observed by the fixed stations. The combined fixed and mobile data is then subjected to interpolation methods to generate more accurate pollution maps. The effectiveness of our approach is demonstrated by experiments conducted on a real-life dataset.

Keywords

Air Quality Monitoring, Opportunistic Mobile Participatory Monitoring, Low-cost sensors, Data integration, Spatial interpolation

1. Introduction

The combination of urbanization and climate change poses a significant threat to the health of urban populations and the environment. It is projected that by 2050, up to 70% of the global population will reside in urban areas, with 75% of Europeans already living in cities. This trend presents a range of interconnected challenges that impact social, economic, and environmental infrastructures, with deteriorating air quality being a particular concern, especially in developing nations. Virtually everyone on Earth is breathing polluted air, according to the World Health Organization (WHO) [1]. Indeed, 99% of the world's population lives in places where air quality exceeds internationally approved limits. WHO estimates show that around 7 million premature deaths per year are attributable to the joint effect of ambient and household air pollution.

The significance of air pollution monitoring has risen in recent years due to its ability to generate the Air Quality (AQ) index for the region under consideration. By aiding policymakers in devising more effective strategies to tackle pollution-induced urbanization challenges, air pollution monitoring can be highly beneficial.

Air pollution monitoring has extensively relied on fixed stations for the last three decades to generate the AQ pollution index. These stations typically record the hourly average of pollution levels in a specific region. Regrettably, the deployment of such stations is financially demanding, and their maintenance is also a significant concern, leading to limited coverage.

Researchers have shown recent interest in using air quality mobile sensing as an alternative method for measuring air pollution.

Mobile sensors for air quality are cost-effective and offer high-resolution pollution measurements while being deployed in high densities, as noted by [2] and [3]. However, calibration is typically necessary for such sensors. In recent years, researchers have attempted to estimate pollution and broaden spatial coverage by combining fixed and mobile measurements.

Various researchers specializing in fixed and mobile monitoring techniques have put forward distinct methods for estimating pollution based on data from fixed stations, mobile air quality sensors, or a blend of both. Fixed stations are capable of producing precise measurements, but they fall short when it comes to the spatial coverage. Conversely, mobile sensing can expand spatial coverage but may also yield some imprecise measurements. Additionally, fixed stations generally maintain continuous temporal coverage at specific locations, while mobile sensors may not have temporal coverage at certain locations.

The GoGreen Routes [1] project is committed to address-

[1] https://gogreenroutes.eu/

Proceedings of the Workshop on Big Mobility Data Analytics (BMDA) co-located with EDBT/ICDT 2023 Joint Conference (March 28-31, 2023), Ioannina, Greece

mohammad.abboud@uvsq.fr (M. Abboud); karine.zeitouni@uvsq.fr (K. Zeitouni); yehia.taher@uvsq.fr (Y. Taher)

https://pages.david.uvsq.fr/kzeitouni/ (K. Zeitouni)

0000-0003-4157-373X (M. Abboud); 0000-0002-5602-6942 (K. Zeitouni); 0000-0002-8706-8889 (Y. Taher)

ing a range of challenges, including monitoring and estimating air pollution. The current research contributes to this project by utilizing fixed and mobile sensor data to broaden air pollution estimates' geographical and temporal coverage.

Researchers have utilized fixed stations and mobile sensor data to estimate pollution maps. Some studies have relied exclusively on fixed stations [4, 5, 6], while others have applied air pollution estimation methods used in fixed stations to low-cost mobile sensor data [7, 8]. However, recent research proposes combining data from fixed and mobile sensors [9, 10], which raises several unresolved questions. Firstly, what are the most effective methods to use deterministic methods, geostatistical methods, or machine/deep learning models? Secondly, what features should be considered during the pollution estimation process? Lastly, how should we address the challenges of merging data from fixed and mobile sensors, considering the differences in their resolution and spatiotemporal coverage?

This paper presents a novel approach to assessing air pollution levels in the city of Versailles by utilizing data from both fixed and mobile sensors. Prior studies that integrate fixed and mobile sensor data or solely rely on mobile sensing typically involve targeted campaigns focused on specific routes or deploying sensors on buses or trams following fixed paths. In contrast, our methodology leverages a mobile crowd-sensing (MCS) approach. MCS [11], as a new paradigm, harnesses data acquired by volunteers using sensor-enhanced mobile devices with GPS capabilities while carrying out their daily routines, resulting in non-persistent data collection and limited outdoor data samples as most activities are indoors. Our research question centers on the incorporation of MCS/MPM[2] data to supplement fixed station data and estimate air pollution levels across the city.

This work proposes a methodology to augment air pollution monitoring stations with randomly collected data from mobile sensing devices (MCS). Our goal is to improve the accuracy of pollution maps by utilizing data from both fixed and mobile sensors, thus increasing spatial and temporal coverage. Our approach is based on the assumption that similar data can be found for fixed stations at different periods. Specifically, we aim to identify clusters of different fixed station data, match them with MCS data at corresponding times, and combine them to generate more data samples and improve the pollution map. This method results in enhanced pollution estimation.

The remainder of this paper is organized as follows: Section 2 reviews related work and different approaches discussed in the literature. Section 3 details and explains our methodology. Section 4 presents the implementation and experimental results. In sections 5 and 6, we summarize our findings and suggest future directions for research.

[2]Please note that MPM (opportunistic mobile participatory monitoring) and MCS (opportunistic mobile crowd sensing) are used interchangeably in this paper

2. Related Work

Researchers have shown interest in the problem of estimating pollution for several years. The problem has been examined in the literature from various perspectives and scales. While meso-scale air quality modeling systems, such as CHIMERE[12], are the most commonly used, urban scale models utilizing Computational Fluid Dynamic (CFD) simulations have also been proposed. However, their computational complexity limits their applicability to a wide area [13]. In addition to these model-driven approaches, data-driven methods have become popular due to the increased use of monitoring stations, including traditional fixed networks, denser networks of low-cost fixed sensors, and low-cost mobile devices. In this discussion, we will focus on data-driven approaches that expand spatial and temporal coverage. This section summarizes the conducted studies on pollution estimation and interpolation for various measurements.

Over the years, numerous techniques have been suggested for approximating or interpolating pollution levels in areas without monitoring stations. Although air quality estimation methods are typically intended for stationary sites, they can also be modified to accommodate information obtained from mobile and stationary sensors. These techniques can be divided into five categories: Land Use Regression (LUR), Dispersion Models, Deterministic Interpolation Methods, Geostatistics, and ML/DL Algorithms.

In their study cited as [4], the authors employed a deep learning method for predicting the concentration of PM2.5 in Beijing, China. Their approach involves using a CNN-LSTM neural network to increase the spatiotemporal coverage by incorporating historical pollutant data, meteorological data, and PM2.5 concentrations from nearby monitoring stations. The proposed approach can capture the spatiotemporal characteristics by combining the convolutional neural network and long-short-term memory network. The study evaluated the proposed approach against other deep learning methods. Notably, this paper focused on predicting future PM2.5 concentrations rather than estimating or interpolating missing values using only fixed monitoring stations and other fixed features in the model.

In [5], the use of LUR methods by Habermann et al. to visualize NO2 pollution concentration distribution is discussed. LUR is employed due to its reliance on air pollutant concentration trends. The authors built a LUR model based on land use, demographic, and geographical

features with NO2 measurements as the dependent variable. Kriging was then used to visualize the LUR-NO2 surface for each point. The model predicted almost 60% of NO2 variability, although the authors note limitations of LUR methods in their paper.

A Multi-AP learning network was introduced in [6] for estimating pixel-wise pollution based on fixed-station measures and features such as land use, traffic, and meteorology. The authors classified features into micro, meso, and macro views and used a fully convolutional network (FCN) to simulate multiple pollutants. The Multi-AP network outperformed other methods in various experiments, although the authors acknowledge data constraints, seasonality, and model extension as potential challenges.

Guo et al. proposed a high-resolution air quality mapping approach for multiple pollutants in [7]. The method uses a dense monitoring network and combines dense networks and machine learning techniques. The authors took advantage of micro-station monitoring systems with multiple sensors, as well as land use and meteorological data. XGBoost algorithm was used to estimate pollution concentration at different grids with fine granularity. However, the monitoring phase relied on dense network data collection.

The paper by Cassard et al. [14] introduces an engine that predicts air quality for PM2.5 and PM10 concentrations in the United States. The authors employed fixed and low-cost sensors near road networks and used traffic data to build features. They utilized the five nearest official monitoring stations, the five closest low-cost sensors, as well as road and traffic features. A convolutional layer was tailored for low-cost sensors, and all features were combined and flattened before being passed through a fully connected layer. The authors considered three prediction models, including using only official stations, only low-cost sensors, or a combination of both. While integrating high-quality data from official monitoring stations with low-cost sensors can improve pollution estimation, the authors acknowledge that more spatial coverage is a potential limitation.

In [15], the authors utilized geostatic methods with data collected from low-cost mobile sensors deployed on top of trams (OpenSense [16]). The study compared kriging and deterministic methods such as IDW, where kriging approaches (simple kriging, ordinary kriging, and kriging with external drift) were found to be superior. Although geostatistical methods do not require external data, machine learning methods that combine different data types have demonstrated better performance for pollution estimation.

In [17], the authors proposed a deep autoencoder model to recover spatiotemporal pollution maps by separating the processes of pollution generation and data sampling using an encoder, decoder, and sampling imitator. The approach utilized mobile sensor data without relying on additional features and incorporated the ConvLSTM structure within the decoder based on a previous study [8].

In [9], the authors introduced HazeEst, a machine learning-based approach that combines sparse fixed stations with dense mobile sensor data to estimate hourly air pollution surfaces. The method utilized air pollution, temporal, and spatial features and merged fixed and mobile data by averaging mobile sensor measurements hourly. The approach implemented several regression methods, such as SVR, DTR, and RFR.

Song et al. proposed the Deep-Maps approach [10] to estimate PM2.5 measures. The method combined mobile sensor data with fixed stations' data to expand spatial coverage and utilized a machine learning framework that adapts gradient-boosting decision trees with local features such as land use and meteorological data. Neighboring features captured spatiotemporal correlations among urban features, while macro features represented pollution measurements from sites outside the study area.

In [18], Zhang et al. proposed machine learning regression models to predict real-time localized air quality, utilizing multiple static and IoT mobile sensors of the same type to monitor air quality effectively. The approach developed gradient boosting, SVR, and RFR regression models to estimate pollution, where the gradient boosting model was most responsive to sudden changes. The results indicated that the hybrid network had better outcomes for all selected dates.

Existing approaches in the literature that use fixed and/or mobile data have typically conducted targeted data collection campaigns on specific roads or outdoor places. However, this work aims to use MCS data to enhance fixed stations' data without relying on directed data collection campaigns or outdoor data collection.

3. Methodology

In this section, we will present our proposed methodology for enhancing fixed station measures with data obtained through mobile crowd sensing. We may have very few samples from various outdoor locations when using mobile crowd sensing. Our proposed solution aims to address the question of how to leverage MCS data to improve fixed station measures and estimate air pollution.

Air pollution levels can vary significantly from one place to another and may change rapidly due to various factors such as meteorological conditions, traffic, and land use. Despite these differences, it is possible to group these changes into clusters that reflect pollution levels during specific time periods.

Our methodology is based on the hypothesis that fixed

station measures that fall within the same pollution cluster could share similar MCS data. To test this hypothesis, we will cluster fixed station measurements and use the dates and periods to identify relevant MCS data. We will then use this data to enrich pollution maps and estimate pollution levels by combining fixed and MCS data.

Algorithm 1: Pollution estimation using Fixed and MCS data

Input: *Hourly Fixed Stations data, Hourly average MCS*

Output: *Enriched pollution estimation map*

1 Create different snapshots of pollution maps based on hourly fixed station data.
2 Apply a clustering algorithm to group those snapshots into clusters.
3 Select the date and periods within each cluster.
4 For each cluster, compute the mean of its pollution maps, and use it as the representative map for that cluster.
5 Select hourly average MCS data matching the periods extracted from each cluster.
6 Enrich each representative map with its corresponding MCS data.
7 Apply the interpolation method to estimate pollution on top of the enriched map.

Our approach is detailed in Algorithm 1, which outlines the following steps. First, we generate snapshots from the fixed station data, considering the data from all fixed stations for each timestamp as the current state of pollution. Next, we apply a clustering algorithm (such as K-means) to identify all similar snapshots. We then calculate each cluster's mean per fixed station, forming a new map representing the cluster. For example, if entries 1 and n in Figure 1 are grouped in one cluster, then the representative vector of this cluster is $([14.25, 16.6, 5.6, 17.95, 5.3, 3.8, 15.1, 17.3])$. These steps are illustrated in Figure 1.

For MCS data, we begin by calculating the hourly average. Then, using each cluster's date and time periods, we extract the relevant MCS data. The selected data enriches the representative map, as shown in Figure 2. Finally, we apply an interpolation technique to generate an air pollution estimation map, as demonstrated in Figure 3.

4. Implementation and Experiments

This section presents the data and methods utilized in our implementation, followed by a discussion of the experiments and results. Our study was conducted in Versailles, within the geographical boundaries of a specified

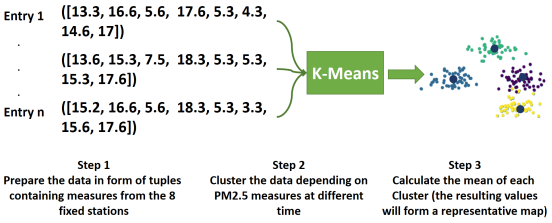

Figure 1: Clustering Fixed Stations data

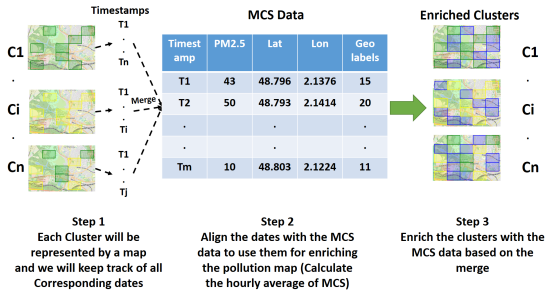

Figure 2: Enriching the representative map with MPM data

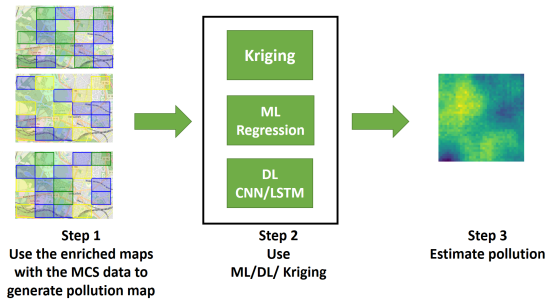

Figure 3: Pollution Estimation

bounding box (2.08170001, 48.79231, 2.1540488, and 48.8283) that covers the Versailles region. The area of this bounding box is approximately 32 square kilometers, and we partitioned the map into grids with varying spatial resolutions.

4.1. Data

Our approach to air quality data collection involves two types: fixed station measures and MPM data. eLichens [3] has deployed eight fixed stations in Versailles, which provide the air quality index and measures of particulate

[3] https://www.elichens.com/

matter ($PM1.0, PM2.5$, and $PM10$), as well as estimates of NO_2, O_3, temperature, and humidity. These sensors provide hourly aggregations, resulting in one representative record per station for each timestamp.

MPM data is collected through the Polluscope [4] project, which conducted multiple campaigns to collect mobile sensory data. Participants are recruited to collect surrounding pollutant concentration and geolocation for one week, 24 hours a day while going about their daily activities. The sensors in their multisensor boxes collect time-annotated measurements of $PM1.0, PM10, PM2.5, NO_2$, $Black\ Carbon\ (BC)$, temperature, and relative humidity. To address typical quality issues with low-cost sensors (outliers, noise, and missing values), the data is preprocessed, and thoroughly screened [19].

We combine fixed station measurements with MCS outdoor data by classifying samples based on previous research to identify micro-environments and selecting only outdoor periods [20, 21]. However, this represents less than 10% of the data due to the majority of time spent indoors. To deal with data scarcity, we propose grouping by similarity after calculating the hourly average of the MCS data to align with the fixed station data.

4.2. Methods

As illustrated in Section 3, our method for estimating pollution involves adapting unsupervised learning and geostatistics techniques. The process consists of three phases. The first phase clusters similar snapshots from fixed stations using PM2.5 measures (but the same process could apply to any other pollutant) where at each timestamp i, the measurements from the eight fixed stations $(S1_i, S2_i, ..., S8_i)$ are considered as a snapshot. Thus, the input of the clustering algorithm is a set of snapshots. The K-means algorithm is applied, and the Elbow method is used to determine the optimal number of clusters. Each cluster is represented by an aggregated record that shows the mean values of the stations in that cluster.

Moving on to the second phase, we select all the date and time values of the different snapshots in each cluster. We use these values to select samples from MPM data, which is then hourly averaged and aggregated over cells. Using each cluster's representative map with the selected MPM data, we generate an enriched map for that cluster.

In the third phase, we estimate pollution in uncovered areas using enriched maps. While many approaches, such as machine and deep learning methods like CNN-LSTM, ConvLSTM, and auto-encoder, have shown promising results, we use geostatistics methods such as ordinary kriging interpolation and deterministic interpolation IDW for simplicity. We note that additional features, such as traffic and meteorology, can also improve performance.

[4] https://polluscope.uvsq.fr/

4.3. Experiments

The experiments were carried out on real-life data collected in Versailles city. Within the context of the GoGreen Routes project, eLichens deployed eight fixed stations in Versailles. Meanwhile, MPM data were collected as part of the Polluscope project. We utilized data between October and December from both fixed and mobile sensors. The fixed stations produced roughly 1087 hourly average records.

Concerning the MPM data, the experiments affirmed the reduction in data when we limited it to an outdoor context. At the start, we had 11500 minutes of outdoor records during the collection period, out of a total of 642200 records, which comprised only 1.7% of the collected data. After filtering only records within the bounding box, we were left with approximately 2538 outdoor records out of 103062 records, which equates to roughly 2.4% of the data collected in Versailles. We observed a rapid increase in pollution levels for all fixed stations after November 26. Furthermore, we noticed that fixed sensor S4 had numerous peaks and missing values, implying that the observations were unreliable. Therefore, we conducted experiments with and without S4 measurements. For both experiments, we applied the same procedure. Firstly, we loaded data from all available stations, specifically the PM2.5 dimension. Secondly, we removed all missing values and kept only records with measurements from all available stations. Finally, we normalized the data using min-max normalization.

4.3.1. First Experiment

Once the data was preprocessed and prepared, we utilized K-means clustering to partition it into distinct clusters. For each record, there were eight measures associated with the eight stations in this particular experiment. By applying the elbow method, we designated $K = 10$, forming 10 clusters.

After grouping the fixed station measures' records, we calculated a representative map for each cluster. The next step is to select all the date and time values in each cluster, which will be used to query MPM data.

On the other hand, for MPM data, we first selected the PM2.5 dimension from the preprocessed data. Then we split the area of interest into grids where each cell has a *1KM X 1KM* granularity. The study area is approximately 32 KM^2; thus, we used nine columns and five rows to split the area of interest into 32 cells. Unfortunately, the eight stations fall in one cell, number 29. This, on the one hand, affects the accuracy as the accurate fixed stations' measurements are all in one cell. However, on the

	1KM X 1KM		*500m X 500m*	
	MAE	RMSE	MAE	RMSE
IDW	4.1	5.2	4	4.8
OK	3.4	3.9	4.5	5.6

Table 1
MAE and RMSE values of the first experiment

other hand, it shows the strength of our approach since the performed MPM enrichment allowed us to estimate pollution even if we have few fixed station measures.

We merged the MPM data for each cluster, sharing the exact date and time values. For clusters 0, 3, and 8, we did not find any MPM data that was collected at the same time as those clusters. We kept only clusters 1, 2, 4, 5, 6, 7, and 9. We merged the Mobile data with the representative map of each cluster (fixed data) to get the enriched maps having MPM data and fixed station data at resolution 1KM x 1KM x 1h.

The final step is to interpolate missing values. We use Inverse distance weighting (IDW) and the Ordinary kriging approach. For each cluster, we applied the two methods. For validation, we use leave-one-out validation, where we try to interpolate the cell's value having the fixed stations, as it is considered the ground truth. Mean absolute error (MAE) and root mean squared error (RMSE) are used as metrics for validation.

We repeated the same experiment while varying the spatial resolution. We split the area of interest into cells of *500m X 500m*. Now the fixed stations fall within two cells. We repeated the same procedure and applied the same approaches. Table 1 reports the results of MAE and RMSE for the different splits.

4.3.2. Second Experiment

We repeated the whole experiment while removing sensor S4. Now for each record, we have seven measures corresponding to the seven available fixed stations. Again, with the help of the elbow method, we set K=8, and we have 8 clusters.

Unfortunately, we did not have MPM data at the same periods in cluster 5, as cluster 5 contains only eight records. We merged MPM data and fixed station data for other clusters 0, 1, 2, 3, 4, 6, and 7, and we got the enriched maps.

The same procedure and methods as in the previous experiment were applied. We have a grid split of *1KM X 1KM* and another split of *500m X 500m*. The results are reported in table 2.

The following results correspond to the second experiment (after removing S4 measures). Figure 4 shows the enriched clusters for the second experiment. We have eight clusters if we count cluster 5. As aforementioned, cluster 5 has only 8 records and no matching MPM data.

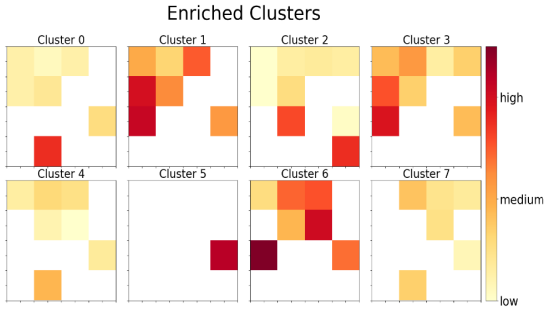

Figure 4: Enriched maps

	1KM X 1KM		*500m X 500m*	
	MAE	RMSE	MAE	RMSE
IDW	3.7	4.8	2	2.6
OK	3.4	3.9	3	4

Table 2
MAE and RMSE values of the second experiment

Therefore, applying interpolation is irrelevant for this cluster. However, as shown in the figure, the MPM data has enriched the other clusters. Initially, all the clusters have located in one *1KM x 1KM* cell. The figure shows the importance of the proposed method and how MPM values change with the change of fixed station measures. The light yellow color shows a low pollution level, while the dark red corresponds to high pollution measures.

Map plot before and after interpolation is shown in figure 5 We chose those 2 clusters to visualize the impact of interpolation when we have a low pollution level map as shown in the top part of figure 5, and a high pollution level in the bottom part. The interpolation is performed with *1KM x 1KM* resolution. The plots show the superiority of the kriging method over the IDW method.

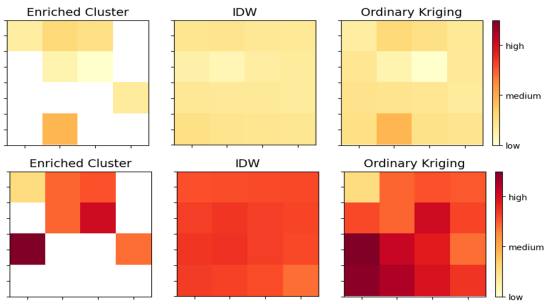

Figure 5: *1KM X 1KM* interpolation – Clusters 4 and 6

Moreover, figure 6 shows the plots of maps before and

after interpolation at *500m X 500m* resolution. Again, clusters 1 and 7 are chosen to show the impact of interpolation on spatial measures with high and low pollution levels. Based on the plots, kriging interpolation preserves the original measurements and estimates pollution levels in uncovered spots.

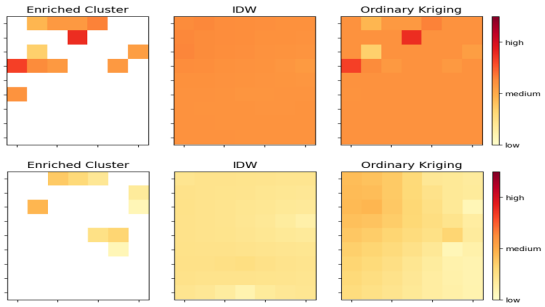

Figure 6: *500m X 500m* interpolation – Clusters 1 and 7

5. Discussion and Future work

The main focus of this study is to enhance air monitoring fixed stations by incorporating mobile sensor data collected from the public. Previous projects have typically conducted targeted mobile sensing campaigns in specific areas or along particular paths. In contrast, our study utilizes opportunistic MPM data to supplement fixed station data.

Our initial challenge was to determine how to integrate MPM data with fixed station data to estimate air pollution. We formulated a hypothesis that periods of air pollution where fixed stations' measurements fall within the same cluster could share similar MPM data. To test our hypothesis, we clustered the fixed stations' data and merged them with MPM data. Our experiments confirmed the validity of our hypothesis, and we believe that this methodology could improve the accuracy of fixed station data.

While we achieved acceptable results using basic interpolation techniques, we anticipate that more advanced geostatistics, machine learning, and deep learning techniques could enhance the performance even further. Convolutional networks could expand spatial coverage, while recurrent neural networks could expand temporal coverage.

For future research, we aim to investigate the use of deep learning for interpolation. To generalize our approach, we plan to collect more MPM data in the context of the GoGreen Routes project and seek out public datasets with community-based data collection, e.g., openaq.org, aircasting.habitatmap.org, and, if available, data from similar projects worldwide [22], [3]. One challenge we face is the distribution of fixed stations, which are all located in small areas. We hope to distribute them better to expand spatial coverage and include additional features such as meteorological data, traffic data, land use features, and other relevant factors that impact air pollution. Overall, we believe incorporating deep learning models will be critical to achieving greater accuracy in our research.

6. Conclusion

Over the past few years, monitoring air pollution through fixed stations and inexpensive and portable sensors has become a popular topic. Due to the constant concern over air quality in urban areas, improving the air quality index has become crucial in dealing with the challenges of urbanization. Several studies have attempted to estimate pollution levels using fixed stations, mobile sensors, or a combination of both, using different methodologies and possibly requiring additional features.

In this study, we present our approach, which involves combining fixed station data with mobile participatory sensing (MPM) data collected by individuals during their daily activities rather than at specific outdoor locations. This type of data collection presents a challenge, as only 10% of the time is spent outdoors, resulting in a scarcity of MPM data. To address this issue, we augment the MPM data by clustering periods with similar pollution maps based on fixed station measurements, which we hypothesize represent the overall conditions. We aggregate identical measurements into a single map for each cluster and use all corresponding periods to select the relevant MPM data, which we merge with the aggregated map to enhance the air pollution monitoring data. Finally, we employ interpolation techniques to estimate pollution levels in uncovered areas.

We tested our approach on a real-life dataset and obtained acceptable results. However, future work can be done to improve the model's accuracy and performance by adopting more advanced interpolation methods and features.

Acknowledgments

This work has been supported by the H2020 EU GO GREEN ROUTES funded under the research and innovation program H2020- EU.3.5.2 grant agreement No 869764, and by the French National Research Agency (ANR) project Polluscope, funded under the grant agreement ANR-15-CE22-0018.

References

[1] Air pollution, world health organization [online]. available:https://www.who.int/health-topics/air-pollution (2023).

[2] P. Kumar, L. Morawska, C. Martani, G. Biskos, M. Neophytou, S. Di Sabatino, M. Bell, L. Norford, R. Britter, The rise of low-cost sensing for managing air pollution in cities, Environment international 75 (2015) 199–205.

[3] C. C. Lim, H. Kim, M. R. Vilcassim, G. D. Thurston, T. Gordon, L.-C. Chen, K. Lee, M. Heimbinder, S.-Y. Kim, Mapping urban air quality using mobile sampling with low-cost sensors and machine learning in seoul, south korea, Environment international 131 (2019) 105022.

[4] A. Bekkar, B. Hssina, S. Douzi, K. Douzi, Air-pollution prediction in smart city, deep learning approach, Journal of big Data 8 (2021) 1–21.

[5] M. Habermann, M. Billger, M. Haeger-Eugensson, Land use regression as method to model air pollution. previous results for gothenburg/sweden, Procedia Engineering 115 (2015) 21–28.

[6] J. Song, M. E. Stettler, A novel multi-pollutant space-time learning network for air pollution inference, Science of The Total Environment 811 (2022) 152254.

[7] R. Guo, Y. Qi, B. Zhao, Z. Pei, F. Wen, S. Wu, Q. Zhang, High-resolution urban air quality mapping for multiple pollutants based on dense monitoring data and machine learning, International journal of environmental research and public health 19 (2022) 8005.

[8] R. Ma, X. Xu, H. Y. Noh, P. Zhang, L. Zhang, Generative model based fine-grained air pollution inference for mobile sensing systems, in: Proceedings of the 16th ACM Conference on Embedded Networked Sensor Systems, 2018, pp. 426–427.

[9] K. Hu, A. Rahman, H. Bhrugubanda, V. Sivaraman, Hazeest: Machine learning based metropolitan air pollution estimation from fixed and mobile sensors, IEEE Sensors Journal 17 (2017) 3517–3525.

[10] J. Song, K. Han, M. E. Stettler, Deep-maps: Machine-learning-based mobile air pollution sensing, IEEE Internet of Things Journal 8 (2020) 7649–7660.

[11] B. Guo, Z. Wang, Z. Yu, Y. Wang, N. Y. Yen, R. Huang, X. Zhou, Mobile crowd sensing and computing: The review of an emerging human-powered sensing paradigm, ACM computing surveys (CSUR) 48 (2015) 1–31.

[12] S. Mailler, L. Menut, D. Khvorostyanov, M. Valari, F. Couvidat, G. Siour, S. Turquety, R. Briant, P. Tuccella, B. Bessagnet, et al., Chimere-2017: From urban to hemispheric chemistry-transport modeling, Geoscientific Model Development 10 (2017) 2397–2423.

[13] X. Jurado, Atmospheric pollutant dispersion estimation at the scale of the neighborhood using sensors, numerical and deep learning models, Ph.D. thesis, Université de Strasbourg, 2021.

[14] T. Cassard, G. Jauvion, D. Lissmyr, High-resolution air quality prediction using low-cost sensors, arXiv preprint arXiv:2006.12092 (2020).

[15] Y. M. Idir, O. Orfila, V. Judalet, B. Sagot, P. Chatellier, Mapping urban air quality from mobile sensors using spatio-temporal geostatistics, Sensors 21 (2021) 4717.

[16] K. Aberer, S. Sathe, D. Chakraborty, A. Martinoli, G. Barrenetxea, B. Faltings, L. Thiele, Opensense: open community driven sensing of environment, in: Proceedings of the ACM SIGSPATIAL International Workshop on GeoStreaming, 2010, pp. 39–42.

[17] R. Ma, N. Liu, X. Xu, Y. Wang, H. Y. Noh, P. Zhang, L. Zhang, A deep autoencoder model for pollution map recovery with mobile sensing networks, in: Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers, 2019, pp. 577–583.

[18] D. Zhang, S. S. Woo, Real time localized air quality monitoring and prediction through mobile and fixed iot sensing network, IEEE Access 8 (2020) 89584–89594.

[19] B. Languille, V. Gros, N. Bonnaire, C. Pommier, C. Honoré, C. Debert, L. Gauvin, S. Srairi, I. Annesi-Maesano, B. Chaix, et al., A methodology for the characterization of portable sensors for air quality measure with the goal of deployment in citizen science, Science of the Total Environment 708 (2020) 134698.

[20] M. Abboud, H. El Hafyani, J. Zuo, K. Zeitouni, Y. Taher, Micro-environment recognition in the context of environmental crowdsensing, in: Workshops of the EDBT/ICDT Joint Conference, EDBT/ICDT-WS, 2021.

[21] H. El Hafyani, M. Abboud, J. Zuo, K. Zeitouni, Y. Taher, B. Chaix, L. Wang, Learning the micro-environment from rich trajectories in the context of mobile crowd sensing, GeoInformatica (2022) 1–44.

[22] E. Bales, N. Nikzad, N. Quick, C. Ziftci, K. Patrick, W. G. Griswold, Personal pollution monitoring: mobile real-time air quality in daily life, Personal and Ubiquitous Computing 23 (2019) 309–328.