How opportunistic mobile monitoring can enhance air quality assessment?

Mohammad Abboud, Yehia Taher, Karine Zeitouni, Ana-Maria Olteanu-Raimond

How opportunistic mobile participatory monitoring can enhance air quality assessment?

Mohammad Abboud

mohammad.abboud@uvsq.fr

Versailles Saint-Quentin-en-Yvelines University

Yehia Taher

Versailles Saint-Quentin-en-Yvelines University

Karine Zeitouni

Versailles Saint-Quentin-en-Yvelines University

Ana-Maria Olteanu Raimond

Université Gustave Eiffel

Research Article

Keywords: Air Quality Monitoring, Opportunistic Mobile Participatory Monitoring, Low-cost Sensors, Data Integration, Spatial Interpolation, Machine Learning

Posted Date: September 25th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3359951/v1

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at GeoInformatica on April 29th, 2024. See the published version at https://doi.org/10.1007/s10707-024-00516-w.

How opportunistic mobile participatory monitoring can enhance air quality assessment?

Mohammad Abboud · Yehia Taher · Karine Zeitouni · Ana-Maria OLTEAN-Raimond

Received: date / Accepted: date

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. The objective of this study is to enhance the data from fixed stations by incorporating opportunistic mobile participatory monitoring (MPM) data. The main research question we are dealing with is: How can we augment fixed station data through MPM? In order to address the challenge of limited MPM data availability, we leverage existing data collected during periods when the pollution maps align with those observed by the fixed stations. By combining the fixed and mobile data, we apply interpolation techniques to produce more accurate pollution maps. The efficacy of our approach is validated through experiments conducted on a real-life dataset.

Keywords Air Quality Monitoring, Opportunistic Mobile Participatory Monitoring, Low-cost Sensors, Data Integration, Spatial Interpolation, Machine Learning

Mohammad Abboud · Karine Zeitouni · Yehia Taher
DAVID Lab
UVSQ-Université Paris-Saclay
Versailles, France
E-mail: firstname.lastname@uvsq.fr

Ana-Maria Olteanu Raimond
IGN, LASTIG Laboratory
Univ Gustave Eiffel
Saint-Mandé, France
E-mail: ana-maria.raimond@ign.fr

1 Introduction

Air pollution has become one of the major concerns of the 21st century, especially in densely populated urban areas. The combination of urbanization and climate change poses a significant threat to the health of urban populations and the environment. The impact of air pollution on human health and the environment has been well-documented, including respiratory and cardiovascular diseases, reduced life expectancy, and ecological damage. By 2050, up to 70% of the global population is projected to 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. Indeed, according to the World Health Organization (WHO), 99% of the world's population lives in places where air quality exceeds internationally approved limits [1]. WHO's estimates show that around 7 million premature deaths per year are attributable to the combined 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. Air pollution monitoring can be highly beneficial by aiding policymakers in devising more effective strategies to tackle pollution-induced urbanization challenges.

Monitoring and estimating air pollution in uncovered spots is essential to take adequate measures to reduce air pollution. Air pollution monitoring involves the measurement of pollutants in the atmosphere, such as particulate matter, nitrogen oxides, sulfur dioxide, carbon monoxide, and ozone. This information can help identify areas with high pollution levels and determine the sources of pollution. With the advancement of technology, air pollution monitoring has become more efficient, accurate, and cost-effective.

Different air pollution monitoring approaches include fixed stations, low-cost fixed sensors, and mobile sensors [6,26,5]. Each monitoring approach has advantages and limitations, and selecting a monitoring approach depends on the specific needs of the study or monitoring program.

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.

On the other hand, low-cost fixed sensors are cheaper and easier to install than fixed stations. They can be placed in various locations, such as street lamps or buildings, and provide real-time air pollution data. However, their accuracy can be limited, and they may only measure some pollutants of interest.

Researchers have shown recent interest in using air quality mobile sensing as an alternative method for measuring air pollution [17]. Mobile sensors, such

as vehicles equipped with sensors, can capture air pollution data in specific areas or along transportation routes. They can provide high spatial resolution data but may not capture long-term trends or variations in 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 [16] and [17]. However, calibration is typically necessary for such sensors.

Fixed stations can produce precise measurements but fall short regarding 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 steady temporal coverage at specific locations, and typically last for a brief period of time.

Air pollution estimation consists in predicting air pollution concentrations at locations without monitoring equipment. This approach is beneficial for regions where monitoring stations are limited or nonexistent.

The GoGreen Routes[1] project is committed to addressing 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. Precisely, to overcome the limitations of individual air pollution monitoring approaches, we advocate that combining different approaches can yield a more comprehensive understanding of air pollution levels in uncovered spots.

Researchers have utilized fixed stations and mobile sensor data to estimate pollution maps. Some studies have relied exclusively on fixed stations [4,10,27], while others have applied air pollution estimation methods used in fixed stations to low-cost mobile sensor data [9,20]. However, recent research proposes combining data from fixed and mobile sensors [11,26]. 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. These studies raise several unresolved questions. Firstly, what are the most effective 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?

However, existing approaches work only for data collected through targeted and synchronized campaigns. Such approaches do not consider opportunistic data acquired from participants performing their real-life activities at different times and places. Nowadays, the concept of opportunistic mobile sensing is rapidly spreading. Smartphones can capture location, motion, environmental and health parameters, etc. In our study, we are trying to use opportunistic mobile data along with fixed sensor data to estimate pollution in uncovered spots. The main problem is the scarcity of opportunistic mobile data matching

[1] https://gogreenroutes.eu/

the fixed sensor measurements, leading to a low enrichment of such opportunistic data to the pollution maps.

In this study, we propose an approach allowing fixed station data enrichment with opportunistic mobile crowd-sensing data (i.e., low-cost hand-held sensors that collect data opportunistically from the crowd) to expand the spatiotemporal coverage of air pollution monitoring. Our research hypothesis is that combining these data sources makes it possible to define enriched maps that capture the spatio-temporal variability of air pollution at a higher resolution than using each source/approach separately.

This paper presents a novel approach to assessing air pollution concentrations using data from fixed and opportunistic mobile sensors. Our methodology leverages a mobile crowd-sensing (MCS) approach. MCS [8], is a new paradigm that 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. Unlike the existing approaches, this schema's main issue is coping with the scarcity of such opportunistic data in the outdoor environment in the enrichment task. From one hand, MCS data do not have a steady temporal coverage, and from another hand the amount of instantaneous data collected outdoors remains very low.

Our research question centers on the possibility to still utilize MCS/MPM[2] data to supplement fixed station data for better estimation of air pollution across the city.

Using deep learning methods, we will then use these enriched maps to quantify air pollution concentrations in uncovered spots. Deep learning methods have shown promise in predicting air pollution concentrations by learning the underlying patterns and relationships.

In order to address the challenge of limited MPM data availability, we merge the MPM data corresponding to similar general pollution conditions. Once the MPM data is aligned with the fixed station map in the same time interval, we look for similar conditions at different periods and harness the MPM data collected in these periods. To do so, we identify clusters of different fixed station data, match them with MPM data at corresponding times, and combine them to generate more data samples and improve the pollution map. This method results in enhanced pollution estimation.

Our contributions are as follows:

- Propose a method to combine fixed station data with mobile participatory monitoring data. We can create enriched maps that capture the spatiotemporal variability of air pollution in uncovered spots. This approach provides a more comprehensive understanding of air pollution levels than individual monitoring approaches and can be used to identify pollution hotspots and sources.

[2] Please note that MPM (opportunistic mobile participatory monitoring) and MCS (opportunistic mobile crowd sensing) are the same. For the rest of this paper, we will use MPM to refer to opportunistic mobile monitoring.

- Using deep learning methods to estimate air pollution levels in uncovered spots, we can expand the spatiotemporal coverage of air pollution monitoring.
- Validating our approach on top of real-life datasets from two cities in France and the USA: Versailles and Chicago.

The remainder of this paper is organized as follows: Section 2 reviews related work and different approaches discussed in the literature. Section 3 details our methodology. Section 5 presents the implementation and experimental results. In sections 6 and 7, we summarize our findings and suggest future directions for research.

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 mesoscale air quality modeling systems, such as CHIMERE [21] and other studies [25,12,29], 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 [14,15,22,24]. Besides 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. This discussion will focus on data-driven approaches that expand spatial and temporal coverage. This section summarizes the conducted pollution estimation and interpolation studies 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.

- **Land Use Regression** in short (LUR) methods that use local environmental characteristics such as land use features, meteorological features, etc., to find a correlation between those features and fixed station data and build a regression model.
- **Dispersion models** use mathematical formulations to characterize the atmospheric processes that disperse a pollutant emitted by a source. A dispersion model can predict concentrations at selected downwind receptor locations based on emissions and meteorological inputs.
- **Deterministic interpolation methods** calculate the value at the unknown location based on created surfaces from measured points. Inverse Distance Weighting is one of the most popular deterministic approaches.

It tries to interpolate the data at a specific location based on the weighted averages of collected data points.
- **Geostatistics** these techniques utilize statistical properties of the measured points. It is known by kriging method we have various types: simple Kriging, ordinary Kriging, etc. The main idea is to determine the spatial covariance of the collected data points. Then, the derived weights from the covariance structure are used to interpolate values of un-sampled points.
- **ML/DL algorithms** Machine learning and Deep learning models try to map the input into the specific output based on features from the training set. Regression models from machine learning are used to build regression models to interpolate data. In addition, CNN and LSTM are used to expand the spatial and temporal coverage.

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 the model's fixed monitoring stations and other fixed features.

In [10], 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.

In [7], authors developed a LUR model to estimate intra-city nitrogen dioxide (NO2) exposure for a Sydney cohort. They compare those estimates from a national satellite-based LUR model (Sat-LUR) and a regional Bayesian Maximum Entropy (BME) model. NO2 and NOx were measured at 46 sites. Based on local knowledge, the sites were categorized a priori: 16 as traffic sites, 28 as urban background sites, and two as regional sites. For the LUR model, the explanatory regression variables were calculated for each geocoded address, and the estimates were made using the NO2 and NOx regression equations. Hold-out validation is considered an improvement on leave-one-out cross-validation (LOOCV) validation.

A Multi-AP learning network was introduced in [27] 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 [9]. 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 and 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. [5] 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, and 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 [30], Zheng et al. proposed a semi-supervised method incorporating temporal and spatial models to estimate pollution. The approach is based on a co-training framework that utilizes an ANN to handle spatial features and a linear chain conditional random field (CFR) to handle temporal features. The authors developed a model that combines historical and real-time data with multiple heterogeneous data sources such as traffic, meteorology, and POIs (points of interest). Compared to other classical approaches, their proposed method exhibits high precision.

In [13], the authors utilized geostatic methods with data collected from low-cost mobile sensors deployed on top of trams (OpenSense [2]). 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 [18], authors developed microscale variables of the urban environment, including Point of Interest (POI) data, Google Street View (GSV) imagery, and satellite-based measures of urban form to use them as features to various pollution estimation models. The idea is to combine the traditional predictor and microscale variables to enhance the models' performance. Different modeling approaches have been adopted, such as Geostatistics and Machine learning (Stepwise Regression + Kriging, Partial Least Square + Kriging, and Machine Learning + Kriging). The authors found that the microscale variables may be a valuable substitute for traditional variables. For example, models using the

microscale variables alone performed similarly to models using the traditional variables.

In [19], 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 [20].

In [11], 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.

In [17], authors adapted mobile sampling low-cost sensors and machine learning to map urban air quality in Seoul, Korea. They collected data by conducting three weeks of campaigns across five routes with ten volunteers sharing seven AirBeams, a low-cost, smartphone-based particle counter. In contrast, geospatial data were extracted from OpenStreetMap. They applied three statistical approaches to constructing the LUR model: linear regression, random forest, and stacked ensemble. The collected air pollution data and the openly available and crowd-sourced geographical data source OpenStreetMap (OSM), were then used to construct LUR models using linear regression and machine learning methods. Notable differences between morning, evening, and night were also observed across the five routes, and the LUR model was sensitive to different segment lengths and buffer radiuses.

Song et al. proposed the Deep-Maps approach [26] to estimate PM2.5 measures. The method combined mobile sensor data with fixed stations' data to expand spatial coverage. It 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 [28], 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. At the same time, SVR and RFR were good at finding overall trends. The results indicated that the hybrid network had better outcomes for all selected dates.

In [6], authors introduce a generic neural attention model, named ADAIN (Attentional Deep Air quality Inference Network), for spatially fine-grained urban air quality inference. They adapted neural networks to model heterogeneous data in a unified way and learned complex feature interactions. Besides the air quality data, the authors use road network data, meteorological data,

POIs, etc. Both monitoring station records and urban data are leveraged, and essential features correlated with air quality are extracted.

Another research [3] proposes a holistic, multi-dimensional approach to gather, monitor, and analyze heterogeneous data sources of air pollutants and noise indicators into an integrated, intelligent computational system. The system will provide high-quality measurements and estimations, relying upon an underlying sensor network consisting of static and mobile sensors. The proposed system will collect data from various subjective and objective air quality and noise monitoring inputs. They proposed a spatial and temporal air-pollutant concentration estimation model based on environmental features.

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 uses MPM data to enhance fixed stations' data without relying on directed or outdoor data collection campaigns.

3 Methodology

This section will present our proposed methodology for enhancing fixed station measures with data obtained through mobile crowd sensing or low-cost sensors. We may have very few samples from various outdoor locations when collecting opportunistic mobile crowd-sensing data. Our proposed solution addresses this data scarcity, allowing us to utilize MPM data and fixed station measures to 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 periods.

This intuition guided our method to overcome MPM data scarcity. Hence, we hypothesize that fixed station measures within the same pollution cluster could share similar MPM data. To test this hypothesis, we will cluster fixed station measurements and use the matching dates and times to identify relevant MPM data. We will then use the union of these MPM data to enrich the pollution maps and estimate pollution using an interpolation or a prediction algorithm with a larger input sample.

The approach involves clustering the air pollution levels based on fixed station data, merging these clusters with MPM data, and applying interpolation using Convolutional Neural Networks and Long Short-Term Memory (CNN-LSTM) to estimate the values at uncovered places.

The methodology for enriching air pollution fixed station data with data collected from mobile crowd sensing or low-cost fixed stations involves clustering, data enrichment, and interpolation. Our approach is detailed in Algorithm 1, which outlines the following steps. First, identify an area of interest to estimate air pollution in that area, and split it into cells using a grid view (Step 1). Then, the available fixed station measures are assigned to their corresponding cells in the map. After that, creating hourly map snapshots based on the

Algorithm 1: Pollution estimation using Fixed and MPM data

Input: *Hourly Fixed Stations measures, MPM measures*
Output: *Enriched pollution estimation map*

1 Split the area of interest into cells using a grid view.
2 Create different snapshots of pollution maps based on **hourly** fixed station measures.
3 Apply a clustering algorithm to group those snapshots into clusters having the same pollution levels.
4 Select the timestamps of measures within each cluster.
5 Calculate the hourly average of MPM data.
6 Select hourly average MPM data matching the timestamps extracted from each cluster.
7 Enrich fixed station snapshot with the available average MPM measures sharing the same timestamps.
8 Adapt an estimation approach to interpolate values in the remaining uncovered spots on top of the enriched map.

pollution levels measured by the fixed stations (Step 2). Clustering the created snapshots is the next step. The algorithm groups the air pollution levels based on the similarity of their values. At this stage, we consider snapshots of different timestamps altogether without distinguishing between rush hours or working and holidays. However, the rush hours or holidays may eventually belong to the same cluster if the corresponding pollution maps are similar. This clustering aims to identify the timestamps sharing the same pollution conditions (Step 3). In the next step, we keep track of timestamps to merge them with MPM data to produce enriched maps (Step 4). Table 1 shows an example of the snapshots representing the measurements of different fixed stations at a timestamp T_i. For instance, the first snapshot corresponds to the pollution map at period T_1 with sensor values (13.3, 16.2, ..., 12.1, 10.9) for the fixed sensors in order S_1, S_2, ..., S_{k-1}, and S_k. The values here represent the same type of measures (e.g., PM2.5); thus, they have the same scale and range. Therefore, we simply used the Euclidean distance between the vectors of sensor values in K-means.

The algorithm output is K clusters where each cluster groups together pollution maps based on their similarity (i.e., the similarity of their fixed sensor vector values). The corresponding timestamps are also returned to match the timestamps of mobile sensors. These steps are described in figure 1.

Time Periods	Vectors of Sensor Values
T_1 = 2023-06-15 18:00:00	(13.3, 16.2, ... , 12.1, 10.9)
.	.
T_i = 2023-06-30 14:00:00	(3.4, 2.2, ... , 1.1, 0.9)
.	.
T_n = 2023-07-14 08:00:00	(6.3, 8.2, ... , 10.1, 5.7)

Table 1: Fixed Stations' Snapshots Example

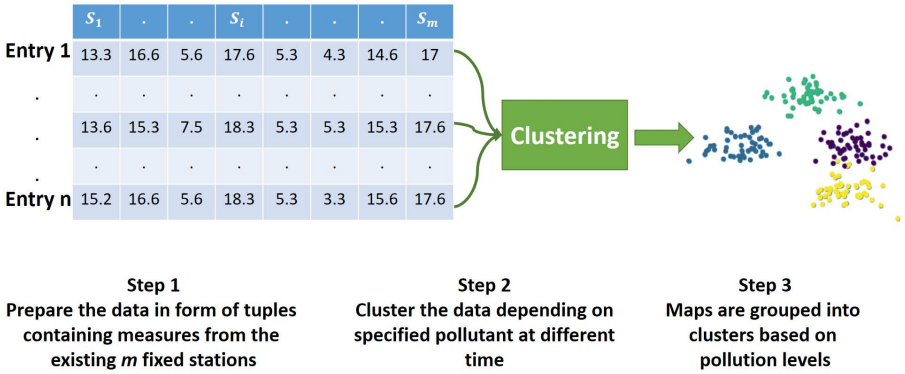

Fig. 1: Clustering Fixed Stations Data

The next step in the methodology is to enrich the data collected from fixed stations with MPM data collected. To do this, we use the date-time values of the measures in each cluster to merge those measures with the mobile crowd sensing or low-cost fixed stations data. We begin by calculating the hourly average of MPM data to have unified timestamps. Then, merging is performed using the date-time values as the common identifier. We add available MPM data to unmonitored cells. Cells containing fixed station measures are not changed. The result is enriched clusters containing data from fixed stations and MPM measures (steps 5 - 7). Figure 2 describes the previous steps.

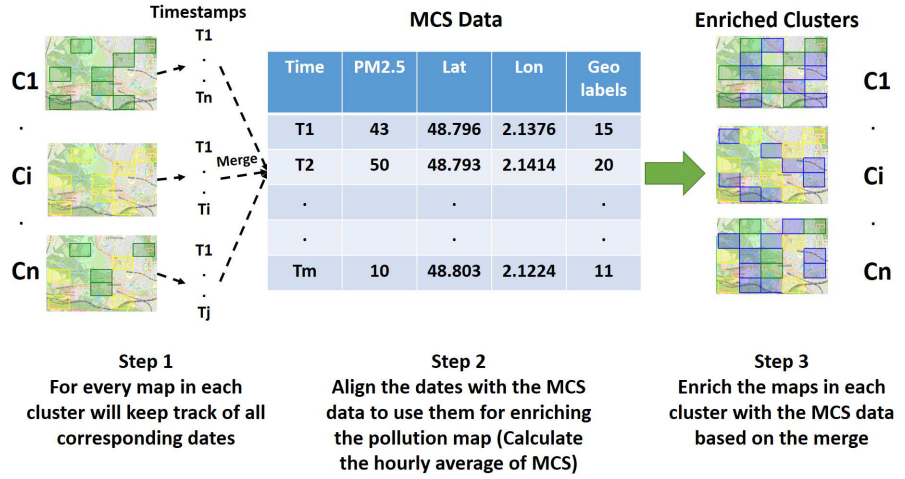

Fig. 2: Enriching the representative map with MPM data

The final step in the methodology is to use interpolation to estimate the air pollution levels at uncovered places. Interpolation is a technique used to

estimate the value of a variable at a point that is not explicitly measured. Based on the available features, we can select the appropriate technique and perform interpolation to estimate pollution levels at uncovered spots (Step 8). As shown in figure 3, the selected model inputs the enriched maps and outputs the estimated map.

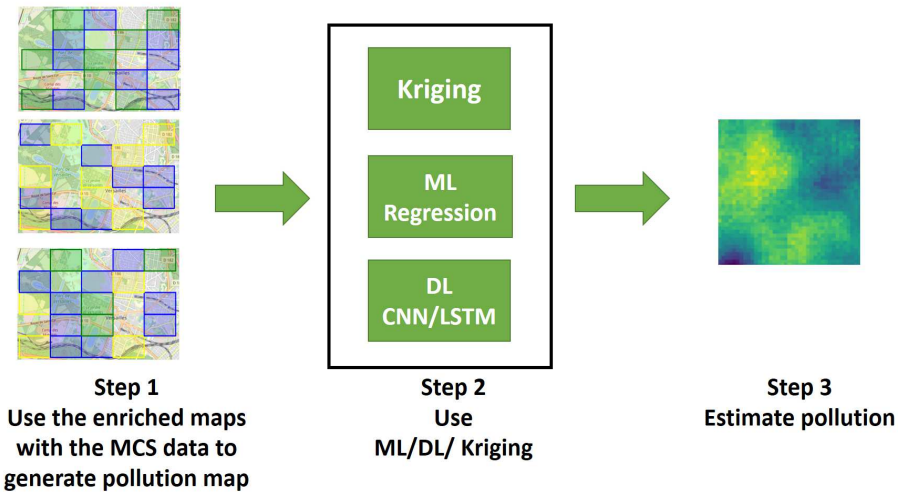

Fig. 3: Approach for Pollution Estimation

Our methodology uses unsupervised machine learning algorithms for clustering and interpolation methods for air pollution estimation. Our approach can be used to create air pollution maps that provide a comprehensive view of air pollution levels in a given area. The maps can be used to identify areas with high pollution levels.

4 Implementation

This section details the implementation pipeline of our air pollution enrichment and estimation approach. Figure 4 shows the implementation pipeline that includes several parts: data collection, data preprocessing, data enrichment, and air pollution estimation.

4.1 Data Collection

This subsection will focus on data collection from two study areas, Versailles and Chicago. We collect data from fixed, low-cost, and mobile sensors.

In **Versailles city**, we collected data from 14 June until 16 July 2023. The collection includes data from 6 fixed stations spread over different spots

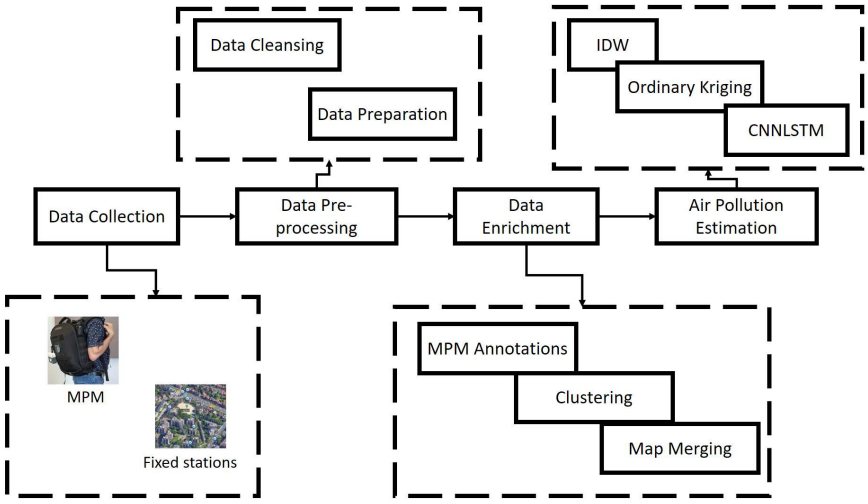

Fig. 4: Implementation Pipeline

in Versailles. We used the AtmoTube Pro sensor for opportunistic MPM data to collect air pollution measures. With the help of ten volunteers, we collected around 4000 minutes of outdoor records. For fixed stations, we had around 700 hourly average records. The study area in this experiment only covers Versailles city with an area around 27 km^2.

Fixed sensors were deployed by eLichens[3] as part of the GoGreen Routes project in Versailles. These sensors monitor air quality and provide data for analysis. Fixed stations provide reliable data over a longer period and are suitable for long-term air quality monitoring. These sensors measure of particulate matter ($PM1.0, PM2.5$, and $PM10$), as well as estimates of Nitrogen dioxide NO_2, Ozone O_3, temperature, and humidity. They provide hourly aggregations, resulting in one representative record per station for each timestamp.

MPM data is collected with the help of ten volunteers. They used AtmoTube Pro sensor which collects $PM1.0$, $PM10$, $PM2.5$, $VOCs$, temperature, and relative humidity. The collection was performed opportunistically, as the participants conducted real-life activities. The data collected from mobile crowdsensing is valuable as it provides a high spatial resolution of air quality data.

To generalize our approach, in the study's second phase, we seek out public datasets with community-based data collection from Aircasting[4] and OpenAQ[5] specifically in **Chicago city**.

OpenAQ is a platform that provides data from various sources, including fixed stations and low-cost sensors. The data from OpenAQ provides a more

[3] https://www.elichens.com/
[4] https://www.habitatmap.org/aircasting
[5] https://openaq.org/

comprehensive picture of air quality by combining data from different sources. This platform helps compare data from different sources and identify patterns in air quality over time. Using the provided API[6], we collected reference grade measures (fixed stations) and low-cost fixed station measures.

On the other side, Aircasting is a platform that provides data from low-cost sensors, which are small, portable, and easy to install. These sensors measure carbon monoxide, nitrogen dioxide, and particulate matter. The data from Aircasting provides valuable information for monitoring air quality in real time. Using AirCasting API[7], we were able to acquire mobile participatory data.

We collected fixed stations, low-cost sensors, and mobile sensor data within a bounding box of 288 km^2 in Chicago. The data collection took place between October and December. The fixed stations produced roughly 1304 hourly average records. For the low-cost fixed sensors, we have 40575 minutes of data. At the same time, the length of MPM data was originally 368276 records at the seconds' timescale, which results in 2515 minutes of data after averaging the measures.

For the sake of simplicity, we restricted our experiments to PM2.5, which is the most available in both fixed and mobile sensors. However, our method can be applied to any environmental measure.

4.2 Data Pre-processing

Preprocessing and data preparation are essential steps in data analysis as they ensure the data is clean, organized, and ready for analysis. This study followed a series of steps to preprocess and prepare the data for analysis.

First, we selected an area of interest for our study. This area could be a city, a neighborhood, or any other geographical region we wanted to analyze. Once we had selected the area of interest, we split it into cells using a grid view. We used a specified granularity of either 1km x 1km or 500m x 500m, depending on the level of detail we wanted in our analysis. This step allowed us to analyze air quality data at a more granular level and identify hotspots or areas of concern.

For MPM data, we filtered the GPS data with the help of scikit-mobility [23]. GPS data often contain noise and outliers that can affect the accuracy of the data. To remove noise and spikes, we used the scikit mobility library. This library is designed for mobile data analysis and provides various data cleaning and preprocessing functions.

The MPM data and low-cost sensor data usually have high-frequency sampling rates. On the contrary, the fixed stations provide an hourly average of pollution levels. Thus, we calculated the hourly averages for mobile participatory monitoring data and low-cost sensor data to have a unified sampling rate.

[6] https://docs.openaq.org/docs/about-api
[7] https://github.com/HabitatMap/AirCasting

This step allowed us to identify air quality patterns over time and compare air quality across different times of day or days of the week.

Finally, after ensuring all the data was cleaned and relevant, we assigned the collected data to their proper cells in the map using the GPS coordinates.

4.3 Data Enrichment

Data enrichment is a crucial step in air pollution analysis. This section provides a more detailed description of the data enrichment process used in our analysis.

We used the data collected from the fixed stations to create clusters of different pollution levels. We grouped the pollution levels based on the fixed stations' data, which allowed us to better understand the spatial distribution of pollution levels over time. This clustering was done using unsupervised machine-learning techniques such as k-means. For each timestamp, we used the vector of air pollution levels from all available fixed stations as the input to the clustering model. The output of this model is the different clusters, where each cluster represents different periods with the same pollution conditions.

After clustering the fixed-station pollution maps, MPM data are merged with the cluster that matches their acquisition time and propagated to the whole validity periods in that cluster. This increases the enrichment power of the MPM data while maintaining their relevance. The output of this step is to augment the data spatial coverage in the same way in each cluster. The resulting coverage can be different from one cluster to another.

We can then integrate this data with the air quality data using timestamp and location as the common variables. This provides valuable insights into the factors contributing to air pollution and informs the development of effective air quality management strategies.

4.4 Air Pollution Estimation

Air pollution estimation estimates the concentration of pollutants in the air at a given location and time. This section discusses interpolation using traditional and deep learning methods (CNN-LSTM). We are using sensory data, and the estimation is conducted on top of pollution maps generated from fixed stations and enriched using opportunistic MPM.

This paper proposes an approach to estimate air pollution using sensory data only. We use the enriched maps created in the previous step 4.3 to do this. We apply three methods for interpolation: Inverse distance weighting (IDW), ordinary kriging, and CNN-LSTM to expand the spatial and temporal coverage.

IDW is a simple and commonly used method for interpolation. It works by assigning a weight to each observation based on its distance to the location is estimated. The weights are then used to calculate the estimated value at the location of interest. The closer the observation is to the location of interest, the higher the weight assigned to that observation.

Ordinary kriging is another interpolation method that considers the spatial autocorrelation of the data. This method assumes that the spatial correlation between the observations decreases with increasing distance between them. It uses this information to estimate the value at the location of interest based on the values of the neighboring observations.

CNN-LSTM is a more complex method that combines convolutional neural networks (CNN) and long short-term memory (LSTM) networks. CNNs are commonly used for image processing, while LSTMs are used for sequence modeling. In this case, we represent each cell by the n nearest stations and their distances. Figure 5 describes the architecture of our CNN-LSTM model. The CNN-LSTM network is then trained to learn the spatiotemporal patterns in the data and estimate the value at the location of interest. In this approach, each cell is represented by the n nearest stations and their distances. The model's output is the air pollution level at the specified cell.

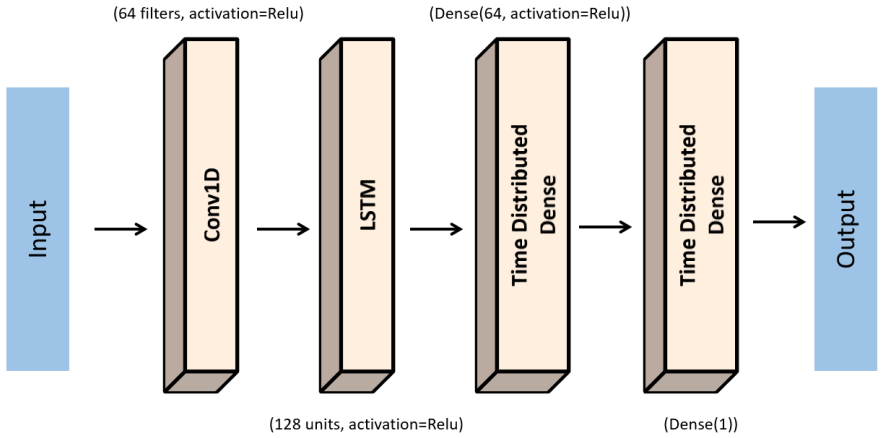

Fig. 5: CNN-LSTM Architecture

5 Experiments and Results

In order to validate our approach, we conducted experiments on two different datasets. The first dataset was collected in Versailles, France, and the second was collected from Chicago, USA. Our approach was applied in all experiments equally, starting with preprocessing and data preparation steps, enriching, and finally, estimation.

Our work aimed to determine the optimal number of clusters (K) in the K-means algorithm for our experiments. We evaluated several commonly used methods to determine K, including the Elbow method, Calinski-Harabasz Index, Davies Bouldin Index, and Silhouette Score.

After applying the mentioned methods to our data, we chose the best K value determined by most methods as our experiment cluster numbers. As for the parameter settings in spatial interpolation methods, that is, the power of distance weight and the variogram, we found through experimentation that linear distance weighting where p = 1 for IDW and the linear variogram for Ordinary Kriging performs best in terms of mean absolute error and mean squared error. Therefore, we applied them in the following experiments.

5.1 Versailles Experiment

The experiment was carried out on real-life data collected in Versailles City as described in 4.1.

Firstly, we loaded data from all available stations, precisely 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. Once the data was preprocessed and prepared, we utilized K-means clustering to partition it into distinct clusters. Using the different approaches for choosing the best K for the clustering, we set $K = 3$ in our experiment, forming 3 clusters. Figure 6 shows the mean of each station per the three clusters. The records with low pollution levels correspond to cluster 0, those with medium pollution levels were grouped in cluster 2, and those with high pollution levels fall in cluster 1.

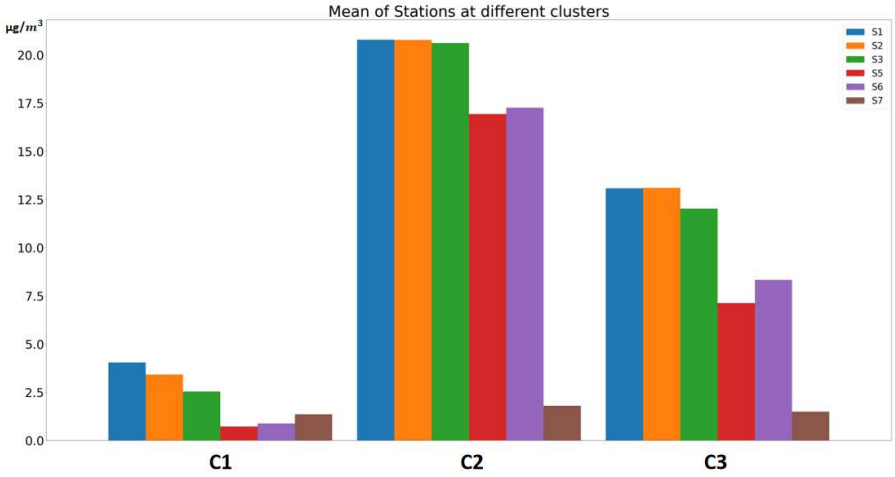

Fig. 6: Mean of fixed stations per clusters - Versailles

Then, we split the map into two granularities to enable a more detailed analysis of our data. The first granularity was set to *1km x 1km*, while the second is *500m x 500m*. The stations were distributed over five cells using the

1km x 1km granularity. While with 500m x 500m granularity, they were spread over six cells. We preprocess and prepare the MPM data and assign it to the proper cells.

For CNN-LSTM, we used a feature vector that included the values of the three nearest neighbors having stations and the distance between the current cell and the nearest neighbors having stations. Specifically, for each cell in the map, we calculated the distance to the nearest neighbors with stations and included their corresponding values in the feature vector.

We use leave-one-out validation (cell containing fixed station "ground truth"), 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.

The distribution of the fixed stations allows experimentation for the comparison between using only fixed stations' maps or using the enriched maps based on our approach. The first part reports the results using only fixed station measures; the second part presents the results of merging both data types based on the proposed approach.

Figure 7 shows the plots of a sample from each cluster at the 1km x 1km granularity while using only fixed station measures to generate the pollution maps. While figure 8 shows the plots at a finer granularity (500m x 500m).

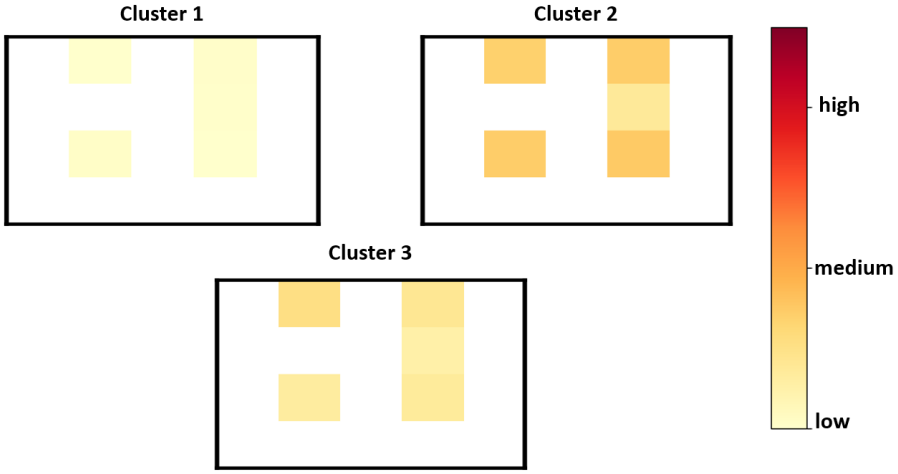

Fig. 7: Fixed Stations' Maps 1km x 1km Granularity - Versailles

Table 2 reports the results of MAE and RMSE for the interpolation using only fixed stations at different granularities. Figures 9 and 10 show the plots of applying IDW and Ordinary Kriging on random samples in the dataset for 1km x 1km and 500m x 500m granularity respectively. On the other side, figures 11 and 12 plots the estimation using the CNN-LSTM method for the two

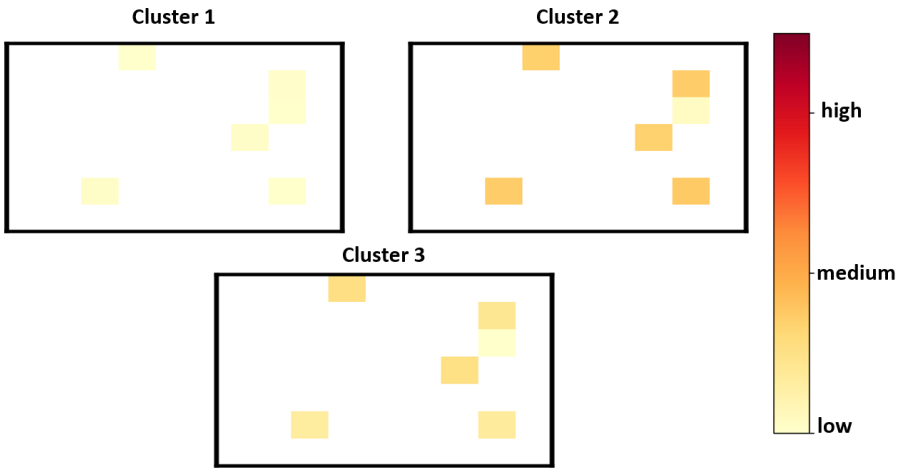

Fig. 8: Fixed Stations' Maps 500m x 500m Granularity - Versailles

granularities 1km x 1km and 500m x 500m respectively. The plots correspond to different samples from different clusters.

	1km x 1km		*500m x 500m*	
	MAE	RMSE	MAE	RMSE
IDW	2.95	4.34	6.45	9.30
Ordinary Kriging	2.85	4.22	6.20	9.09
CNN-LSTM	2.95	3.75	6.36	9.41

Table 2: MAE and RMSE (Fixed Stations) - Versailles

After enriching the pollution maps with the opportunistic MPM data following the proposed approach, we repeated the same experiments. Figure 13 shows the plots of a sample from each cluster at the 1km x 1km granularity after enriching the pollution maps with the opportunistic MPM data. While figure 14 shows the plots at a finer granularity (500m x 500m).

Table 3 reports the results of MAE and RMSE for the interpolation using enriched maps at different granularities. Figures 15 and 16 show the plots of applying IDW and Ordinary Kriging on random samples in the dataset for 1km x 1km and 500m x 500m granularity respectively. On the other side, figures 17 and 18 plots the estimation using the CNN-LSTM method for the two granularities 1km x 1km and 500m x 500m respectively. The plots correspond to different samples from different clusters.

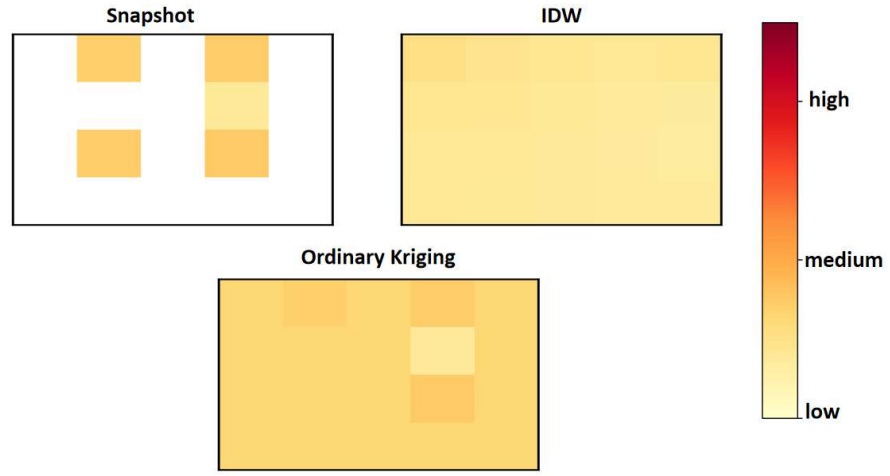

Fig. 9: IDW and Ordinary Kriging 1km x 1km (Fixed Stations) - Versailles (Cluster 2)

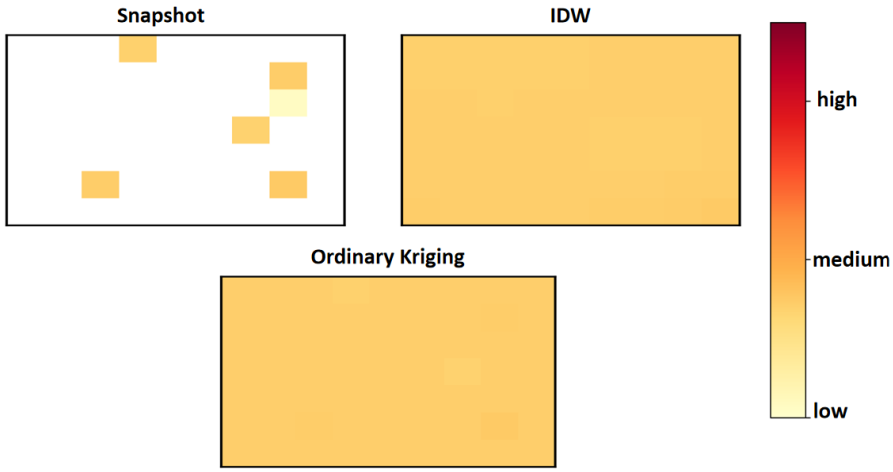

Fig. 10: IDW and Ordinary Kriging 500m x 500m (Fixed Stations) - Versailles (Cluster 2)

5.2 Chicago Experiments

In order to validate our approach, we applied our methodology to data collected from open datasets such as OpenAQ and Aircasting as described in 4.1 section.

This experiment has two types of opportunistic sensing data besides the fixed station measures. We have low-cost fixed sensors that provide measures

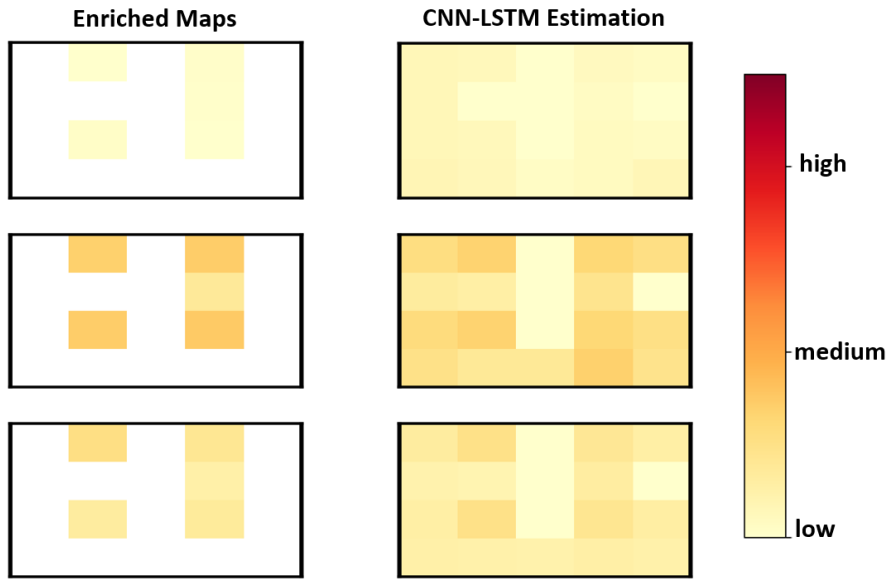

Fig. 11: CNN-LSTM 1km x 1km (Fixed Stations) - Versailles

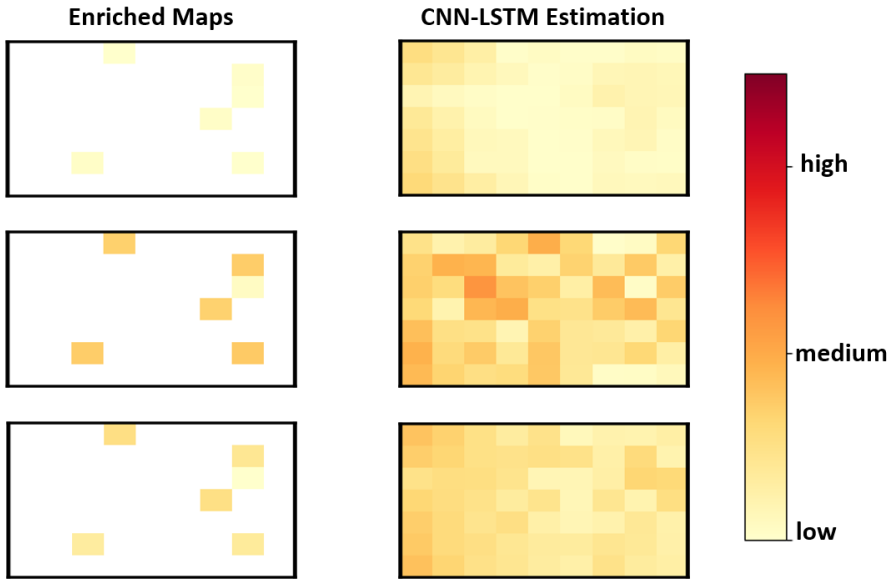

Fig. 12: CNN-LSTM 500m x 500m (Fixed Stations) - Versailles

at a specific place but only sometimes provide measures. In addition, we have the opportunistic MPM data as in the previous experiment.

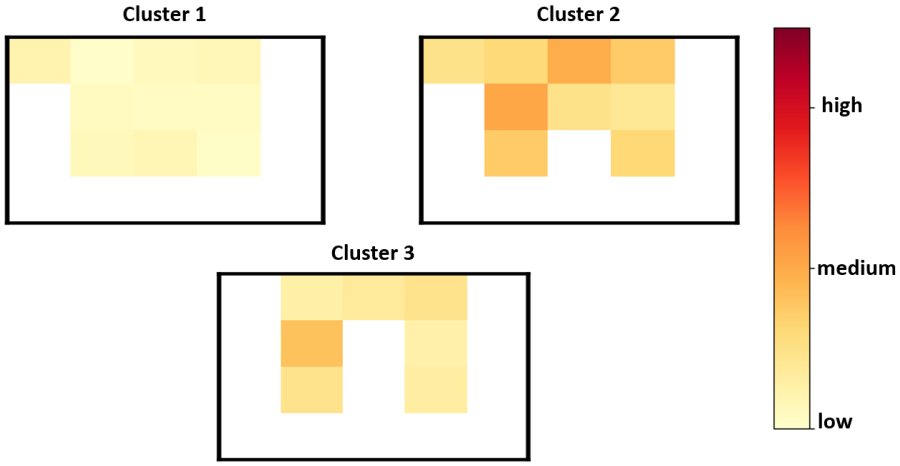

Fig. 13: Enriched Maps 1km x 1km Granularity - Versailles

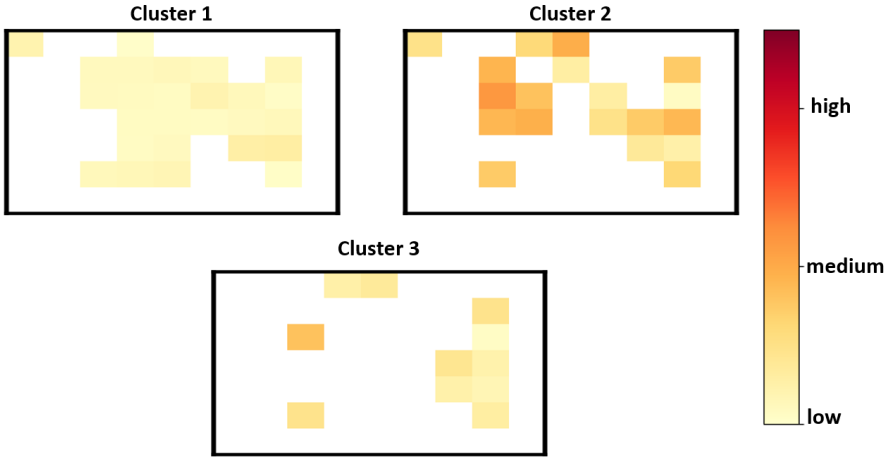

Fig. 14: Enriched Maps 500m x 500m Granularity - Versailles

Once the data was preprocessed and prepared, we utilized K-means clustering to partition it into distinct clusters. For each record, three measures were associated with the three reference grade stations in the area of interest. Using the different approaches for choosing the best K for the clustering, we set $K = 4$ in our experiment, forming 4 clusters. Figure 19 shows the mean of each station per the four clusters. Records with low pollution levels correspond to cluster 2, records between low and medium pollution levels are grouped in cluster 1, records with medium pollution levels correspond to cluster 3, and records with high pollution levels fall in cluster 0.

	1km x 1km		*500m x 500m*	
	MAE	RMSE	MAE	RMSE
IDW	0.63	0.65	5.25	7.34
Ordinary Kriging	1.03	1.22	5.75	7.82
CNN-LSTM	0.20	0.39	3.24	5.51

Table 3: MAE and RMSE - Versailles

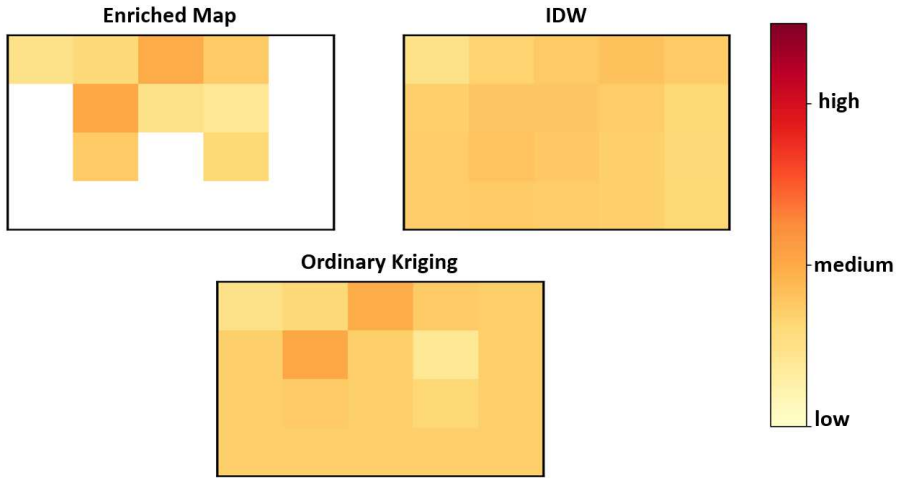

Fig. 15: IDW and Ordinary Kriging 1km x 1km - Versailles (Cluster 2)

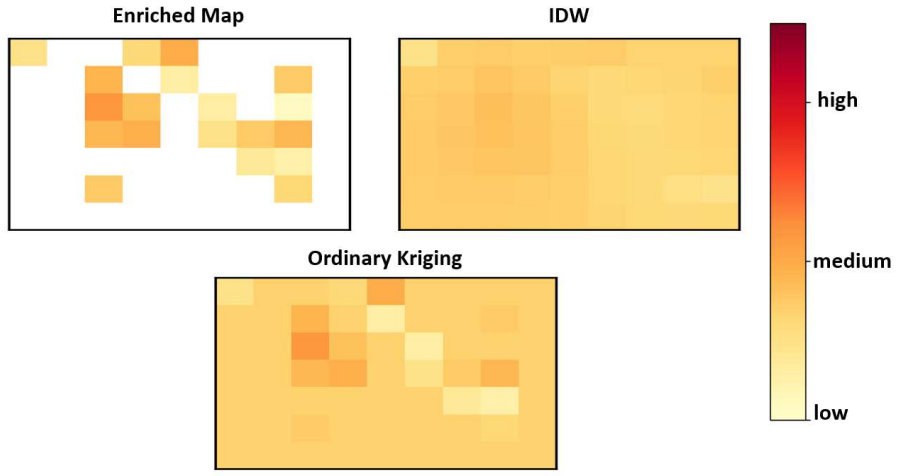

Fig. 16: IDW and Ordinary Kriging 500m x 500m - Versailles (Cluster 2)

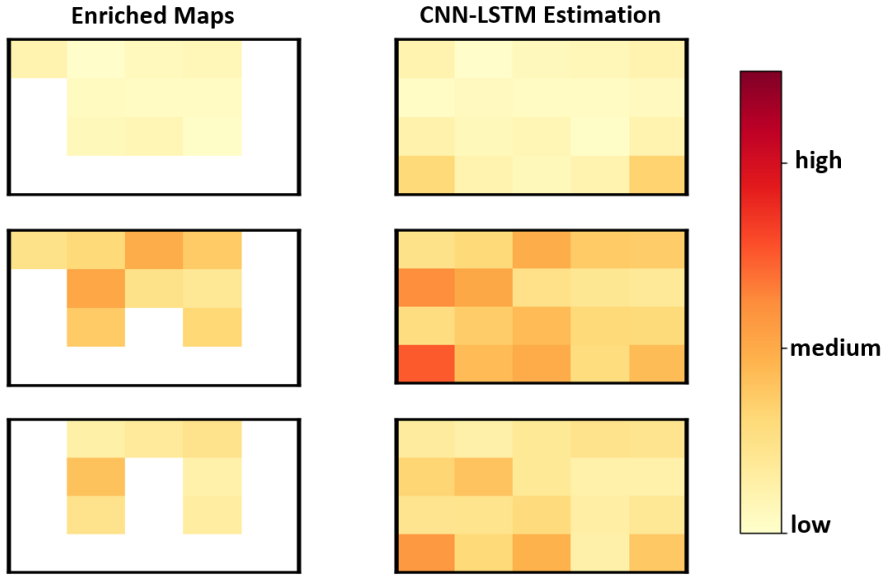

Fig. 17: CNN-LSTM 1km x 1km - Versailles

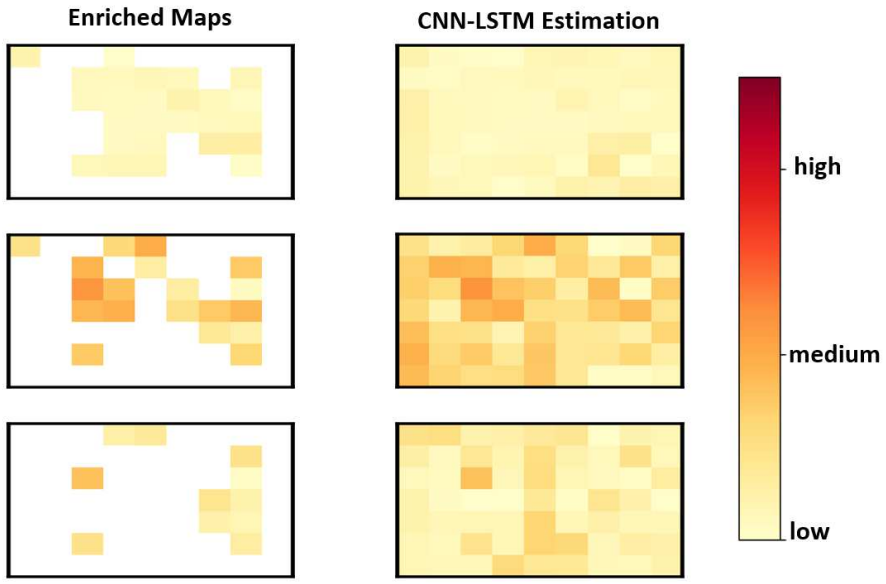

Fig. 18: CNN-LSTM 500m x 500m - Versailles

We applied the same settings as the previous experiment for the opportunistic data. We first selected the PM2.5 dimension from the preprocessed data. Then, we split the map into two granularities to enable a more detailed

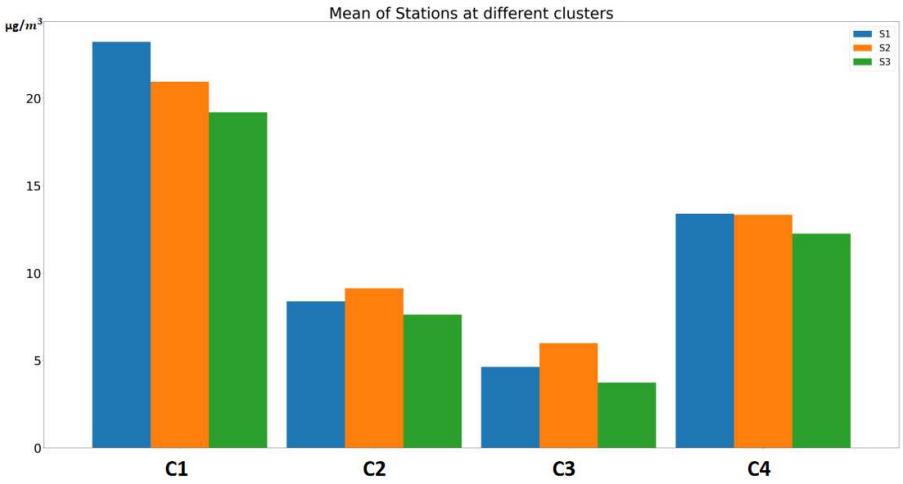

Fig. 19: Mean of fixed stations per clusters - Chicago

analysis of our data. The first granularity was set to *1km x 1km*, while the second is *500m x 500m*.

Figure 20 shows the plots of a sample from each cluster at the 1km x 1km granularity after we enriched the fixed stations' data with MPM data. While figure 21 shows the plots at a finer granularity (500m x 500m).

Fig. 20: Enriched Maps 1km x 1km Granularity - Chicago

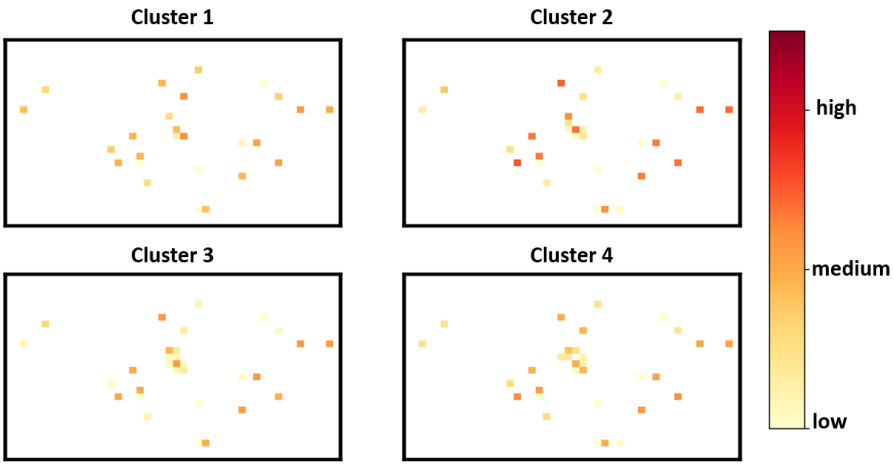

Fig. 21: Enriched Maps 500m x 500m Granularity - Chicago

We applied the same methods in experiments conducted in Versailles 5.1. Also, the same validation metrics, MAE and RMSE, apply here. The validation is also performed using leave-one-out validation.

Table 4 reports the results of MAE and RMSE for the different splits. The results show a significant improvement in using CNN-LSTM to estimate pollution levels. Figures 22 and 23 show the plots of applying IDW and Ordinary Kriging on random samples in the dataset for 1km x 1km and 500m x 500m granularity respectively. On the other side, figures 24 and 25 plots the estimation using the CNN-LSTM method for the two granularities 1km x 1km and 500m x 500m respectively. The plots correspond to different samples from different clusters.

	1km x 1km		*500m x 500m*	
	MAE	RMSE	MAE	RMSE
IDW	7.381	8.211	6.307	6.893
Ordinary Kriging	7.979	8.651	7.389	7.993
CNN-LSTM	0.793	0.917	0.804	1.017

Table 4: MAE and RMSE - Chicago

6 Discussions

Our primary goal in this study is to expand the spatiotemporal coverage. We are enhancing air monitoring fixed stations by incorporating mobile sensor data collected from the public. Previous projects have typically conducted

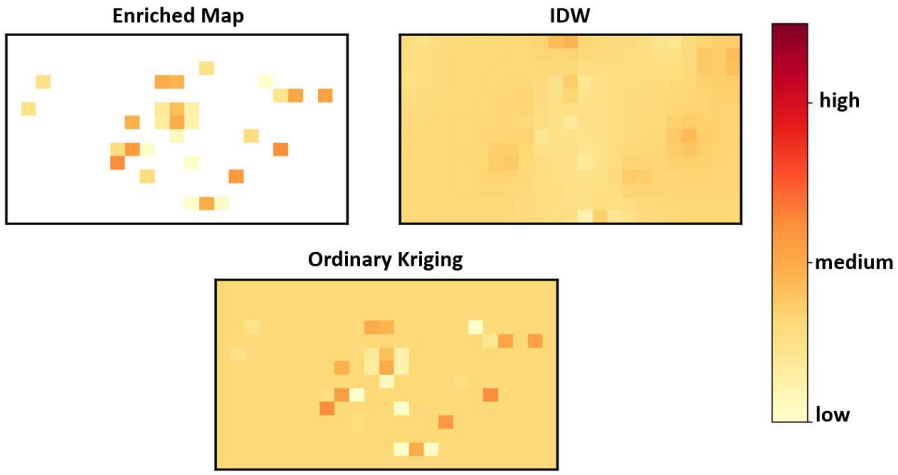

Fig. 22: IDW and Ordinary Kriging 1km x 1km - Chicago (Cluster 4)

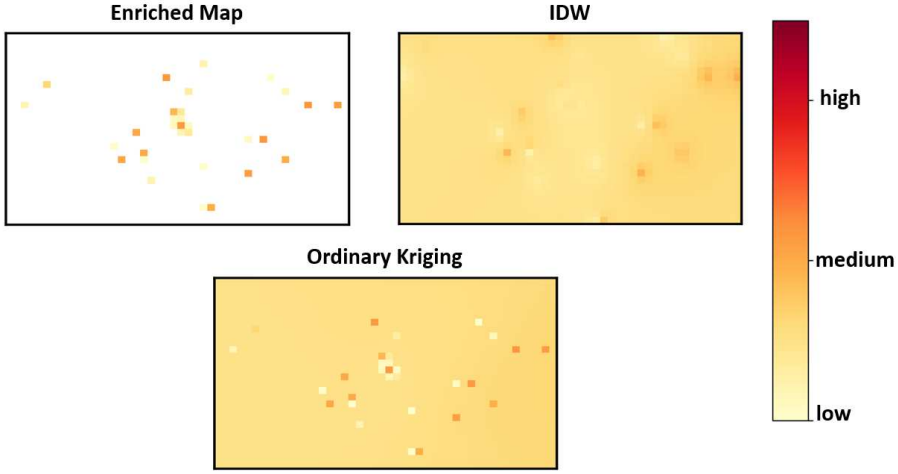

Fig. 23: IDW and Ordinary Kriging 500m x 500m - Chicago (Cluster 4)

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 figure out how to combine opportunistic MPM data with fixed-station data for estimating air pollution. We hypothesized that periods of air pollution where fixed stations' measurements fall within the same cluster could share similar MPM data. We clustered the fixed stations' data to test our hypothesis and merged them with MPM data. Our experiments

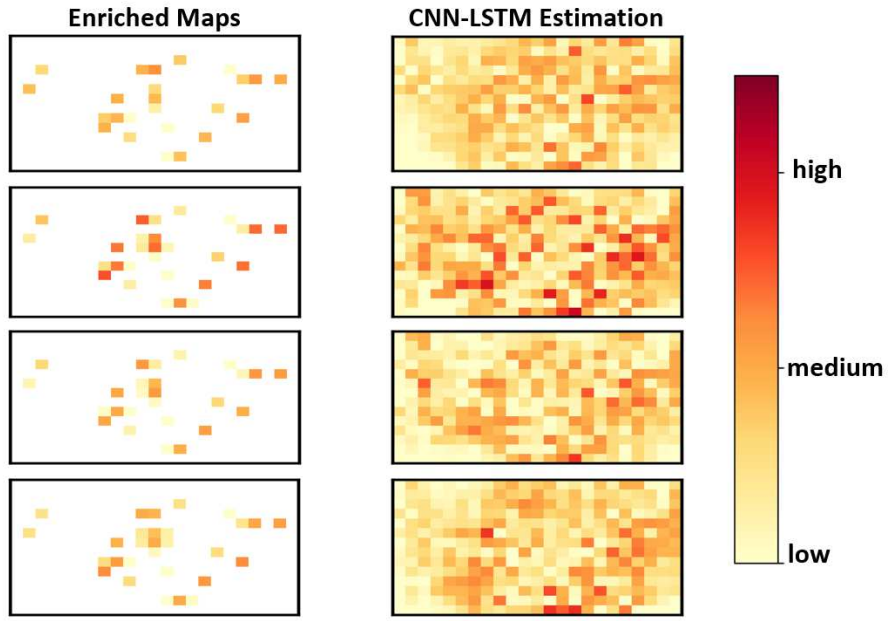

Fig. 24: CNN-LSTM 1km x 1km - Chicago

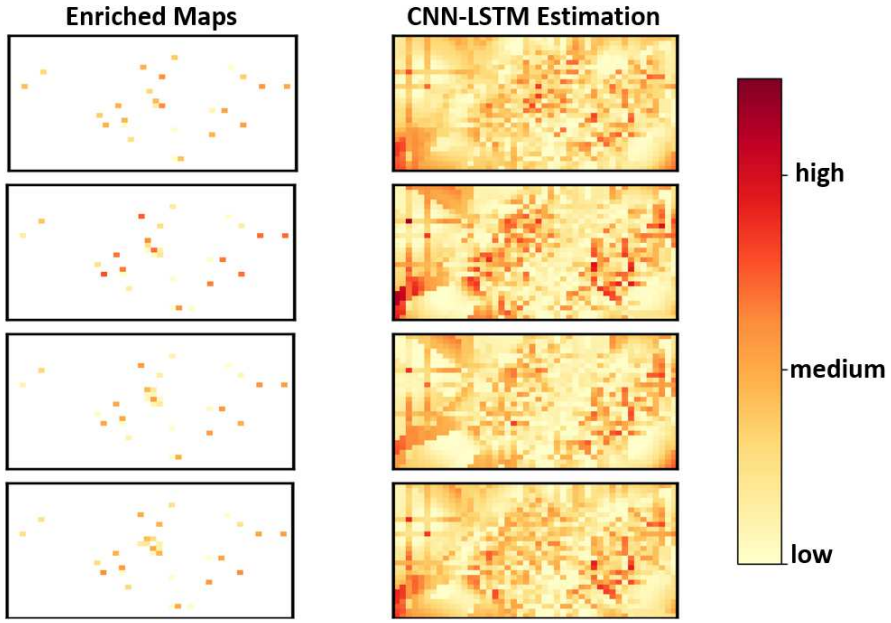

Fig. 25: CNN-LSTM 500m x 500m - Chicago

	Min MPM	Max MPM	50 - Percentile MPM	75 - Percentile MPM	90 - Percentile MPM	Coverage Before Enrichment	Coverage After Enrichment
Versailles Experiment	0	6	1	2	2	41.6 %	58.3 - 83.3 %
Chicago Experiment	0	24	18	19	19	1.39 %	8.3 - 10.1 %

Table 5: Monitoring Coverage Before and After Enrichment

confirmed the validity of our hypothesis, and this methodology could improve the accuracy of fixed station data.

This section will analyze and interpret the experiments' results of estimating air pollution using fixed and opportunistic Mobile Participatory Monitoring (MPM) data. The experiments aimed to evaluate the effectiveness of the proposed approach of enriching the fixed stations generated air pollution maps with opportunistic MPM data. We used two basic interpolation techniques, Inverse Distance Weighting (IDW) and Ordinary Kriging; we then utilized CNN-LSTM to enhance the estimation accuracy.

In **Versailles experiment**, we compared the results using fixed stations alone for estimation versus the combination of fixed stations and opportunistic MPM data following the proposed approach. The results interpreted in the experiment section under Versailles experiments 5.1 show the efficiency of our proposed approach. When using the enriched air pollution maps to estimate air pollution, we had better results in terms of MAE and RMSE using all the interpolation techniques. Table 2 and table 3 summarizes the error of air pollution estimation using different techniques in Versailles city while using only fixed stations data and while using the combination of fixed stations and opportunistic MPM data respectively.

In table 2, we notice that the MAE and RMSE for all the methods are similar. Even using advanced techniques such as CNN-LSTM, we still had a high MAE and RMSE. Hence, the model couldn't learn the correct patterns while using only fixed stations. On the other side, in table 3, MAE and RMSE decreased significantly for all the used methods, and CNN-LSTM has shown the best results among the used techniques.

Moreover, we used data from open datasets such as OpenAQ and Aircasting to better validate our approach **Chicago experiment**. We ran our approach on top of the available data in the area of interest. The reported results in section 5.2 illustrate the effectiveness of our approach in better-estimating air pollution when following the proposed approach. Results in table 4 show acceptable results in terms of MAE and RMSE, especially for the CNN-LSTM model.

Table 5 summarizes the Mobile Participatory Monitoring data statistics in the two experiments for 1KM x 1KM granularity. Min MPM and Max MPM columns refer to the minimum and maximum number of sensors available in a time period. The percentiles show that we have a very low contribution of MPM data to the map. Coverage Before Enrichment column shows the percentage of the monitored area. Coverage After Enrichment shows the percentage of the monitored area after we applied our enrichment process. Each

cluster has a percentage as the number of MPM sensors varies between clusters. In *Versailles Experiment*, the coverage after enrichment is 83.3%, 75%, and 58.3% for clusters one, two, and three, respectively. While for *Chicago Experiment* 8.3%, 9.7%, 9%, 10.1% are the coverage percentages in clusters one, two, three, and four, respectively.

Moreover, figure 26 plots the map of the monitoring coverage before and after enrichment for one sample of the maps in Versailles. Black squares represents the original averaged MPM data collected at that time. Red squares denote the averaged MPM data after enrichment. After enrichment if we have MPM and fixed stations data in one cell we take the measurements of the fixed stations only as they are more precise. The map after enrichment refers to first cluster where the coverage after enrichment reaches around 83%. It is clear that using our approach, we can expand the spatiotemporal coverage. The output of the enrichment phase is an enriched map with better observations. Thus, we can have better estimation when applying interpolation.

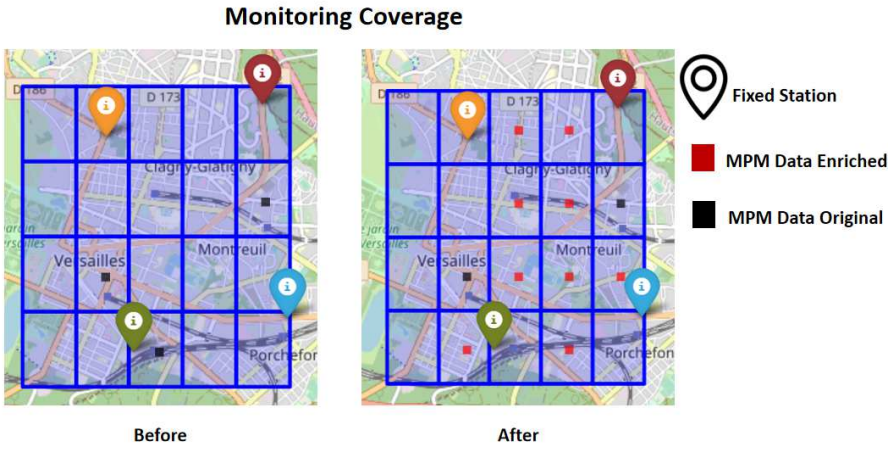

Fig. 26: Monitoring Coverage Before and After Enrichment

The experiments' results indicate that the proposed method performs well even when using only sensory data. This finding is valuable, suggesting that the approach can be utilized in areas with limited access to extensive supplementary data.

IDW and Ordinary Kriging may have exhibited suboptimal results when the density of observations varies significantly across various sites. When certain places have many measurements while others have none, IDW and Ordinary Kriging tend to over-smooth the interpolated values, resulting in an unsatisfactory depiction of local variations and sudden changes in the data. Advanced interpolation techniques, such as machine and deep learning-based approaches that account for the spatial characteristics of the data and the un-

derlying processes, may yield more accurate and reliable results. For example, the CNN model can handle the spatial characteristics of the data. As for the temporal characteristics, recurrent neural networks such as LSTM can perform well. That's why we have used the CNN-LSTM interpolation to handle both spatial and temporal characteristics of the data and have a better estimation of pollution. However, even while using advanced techniques, we still have some concerns. Indeed, in some areas where no observations are found, the model tends to overestimate values, such as in the left corner of the plots in figure 25. This suggests to investigate the explainability of these models. A possible solution to work around the estimation near the border is to retain only the results within the convex hull of the dataset.

Moreover, there are still opportunities for further improvement. Augmenting the approach with additional features can result in better estimation. We believe that enrichment with air pollution features such as land use, traffic, and meteorological data has significant improvements. These features can give more insights and provide valuable context into the factors influencing air pollution levels. Integrating such data could refine the accuracy of the estimation and provide more comprehensive predictions.

The experiments show that the proposed method successfully estimates air pollution measurements, especially when incorporating opportunistic MPM data and leveraging deep learning models like CNN-LSTM. The technique shows promise for further improvement in air quality estimate with potential advancements by adding new significant features.

7 Conclusion

Air pollution monitoring using fixed stations and low-cost and mobile sensors has been a trendy topic over the last few years. Air quality is a permanent concern in urban areas, as improving the air quality index can help face urbanization challenges. Several studies tried to interpolate pollution measures from fixed stations, mobile sensors, or a combination. They use different methods and may require additional features.

In this study, we present an approach that combines fixed station data with mobile participatory sensing 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. The mobile sensing data, in our case, have different characteristics than the data used in previous approaches. MPM data is not collected initially to monitor air pollution outdoors and in specific places. We were interested in using the opportunistic MPM data in enriching fixed stations' data to better estimate air pollution in uncovered spots.

The primary objective was to leverage the opportunistic MPM data and utilize it to enhance the fixed stations' measures to generate enriched air pollution maps to estimate air pollution better. Our approach has been validated using two real-world datasets from Versailles and Chicago cities. The reported

results show our proposed approach's applicability and efficiency in estimating air pollution in unmonitored spots. For now, we have validated the feasibility of our approach using sensory data. However, we believe involving more air pollution-related features such as land use, meteorological, and traffic features can significantly improve the estimation performance, especially when utilizing deep learning models such as the CNN-LSTM model.

Acknowledgements This work has supported by the H2020 EU GO GREEN ROUTES funded under the research and innovation programme H2020- EU.3.5.2 grant agreement No 869764. We would like to express our sincere gratitude to all the participants who contributed in the data collection in Versailles. We are thankful to all volunteers as their invaluable contributions were instrumental in enriching the quality and depth of our research.

References

1. Air pollution, world health organization [online]. available: https://www.who.int/health-topics/air-pollution (2023)
2. Aberer, K., Sathe, S., Chakraborty, D., Martinoli, A., Barrenetxea, G., Faltings, B., Thiele, L.: Opensense: open community driven sensing of environment. In: Proceedings of the ACM SIGSPATIAL International Workshop on GeoStreaming, pp. 39–42 (2010)
3. Bardoutsos, A., Filios, G., Katsidimas, I., Krousarlis, T., Nikoletseas, S., Tzamalis, P.: A multidimensional human-centric framework for environmental intelligence: Air pollution and noise in smart cities. In: 2020 16th International Conference on Distributed Computing in Sensor Systems (DCOSS), pp. 155–164. IEEE (2020)
4. Bekkar, A., Hssina, B., Douzi, S., Douzi, K.: Air-pollution prediction in smart city, deep learning approach. Journal of big Data **8**(1), 1–21 (2021)
5. Cassard, T., Jauvion, G., Lissmyr, D.: High-resolution air quality prediction using low-cost sensors. arXiv preprint arXiv:2006.12092 (2020)
6. Cheng, W., Shen, Y., Zhu, Y., Huang, L.: A neural attention model for urban air quality inference: Learning the weights of monitoring stations. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32 (2018)
7. Cowie, C.T., Garden, F., Jegasothy, E., Knibbs, L.D., Hanigan, I., Morley, D., Hansell, A., Hoek, G., Marks, G.B.: Comparison of model estimates from an intra-city land use regression model with a national satellite-lur and a regional bayesian maximum entropy model, in estimating no2 for a birth cohort in sydney, australia. Environmental research **174**, 24–34 (2019)
8. Guo, B., Wang, Z., Yu, Z., Wang, Y., Yen, N.Y., Huang, R., Zhou, X.: Mobile crowd sensing and computing: The review of an emerging human-powered sensing paradigm. ACM computing surveys (CSUR) **48**(1), 1–31 (2015)
9. Guo, R., Qi, Y., Zhao, B., Pei, Z., Wen, F., Wu, S., Zhang, Q.: 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**(13), 8005 (2022)
10. Habermann, M., Billger, M., Haeger-Eugensson, M.: Land use regression as method to model air pollution. previous results for gothenburg/sweden. Procedia Engineering **115**, 21–28 (2015)
11. Hu, K., Rahman, A., Bhrugubanda, H., Sivaraman, V.: Hazeest: Machine learning based metropolitan air pollution estimation from fixed and mobile sensors. IEEE Sensors Journal **17**(11), 3517–3525 (2017)
12. Hu, Z.: Spatial analysis of modis aerosol optical depth, pm2. 5, and chronic coronary heart disease. International journal of health geographics **8**(1), 1–10 (2009)
13. Idir, Y.M., Orfila, O., Judalet, V., Sagot, B., Chatellier, P.: Mapping urban air quality from mobile sensors using spatio-temporal geostatistics. Sensors **21**(14), 4717 (2021)

14. Jurado, X.: Atmospheric pollutant dispersion estimation at the scale of the neighborhood using sensors, numerical and deep learning models. Ph.D. thesis, Université de Strasbourg (2021)
15. Jurado, X., Reiminger, N., Benmoussa, M., Vazquez, J., Wemmert, C.: Deep learning methods evaluation to predict air quality based on computational fluid dynamics. Expert Systems with Applications **203**, 117294 (2022)
16. Kumar, P., Morawska, L., Martani, C., Biskos, G., Neophytou, M., Di Sabatino, S., Bell, M., Norford, L., Britter, R.: The rise of low-cost sensing for managing air pollution in cities. Environment international **75**, 199–205 (2015)
17. Lim, C.C., Kim, H., Vilcassim, M.R., Thurston, G.D., Gordon, T., Chen, L.C., Lee, K., Heimbinder, M., Kim, S.Y.: Mapping urban air quality using mobile sampling with low-cost sensors and machine learning in seoul, south korea. Environment international **131**, 105022 (2019)
18. Lu, T., Marshall, J.D., Zhang, W., Hystad, P., Kim, S.Y., Bechle, M.J., Demuzere, M., Hankey, S.: National empirical models of air pollution using microscale measures of the urban environment. Environmental Science & Technology **55**(22), 15519–15530 (2021)
19. Ma, R., Liu, N., Xu, X., Wang, Y., Noh, H.Y., Zhang, P., Zhang, L.: 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, pp. 577–583 (2019)
20. Ma, R., Xu, X., Noh, H.Y., Zhang, P., Zhang, L.: Generative model based fine-grained air pollution inference for mobile sensing systems. In: Proceedings of the 16th ACM Conference on Embedded Networked Sensor Systems, pp. 426–427 (2018)
21. Mailler, S., Menut, L., Khvorostyanov, D., Valari, M., Couvidat, F., Siour, G., Turquety, S., Briant, R., Tuccella, P., Bessagnet, B., et al.: Chimere-2017: From urban to hemispheric chemistry-transport modeling. Geoscientific Model Development **10**(6), 2397–2423 (2017)
22. Murga, A., Sano, Y., Kawamoto, Y., Ito, K.: Integrated analysis of numerical weather prediction and computational fluid dynamics for estimating cross-ventilation effects on inhaled air quality inside a factory. Atmospheric Environment **167**, 11–22 (2017)
23. Pappalardo, L., Simini, F., Barlacchi, G., Pellungrini, R.: scikit-mobility: A python library for the analysis, generation, and risk assessment of mobility data. Journal of Statistical Software **103**(1), 1–38 (2022). DOI 10.18637/jss.v103.i04. URL https://www.jstatsoft.org/index.php/jss/article/view/v103i04
24. Santiago, J.L., Martín, F., Martilli, A.: A computational fluid dynamic modelling approach to assess the representativeness of urban monitoring stations. Science of the total environment **454**, 61–72 (2013)
25. Simpson, D., Benedictow, A., Berge, H., Bergström, R., Emberson, L.D., Fagerli, H., Flechard, C.R., Hayman, G.D., Gauss, M., Jonson, J.E., et al.: The emep msc-w chemical transport model–technical description. Atmospheric Chemistry and Physics **12**(16), 7825–7865 (2012)
26. Song, J., Han, K., Stettler, M.E.: Deep-maps: Machine-learning-based mobile air pollution sensing. IEEE Internet of Things Journal **8**(9), 7649–7660 (2020)
27. Song, J., Stettler, M.E.: A novel multi-pollutant space-time learning network for air pollution inference. Science of The Total Environment **811**, 152254 (2022)
28. Zhang, D., Woo, S.S.: Real time localized air quality monitoring and prediction through mobile and fixed iot sensing network. IEEE Access **8**, 89584–89594 (2020)
29. Zhang, Y., Zhang, X., Wang, L., Zhang, Q., Duan, F., He, K.: Application of wrf/chem over east asia: Part i. model evaluation and intercomparison with mm5/cmaq. Atmospheric Environment **124**, 285–300 (2016)
30. Zheng, Y., Liu, F., Hsieh, H.P.: U-air: When urban air quality inference meets big data. In: Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 1436–1444 (2013)