On the Data Analysis of Participatory Air Pollution Monitoring Using Low-cost Sensors

Mohamed Anis Fekih, Walid Bechkit, Hervé Rivano

On the Data Analysis of Participatory Air Pollution Monitoring Using Low-cost Sensors

Mohamed Anis Fekih
CITI Laboratory EA 3720
Univ Lyon, INSA Lyon, Inria
Villeurbanne, France
mohamed-anis.fekih@insa-lyon.fr

Walid Bechkit
CITI Laboratory EA 3720
Univ Lyon, INSA Lyon, Inria
Villeurbanne, France
walid.bechkit@insa-lyon.fr

Hervé Rivano
CITI Laboratory EA 3720
Univ Lyon, INSA Lyon, Inria
Villeurbanne, France
herve.rivano@insa-lyon.fr

***Abstract*—Participatory sensing leverages population density and involves citizens in the collection of extensive data in multiple fields such as air pollution monitoring, enabling large-scale deployments and improving the knowledge of air quality. This study highlights the potential of low-cost sensors through a data analysis of pollutant concentrations collected during multiple sensing campaigns we co-organized using a participatory sensing platform we designed. We first compare the estimation quality of four statistical models and investigate the impact of sampling frequency on the quality of estimation and energy consumption of the nodes using an energy model based on the sensing duty cycle. In addition, we evaluate the capacity of regression models to recover missing data of one sensor based on the other sensors. Results are satisfactory and reveal that a small decrease in the sampling frequency slightly reduces the estimation quality, but in contrast, allows the nodes to operate on a longer period.**

***Index Terms*—Air quality, low-cost sensors, participatory monitoring.**

I. Introduction

Air pollution is one of the main problems facing many urban cities around the world. Indeed, despite countries' efforts, air pollution remains a serious concern, especially in and around big cities due to the massive growth of urbanization and industrialization. In 2016, 7 millions deaths were reportedly linked to indoor and outdoor pollution, according to the World Health Organization (WHO) [1]. In fact, exposure to high concentrations of pollutants over an extended period of time poses serious health problems such as heart disease, reduced lung function and respiratory infections [2], [3]. Among pollutant chemicals and particles, $PM_{2.5}$ (particulate matter with a diameter of 2.5 μm or less) is a wide range of solid or liquid particles that are small enough to be inhaled by people and travel deep in the lungs [4]. According to the European Environment Agency (EEA), an increase in $PM_{2.5}$ concentration of 10 $\mu g/m^3$ increase in the risk of mortality of 6.2 % [2]. Moreover, many natural and human-made sources contribute in the creation of $PM_{2.5}$ making it significantly hard to control compared to other pollutants.

Various solutions have been adopted by many cities around the world in order to mitigate the adverse effects of air pollution, such as greening public transport and building bike lanes. However, these solutions have to be supported by a fine-grained knowledge of air quality. Air pollution monitoring is traditionally performed using networks of fixed sensing stations equipped with various sensing probes that can accurately measure a plethora of environmental parameters such as meteorological conditions, ozone (O_3) and Particulate Matter (PM) [5]. However, despite being accurate, these monitoring stations are extremely expensive and too big to be deployed anywhere and in large numbers. Therefore, these networks are sparse and deployed in small numbers even in big metropolitan cities, which significantly limits the spatial knowledge of the phenomena given the dynamic nature of air pollution [5]. Thus, to help cut air pollution and adopt appropriate policies, local authorities need to achieve a solid knowledge of the phenomenon.

Recent advances in environmental sensing technologies and communication protocols have paved the way for the emergence of small, energy efficient and low-cost air quality sensors which provide new opportunities and unlock new capabilities compared to conventional stations [6]. In fact, as their name implies, these sensors have a remarkably reduced operational and maintenance cost, hence, opening the door for large scale deployments. In addition, owing to their small size, low-cost sensors can be mounted on fixed or mobile platforms and even carried by people. However, despite all these positive aspects, these sensors present low accuracy characteristics and stability problems along with a frequent need for calibration. For instance, PM sensors are based on light scattering technique and their measurements highly depend on the shape and density of particles, which presents a challenge when converting to mass count [7].

The recent and rapid development in Internet of Things has paved the way for a new sensing model known as participatory sensing. This paradigm has gained a lot of attention in recent years due to its potential to enable the collection of extensive data by leveraging the population density [8]. In air quality monitoring, participatory sensing presents a great advantage as it involves citizens in the process of monitoring the air they breathe and therefore increase their awareness of the subject. Following this logic, we have developed a participatory air quality sensing platform based on small and low-cost nodes

This work has been supported by the "LABEX IMU" (ANR- 10-LABX-0088) of Université de Lyon, within the program "Investissements d'Avenir" (ANR-11-IDEX-0007) operated by the French National Research Agency (ANR).

that use long-range communication technology and do not require any special training for participants [9]. Therefore, after the first measurement campaigns that we have conducted, we present in this paper the analysis of the collected data and its potential.

The focus of this work is to highlight the potential benefits of using low-cost sensors in estimating air pollution using regression models. We will investigate the impact of the sensing frequency of sensors on the performance of air pollution estimation and the possible advantage or disadvantage of lowering the sampling frequency and what it can add to the system in terms of power consumption and estimation accuracy. In addition, we will evaluate the capacity of regression models to recover/fill missing data of one sensor based on the other sensors of the network, which will give an indication of the overall performance of the system and the degree of correlation between the different sensors.

The remaining of this paper is organized as follows. Section II presents related research work. The area of interest and the data set used in this work are introduced in IV. The regression models used in this study are described in Section V. Section VI discusses the impact of the sampling frequency on the pollution estimation quality. We evaluate the capacity of the system to predict a sensor's measurement using other sensors in Section VII. Finally, Section VIII concludes the paper.

II. Related Work

Air quality sensing using low-cost sensors has gained great attention in the recent years. Over the last decade, multiple low-cost pollution monitoring systems have been developed and their contribution in improving air quality estimation tested. For instance, the Citi-Sense-MOB monitoring system [10] featured mobile sensing nodes measuring multiple parameters such as NO_2, O_3, and temperature and that can be mounted on electrical bicycles. The solution offered a web-based visualization interface and aimed to increase the global awareness around air pollution. Airsense [11] proposed personal battery-powered nodes for air quality monitoring integrating $PM_{2.5}$, temperature/humidity sensors. The nodes do not implement a wireless transmission, but store the measurements on an SD card that the user can read.

A mobile air quality sensing was performed in [12] across five predefined routes in Seoul, South Korea using seven AirBeams, a low-cost and smartphone-based $PM_{2.5}$ sensor. The collected data along combined with geospatial information were used to compared three statistical models: Land-Use Regression, Random Forest, and Stacked Ensemble which combines predictions of multiple machine learning algorithms. Results showed good performance across all models with stacked ensemble achieving the lowest Root Mean Squared Error (RMSE= 5.22 $\mu g/m^3$), outperforming both random forest (6.2 $\mu g/m^3$) and linear regression (7.01 $\mu g/m^3$).

Synthetic measurements were used in [13] to build Nitrogen Dioxide (NO_2) concentration map considering up to 4500 bike tracks randomly generated across the city of Marseille, France. The simulated observations were generated using a numerical model and the geographical information was collected from Open Street Map. The two were combined to train and compare the performance of three statistical models: Kriging, Land-Use Regression, and a Neural network. The metrics of the comparison demonstrated that Kriging offered the best performance in terms of Pearson correlation coefficient, Mean Absolute Error (MAE), and RMSE. The comparison also revolved around the impact of the number of bikes and the sampling distance.

Data from satellite acquisition combined with air temperature and relative humidity measurements collected during four mobile sensing campaigns were used in [14] to estimate air temperature in the city of Lyon, France. The estimations were performed with three regression approaches: Multiple Linear Regression, Partial Least Square Regression, and Random Forest. The outcome of the study revealed a superiority of the Random Forest over the Multiple Linear Regression and Partial Least Square Regression.

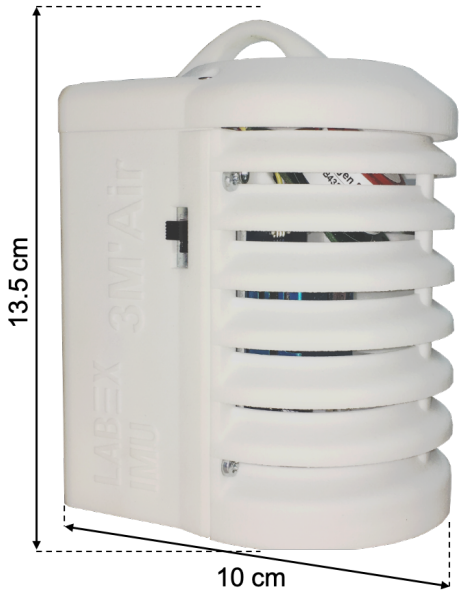


Fig. 1. 3M'Air sensing node

III. Scope of this work and presentation of our platform

This study is part of 3M'Air ("Mobile Citizen Measurements and Modeling: Air Quality and Urban Heat Islands"), a multidisciplinary project that explores the potential of participatory sensing to increase local understanding of air quality and urban heat islands. In our previous work [9], we have designed a participatory air quality and urban heat islands monitoring system based on a four-layer architecture and features small, low-cost, battery-powered, and portable air pollution sensing nodes (see Fig 1). The nodes are driven by an Arduino MKR WAN1300 and are equipped with a Nitrogen Dioxide (NO_2) sensing probe, a temperature/relative humidity sensor, and a low-power laser dust sensor measuring

three sizes of particulate matters (PM_1, $PM_{2.5}$ and PM_{10}). In addition, the designed nodes have a low-power GPS receiver to geolocate measurements, a local storage in case of communication failures, and an anti-solar radiation shield to protect the sensing probes. The measurements are performed at a configurable sampling rate (20 seconds by default) and sent over LoRaWAN which is a prominent Low-Power Wide-Area Network (LPWAN) technology for Internet of things. All this powered by a small lithium-ion polymer battery.

Following the same participatory principle, the developed solution relies on "The Things Network" LoRaWAN infrastructure, which is a global and open data network that provides numerous free-to-use LoRaWAN gateways mainly deployed by volunteers. For measurement visualization, our solution offers a web-based user interface on which it is possible to visualize measurements and get other statistics about the platform (further details on the design and validation of the platform can be found in [9]).

IV. Area of interest and Data source

We have conducted four sensing campaigns in the agglomeration of Lyon, which is located in the south-eastern region of France. It comes third in the ranking of the largest metropolis in France with over 1.4 million people over 533.6 km^2 [15]. Our work focuses mainly on the "Presqu'île" peninsula located in the heart of the city of Lyon and is bordered by two rivers, the Saône on the west and the Rhône on the east. The sensing campaigns took place during June and October 2019 and gathered on average 10 participants from different backgrounds (students, non-scientific participants, etc.). The measurement were performed across pre-defined routes, and for performance evaluation purposes, some routes were affected to more than one sensor/participant.

Although low-cost sensors allow for large-scale deployments, some areas of interest might be challenging to sample enough. Estimation models allow to approximate the air pollution concentration levels where there is no measurement in the area of interest. These models require extra features to explain the phenomenon and complement the measurements obtained by fixed or mobile sensors. Such features, or explanatory variables may include information about meteorology conditions (temperature, humidity, etc.), traffic network (number and length of roads, etc.), land-use (number of buildings, vegetation, etc.), and population density.

The data set used in this work comes from multiple sources. The $PM_{2.5}$ concentrations were collected using our low-cost sensing nodes during the sensing campaigns. Meteorological conditions were provided by "Météo-France" [1]. Traffic and land-use information were obtained from Data Grand Lyon [2] and Open Street Map. A pre-processing on these variables has been performed to exclude redundant and non-significant ones, resulting in more than 40 explanatory variables. It is worth mentioning that the explanatory variables are available all over the city of Lyon and with a spatial resolution of 20 meters. We have assigned to each measured point by our nodes the explanatory variables of the nearest point on the map.

[1]https://donneespubliques.meteofrance.fr/

[2]https://data.grandlyon.com/

V. Comparison of regression models

Although the emergence of low-cost WSNs opened the doors for dense networks deployment, monitoring systems still cannot have measurement at each point of the area of interest due to infrastructure and budget limitations. Therefore, there is a constant need to preform interpolation to be able to generate pollution maps. Spatial interpolation methods give an estimation of unmeasured data based on available samples of the same studied variable [16]. In order to get insight on which model is more suitable to use with our data, we have compared the performance of four models: Land-Use Regression (LUR), K-Nearest Neighbors (KNN), Extreme Gradient Boosting (XGBoost), and a multi-layer perceptron neural network (MLP).

Land-Use Regression builds a relationship between the studied phenomenon and the surrounding geographical parameters to predict the concentration values at points where no data was collected. Generally, LUR models rely on these explanatory variables such as meteorological parameters (i.e., temperature, humidity, wind direction and speed, etc.), land-use data (type of routes, distance to main routes), and traffic information to explain the studied phenomenon [16], [17].

K-Nearest Neighbors applies a distance metric to explanatory variables in order to find the K nearest points in terms of similarity, then it averages these data to estimate the concentration at the point where no data is collected. The parameter K designates the number of points to take into consideration in the estimation process. This value has to be set wisely, as a large value means including more points in the estimation, causing high sensibility to noise. In contrast, fixing a small K reduces the number of points used in the regression, leading to overfitting problems [18], [19].

XGBoost is an ensemble learning method and one of the most powerful machine learning algorithms [20]. It is based on gradient boosting, which builds iteratively a model on the data made of ensemble decisions trees. Each new tree is built based on the error made by the previous one, i.e., it boosts the attributes that led to estimation errors of the previous trees that are already part of the model. XGBoost follows the logic of taking lots of small steps in the right direction gives a better prediction [18], [19].

Multi-layer Perceptron is a type of artificial neural network which has shown good results in air quality monitoring [21], [22]. It consists of a system of interconnected layers of nodes, with an input layer, an output layer and at least one hidden layer. Layers are connected by weights

In order to get an insight on which regression model offers the best performance with our data, we have run multiple iterations of estimation for each measurement campaign, with 80% of data randomly selected and the remaining 20% for testing. Results in Fig 2 show the Mean Absolute Error (MAE) for each model applied to all sensing campaigns. We observe

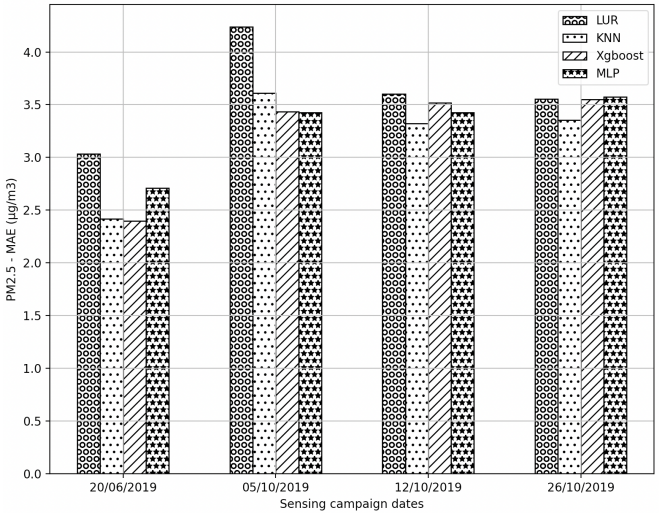


Fig. 2. MAE of $PM_{2.5}$ concentrations estimation for each sensing campaign

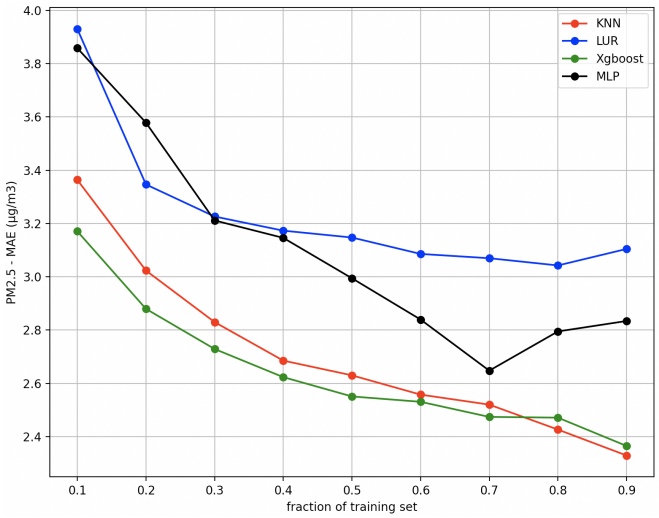


Fig. 3. MAE of $PM_{2.5}$ concentrations estimation in function of the training set fraction

that land-use regression has the worst performance overall while KNN, MLP, and XGBoost give approximately the same estimation error with a slight advantage for KNN.

In addition, we have evaluated the estimation error of the four models in function of the size of the training set. We have varied the fraction of the training set from 10% to 90%. For each fraction, we randomly construct the training set and evaluate the performance of the four models. This process was repeated 40 times for each fraction. In Fig 3, the MAE of each model in function of the size of the training set. Results reveal that the size of training set could have a different impact depending on the model. For instance, LUR is less sensitive to the fraction of training set after 30% while the MLP shows more variation. KNN and XGBoost have globally the same behavior with respect to the size of the training set.

Based on these observations and the information from Table I that show the average execution time of the four models for one iteration of estimation using 80% of data for training and 20% for testing, we have chosen to use KNN for the coming tests with $K = 4$. It is to be noted that we have evaluated the impact of the parameter K on the MAE and results showed very low impact. Nevertheless, the lowest MAE was achieved with $K = 4$.

TABLE I
AVERAGE EXECUTION TIME OF 1 ITERATION FOR KNN, LUR, XGBOOST, AND MLP

Model	KNN	LUR	XGBoost	MLP
Average execution time (seconds)	0.056	0.035	1.86	7.6

VI. ENERGY GAIN VS SAMPLING RATE

Energy consumption is of utmost importance when dealing with low-cost WSNs. In fact, sensing nodes and especially portable ones are equipped with small batteries to meet multiple requirements in terms of size and budget, which can result in a greatly limited operating time. To cope with that, sensing nodes manufacturers often have to reduce the sampling frequency to extend the lifetime of the sensor nodes. However, this often comes at the cost of spatio-temporal resolution, which impacts the knowledge on the phenomenon.

In our previous work [9] we have evaluated the power consumption of our sensing nodes with multiple configurations. Results revealed that with sampling and transmission frequencies fixed at 20 seconds and 1 minute respectively, the average energy consumption of our sensors is 231 mA. However, turning off the PM sensor makes the energy consumption remarkably drops to 115 mA. The PM sensor alone is responsible for half of the node's energy consumption (116 mA) mainly because of its integrated fan. By carefully turning the PM sensor off, one could considerably increase the operating time of the sensor.

Based on these data, we have estimated the energy consumption for different configurations of sensing frequency using a simple energy model. We introduce a sensing duty cycle D that represents the fraction of time during which the PM sensing probe is active over the sensing window. For example, if the sensor is active for 30 seconds and off for 10 seconds, then $D = 3/4$. The sensing duty cycle has to take into account the convergence time of the sensor, which is the time needed for a sensor to provide reliable measurements. This value varies from a sensor to another [23]. The formula of the energy consumption is described in (1).

$$I_{average} = I_{PM_{ON}} * D + I_{PM_{OFF}} * (1 - D) \qquad (1)$$

where $I_{average}$ is the average operating current of the 3M'air node, $I_{PM_{ON}}$ is the operating current of the node when the PM sensor is ON (231 mA in our case), and $I_{PM_{OFF}}$ is the operating current when the PM sensor is turned off (115 mA in our case).

TABLE II
3M'AIR NODE'S OPERATING TIME IN FUNCTION OF THE SAMPLING RATE

Sampling rate (seconds)	20	40	60
Average operating current (mA)	231	202	173
Average operating time (hours)	22.07	25.24	29.48

In addition to the energy consumption of our nodes using a sensing frequency of one sample every 20 seconds, we have estimated their energy consumption for a sampling rate of 40, and 60 seconds using (1) and a convergence time for the PM sensor of 30 seconds as indicated in its datasheet [24]. For a rate of 20 seconds, the PM sensor cannot be turned off because it does not have enough time to reach a steady state ($D = 1$). In contrast, the PM sensor can be turned off for 10 seconds and 30 seconds when the sampling rate is set to 40 ($D = 3/4$) and 60 seconds ($D = 1/2$) respectively. Table II shows the estimated operating current and time of the nodes for the three configurations using a 5100 mAh battery. It can be observed that a small change in the sampling rate can have significant impact on the node's autonomy. It is worth mentioning that we do not turn off the other sensing probes because their energy consumption is much lower compared to the PM sensor. The GPS receiver is not turned off either because it needs a longer time to get to acquire satellite signals

We evaluate in this work the performance of PM concentrations estimation when reducing the sampling frequency. For the first test, we have randomly picked 80% of our sensors' data for each sensing campaign to train the model and three configurations were tested on the training set. In the first configuration, we kept the initial sampling rate (i.e., 20 seconds). The sampling rate was then reduced to 40 seconds (by considering one measurement every two measurements) for the second configuration. Similarly, we lowered the rate to 60 seconds for the third configuration (i.e., taking one measurement every 3 measurements). The temporal resolution for the testing set was kept at 20 seconds and the process was repeated 40 iterations.

Fig 4 shows the Mean Absolute Error (MAE) of the estimation model for four sensing campaigns in function of the sensing frequency. We observe that even when reducing the sampling frequency by a factor of two to three, the estimation models can still achieve acceptable results compared to using the initial frequency. Indeed, reducing the sampling rate to 40 seconds results in an error 1.57% larger while achieving 14.36% longer operating time. Moreover, by lowering the frequency to one sample every 60 seconds the performance of estimation decreases by around 4.64%, but the node manages to save 33.57% more energy. Therefore, the gain in power outweighs the loss in estimation accuracy. Depending on the application, this extra battery autonomy may allow sampling more locations and hence, reduce the estimation error even further.

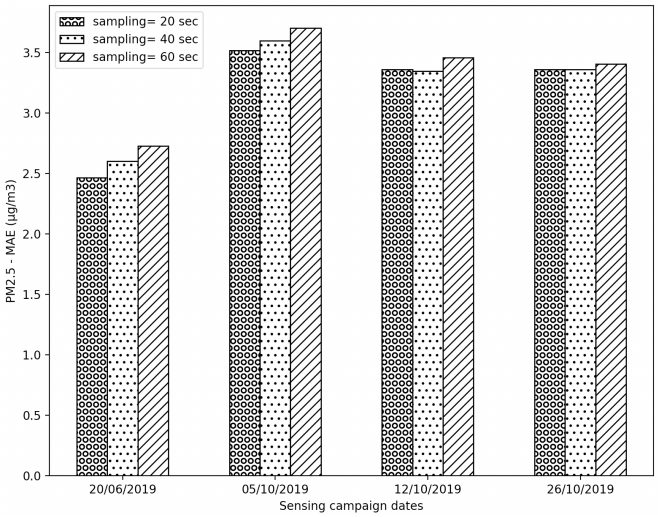


Fig. 4. MAE of $PM_{2.5}$ concentrations estimation with three different sampling frequencies using 80% of data for training

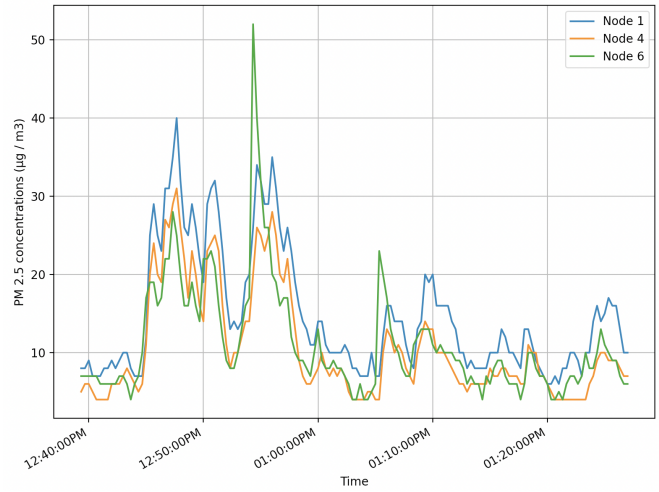
TABLE III
AN EXAMPLE OF PEARSON'S COEFFICIENT OF CORRELATION FOR NODES SAMPLING SAME ROUTES

Date	Group of nodes	Pearson's coefficient
October, 5th 2019	nodes 1 and 4	0.976
	nodes 4 and 6	0.813
	nodes 1 and 6	0.807
October, 12th 2019	nodes 5 and 9	0.783

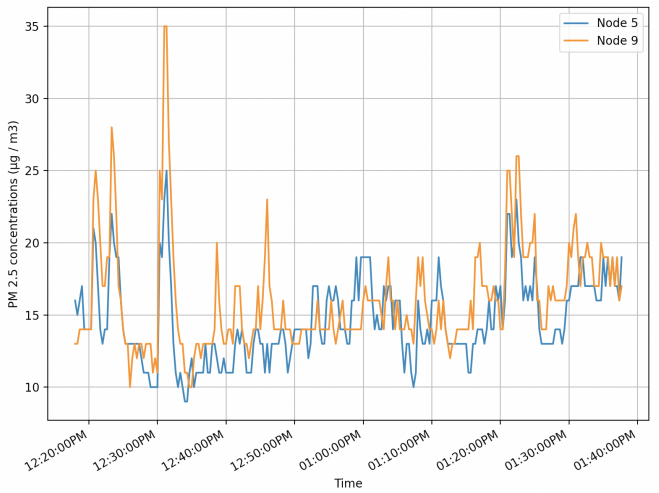
VII. SENSOR PREDICTABILITY

Deploying low-cost WSNs often involves multiple challenges such as loss of connection, sensor failures, etc. In air quality monitoring applications, these events can have a huge impact on the performance of the application, whether it is a simple loss of communication or complete failure of the sensing probe. Thus, it is important to evaluate the capacity of reproducing a sensor's measurements based on the other available sensors of the network. We have plotted in Fig 5 (a) and (b) $PM_{2.5}$ concentrations for two different routes of two sensing campaigns. Fig 5 (a) shows measurements from sensors 1, 4, and 6 which were sampling the same route during the campaign of October 5th, while Fig 5 (b) depicts measurements from sensors 5 and 6 which were sensing the same route during the campaign of the 12th of October. It can be seen from the plots that sensor readings follow globally the same trend, except for some differences that could be related to incorrect handling of the sensors. The Pearson coefficient of correlation has been calculated between these sensors, and it confirmed the visual observation as indicated in Table III.

In order to evaluate the possibility of predicting faulty sensor's data based on the remaining operational ones, we have imagined a scenario in which the system receives no measurements from a node due to an operation problem. For this, we have selected four PM2.5 campaigns which took place between the 20th of June and the 26th of October 2019 with a minimum of 8 sensing nodes each, then performed



(a)



(b)

Fig. 5. Example of $PM_{2.5}$ concentrations measured by our sensing nodes during the campaign of (a) October, 5th 2019 (b) October, 12th 2019

a cross validation by taking one sensor's measurements for testing while using the other sensors to train the model. This process has been repeated for each measurement campaign. The bar chart depicted by Fig 6 presents the MAE of predicting PM2.5 concentrations for each sensor during four sensing campaigns. The error of prediction varies depending on the faulty sensor and the sensing campaign with the largest MAE (10.31 $\mu g/m^3$) reached during the campaign of October, 12th for sensor number 2 while sensor number 10 gets the lowest error (2.18 $\mu g/m^3$) during the same campaign.

These results help to evaluate the performance of the sensors as low-cost sensors are likely to present some divergence in sensing despite being of the same type, hence the need for frequent calibration. Moreover, it can be noted from Fig 6 that sensors number 10 and 11 for example have in general the lowest MAE across four campaigns, meaning that they are well represented by the other sensors. These observations give insights on which approach to adopt with a dense network of sensors in function of the predictability of each sensor, such as fixing different sampling frequencies for different sensors or choosing a scheduling approach in which one sensor stops measuring when it is located in the vicinity of another sensor.

VIII. Conclusion

In the recent years, air pollution has become a serious concern and the public awareness surrounding it has considerably increased. Millions of people live in cities where the standard limits are far exceeded, which is why multiple efforts from public and governments have been made to help cut it. Today, low-cost air quality monitoring platforms represent a major asset in improving the local knowledge of air pollution. The goal of this work is to point the potential of estimating environmental parameter using low-cost sensors with $PM_{2.5}$ measurements collected during four sensing campaigns. This study evaluates the performance of KNN, LUR, XGBoost, and MLP in estimating $PM_{2.5}$ concentrations. It shows the capacity of the statistical model to achieve acceptable performance despite lowering the sampling frequency of the sensors, which results in a lower spatial resolution. Moreover, we have evaluated the correlation between the different sensors and the ability to recover or predict a sensor's readings in case the sensing probe fails to perform measurements. The results are satisfactory and help to imagine new approaches of mobile sensing, such as adopting different sampling frequencies for different sensors or implementing a scheduled sensing mechanisms.

Acknowledgment

We would like to thank all volunteers of the participatory sensing campaigns for their time and efforts. We also thank all 3M'Air project members, especially our colleagues from EVS laboratory and "La Métropole de Lyon".

References

[1] World Health Organization, "Burden of disease from the joint effects of household and ambient air pollution for 2016," 2018. [Online]. Available: https://www.who.int/airpollution/data/AP_joint_effect_BoD_results_May2018.pdf

[2] European Environment Agency,, "Air quality in europe — 2020 report," 2020. [Online]. Available: https://www.eea.europa.eu/publications/air-quality-in-europe-2020-report

[3] G. Cannistraro, M. Cannistraro, A. Cannistraro, A. Galvagno, and F. Engineer, "Analysis of air pollution in the urban center of four cities sicilian," *Int. J. Heat Technol*, vol. 34, pp. S219–S225, 2016.

[4] Z. Dagher, G. Garçon, P. Gosset, F. Ledoux, G. Surpateanu, D. Courcot, A. Aboukais, E. Puskaric, and P. Shirali, "Pro-inflammatory effects of dunkerque city air pollution particulate matter 2.5 in human epithelial lung cells (l132) in culture," *Journal of Applied Toxicology: An International Journal*, vol. 25, no. 2, pp. 166–175, 2005.

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

[6] L. Morawska, P. K. Thai, X. Liu, A. Asumadu-Sakyi, G. Ayoko, A. Bartonova, A. Bedini, F. Chai, B. Christensen, M. Dunbabin *et al.*, "Applications of low-cost sensing technologies for air quality monitoring and exposure assessment: How far have they gone?" *Environment international*, vol. 116, pp. 286–299, 2018.

[7] D. Liu, Q. Zhang, J. Jiang, and D.-R. Chen, "Performance calibration of low-cost and portable particular matter (pm) sensors," *Journal of Aerosol Science*, vol. 112, pp. 1–10, 2017.

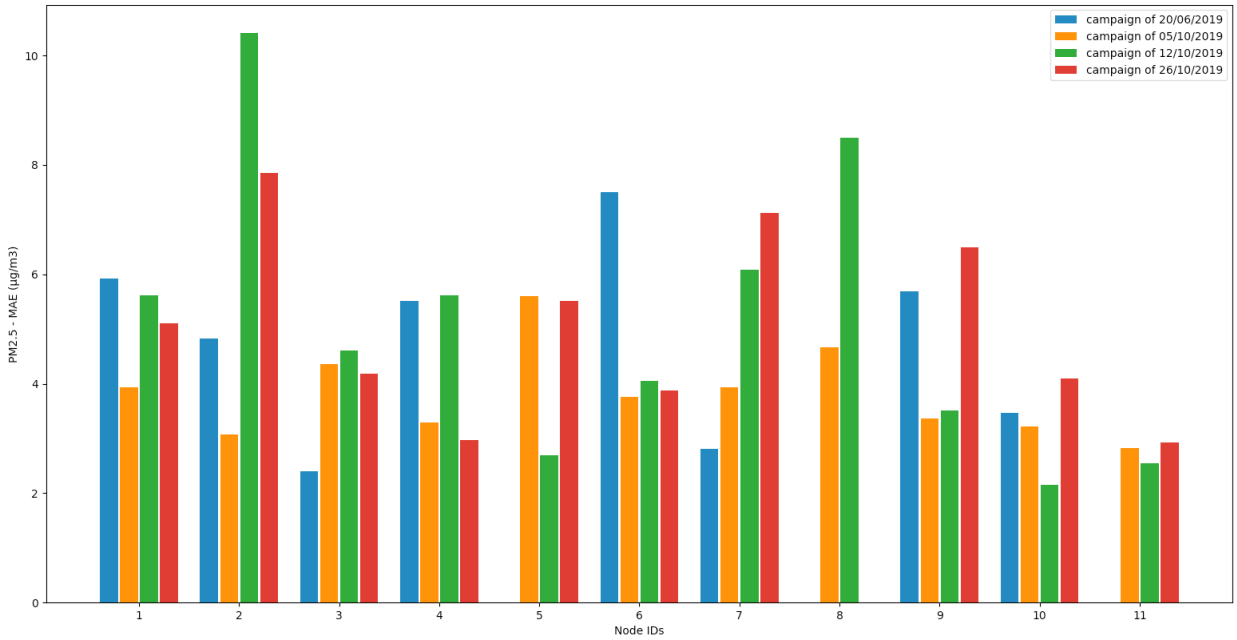


Fig. 6. $PM_{2.5}$ concentration estimations MAE for cross-validation

[8] B. Guo, Z. Yu, X. Zhou, and D. Zhang, "From participatory sensing to mobile crowd sensing," in *2014 IEEE International Conference on Pervasive Computing and Communication Workshops (PERCOM WORKSHOPS)*. IEEE, 2014, pp. 593–598.

[9] M. A. Fekih, W. Bechkit, H. Rivano, M. Dahan, F. Renard, L. Alonso, and F. Pineau, "Participatory air quality and urban heat islands monitoring system," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–14, 2020.

[10] N. Castell, M. Kobernus, H.-Y. Liu, P. Schneider, W. Lahoz, A. J. Berre, and J. Noll, "Mobile technologies and services for environmental monitoring: The citi-sense-mob approach," *Urban climate*, vol. 14, pp. 370–382, 2015.

[11] Y. Zhuang, F. Lin, E.-H. Yoo, and W. Xu, "Airsense: A portable context-sensing device for personal air quality monitoring," in *Proceedings of the 2015 Workshop on Pervasive Wireless Healthcare*, 2015, pp. 17–22.

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

[13] C. Bertero, J.-F. Léon, G. Trédan, M. Roy, and A. Armengaud, "Urban-scale no2 prediction with sensors aboard bicycles: A comparison of statistical methods using synthetic observations," *Atmosphere*, vol. 11, no. 9, p. 1014, 2020.

[14] L. Alonso and F. Renard, "Compréhension du microclimat urbain lyonnais par l'intégration de prédicteurs complémentaires à différentes échelles dans des modèles de régression," *Climatologie*, vol. 17, p. 2, 2020.

[15] Data Grand Lyon, "Métropole de lyon," 2020. [Online]. Available: https://www.grandlyon.com/metropole/59-communes.html

[16] X. Xie, I. Semanjski, S. Gautama, E. Tsiligianni, N. Deligiannis, R. T. Rajan, F. Pasveer, and W. Philips, "A review of urban air pollution monitoring and exposure assessment methods," *ISPRS International Journal of Geo-Information*, vol. 6, no. 12, p. 389, 2017.

[17] A. Larkin, J. A. Geddes, R. V. Martin, Q. Xiao, Y. Liu, J. D. Marshall, M. Brauer, and P. Hystad, "Global land use regression model for nitrogen dioxide air pollution," *Environmental science & technology*, vol. 51, no. 12, pp. 6957–6964, 2017.

[18] X. Ren, Z. Mi, and P. G. Georgopoulos, "Comparison of machine learning and land use regression for fine scale spatiotemporal estimation of ambient air pollution: Modeling ozone concentrations across the contiguous united states," *Environment International*, vol. 142, p. 105827, 2020.

[19] M. A. Fekih, I. Mokhtari, W. Bechkit, Y. Belbaki, and H. Rivano, "On the regression and assimilation for air quality mapping using dense low-cost wsn," in *International Conference on Advanced Information Networking and Applications*. Springer, 2020, pp. 566–578.

[20] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.

[21] A. Rahimi, "Short-term prediction of no 2 and no x concentrations using multilayer perceptron neural network: a case study of tabriz, iran," *Ecological Processes*, vol. 6, no. 1, pp. 1–9, 2017.

[22] S. M. S. Cabaneros, J. K. S. Calautit, and B. R. Hughes, "Hybrid artificial neural network models for effective prediction and mitigation of urban roadside no2 pollution," *Energy Procedia*, vol. 142, pp. 3524–3530, 2017.

[23] M. Mead, O. Popoola, G. Stewart, P. Landshoff, M. Calleja, M. Hayes, J. Baldovi, M. McLeod, T. Hodgson, J. Dicks *et al.*, "The use of electrochemical sensors for monitoring urban air quality in low-cost, high-density networks," *Atmospheric Environment*, vol. 70, pp. 186–203, 2013.

[24] Shenzhen Co., Ltd, "Hm-3300/3600," 2018. [Online]. Available: https://github.com/SeeedDocument/Grove-Laser_PM2.5_Sensor-HM3301/raw/master/res/HM-3300\%263600_V2.1.pdf