Scalable Measurement of Air Pollution using COTS IoT Devices

Varun Jain*
IIIT Delhi
New Delhi, India
varun14170@iiitd.ac.in

Mansi Goel*
IIIT Delhi
New Delhi, India
mansi14062@iiitd.ac.in

Mukulika Maity
IIIT Delhi
New Delhi, India
mukulika@iiitd.ac.in

Vinayak Naik
IIIT Delhi
New Delhi, India
naik@iiitd.ac.in

Ramachandran Ramjee
Microsoft Research
Bangalore, India
ramjee@microsoft.com

Abstract—**Air pollution levels have been rising at an alarming rate for the past ten years. The situation is considerably worse in developing nations, such as India. The average concentration of PM10 in Delhi has increased by over 66% between the years 2007 and 2010 and continues to increase further. Rising air pollution has been shown to have a detrimental effect on human health. The first line of action is to sensitize people about the problem by informing them about the quality of air that they are breathing in their immediate vicinity. Unfortunately, India still lacks the infrastructure required to measure pollution at a granular scale. Most of the pollution monitoring stations are placed in regions of low population density, and hence, it is difficult to calculate the personal exposure to air pollution for most of the population. It is also not economically viable to add pollution monitoring devices at such a scale in a short period of time.**

We propose a framework to estimate air pollution for a given locality by leveraging the existing infrastructure of monitoring stations and looking at factors, such as traffic conditions and greenery. We evaluate our framework by estimating the pollution exposure for long trips undertaken by users, given the seed pollution values at a few spots. Our framework incurs a reasonable accuracy. We find that greenery has more impact on pollution than traffic conditions.

I. Introduction

Delhi, the national capital of India, has been declared the 11th most polluted city by Particulate Matter 2.5 (PM2.5) and the 25th by PM10 concentrations around the world. Several other metropolitan cities in north India, e.g. Gwalior and Allahabad figure in the top five of the list [11]. Road transport is the primary source of PM2.5 [8]. This has led to numerous cases of ill-health causing asthma and chronic breathing ailments among the residents of these cities [1]. Air pollution in a given locality not only affects the overall health of its residents but also poses a threat to people who are exposed to the polluted environment while traveling through it on a regular basis. The risk is particularly alarming for people traveling by two-wheeler or open vehicles, such as rickshaws and motor-bikes. The Delhi government has taken steps in the past, e.g. odd-even scheme, to control the number of vehicles on road. The government is also experimenting with the installation of mist fountains and air purifiers to control the situation.

However, until the time a permanent solution is found to control air pollution levels, sensitizing people about the quality of the air they breathe and its ill effects on their personal health is crucial. For this, Delhi currently has 20 pollution monitoring stations across the city [2]. These stations measure the concentration of different air pollutants and make this data available publicly. However, given the vast geographic area of the city, these measurements do not give an accurate estimate of pollutants inhaled by people at their locations.

An apposite temporal-spatial coverage of air pollution monitoring is crucial for providing residents an accurate measurement of their local air quality. This, in turn, can help them make decisions that minimize the pollution's impact on their health. The concentration of pollutants at a given location depends on many factors. We study the following among them:

1) *Traffic*: The volume and the type of traffic affect the local air pollution. Areas with a high volume of commercial traffic have shown to be much more polluted.
2) *Greenery*: High density and area of green cover around a locality have a strong negative impact on pollution.
3) *Quality of Locality*: The presence of industries, waste disposal system and construction sites have a positive co-relation with pollution.
4) *Time of the day*: Time of day as classified into Morning and Evening affects pollution in a given area.

We drove around the Delhi NCR region and collected data for various trips at different times. We measure PM2.5 and PM10 concentrations using 3 Airveda sensors[1], which are commercial off-the-shelf (COTS) IoT devices. Our solution can use any of the existing sensors in the spirit of participatory sensing. Given actual measured pollution levels at a few locations, we use the above-mentioned factors to interpolate pollution at adjoining locations. We have collected approximately 56 hours worth of data for a total of 1600 Kms. We evaluate our framework on three trips lasting a cumulative of four hours. We incur an overall error of 16.9%. We also evaluate our model on a trip taken just after the festival of Diwali and demonstrate that our model is robust to factors that might cause sudden changes in the average trends of pollution. We find that among the factors discussed before, greenery around a locality has the maximal impact on pollution.

II. Related Work

The literature on the field of air pollution has focused on the following aspects: characterization of ambient PM ([9]), analyzing air quality trends ([4]), health impacts of

[1] http://www.airveda.com/ *These authors contributed equally to this work.

air pollution ([10]), and estimating personal exposure of air pollution ([6], [7]). In this paper, we focus on prior work related to personal exposure to air pollution.

PEIR [6], is a participatory sensing application. It calculates personalized estimates of the impact on environment and exposure to it. It uses GPS and cell tower based location from mobile phones to determine the trajectory of a person. Next, it looks up traffic, weather and Geographic Information Systems (GIS) to come up with a model that provides estimates of impact and exposure. The processing is done at a server, which produces a web-based personalized report. PEIR had been evaluated with thirty users, where the users get to know their exposure and impact after uploading the location data of their trip. However, PEIR did not study the accuracy of their estimates by comparing it with the ground truth data. Additionally, we lack detailed GIS data in developing countries, such as India.

Gao et al. propose Mosaic [3]: PM2.5 measuring devices that minimize airflow disturbance while collecting data on moving vehicles. To maximize coverage while keeping the number of sensors minimal, they introduce a greedy vehicle-selection algorithm. Mosaic uses a simple Gaussian inference model to calculate the air quality of locations without direct measurements. We incorporate learnings from their work in our data collection methodology.

Hazenfratz et al. [5] presented a novel approach which incorporates previous pollution meta-data to build pollution maps. The authors collected over 25 million measurements throughout a year using public transport system in Zurich for ultra-fine particles. They collected meta-data from various agencies for 12 variables such as population, industry, elevation, slope and traffic volume. They developed land-use regression model for spatio-temporal analysis and further optimized it using past measurements of annotated data.

Pant et al. [7] present a case study of personal exposure to pollutants in Delhi. The authors collected data over two sessions during summers and winters. This study reports that the average concentration of PM2.5 was $53.9 \pm 136 \mu g/m^3$ and $489.2 \pm 209.2 \mu g/m^3$ and that of black carbon were 3.71 ± 4.29 and 23.3 ± 14.9 during summers and winters respectively. This study found out that cooking and indoor cleaning were the highest exposure activities for PM2.5 and commuting for black carbon. Further, auto-rickshaws were found to be causing highest exposure to pollution when compared to other means of transport.

Zuurbier et al. [12] studied how the commuters' exposure to particulate matter is affected by the mode of transport, fuel type used and route. The authors collected data for 47 days over a year in Arnhem (the Netherlands) for a 22 Km road stretch. They collected data on three different types of fuels and transport (cars, buses, and bicycles). The cars were fueled by diesel or gasoline, and buses were either diesel or electric. The authors found that the exposure was highest in diesel buses and the lowest in electric buses. For bicycles, they collected data for both high and low traffic intensity routes. The cyclists on high traffic intensity routes had more exposure than people on diesel buses. Our solution is aimed to eliminate the need for expensive sensors and extensive meta-data annotations used by the above studies.

TABLE I
AQI AND 24-HOUR AVERAGE CONCENTRATIONS ACCORDING TO THE NATIONAL AMBIENT AIR QUALITY STANDARDS

Air Quality Index (AQI)	**PM10** ($\mu g/m^3$)	**PM2.5** ($\mu g/m^3$)
Good (0-50)	0-50	0-30
Satisfactory (51-100)	51-100	31-60
Moderate (101-200)	101-250	61-90
Poor (201-300)	251-350	91-120
Very Poor (301-400)	351-430	121-250
Severe (401-500)	431+	251+

III. Data Collection

We install pollution monitoring sensors in moving vehicles to record the ground truth pollution values. In this paper, we focus on the two major air pollutants i.e., PM2.5 (particulate matter of diameter $< 2.5\mu m$) and PM 10 (particulate matter of diameter $< 10\mu m$). We collected measurements of PM2.5 and PM10 concentrations along with meta-data of traffic conditions, greenery around the locality, and the quality of locality. We use these sample factors to train our model for air quality prediction.

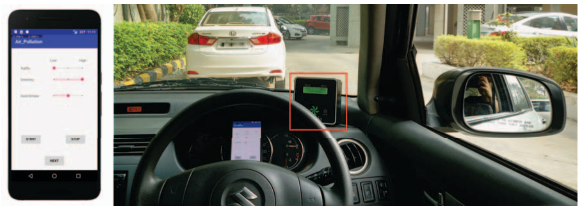

Fig. 1. Airveda sensor is placed on the dashboard with all four windows open. Inset left: An Android application used to collect the meta-data.

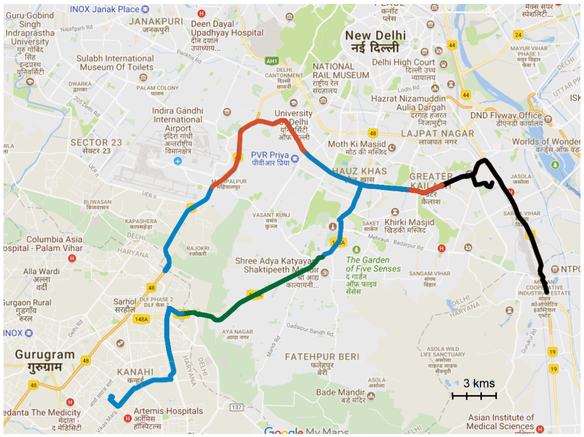

Fig. 2. Regions of Delhi NCR where data was collected and the long-term PM10 averages along segments. Green lines represent areas with Good AQI, Blue with Satisfactory, Red with Moderate and Black with Poor.

We use three Airveda sensors and place them on the dashboard of three different cars as shown in Figure 1. We keep all four windows of the car fully open and switch off Air Conditioning. This ensures that we get an accurate pollution exposure for somebody driving in an open vehicle. Airveda sensor records PM2.5 and PM10 concentrations every 60 seconds and stores them in the cloud. The device is connected to the Internet via a smartphone. We have built a user-friendly mobile application for the driver to collect meta-data. This includes information of traffic conditions, greenery levels and quality of locality for the entire route. It also records the GPS coordinates and queries the Google Maps API to get automatically generated GIS annotations for the route.

We restrict ourselves to the National Capital Region and Delhi. We collect data on the Gurgaon-South Delhi and Delhi airport-Faridabad stretches, each extending for about 31 Km. The areas cover a diverse set of localities ranging from city forests to industrial estates. Figure 2 shows the long-term average PM10 indices as defined in Table I. Green segments represent areas with Good Air Quality Index (AQI) as defined by the National Ambient Air Quality Standards. Black regions represent regions with Poor AQI averages which coincide with the industrial areas of Delhi NCR.

We have collected approximately 56 hours worth of annotated data for a total of 1600 Km. We represent the entire data as a collection of smaller 'segments'. Each segment is constructed such the intra-variance of PM2.5 and PM10 for a segment is minimized. This gives us approximately 800 segments in total.

IV. Data Preprocessing

In addition to the manual annotations provided by the driver, we make use of the Google Maps API to determine the traffic conditions. Our Android app records GPS coordinates, distances, and time taken for travel over the entire trip. Now, we assume that traffic is minimum at 0200 hrs and call this the 'minimum traffic time'. This parameter is tunable given a locality. For distances covered by the driver each minute, we query Google Maps API at the 'minimum traffic time' to determine the time it would take to travel the same distance without any traffic. Comparing these values with the actual time taken by the driver provides us a metric to decide the traffic condition of each region. These computed traffic levels are used to augment the annotations given by the driver.

We utilize manual annotations provided by the driver to determine greenery level and the quality of locality. In future, this can be replaced by GIS data from Google Maps API.

V. Methodology

While collecting annotations for the meta-data, each of the metrics was recorded on the following scale:

1) Traffic: Stagnant traffic or red light (2), traffic with frequent stops (1), and free-flowing traffic (0).
2) Greenery: Very green or dense vegetation nearby (2), average green cover with some plants along the road (1), and complete absence of greenery (0).
3) Quality of Locality: Industrial or very dusty (2), average amount of dust (1), and very good and clean locality (0).

We focus on four factors that can contribute towards pollution in a given area: time of the day, vehicular pollution (traffic), greenery, and dust/smoke (quality of locality). We analyze these parameters to estimate the change in pollution levels from the previous segment to the current segment.

Our feature set contains 800 segments containing information of PM2.5, PM10, Traffic, Greenery, Dust/Smoke, and Time of the Day. For each of PM2.5 and PM10, we calculate the difference between consecutive values in a trip and normalize it with the PM value of the previous segment. This provides us with the trend across the trip.

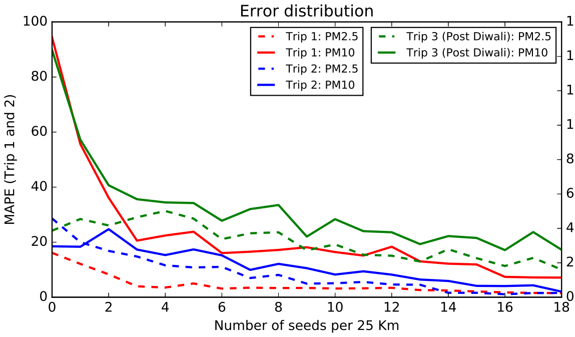

Fig. 3. The effect of the number of seed values chosen on the error. We select an optimal seed value of 5 for every 25 Km.

We split our data into training and testing set with 90:10 ratio and evaluate our model on the testing data which constitutes 2 trips of 80 segments. We train separate Linear Regression models for PM2.5 and PM10 using the meta-data as the feature set to predict these trends. A negative value of weights would imply a decrease in pollution level whereas a positive one would imply an increase. The absolute value represents the factor of change. We take the recorded PM value of a segment and predict the value of next segment using its meta information over our testing data.

We further evaluate our model by setting some reset points in our trip which we call as 'seeds'. We now predict the PM value of each segment using the predicted value of the previous segment until we reach a seed point. Hence, the value for a given segment depends only on the last seen seed point and the collection of meta-data since the seed point. These seeds help us to be consistent with the magnitude of pollution levels at different localities wherein the average PM value may differ.

VI. Results and Evaluation

We test our models on two trips not used in training. The trips are 28.5 Km and 27 Km long respectively, lasting for an average of 58 minutes. They cover cross-city trips between Gurgaon, Delhi, and Faridabad.

We first determine the optimal number of seed values that are required for accurate predictions as seen in Figure 3. For this, we run our model with different seed values and record the Mean Absolute Percentage Error (MAPE). We observe that using one seed value every 5 Km yields an error which is less than 20% for trips 1 and 2 for PM2.5.

As discussed in the previous section, we predict PM2.5 and PM10 values for each segment in two ways. In the first, we predict the change over recorded value for each segment over the entire trip. In the second scenario, we keep making predictions from our previously predicted segment values until a seed point is reached. The predictions start afresh from each seed point. Figure 4(a) and (b) show the predictions using a seed value at every 5 Km as shown by lines having markers representing the seed points. We also

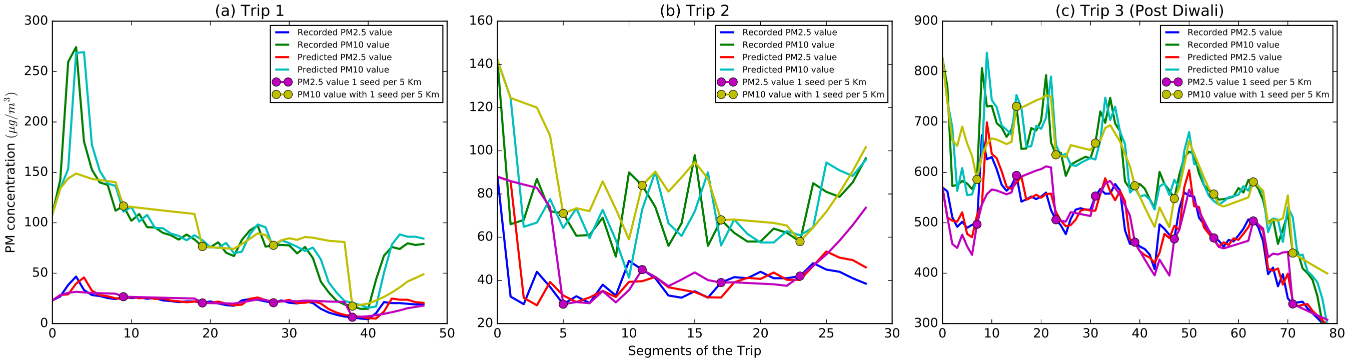

Fig. 4. Actual PM values as recorded by the Airveda sensor and the predicted values. (a) for a Gurgaon-South Delhi 28.5 Km long trip with an error of 3.5%, (b) for a Delhi Airport-Faridabad 27 Km long trip with an error of 11.5%, and (c) for a trip taken post-Diwali with an error of 27.3%. Our method is robust to factors that may cause sudden changes in the air pollution.

TABLE II
RELATIVE IMPORTANCE OF FACTORS

Parameters Used for Prediction	MAPE for the Three Trips
Greenery	76.5%
Quality of Locality	87.9%
Traffic	101.5%

show our predictions when we use the immediately previous segment for the prediction of each new segment.

We observe that the errors in the former are less and hence, we claim that our model would perform incrementally better as the density of pollution measuring devices increases across the city. We observe a MAPE of 3.5% and 11.5% on PM2.5 for the two trips respectively.

We also test our model on a trip that was taken in the morning after the festival of Diwali in 2017. It lasted for 2.3hrs. Figure 4(c) shows that there was a sharp increase in the PM concentrations when compared to the trips taken before the festival. In this case, we observe a MAPE of 27.3% on PM2.5. In this case also, our model performs equally well. Therefore, our model is robust to factors that might cause sudden changes in the air pollution.

Table II shows the contribution of the three features on the prediction error for a given time of the day. We observe that the greenery around a locality maximally impacts the pollution levels as it gives the lowest error. Our initial hypothesis behind this is that macro-level sources, such as greenery and dust, affect pollution more than micro-level, such as traffic.

VII. FUTURE WORK

We currently rely on manual annotations provided by a driver to determine the greenery levels for a locality and the quality of locality. We are working on using data from Google Maps API to build classifiers that can classify localities by looking at the places nearby. We plan to collect extensive data and build a baseline framework with average PM values for each locality in order to predict the change accurately and will test it rigorously using cross-validation. We are also looking at deep neural networks and attention mechanisms to better leverage the meta information that we collect.

VIII. CONCLUSION

Air pollution levels are rising at an alarming rate in developing countries, such as India. In this paper, we propose a method to estimate the particulate matter concentrations for a given locality by leveraging the existing infrastructure of monitoring stations and looking at factors such as traffic conditions, greenery, quality of the locality, and time of the day. We train and evaluate our model on data collected from various parts of the Delhi-NCR region. We have collected approximately 56 hours worth of data for a total of 1600 Km. Using only one seed pollution value for every five kilometres on the trips, along with meta-data like time of the day, and traffic conditions, we incur an overall error of 16.9%. We find that macro-level sources affect pollution more.

REFERENCES

[1] S. K. Chhabra, P. Chhabra, S. Rajpal, and R. K. Gupta. Ambient air pollution and chronic respiratory morbidity in delhi. *Archives of Environmental Health: An International Journal*, 2001.

[2] CPCB. Map of air quality monitoring stations across India: http://www.cpcb.gov.in/CAAQM/mapPage/frmindiamap.aspx.

[3] Y. Gao, W. Dong, K. Guo, X. Liu, Y. Chen, X. Liu, J. Bu, and C. Chen. Mosaic: A low-cost mobile sensing system for urban air quality monitoring. In *Computer Communications, IEEE INFOCOM 2016-The 35th Annual IEEE International Conference on*, pages 1–9. IEEE, 2016.

[4] S. K. Guttikunda and G. Calori. A gis based emissions inventory at 1 km× 1 km spatial resolution for air pollution analysis in delhi, india. *Atmospheric Environment*, 2013.

[5] D. Hasenfratz, O. Saukh, C. Walser, C. Hueglin, M. Fierz, T. Arn, J. Beutel, and L. Thiele. Deriving high-resolution urban air pollution maps using mobile sensor nodes. *Pervasive and Mobile Computing*, 16:268–285, 2015.

[6] M. Mun, S. Reddy, K. Shilton, N. Yau, J. Burke, D. Estrin, M. Hansen, E. Howard, R. West, and P. Boda. Peir, the personal environmental impact report, as a platform for participatory sensing systems research. In *Proceedings of the 7th international conference on Mobile systems, applications, and services*, pages 55–68. ACM, 2009.

[7] P. Pant, G. Habib, J. D. Marshall, and R. E. Peltier. Pm 2.5 exposure in highly polluted cities: A case study from new delhi, india. *Environmental Research*, 2017.

[8] P. Pant and R. M. Harrison. Estimation of the contribution of road traffic emissions to particulate matter concentrations from field measurements: a review. *Atmospheric Environment*, 2013.

[9] P. Pant, A. Shukla, S. D. Kohl, J. C. Chow, J. G. Watson, and R. M. Harrison. Characterization of ambient pm2. 5 at a pollution hotspot in new delhi, india and inference of sources. *Atmos. Environ*, 2015.

[10] S. Siddique, M. Banerjee, M. R. Ray, and T. Lahiri. Air pollution and its impact on lung function of children in delhi, the capital city of india. *Water, Air, & Soil Pollution*, 2010.

[11] WHO. List of most polluted cities by particulate matter concentration https://en.wikipedia.org/wiki/List_of_most_polluted_cities_by_particulate_matter_concentration.

[12] M. Zuurbier, G. Hoek, M. Oldenwening, V. Lenters, K. Meliefste, P. van den Hazel, and B. Brunekreef. Commuters exposure to particulate matter air pollution is affected by mode of transport, fuel type, and route. *Environmental health perspectives*, 2010.