MY-AIR: A Personalized Air-quality Information Service

Jane Lin
Department of Civil, Materials, and Environmental Engineering & Institute for Environmental Science and Policy
University of Illinois at Chicago
Chicago, IL, U.S.A.
janelin@uic.edu

Ouri Wolfson
Department of Computer Science
University of Illinois at Chicago
Chicago, IL, U.S.A.
wolfson@uic.edu

***Abstract*— This paper describes an information service that personalizes air pollution monitoring by considering the fine grained user location, her microenvironment, and her activity. Personalization is obtained by integrating a large number of information sources including the Environmental Protection Agency (EPA) monitoring stations, traffic, weather, portable air pollution data from sensors carried by a small fraction of the population, smartphone sensors, vehicle sensors data captured via on-board diagnostics.**

Keywords: Air pollution, activity recognition

I. INTRODUCTION

This paper describes an ongoing project whose objective is to provide personalized air pollution information. The project is conducted in collaboration with the EPA.

Background

Air pollution has been identified by the World Health Organization (WHO) as the world's largest single environmental health risk. In the year 2012 alone around 7 million people died - one in eight of total global deaths - as a result of air pollution exposure [1]. In India pollution levels often exceed 20 times the maximum indicated by the WHO [2]. In the U.S., 166 million people live in areas with unhealthy air [3]. Pope III et al. [4] found that each 10 $\mu g \cdot m^{-3}$ elevation in $PM_{2.5}$ air pollution was associated with approximately a 4%, 6%, and 8% increased risk of all-cause, cardiopulmonary, and lung cancer mortality, respectively. In the context of COVID-19, air pollution is important for two reasons:

1. The virus may attach to air-pollution particles, which turn into vehicles for its spread (see [5]).
2. Patients who live in air-polluted areas are more vulnerable to the disease (see [6]-[8]).

In the U.S., the Environmental Protection Agency's AirNow program has been providing hourly air quality data and daily forecasts to the public since 1998 [9]. The data source for AirNow is the ambient air quality monitoring data obtained from the U.S. EPA's air monitoring station network. AirNow uses the EPA ***Air Quality Index (AQI)*** (see sec. 2.1 for further AQI details) to ensure that the data is presented with human health in mind [10]. The AQI is often displayed by weather apps, as shown in Figure 1.

The problem addressed in this project is that the AQI is not personalized to an individual, and by this we mean that the AQI is not specific to the user location, breathing rate, and microenvironment (i.e. indoors[1], outdoors[2], in-vehicle). The microenvironment can be automatically detected by the smartphone activity recognition API (e.g. [11]), and other methods (e.g [12]).

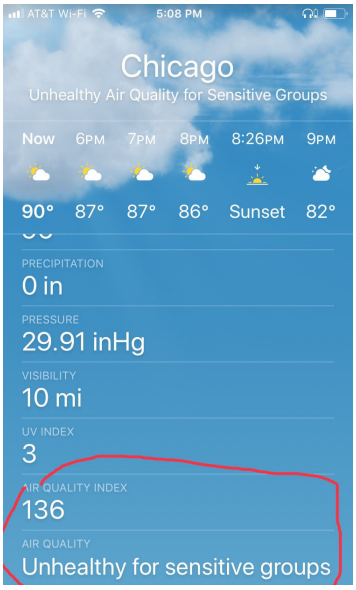

Figure 1: AQI in Chicago

Let us consider first the user location. Currently 42 million people reside in populated places farther than 40 km from the nearest EPA AirNow station. The reason is that an air monitoring station costs over $200,000 and over $30,000/year to maintain, so even densely populated areas have very few of them. For example, the city Chicago has only six stations covering an area of about 240 mi^2. Considering that some sources of pollution such as traffic and industrial plants affect mainly their immediate surrounding, the EPA stations provide only a very coarse estimate for the exposure of a particular individual in the city. More specifically, a station does not capture the significant spatial variations in pollution concentration levels.

[1] The indoor air quality depends on many factors including Heat, Ventilation, and Air Conditioning (HVAC), building material, air filtering, window status, cooking activities. Thus indoor air quality is outside the scope of this project.

[2] The outdoor microenvironment of a person refers to the immediate neighborhood, e.g. within 1km of her location, as opposed to the location of the closest EPA station which may be many kilometers away.

Now consider the user's microenvironment and breathing rate. The EPA-published AQI, or pollutant concentration values, do not reflect the actual pollutant intake of individuals. This is because the personal pollution intake also depends on whether the person is indoor, outdoor, in-vehicle, on the user- physiology (age, gender, health condition) and activity (sitting, sleeping, running, walking, biking etc.). For example, the breathing rate of a healthy adult at rest is approximately 12 breaths per minute, whereas the same adult may exceed 60 breaths per minute when running. Clearly, with a higher breathing rate comes a higher pollution intake.

Relevant Work

Generally, there are two purposes for statistical analysis of air quality data, interpolation and forecasting. Interpolation refers to estimating the air quality at a location other than that of an EPA air monitoring station, or an available measurement site, at the time of querying. Forecasting refers to predicting the air quality in the future. In the existing literature this prediction is usually performed at the location of an EPA station, and the prediction pertains to that location.

In terms of methodologies, broadly speaking, there are two approaches 1. Atmospheric physical models that use air dispersion mechanisms and spatial statistics for interpolation and forecasting (e.g. [13]), and 2. Data driven models that rely on (deep) machine learning (ML) techniques (e.g. [14]). The data driven models have been shown superior to the physical ones in recent literature (see [15]). However, even in the data driven models the reported accuracy of the interpolation is not very high. For example, for $PM_{2.5}$, ADAIN obtains an accuracy of 0.6, and this beats the other existing methods [15]. Obviously, such accuracy leaves a lot to be desired.

Another approach to provide air-quality data is reflected in experimental systems that use mobile pollution sensors. For example, a project involving bicycle mounted pollution sensors in Copenhagen [16], buses-mounted sensors in Switzerland [17], Google's Air View project in which Google Street View cars are equipped with pollution sensors [18]. These projects simply use, serve or display the readings of the mobile sensors, so they can be used by citizens in spatio-temporal vicinity. However, the value of these readings decays with both time and distance from the reading location and time. Furthermore, these projects do not personalize the readings of the mobile sensors by considering users' activities and intensity of these activities, physiology, etc.

Approach and Challenges

We hypothesize that the accuracy of existing models can be improved using a very recent development, i.e., the availability of personal portable air pollution sensors[3]. For example, the Flow air pollution sensor (Figure 2) is a small and light (0.1 lbs) particle sensor which is self-calibrating and has a 90-95% accuracy [19]. It is designed to be carried by a user throughout the day and connects to her smartphone by Bluetooth Low Energy. Miniature mobile sensors for air pollutants are currently in the research and development stage, according to a recent EPA report [20].

Figure 2: The Flow air pollution sensor [19]

Furthermore, some other data sources affect local pollution (e.g., wildfires, major transportation activities, weather conditions, locations of emission sources such as factories). We propose to also take advantage of these sources in order to address the problem of the sparsity of the U.S. EPA stations. More specifically, the first challenge of this project is to generate fine-grained spatial estimates of pollutant levels, by integration of the following data sources:

- the AirNow ground measurements at the EPA air monitoring stations,
- local conditions affecting pollution (e.g., traffic and weather conditions), and
- current and historical samples obtained by the portable sensors.

The second challenge is to adjust the fine-grained spatial estimates of pollutant levels to the microenvironment and breathing rate of a user. This involves two sub-challenges. First is to automatically recognize the micro-environment and breathing rate, and second is to adjust the air pollution intake accordingly.

Finally, let us mention that the approach in this project is general and applies to other problems (e.g. noise pollution). Abstractly, consider a geographic area with a few fixed sensors, and a number of mobile sensors. The readings of the sensors are spatially and temporally auto-correlated, and also correlated with other variables such as traffic, weather, and Land Use (e.g. industrial, residential, commercial, etc.). The first challenge is to devise a method of estimating the current reading at a location that is not covered by any sensor. This will be done by integration of readings from the existing sensors, as well as data sources providing values for the other variables. The second challenge is to personalize the estimates to a user- microenvironment and activity.

[3] We assume that only a small fraction of the population, perhaps mainly the sick and vulnerable, will continuously carry such sensors.

Project Objectives

The objectives of the project are as follows.

1. improve the accuracy of current pollution-estimates at fine spatio-temporal granularities by data integration (Task 1);
2. provide personalized pollution inhalation (or intake) estimates (Task 2); and
3. build the Monitor Your Air-pollution Intake and Risk (MY-AIR) app to validate the approach, and experimentally quantify the improvement and the remaining errors (Task 3).

The rest of this paper describes the tasks to achieve the above objectives in sec. II, and concludes in sec. III.

II. TASKS TO ACHIEVE THE OBJECTIVES

II.1 Task 1: Improve the interpolation of EPA air-pollution stations

To improve the interpolation of EPA stations readings we integrate these with mobile sensor readings and local conditions (e.g. traffic and land use information). Observe that for most locations, the only available local measurements are outdated (to various degrees) samples of the mobile sensors. These samples are obtained opportunistically. The proposed methodology of integrating data from EPA stations, mobile sensors and local conditions outlets is to estimate the value of the pollutant $PM_{2.5}$ as follows[4]:

1. Divide an area, initially Chicagoland, into a grid of cells, say 1 km x 1 km each. Similarly, time is divided into intervals, e.g. of 1 hour each. In this case, $PM_{2.5}$ is estimated for each spatio-temporal prism of $1km^2$ x 1hour.
2. For each spatial cell c, the numeric value *v* of $PM_{2.5}$ in c is estimated every hour with an error called *age*, where *age* is the age of the latest $PM_{2.5}$ reading in c. If c contains an EPA station, then it is called *strongly labeled*, and the current value *v* of $PM_{2.5}$ is read from the station and given an *age* of 0. If c does not contain an EPA station, then it is called *weakly labeled*, and its value *v* is the average of the mobile-sensor readings in c in the latest time interval in which such readings exist[5]; and *age* is the difference between the current hour and the hour of the reading; older readings are less accurate than more recent ones[6].
3. Refine the estimate *v* of each cell by interpolation considering the ages of the weakly labeled cells, traffic, weather, and land use information including pollution sources (available for the Chicagoland area at [21]).

Steps 2 and 3 above will use variants of existing interpolation and forecasting methods for air pollution. Since existing methods do not use or account for different confidence levels for different cells, they will have to be adjusted to account for lower confidence in the values as their ages increase.

We will start with the methods of interpolation and forecasting that we developed in [22]. The Land Use Data were obtained from [23] and projected into the Chicago map. For each 1km x 1km grid cell we calculated the percentage of each land use class that is contained inside it. In total eleven land uses (features) were extracted: % Agriculture Land, % Commercial Land, % Industrial Land, % Institutional Land, % Non-Parcel Land, % Open Space, % Transportation Land, % Unclassified Land, % Urbanized Land, % Vacant Land, and % Water. Additionally, traffic data were retrieved from the Chicago Traffic Tracker (https://webapps1.cityofchicago.org/traffic). By using the GPS traces retrieved from the CTA buses, real time hourly traffic congestion on Chicago' s arterial streets was estimated by the Chicago Traffic Tracker. Communities with similar traffic conditions are typically grouped together to form a region.

The resulting data driven model will provide local estimates of ambient outdoor and in-vehicle air pollution. The estimates can be accessed in real-time, without the burden and expense of carrying a Flow sensor. Thus, in the project we will determine: (a) the data driven interpolation and forecasting methods to use, (b) the prism granularity[7], (c) to what extent the portable sensors can improve the existing data driven models, and (d) how the improvement depends on the density of portable air pollution sensors.

II.2 Task 2: Personalization of AQI Information

In this section we first explain the concept of AQI (sec. 2.1), and then its personalization (sec. 2.2). Then we discuss the work that will be performed in this task (2.3).

2.1 The Air Quality Index

Each EPA station monitors concentration levels of pollutants. Instead of concentration values of the various pollutants, which would be meaningless to the average person, the EPA publishes an Air Quality Index (AQI) for each pollutant *p*. The AQI is standardized to be common to

[4] The proposed approach is also applicable to other pollutants.

[5] Example: suppose that the value for 9:00a is currently calculated, and the latest hour for which there are mobile-sensor readings in c is 6-7am; and then there are 2 readings 15 at 6:25a and 25 at 6:40. Then the value *v* of c is taken to be 20 with an age of 2 (hours).

[6] An alternative is to consider a time series of the latest x hours, based on which the value of the current time interval (e.g. hour) is predicted using an air-quality forecasting method. Each member of the time series is the average of the mobile sensor readings in the corresponding hour. The standard error of the forecast depends on the ages of the times series members, and it can be computed by the method proposed in [L. Zhu, N. Laptev, "Deep and Confident Prediction for Time Series at Uber", 2017 IEEE International Conference on Data Mining Workshops]

[7] An additional option that will be explored is dividing the geography into uniform areas in terms of some semantic property, instead of equal area-sizes. For example, each cell may have an equal population size. For this purpose zip codes or traffic analysis zones (see https://datahub.cmap.illinois.gov/dataset/cmap-modeling-zone-systems) may be used. The average area of a traffic analysis zone is $1.77mi^2$.

Table 1. EPA's $PM_{2.5}$ standards (μg/m^3) ([24])

$PM_{2.5}$	Air Quality Index	$PM_{2.5}$ Health Effects	Precautionary Actions
0 to 12.0	Good 0 to 50	Little to no risk.	None.
12.1 to 35.4	Moderate 51 to 100	Unusually sensitive individuals may experience respiratory symptoms.	Unusually sensitive people should consider reducing prolonged or heavy exertion.
35.5 to 55.4	Unhealthy for Sensitive Groups 101 to 150	Increasing likelihood of respiratory symptoms in sensitive individuals, aggravation of heart or lung disease and premature mortality in persons with cardiopulmonary disease and the elderly.	People with respiratory or heart disease, the elderly and children should limit prolonged exertion.
55.5 to 150.4	Unhealthy 151 to 200	Increased aggravation of heart or lung disease and premature mortality in persons with cardiopulmonary disease and the elderly; increased respiratory effects in general population.	People with respiratory or heart disease, the elderly and children should avoid prolonged exertion; everyone else should limit prolonged exertion.
150.5 to 250.4	Very Unhealthy 201 to 300	Significant aggravation of heart or lung disease and premature mortality in persons with cardiopulmonary disease and the elderly; significant increase in respiratory effects in general population.	People with respiratory or heart disease, the elderly and children should avoid any outdoor activity; everyone else should avoid prolonged exertion.
250.5 to 500.4	Hazardous 301 to 500	Serious aggravation of heart or lung disease and premature mortality in persons with cardiopulmonary disease and the elderly; serious risk of respiratory effects in general population.	Everyone should avoid any outdoor exertion; people with respiratory or heart disease, the elderly and children should remain indoors.

all pollutants; thus an AQI of 152 is unhealthy, regardless of whether it is the AQI for $PM_{2.5}$, or O_3, or SO_2. And the lower the AQI, the better. Furthermore, the range of possible AQI values is divided into color-coded intervals, so each AQI value falls in an interval. There are six intervals, ranging from good to hazardous. For example, Table 1 shows the $PM_{2.5}$ average concentration values and their corresponding AQI published by EPA [24][8]. Specifically, the EPA uses Eq.(1) below (see [10]) to convert pollutant p concentration to AQI_p, i.e. the AQI value for pollutant p (which is $PM_{2.5}$ in our case):

$$AQI_p = \frac{AQI_{Hi}-AQI_{Lo}}{BP_{Hi}-BP_{Lo}}\left(C_p - BP_{Lo}\right) + AQI_{Lo} \qquad (1)$$

where, C_p is the concentration value of pollutant p; BP_{Hi} is the upper bound of the concentration-interval in which C_p falls (as given by Table 1 for $PM_{2.5}$), and BP_{Lo} is the lower bound of the same concentration interval. AQI_{Hi} is the AQI value corresponding to BP_{Hi}; and AQI_{Lo} is the AQI value corresponding to BP_{Lo}. Essentially Eq.(1) represents the linear interpolation of AQI between AQI_{Hi} and AQI_{Lo}. For example, if the $PM_{2.5}$ concentration C_p is 9 μg/m^3, then the corresponding AQI_p value is 37.5[9].

[8] Similar tables are published for other pollutants.

[9] Concentration and AQI are values that pertain to an instance of time. The color-codes indicate the health implications of exposure to the corresponding concentration over a 24 hour period. So for example, it is "unhealthy (red)" if a person is exposed to a $PM_{2.5}$ concentration of 60 μg/m^3 over a 24-hour period.

Currently, the AQI value is calculated based on an area-wide ambient concentration value C_p measured at an EPA station. As indicated, this often does not represent the actual local pollutant concentration an individual is exposed to in his/her microenvironment (outdoors or in vehicle). Nor does it take into account the possible elevated intake levels due to the person's intensity of an activity such as running. In other words, AQI_p is the same regardless of the individual, her physiology, and her activity. Hence, the AQI is informative but not representative of the actual inhalation level of an individual.

2.2 Personalization of the AQI

We will personalize the AQI by taking into consideration the location, microenvironment, physiology, and activity of a person *i*. We will do so as follows. First, the user's intake will be adapted to her microenvironment; for example, vehicles usually use air filters which reduce pollution. Second, the user's physiology, including age, gender, health conditions will be taken into consideration. And finally, the activity, e.g. running, and its intensity also clearly affect breathing rate, and in turn pollution inhalation, thus will be taken into consideration.

More specifically, we will personalize the AQI of Eq (1) to produce the personalized AQI for individual *i,* denoted $PAQI_{p,i}$, and defined as follows:

$$PAQI_{p,i} = \frac{AQI_{Hi}-AQI_{Lo}}{BP_{Hi}-BP_{Lo}}\left(\bar{C}_{p,i} - BP_{Lo}\right) + AQI_{Lo} \qquad (2)$$

The only difference between formulas 1 and 2 is that Eq (2) uses the modified C_p for individual *i,* denoted $\bar{C}_{p,i}$ rather than C_p. In other words, $PAQI_{p,i}$, which is updated hourly on the user's smartphone, is a color coded value representing the pollutant p (i.e. $PM_{2.5}$ in our case) intake outdoors and in-vehicle of individual *i* during the past hour. The MY-AIR app also displays the number of minutes during the last hour that the user spent in these two microenvironments.

So, for example, MY-AIR will update its value at 9:00a to the pair (75, 45), where 75 will be yellow coded. It indicates that between 8am and 9am the user spent in-vehicle or outdoors 45 minutes, and during this time her inhalation of $PM_{2.5}$ was equivalent to her inhalation of $PM_{2.5}$ while at rest in an area where the AQI is 75.

For the rest of this subsection we explain how to calculate $\bar{C}_{p,i}$ used in Eq (2). Observe that during the past hour the user may have been in multiple locations, and furthermore, multiple spatial grid cells. Specifically, $\bar{C}_{p,i}$ is given by the following formula:

$$\bar{C}_{p,i} = \frac{\sum_{c\in G}\left(\frac{Ahr_c^{out}}{RHr}\times C_c^{out}\times t_c^{out} + C_c^{veh}\times t_c^{veh}\right)}{\sum_{c\in G}\left(t_c^{out}+t_c^{veh}\right)} \qquad (3)$$

Since all the variables on the right hand side of Eq (3) pertain user *i* and pollutant $PM_{2.5}$ the subscripts *p* and *i* are omitted.

Intuitively, formula 3 gives the average (over all the minutes of the last hour) intake of $PM_{2.5}$ by the user *i*. In order to obtain this intake, during each outdoor minute the ambient $PM_{2.5}$ concentration is multiplied by the ratio of the user's breathing rate and her resting breathing rate; and during each in-vehicle minute the ambient $PM_{2.5}$ concentration is multiplied by a factor that depends on the status of the vehicle's windows (open or closed).

Now we explain formula 3 more precisely and define each one of its variables. Assume that the user visited the set of spatial cells G during the last hour. The formula averages, over all the minutes spent during the hour outdoors and in vehicle, two terms per cell, A and B. These terms are defined for each cell $c \in G$ as follows:

1). A = The concentration of $PM_{2.5}$ inhaled during the time (in minutes) outdoors in c, multiplied by the time outdoors in *c*, t_c^{out}; and

2). B = The concentration of $PM_{2.5}$ inhaled during the time in-vehicle, multiplied by the time in-vehicle t_c^{veh}.

Term A is obtained by multiplying the ambient pollutant concentration in the cell, computed by task 1 and denoted C_c^{out}, by the ratio between the average heart rate of the user while outdoor in the cell, denoted AHr_c^{out}, and the resting heart rate of the user[10] denoted *RHr*.

Term B is obtained using the ambient pollutant concentration in a vehicle C_c^{veh}. In turn, C_c^{veh} is obtained by assuming that the user's smartphone is connected to her vehicle's onboard diagnostics (OBD). Then the smartphone can determine car-windows open/closed status, air conditioning/ventilation on/off status, and air recycling on/off status; and based on these use the appropriate conversion ratios between outdoor and in-vehicle concentrations of p (see [25-28]). The ratio between C_c^{out} and C_c^{veh} given in these references ranges between 0.76 and 2.68 depending on these factors.

2.3 Model Development and Its Evaluation

The objective of the work performed in this task is to automatically and seamlessly, i.e. without user intervention, determine the variables used in Eq (3) in order to calculate $\bar{C}_{p,i,t}$ and produce the output of MY-AIR according to Eq (2).

This will be done as follows. We assume that the user will enter her physiology information when installing the MY-AIR app, and only then. This information includes age, gender, categorical health condition (i.e. excellent, good, mediocre, poor) and resting heart rate RHr.

Using existing work on activity and microenvironment recognition [29-31], MY-AIR can determine when the user is outdoor and in-vehicle, and thus the periods of time t_c^{out} and t_c^{veh} for each geographic cell that the user visits.

Now we discuss our proposed approach to determine the heart rate AHr_c^{out} during outdoor activities. Obviously, we would like to do so continuously and unobtrusively, i.e. without burdening the user to manually enter the value. There are devices such as smart watches that continuously monitor and report the heart rate, and our prototype will be able to ingest these reports. However, the majority of the population do not continuously wear such devices. Thus, for the cases where a feed of the heart rate is not available, we will construct a model that machine learns the heart rate as follows. First we will determine the activity using the ios/android activity recognition; and if the activity is walking, cycling, or running, and if the microenvironment is outdoors (see [29,30]), then we will determine the corresponding heart rate.

Observe that the breathing rate associated with an activity depends not only on the activity, but also on its intensity. For example, running at x miles/hr the heart rate is a, whereas running at y the heart rate is b. Thus the heart rate will be determined based on sensors such as GPS, accelerometer, compass, meteorological features such as wind speed and direction (since the effort of running at the same speed with the wind is higher than that against it), and physiological features such as age, gender, health-condition. Labeled data will be collected by having subjects (e.g. students and faculty) wear a heart rate monitor (see e.g. [32]). The labeled data will consist of features (moving speed, moving direction, wind direction, wind speed, age, gender, health condition), and the label will give the corresponding heart rate. Based on the labeled data, the heart rate will be machine-learned. The most appropriate machine learning method will be determined as part of this project. Our preliminary work on activity recognition with power consumption constraints will serve as a starting point for this work [33][34].

The evaluation of the model will include a ten-fold cross validation using the collected data. Additionally, we will consider an individual who did not participate in the training of the model. The evaluation will compare the prediction of the machine learning model with the actual heart rate measured by a monitor. Obviously, the purpose of the study will be to determine feasibility, rather than high-accuracy for demographic groups and weather conditions that did not participate in the training. For example, we expect that the accuracy for an 8 year old boy running in unusual wind conditions will be lower since the training data did not have such demographic; however, adding similar labeled data to the training stage will improve the accuracy. Indeed, we expect the initial machine learning model to serve as a basis for further training after the end of the project; perhaps as part of a commercialization effort.

III.3 Task 3: The MY-AIR App and Evaluation

[10] The formula uses the heart rate as a proxy for the breathing rate (see [43] for the empirical linear relationship between heart rate and oxygen consumption).

In this section we first discuss the application (Section 3.1), and then our experimental evaluation and validation (Section 3.2). Additional considerations including privacy and incentives for crowd-sourcing will be discussed in Sections 3.3 and 3.4, respectively.

3.1 The MY-AIR application

The MY-AIR app will include prototype software to collect data and train the interpolation (Task 1) and personalization (Task 2) models. The data collection will be performed on smartphones, and the training will be performed in the cloud. The collected data will include the features indicated in Tasks 1 and 2, and the labels will be provided by the $PM_{2.5}$ reading on the mobile pollution sensor, and by the wearable heart rate monitor.

Now we discuss the MY-AIR architecture. The interpolation engine that estimates for every hour (or alternate time unit) the AQI in each geographic cell will be deployed at a server or in the cloud. The personalization engine that determines each minute the user's microenvironment and heart rate and computes the PAQI by Eq (2) will reside in the smartphone[11]. The architecture is illustrated in Figure 3.

The AQI in each geographic cell is computed every hour, but the PAQI can be calculated and displayed by the smartphone more frequently. For example, if in the middle of the hour the PAQI moves into the red zone, then it is displayed, and optionally triggers an alarm. Otherwise, it is displayed (and obviously recorded for further analysis) every hour.

Of course, when it comes to health impact, cumulative effects are more important than instantaneous pollution readings. And MY-AIR will be built with extension capabilities, e.g., producing daily, weekly, monthly inhalation reports; or making suggestions in terms of alternative routing (e.g. walk on Franklin rather than Roosevelt) or activity (wearing an air-pollution mask, circulating in-vehicle air, etc.). Furthermore, observe that although we use forecasting for the purpose of interpolation, this project does not address pollution forecasting per se. However, when designing MY-AIR we will ensure that it can be extended with forecasting capabilities. This will be useful, for example, in making alternative routing suggestions.

Battery consumption on the smartphone will be considered as follows. First observe that the critical power hungry sensor is the GPS. The other sensors used are very efficient in terms of power consumption. GPS power consumption will be reduced by Geofencing. This will ensure that the GPS sensor will not be activated if the user is stationary or does not deviate significantly from the last-transmitted location. Our target battery consumption for MY-AIR is 7% or less of battery capacity.

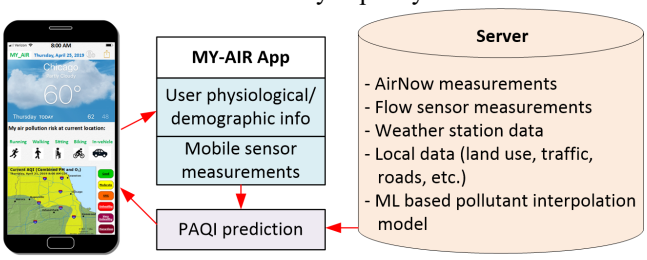

Figure 3: The architecture and User Interface of the MY-AIR app

3.2 Research Results Validation Using MY-AIR

Recall that the main novelties of the project are: 1. the use of mobile air pollution sensors for improving air-pollution interpolation, and 2. the personalization of AQI information. We will consider our approach as success if with 0.1% of the population of an area carrying the mobile sensors and sharing their generated data, the accuracy of the air pollution interpolation in the area will increase from 60% reported in [15] to at least 80%. In terms of personalization, we will consider our method a success if based on training data generated by 0.1% of the population, 95% of the heart-rate predictions are within 15% of the actual readings of the heart rate monitors.

The project is conducted in the Chicagoland area. Of course, the scope of this project does allow us to provide sensors to 0.1% of Chicagoland population. Thus we will restrict attention to the UIC area covering several square kilometers, and 0.1% of the UIC population which consists of about 30,000 students and faculty. Thus the study involves about 30 UIC students and faculty. These are called the experiment participants. They will carry a Flow sensor and will wear a heart rate monitor. Data collection by the participants will start after the app has been developed, and will continue throughout the project. We will measure battery consumption, accuracy improvement of $PM_{2.5}$ interpolation due to Flow air pollution sensors, heart rate accuracy prediction, and user-privacy sensitivity.

We will also determine the model-accuracy improvement as a function of the number of experiment participants. This will be achieved by elimination from the training set of the data of 1, 2, 3,…, 20 random participants.

We are collaborating on this project with the EPA. If MY-AIR becomes available for download, e.g. as part of EPA's TracMyAir [12]app [35], the interpolation and personalization models built in this project can be continuously improved even after the end of the proposed project. The interpolation

[11] An alternative is that both engines reside in the smartphone. This increases privacy, but also increases power and computation consumption; this architecture will be considered. Another alternative is that both engines reside in the cloud. This decreases privacy, but also decreases power and computation consumption

[12] Currently TracMyAir takes the air pollution readings from the closest EPA station to be the local pollution-value; as indicated above, this may be highly inaccurate since the closest station may be many miles away. Also, automatic personalization is not performed in TracMyAir.

model can be improved if wearers of portable pollution monitors contribute their labeled data. Similarly, the personalization model can be improved if wearers of heart rate monitors contribute their data. Incentives for such crowdsourcing are discussed next.

3.3 Incentives for providing sensor readings

Why would carriers of air-pollution monitors send their readings along with their location to enhance a machine learning model? The answer is that MY-AIR will provide the personalization component, indicating the individual intake of $PM_{2.5}$. In our experiment we will test the hypothesis that personalization provides a sufficient incentive for data sharing. Specifically, 20 months into the experiment, we expect that the heart rate machine learning model will be complete. By that time we will recruit a new set of experiment participants, but provide them only with the Flow sensors, not the heart rate monitor. We will test their willingness to share the Flow sensor data in exchange for the personalization provided by the MY-AIR app.

A similar question arises for wearers of heart rate monitors, and the answer is symmetric: If wearers of heart rate monitors provide their data, then MY-AIR will provide them with more accurate personalized air-pollution inhalation information, based on an interpolation that uses mobile sensors.

An additional possible incentive for sensor data sharing is inspired by recent work on a Marketplace for personal spatio-temporal data [36]. In this vision, personal sensor data collected by mobile users is offered for sale through micro-transactions. In other words, carriers of air-pollution monitors sell their data for dollars or cents. The buyers may be integrators of pollution data that supply precise, accurate, spatially fine-grained information to health care facilities, nursing homes, contact tracers, epidemiologists, and real-estate developers. Furthermore, climate change and COVID-19 may produce significant business opportunities for such integrators.

3.4 Privacy considerations

Users carrying mobile devices that periodically transmit time- and location-stamped information may be concerned about their location privacy. Location privacy has been a topic of active research conducted by us (e.g. [37]) and others [38]. However, [39] reveals that by and large people are not particularly sensitive to location privacy. Nevertheless, privacy can be addressed as follows:

a) Allowing users to configure My-Air for privacy (this will indicate the sensitivity of users to privacy concerns in the context of air-pollution). We discuss this option for two types of users:

a.1) The users that do not wear heart rate monitors nor pollution sensors and use the model that has been trained previously. This means that their data is not used to train the model, but they provide their location to the MY-AIR cloud to receive interpolated pollution information specific to this location. In this case, configuration of MY-AIR for privacy means that the user will be allowed to download the interpolation model to their smartphone, thus avoiding the transmission of potentially sensitive information to the server. Obviously, this privacy protection comes at the expense of higher power and compute consumption. Another option for these users is to keep the interpolation engine on the server, but request from the server the interpolated information for multiple geographic cells around them, thus cloaking their exact location.

a.2) The users that wears sensors, pollution, or heart rate, or both. Then collection of heart rate or pollution information is used to refine the model. In this case, configuration of MY-AIR for privacy also means that the user will be allowed to download the interpolation and personalization models to their smartphones. But in this case, it also means that refining of the model based on local sensor information will occur on the smartphone. And in this case there are again two options:

i. The user chooses to keep the refinement private.
ii. The user is willing to share the revised model (observe that since the user is not sharing raw sensor data, the risk to privacy violation is drastically reduced). If so, then a mechanism such as Google's federated learning [40][41] can be used in order to combine the refined models of different users. Obviously, a federated learned model will be less accurate than a centralized trained model that uses the raw data. We will quantify the loss of accuracy that is a result of privacy protection enabled by local refinement of the model.

b) Changing pseudonyms [42], and
c) Specifying a variable level of location privacy, including increasing obfuscation for certain locations such as home and work, and studying the tradeoffs between privacy and accuracy when location obfuscation is used.

If and when MY-AIR becomes available for use by the general public, a combination of the three methods above can be used. In this project we will only explore method a).

III. Conclusion

This paper described a project that personalizes air-quality information. The project has three novel aspects: 1. incorporation of opportunistic air-quality readings obtained from portable sensors to improve estimation of the Air Quality Index (AQI) at the user location, 2. personalization of the AQI based on the user activity, intensity, and microenvironment, and 3. estimation of user's heart rate to enable determination of the intensity of an activity such as running. These novel aspects are integrated with existing work on AQI forecasting and interpolation, and with existing work on activity recognition.

Acknowledgement: This work was partially supported by NSF grant IIP1534138 and UIC grant MY-AIR.

[1] World Health Organization, "7 million premature deaths annually linked to air pollution" March 25, 2014,

[2] https://www.bbc.com/news/world-asia-india-50258947

[3] American Lung Association, 2016. "State of the Air 2016 Report". Last accessed November 29, 2016.

[4] Pope III, C.A., Burnett, R.T., Thun, M.J., Calle, E.E., Krewski ,D., Ito, K., Thurston, G.D. Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. *Jama*. 287(9):1132-41, 2002.

[5] Liu, Y., Ning, Z., Chen, Y. *et al.* Aerodynamic analysis of SARS-CoV-2 in two Wuhan hospitals. *Nature*, 2020.

[6] Cao C, Jiang W, Wang B, Fang J, Lang J, Tian G, Jiang J, Zhu TF. Inhalable Microorganisms in Beijing's PM2.5 and PM10 Pollutants during a Severe Smog Event, Environmental Science & Technology 2014, 48: 1499-1507.

[7] Chen G, Zhang W, Li S, Zhang Y, Williams G, Huxley R, Ren H, Cao W, Guo Y. The impact of ambient fine particles on influenza transmission and the modification effects of temperature in China: A multi-city study, Environment International, 2017, 98: 82-88.

[8] Myatt TA, Kaufman MH, Allen JG, MacIntosh DL, Fabian MP, McDevitt JJ. Modeling the airborne survival of influenza virus in a residential setting: the impacts of home humidification, Environmental Health, 2010, 9:55.

[9] U.S. Environmental Protection Agency, AirNow program, https://www.airnow.gov/.

[10] U.S. EPA, 2016. Technical Assistance Document for the Reporting of Daily Air Quality – the Air Quality Index (AQI), report No. EPA-454/B-16/002.

[11] https://developers.google.com/location-context/activity-recognition

[12] Wang W, Chang Q, Li Q, Shi Z, Chen W. Indoor-Outdoor Detection Using a Smart Phone Sensor, *Sensors (Basel)*. 2016;16(10):1563. Published 2016 Sep 22. doi:10.3390/s16101563

[13] AERMOD model, https://www.epa.gov/scram/air-quality-dispersion-modeling-preferred-and-recommended-models.

[14] Li, C., Hsu, N. C., and Tsay, S.-C.: A study on the potential applications of satellite data in air quality monitoring and forecasting. Atmospheric Environment, 45(22):3663–3675, 2011.

[15] Cheng, W., Shen, Y., Zhu, Y., and Huang, L.: A neural attention model for urban air quality inference: Learning the weights of monitoring stations. In Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

[16] http://news.mit.edu/2009/ratti-copenhagen-1216

[17] K. Aberer, S. S. Chakraborty, D., Martinoli, A. G. Barrenetxea, B. Falting, and L. Thiele, "OpenSense: open community driven sensing of environment," Proceedings of the ACM SIGSPATIAL International Workshop on GeoStreaming, 2010. pp. 39-42.

[18] https://www.bbc.com/news/technology-49986740.

[19] Flow air pollution sensor, https://plumelabs.com/en/flow/.

[20] U.S. Environmental Protection Agency, Research and Development Highlights: Mobile Sensors and Applications For Air Pollutants, Report No.EPA/600/R-14/051 Research Triangle Park NC, 2013.

[21] Chicago data portal, https://data.cityofchicago.org/

[22] Miglionico, M. *A deep learning framework for air quality monitoring*, MS thesis, Department of Computer Science, UIC, 2018.

[23] https://catalog.data.gov

[24] U.S. EPA National Ambient Air Quality Standards, https://www.epa.gov/naaqs.

[25] Du, X., Wu, Y., Fu, L., Wang, S., Zhang, S., Hao, J., 2012. Intake fraction of $PM_{2.5}$ and NO_x from vehicle emissions in Beijing based on personal exposure data. *Atmospheric Environment* 57, 233-243.

[26] Zhang K., Batterman, S., 2009. Time allocation shifts and pollutant exposure due to traffic congestion: An analysis using the national human activity pattern survey. *Science of the Total Environment* 407(21), 5493-5500.

[27] Marshall. J.D., Riley. W.J., McKone. T.E., Nazaroff. W.W., 2003. Intake fraction of primary pollutants: motor vehicle emissions in the South Coast Air Basin. *Atmospheric Environment* 37, 3455-3468.

[28] Riediker, M., Williams, R., Devlin, R., Griggs, T., Bromberg, P., 2003. Exposure to particulate matter, volatile organic compounds, and other air pollutants inside patrol cars. *Environmental Science and Technology* 37(10): 2084-2093.

[29] Radu, V, Katsikouli, P, Sarkar, R & Marina, MK 2014, "Poster: am i indoor or outdoor?" in The 20th Annual International Conference on Mobile Computing and Networking, MobiCom'14, Maui, HI, USA, Sept. 2014.

[30] W. Wang, Q. Chang, Q. Li, Z. Shi, W. Chen "Indoor-Outdoor Detection Using a Smart Phone Sensor", MDPI Sensors, Sept. 2016, doi: 10.3390/s16101563.

[31] https://developers.google.com/location-context/activity-recognition

[32] https://www.cnet.com/news/7-of-the-best-chest-strap-heart-rate-monitors-for-running-according-to-amazon-reviews/

[33] S. Ma, O. Wolfson, B. Xu, "UPDetector: Sensing Parking/Unparking Activities Using Smartphones", Proc. of the 7th ACM SIGSPATIAL International Workshop on Computational Transportation Science, Dallas, TX, Nov. 2014, pp. 1-10.

[34] L. Stenneth, *O. Wolfson*, P. Yu, B. Xu, "Transportation Mode Detection using Mobile Devices and GIS Information", *Proc. of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (ACM GIS),* Chicago, IL, Nov. 2011, pp. 54-63.

[35] Breen, M., Seppanen, C., Isakov, V., Arunachalan, S., Breen, M., Samet, J., Tong, H. Development of TracMyAir Smartphone Application for Modeling Exposures to Ambient PM2.5 and Ozone, *International Journal of Environmental Research and Public Health* 2019, 16, 3468; doi:10.3390/ijerph16183468.

[36] K. D. Nguyen, G. Ghinita, M. Naveed, C. Shahabi, "A Privacy-Preserving, Accountable and Spam-Resilient Geo-Marketplace", *Proc. of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (ACM GIS),* Chicago, IL, Nov. 2019.

[37] L. Stenneth, P. Yu, O. Wolfson Mobile Systems Location Privacy: "MobiPriv" A Robust K Anonymous System. 2010 IEEE 6th Intemational Conference on Wireless and Mobile Computing, Networking and Communications.

[38] J. Krumm, "A survey of computational location privacy", Personal and Ubiquitous Computing, August 2009, Volume 13(6).

[39] Kaasinen, E, "User needs for location-aware mobile services", Personal and Ubiquitous Computing, 2003, 7(1) pp. 70-79.

[40] https://www.tensorflow.org/federated

[41] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, B. Aguera y Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data", Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS) 2017, Fort Lauderdale, Florida, USA.

[42] A. Beresford and F. Stajano, "Location Privacy in Pervasive Computing", in IEEE Pervasive Computing Magazine. 2003, IEEE, pp. 46-55.

[43] Reybrouk, T., Mertens, L., Brusselle, S., Weymans, M., Eyskens, B., Defoor, J., and Gewillig, M., 2000. Oxygen uptake versus exercise intensity: a new concept in assessing cardiovascular exercise function in patients with congenital heart disease, Heart 84: 46-52.