Air Pollution Exposure Estimation and Finding Association with Human Activity using Wearable Sensor Network

Ke Hu, Timothy Davison
UNSW, Australia
{ke.hu,z3288508
}@student.unsw.edu.au

Ashfaqur Rahman
CSIRO Computational Informatics
Hobart Tasmania, Australia
ashfaqur.rahman@csiro.au

Vijay Sivaraman
UNSW, Australia
vijay@unsw.edu.au

ABSTRACT

Air quality and pollution monitoring services are provided by many countries and cities. However, individuals are more concerned about personal exposure and dosage, which can rarely be estimated due to the low spatial resolution of air pollution data and lack of personal data. In recent years, an increasing number of research groups, including ours, have focused on increasing the spatial resolution of air pollution data using ubiquitous sensor networks. These works did raise the spatial granularity compared with data from fixed air pollution monitoring sites. In this paper, we combine air pollution and human energy expenditure data to give individuals real-time personal air pollution exposure estimates. In particular, this paper describes our experiences with developing a personal air pollution exposure estimation system utilising participatory air pollution monitoring system and energy expenditure data collected from wearable activity sensors. Our system and applications will benefit the understanding of the relationship between air pollution exposure and personal health. We also conducted a trial to get a full day's air pollution inhalation dosage for one participant, and applied multiple data mining techniques to find out associations between activity mode, location, and the inhaled pollution. Results show that sleep, having meals, working in a campus, and general home activities like reading books will lead to a low air pollution dosage, while working out, walking and driving will cause higher inhaled dose. Furthermore, classification results in our study based on activity modes, locations and dosage data which is collected in the trial show that up to 94% classification accuracy can be achieved.

Categories and Subject Descriptors

C.2.4 [**Computer-Communication Networks**]: Distributed Systems - Distributed applications; H.2.8 [**Database Applications**]: Data mining

MLSDA'14, December 2, 2014, Gold Coast, Australia

Keywords

Human Energy Expenditure, Pollution Monitor, Classification, Cluster, Mobiles Applications

1. INTRODUCTION

Air pollution is a world-wide concern both in developing and developed countries. Consensus is that atmospheric pollutants can cause cardiovascular diseases and effect cerebral function. The World Health Organization (WHO) reports that air pollution kills about 7 million people a year and is linked to 1 in 8 deaths worldwide. It also says that air pollution is the world's largest single environmental health risk [27]. Furthermore, air pollution is believed to contribute to serious environmental issues, such as global warming. Hence, air pollution monitoring and control is of great interest to the population.

To date, air pollution is monitored by fixed-site stations operated by government agencies. These sites generate periodic readings that report on a range of pollutants. However, the high cost and space required for such sites limits their installation. As a result, the density of air pollution data is sparse, which leads to low accuracy in spatial air pollution maps and inaccurate health inferences. For instance, authors in [18] evaluated the association between new-onset asthma and traffic-related air pollution near schools and homes. But the air pollution data they used was just from 13 central sites in 13 communities, which may not be representative of the real air pollution exposure of each child, and can lead to biased conclusions. The idea of crowdsourcing wireless sensor networks to collect air pollution data has been applied by many research groups, including ours [9, 10, 12, 13, 24, 29]. The use of mobile sensor nodes can increase spatial resolution of air pollution maps without installing a large number of fixed monitoring sites.

Individuals are more concerned about their personal air pollution dosage rather than general pollution concentrations. Lack of personal data such as age, weight, physical activity, etc. turns personal inhalation dose estimation into an arduous task. In recent years, wearable technologies have become commercially viable and commonplace, which makes personal dose estimation achievable. In this paper, we utilize air pollution data from a ubiquitous sensor network, along with human energy expenditure information from wearable sensors, to make better medical inferences. For example, two individuals may be exposed to the exact same air pollution concentration, but one individual may be sedentary while the other is active (e.g. jogging). The dosage of inhaled pollutants may vary greatly between these individuals

because of their different respiratory rates. In this context our specific contributions are:

1. We build a novel mobile application that estimates personal air pollution dosage using human energy expenditure and other personal data from wearable activity sensor devices. Further, we show that users with wearable activity sensors, and those without, can all benefit from our application.

2. We conduct trials to collect a full day's pollution and activity data for one participant, and then use cluster and classification data mining techniques to find the relationships between activity type, location and air pollution inhalation dosage. We compare the performance of seven different classification techniques on our data; the results show that the achievable classification accuracy is as high as 94%, when using the J48 classifier [21].

The rest of this paper is organized as follows: §2 discusses prior work relevant to this paper. In §3 we describe the dosage estimation method and the mobile application that we developed in our study. §4 presents the full day's trial that we conducted and the collected data. In §5 we introduce several data mining methods to cluster and classify the data, and compare the performance between different techniques. The paper is concluded in §6.

2. RELATED AND PRIOR WORK

Several prior studies have included activity information in estimating air pollution exposure [9, 25]. Most of them are using respiratory rate measurements (or estimated values) as the activity parameters. For instance, researchers in [20] compared vehicle exhaust air pollution exposure between car passengers and bikers. PM_{10} and $PM_{2.5}$ were the pollutants considered, and the data was collected using portable optical dust monitors. Minute ventilation (VE), which was obtained by a portable cardiopulmonary indirect breath-by-breath calorimetry system, was used to calculate inhaled dose, and they concluded that inhaled particular matter (PM) was significantly higher while riding a bicycle compared to driving a car.

Other researchers use energy expenditure as the parameter to estimate exposure levels. In [4], the authors designed a trial to determine the level of energy expenditure and exposure to air pollution for cyclists. This study consisted of laboratory measurements and field measurements. In the laboratory part, the relationship between heart rate and pulmonary ventilation were established. In the field measurements part, heart rate was measured by heart rate monitors, while PM_{10} and NO_2 were recorded by dust monitors. In contrast to this study, the authors of [19] assessed personal exposure to $PM_{2.5}$ and physical activity energy expenditure rate for transportation by car, subway, or walking. Twenty participants who each carried an air quality monitor and a GPS receiver travelled on intended appropriate routes by car, subway and walking on 3 different days. Energy expenditure rates were calculated by activity metabolic equivalent (MET), speed and body weight. These two studies, however, lacked personal inhalation dose despite acquiring the energy expenditure data.

A research group in Spain [8] has built a model for the analysis of competing risks associated with the built environment and its transformation to be more pedestrian friendly.

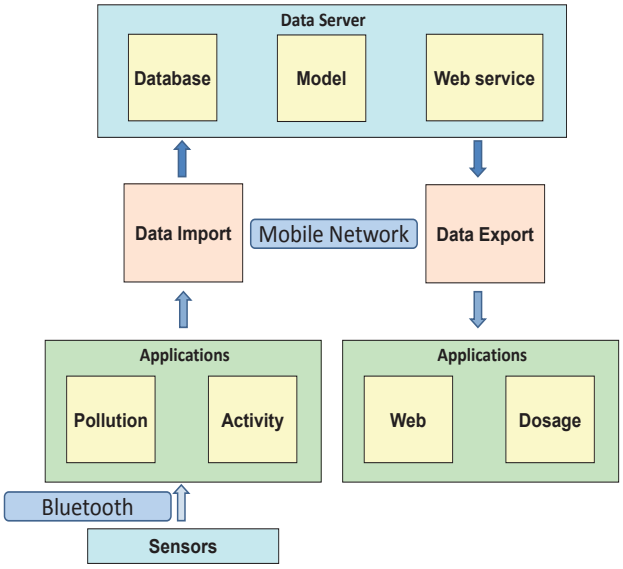


Figure 1: System architecture

The model used activity pattern, location, and travel mode to derive energy expenditure data, and then modeled air pollution data was used to estimate inhalation dose. The results indicated a pedestrian-friendly environment would cause lower average exposure while increasing energy expenditure overall, hence increasing inhalation dosage.

Our prior work [14] had demonstrated a novel personal inhalation dose estimation system as shown in Fig. 1. The system consists of sensors, applications and data server. **Sensors** include various wearable sensors, which can record activity and air pollution data. **Applications** are made of two parts: The data upload part and the user part. Data upload part applications can communicate with sensors via Bluetooth and upload air pollution or activity data to the server with a mobile network. With the user part applications, users can request air pollution maps though a web service, or obtain personal inhalation dosage data from mobile applications. A **Data Server** stores the air pollution data and/or activity data (the latter is not made visible publicly to protect personal privacy). Interpolation models and web-based maps are also be supplied by the server. Wahoo heart rate monitor and air pollution sensor Node device [26] are used as activity sensor and air pollution monitor respectively. We used heart rate values to evaluate respiratory rate, and combined with air pollution concentrations to calculate personal dosage. We also conducted a trial to compare the dosage between driving, cycling and jogging.

In terms of studies using cluster and classification data mining methods to estimate air pollution exposure levels, most of them only took air pollution concentrations into consideration other than personal exposure [3, 17, 28]. To the best of our knowledge, we are the first research group which uses cluster and classification techniques to find relationship between activity, location and personal air pollution inhalation dosage.

3. DOSAGE ESTIMATION APPLICATION

This section illustrates our dosage estimation application.

The principal system architecture is discussed in [14]. Other than the last paper, all the users can get personal dosage estimates values with or without activity sensors in this study.

3.1 Activity Sensors

The chosen activity wearable sensor for acquiring energy expenditure rate is the Fitbit Flex. We selected two activity sensors which can record energy expenditure data in the first place. One was Jawbone UP and the other was Fitbit Flex. Both sensors were desirable because they were built around a social platform, which encourages the user to continuously wear the sensor. While the battery life of UP was longer than that of Flex, the battery life of Flex was also over a week and thus considered acceptable. Both sensors did not provide a way to access accelerometer data directly, and they could offer information stored on their server via an open API. The single aspect that made UP unattractive to this study was that to sync the device, it needed to be connected to the audio jack of the iPhone. However, with Fitbit Flex, the users can sync the device with Bluetooth. The Jawbone UP released a new product named UP24 recently which can also support Bluetooth.

Fitbit Flex trackers use a 3-axis accelerometer which can convert movements of a person into digital data to record personal motions. A tuned algorithm is also added to the device to distinguish the motion patterns.

3.2 Dosage Calculation algorithms

The algorithm that we used to convert energy expenditure rate to inhalation dose is discussed in [16] and its appendices and shown in Table 1 for convenience. The whole algorithm comprises of two parts, which enables both activity sensor users and non-activity-sensor users to benefit from our system. Energy expenditure estimation part is not necessary for activity sensor user as energy expenditure data can be acquired from activity sensor server.

Table 1: Estimate inhalation dosage algorithm

1.Estimate Energy Expenditure
1.1. Calculate Resting Metabolic Rate (RMR):

$$RMR = (0.166) * [a + b * (BM) + e], \quad (1)$$

where a and b are constant appropriate regression parameters and determined by age and gender, and BM is body mass (kg) while e is a randomly selected value from a normal distribution with mean equal to zero. 0.166 converts the unit of RMR from MJ day^{-1} to Kcal min^{-1}.
1.2. Calculate Energy Expenditure (EE):

$$EE = MET * RMR, \quad (2)$$

in which MET is Metabolic Equivalent. The MET would be different based on various physical activities and listed in [1]. EE is also expressed in Kcal min^{-1}.
2.Calculate Inhalation Dosage (ID):
2.1. Calculate Oxygen Uptake Rate (VO_2):

$$VO_2 = ECF * EE, \quad (3)$$

in which ECF represents Energy Conversion Factor defined as the volume of oxygen required to produce one kilocalorie of energy and has unit of liters oxygen $Kcal^{-1}$. ECF is unique to diverse people and a random variable of uniform distribution between 0.20 and 0.21. 0.205 is applied in our application. VO_2 has the unit of liters oxygen min^{-1}.
2.2. Calculate Ventilation Rate (VR):

$$VR = BM * e^{c+d*\ln \frac{VO_2}{BM}}, \quad (4)$$

where VR is measured in liter min^{-1} and constants c and d are also determined by age and gender.
2.3. Calculate Inhalation Dosage (ID):

$$ID = VR * PC, \quad (5)$$

in which PC represents air pollution concentrations in μg $liter^{-1}$.

3.3 Mobile Application

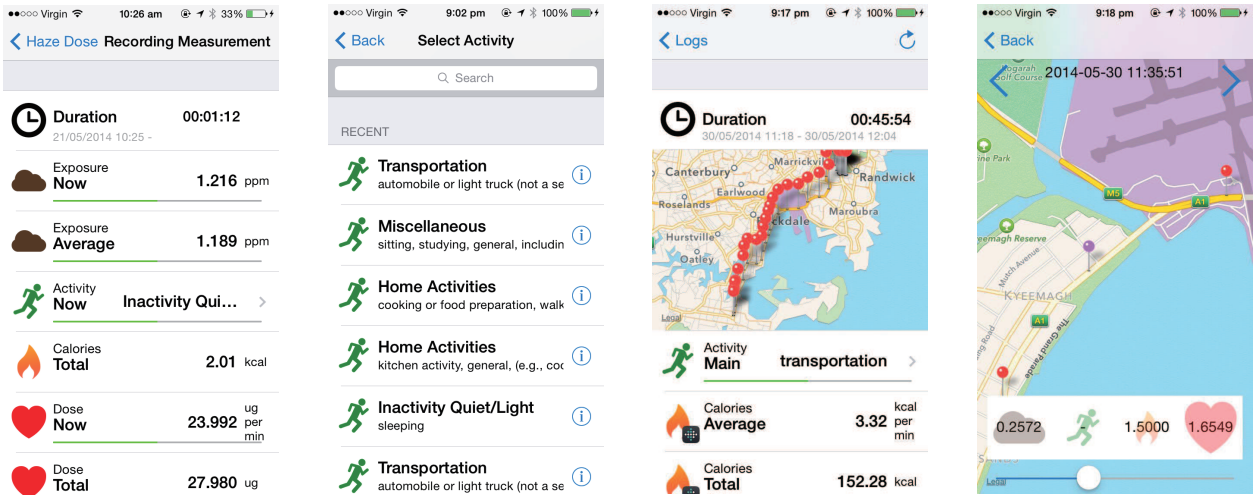
3.3.1 Development Platform

Our exposure estimates mobile application was developed in the iOS platform. Development on Android platform requires that the application should be designed for multiple screen sizes, and to reach the full market, it must support a spectrum of operating system versions. The android version of our mobile application is possible in future work.

3.3.2 Application Demonstration

On first time application launch, the first step is setup, in which Fitbit users would be asked to log-in their Fitbit account to synchronize the age, body mass and gender from Fitbit server. Alternatively if the user does not use Fitbit these details can be obtained in the set up process. It must be pointed out that an accurate value for personal information like body mass is not guaranteed as it is dependent on the user diligently updating this details in the application or through Fitbit. In the case of a negligent user, if the error between the recorded body mass and actual body mass is 5 kg, it will result in an error of up to 10% in the final calculated dose. Such a large error is undesirable and cannot be corrected for within our application.

When the setup has been done, users can just tap Start-Measurement icon to start recording. The RecordingMeasurement visualization is shown in Fig. 2(a). Here statistics about the user based on data collected from the beginning of recording are displayed. Meanwhile, users without Fitbit can select activity mode to get energy expenditure data in recording page. Activity mode selection visualization is shown in Fig. 2(b). Also, user must open the GPS function of mobile phone to record location information, as the application will acquire the air pollution data from our air pollution server with time stamp and location. Users with Fitbit cannot get real-time dosage estimation in the measurement page as the Fitbit Flex must sync through its app before energy expenditure rate data can be retrieved from the Fitbit servers. After a measurement recording is finished, users can view all the recorded data in log menu. Each log file



(a) (b) (c) (d)

Figure 2: Mobile application interface of (a)measurement, (b)activity mode selection, (c)log, and (d)log map

can present information like recorded duration, route tracking, calories burned, total dosage, etc. as shown in Fig. 2(c). A non-Fitbit-user can just get personal dosage data from the log interface, while a Fitbit user has to synchronize with Fitbit server using his/her mobile application to get past energy expenditure data by Kcal per minute. The map in the Log view can be tapped to transition to MeasurementMap view as shown in Fig. 2(d). In MeasurementMap view, user is provided a large map, which, at first, shows the user's location at the beginning of each minute of the recording. Pins represent these locations. If a pin is tapped, it becomes selected, turning purple and causing a box to appear that displays air pollution concentration level, activity level in Metabolic Equivalent (METS), calories burned and dosage of that minute. All the measurements data includes personal information, activity modes, calories burned, locations, dosages can be sent as attachments by e-mail.

A detailed description of the application design and implementation can be found in our report [7].

4. EXPERIMENT AND RESULT

4.1 Air pollution sensors and data source

Carbon Monoxide (CO) is selected as the pollutant in this study because of its most well characterised effect on the human body, reducing blood's oxygen holding capacity which in turn causes the heart to work harder to deliver needed oxygen to tissue and organs.

CO data was contributed by two sources: Our air pollution sensors and government fixed monitoring sites. First, Node devices [26] was used as air pollution sensors. Node device is a commercial air pollution sensing device which is designed with plug-in modules mode. With changing headers, Carbon Monoxide (CO), Nitric Oxide (NO), Sulfur Dioxide (SO_2) and other pollutants can be measured. With our data upload application which is described in Section 2, CO data can be received from sensors via Bluetooth and uploaded every 5 seconds to the server over cellular networks. Second, there are four fixed monitoring sites which also contributed the CO data. All the data from government sites is updated hourly. We used inverse distance weight (IDW) interpolation model to estimate the spatial distribution of CO concentrations.

Table 2: Participant general information

Gender	Body mass (kg)	Age	Stature(cm)
Male	67.8	30	180

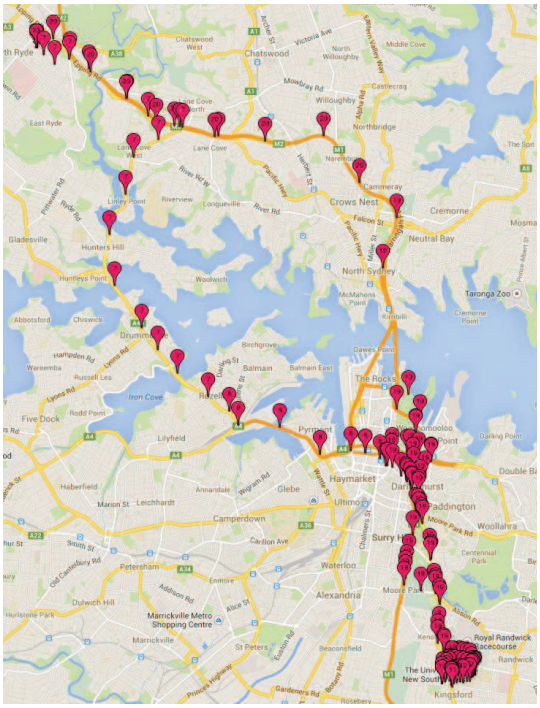


Figure 3: Trial route

4.2 Trial design and route

One participant is involved in this trial. Table 2 gives the general information of this participant, who lives in North Ryde and works in UNSW Australia. Trail route is shown in Fig. 3, where number on the plot indicates time by hours. The route which covers a urban area of 100km^2 contains motorway, foot way, campus and residential areas.

The participant was asked to wear the Fitbit sensor and carry air pollution sensor during a 24 hours period and keep recording the personal dosage data with our application. In the mean time, CO concentrations data from the Node sensor was also uploaded to the server. During the recording period, participant had to change activity modes manually to get MET data for different activity types.

4.3 Result

Table 3 gives a sample of what we have collected during the 24 hours period. There are two data sets which are distinguished by energy expenditure data source.

The whole day CO concentrations around the participant is shown in Fig. 4(a). In general, the concentrations kept low (below 0.2 ppm) during the whole day except 8am and 8pm, when the participant was driving along the motor way. CO concentration peaks at 1.4252 ppm at 7:44 in the morning and 6.835 ppm at 19:44 in the afternoon.

A summary of calories burned information is shown in Fig. 4(b). Two immediate observations can be made - first, data from MET model does change from time to time, however, it is flat while engaging in one particular activity mode. Unlike this, calorie data from the Fitbit sensor shows significant variation even in the same activity mode. For instance, calories data from jogging between 17:32 to 19:02 stays at 8.03 Kcal per minute from MET mode when it can range from 3.48 to 14.18 Kcal per minute from the Fitbit sensor. Another observation emerged from this plot is that calories burned data is assembled in few time slot in a day, mostly at 8am, 12pm and 7pm.

Inhaled dosage shown in Fig. 4(c) indicates the different CO dose levels. As inhaled dosage is computed by combining calories and CO concentrations data, the trend of this plot is similar to the two figures above. We can observe that dosage data with or without Fitbit correlate with each other well, although they are distinct during jogging as a result of activity MET estimation deviation. This observation proves that all users, including Fitbit users and non-Fitbit users, can benefit from our personal dosage estimation method and application.

Dosage percentage of different activities during the whole day is concluded in Fig. 4(d). We found that, dosage can be various on account of different activities, and jogging can occupy up to 42.9% of the whole day dosage, while driving, which exposes the largest amount of Carbon Monoxide, only seize 14.2%, the same as working. It indicates that doing sports or fitness may not be as healthy as people think, because increasing respiratory rate can lead people to further inhaled dose, even if people are under low air pollution exposure. However, this conclusion cannot be confirmed as it is unclear what level of exposure leads to health risk.

5. DATA EVALUATION AND DISCUSSION

We used two steps to evaluate the trial data set which is based on Fitbit energy expenditure data and try to find out the association between activity, location and dosage: **(1)** Applied K-means method to cluster the whole data points into three groups based on the dosage levels. **(2)** Applied several classification techniques to classify the data in terms of activity modes, locations and dosage levels which was clustered above. We assumed activity mode feature and location feature are equally important [23]. We also compared the performance between these classification methods. These techniques were running in an open source data mining software WEKA [11] on a HP computer with Intel dual-core processors 3.2GHz and 8GB RAM.

5.1 K-means cluster method

K-means cluster [2] method is often used to partition one data set into k groups. It firstly selects an initial set of k cluster centres, then assigns every instance to these clusters in which each instance belongs to the cluster with the nearest mean, and each cluster center will be updated to be the mean value of all the instances in this cluster.

In this study, we distinguish 1440 dosage data points into three categories. In fact, we tried a number of categories as well. However, there was some missing information using two categories, while classification algorithms can be easily over-trained when used numbers of categories that larger than three. When maxIterations is selected as 500, the initial cluster centres are 0.995703, 1.13702 and 20.642867, and partition result is shown in Table 4.

Table 4: Clustered dosage range

	range (μg per min)	No. of instances
Low	4.23 - 0.57	1261
Medium	13.69 - 4.31	101
High	63.85 - 14.58	78

5.2 Classification method

We split the data set which contains activity modes, locations and dosages as attributes into a training (66%) and a test set (34%). There are 950 data points in the training set and 490 data points in the test set separately. Seven classification approaches are then used to classify the data set.

5.2.1 ZeroR

ZeroR is a simple classification method which can predict majority class in training set without constraints on attributes. It is often used as a base-line for comparing classification performance.

5.2.2 Naive Bayes

Naive Bayes classifiers [15] is one of the most practical learning methods. It assumes that there is no hidden relationship between different attributes, and all the attributes are independent.

5.2.3 BayesNet

Bayesian network allows prior knowledge to be involved in data distribution. It can distribute a set of associated conditional attributes into a directed acyclic graph. In this study, we used simple estimator as classification estimator and K2 algorithm as the search algorithm which uses a hill climbing algorithm restricted by an order on the variables. ADTree is not included here.

Table 3: Samples of recorded data

	Time	Activity modes	Latitude	Longitude	Calories (Kcal per min)	CO Concentrations (ppm)	Dosage (μg)
Non-Fitbit	19:22	Drive	-33.8880822904876	151.219555087863	2.866571	1.6621	23.74718802
With Fitbit	19:22	Drive	-33.8880822904876	151.219555087863	3.227280	1.6621	27.36695547

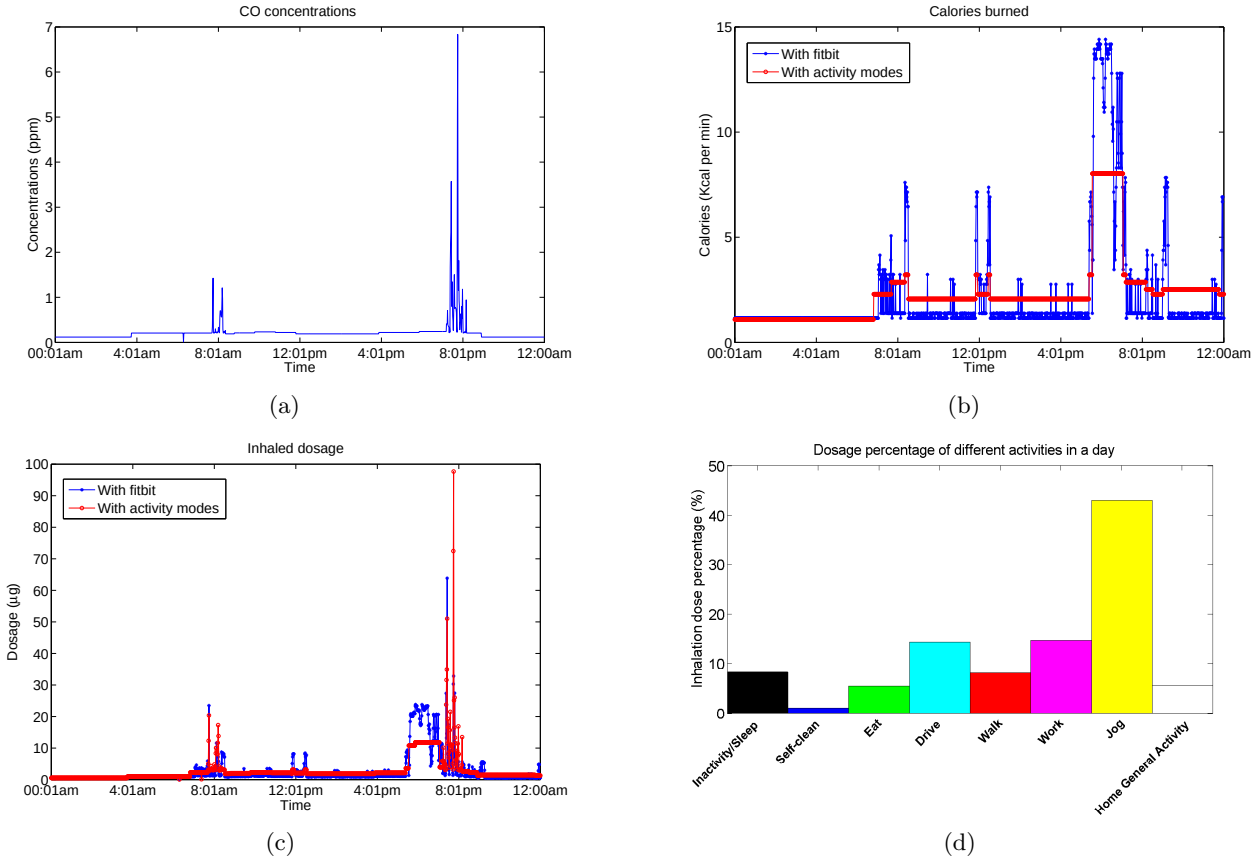


Figure 4: Whole day (a)CO concentrations, (b)calories burned, and (c)personal dosages

5.2.4 *LibSVM*

LibSVM [5] is a library for support vector machine binary classification algorithms which learns linear classifiers avoiding over-fitting with learning a form of decision boundary called the maximum margin hyperplane. Data points closest to maximum margin hyperplane are called support vectors. In this study we evaluate LibSVM classifier in WEKA with C-SVC SVM type and the parameters are: Degree = 3 and KernelType = radial basis function.

5.2.5 *MLP*

Multilayer perceptron (MLP) classifier is feed-forward artificial neural network model. It comprises of perceptrons which are organized into layers. Each perceptron in one layer connects to every perceptron in the next layer with a certain weight. MLP also uses a back-propagation supervised learning technique to train the network. We used 2 hidden units, 1 thread and 1 pool in this study.

5.2.6 *JRip*

JRip which is also known as repeated incremental pruning to produce error reduction (RIPPER) is proposed in [6]. Classes are examined in ascending size and an initial set of rules is generated with incremental Reduced error pruning (REP), then this set of rules is repeatedly simplified by applying one of a set of pruning operators. The amount of data used for pruning is 3, and the minimum total weight of the instances in a rule is 2.

5.2.7 *J48*

J48 is a Java reimplementation of the C4.5 classifier [22] which is the most used algorithm to generate decision trees. The minimum instances number of each leaf is selected as 2, and the mount of data which is used for reduced-error pruning is 3. We also consider the raising operation when pruning.

5.3 Results discussion

The summary of training results is concluded Table 5 and Table 6. The important observations are:

1. The best accuracy achieved in this work is 94.898% with JRip and J48 algorithms. Nevertheless, J48 has

Table 5: Different performance attributes running

Classifier	TP rate	FP rate	Precision	Recall	F-measure	ROC area
ZeroR	0.861	0.861	0.742	0.861	0.797	0.500
Naive Bayes	0.912	0.444	0.919	0.912	0.882	0.969
BayesNet	0.916	**0.119**	0.925	0.916	0.920	**0.970**
LibSVM	0.941	0.254	0.941	0.941	0.933	0.843
MLP	0.941	0.254	0.941	0.941	0.933	0.969
JRip	**0.949**	0.129	**0.946**	**0.949**	**0.946**	0.916
J48	**0.949**	0.129	**0.946**	**0.949**	**0.946**	0.965

Table 6: Measurements for different classifiers

Classifier	Accuracy	Mean absolute error	Root mean squared error	Relative absolute error	Root relative absolute error	Kappa statistic	Time (Seconds)
ZeroR	86.1224%	0.1550	0.2881	100.0000%	100.0000%	0.0000	**0.00**
Naive Bayes	91.2245%	0.0643	0.2279	41.5029%	79.0845%	0.5448	0.02
BayesNet	91.6327%	0.0545	0.2194	35.1307%	76.1476%	0.6835	0.03
LibSVM	94.0816%	**0.0395**	0.1986	**25.4527%**	68.9389%	0.7288	0.27
MLP	94.0816%	0.0540	0.1801	34.8591%	62.4953%	0.7288	1.59
JRip	**94.8980%**	0.0500	0.1760	32.2426%	61.0965%	**0.7851**	0.08
J48	**94.8980%**	0.0471	**0.1748**	30.3831%	**60.6660%**	**0.7851**	0.03

a lower mean absolute error and running time which makes J48 performs the best among these classification algorithms with trial dosage data set.

2. Algorithm LibSVM has the lowest mean absolute error rate and relative absolute error, which is 0.0395 and 25.4527% respectively. The rest mean absolute errors range from 0.0471 to 0.155, whilst the remaining relative absolute errors range from 30.3831% to 100%.

3. The longest training time is taken by MLP algorithm, and the time is 1.59 seconds, which is much longer than the other algorithms.

4. Despite the classification accuracy is same for LibSVM and MLP, the performance of LibSVM is better because its mean absolute error is low and it takes less running time.

Confusion Matrix and pruned tree of J48 is shown in Table 7 and Table 8. We can observe that location will not effect dosage levels in the same activity mode except driving. Dosage levels can be various from low to high during driving from place to place. Other than this, dosage level of sleep, self-clean, eating, working, and home general activity is low, while walking and jogging is medium and high respectively.

Table 7: J48 pruned tree

```
Activity = Inactivity/Sleep: Low (409.0)
Activity = Self-clean: Low (37.0/4.0)
Activity = Eat: Low (91.0/3.0)
Activity = Drive
| Latitude <= -33.877125
| | Longitude <= 151.220242
| | | Longitude <= 151.217461: Medium (14.0/1.0)
| | | Longitude > 151.217461
| | | | Latitude <= -33.886577
| | | | | Longitude <= 151.219215: Low (2.0)
| | | | | Longitude > 151.219215: High (2.0)
| | | | Latitude > -33.886577
| | | | | Longitude <= 151.218538: High (2.0)
| | | | | Longitude > 151.218538: Medium (5.0/2.0)
| | Longitude > 151.220242: Low (20.0/2.0)
| Latitude > -33.877125: Low (54.0/7.0)
Activity = Walk: Medium (39.0/1.0)
Activity = Work: Low (489.0)
Activity = Jog: High (91.0/21.0)
Activity = Home General activity: Low (185.0/12.0)
```

Table 8: Dosage confusion matrix generated from J48

	Low	Medium	High	Total
Low	419	3	0	422
Medium	9	23	8	40
High	1	4	23	28
Total	429	30	31	490

6. CONCLUSION

This paper has presented a method and application that we developed to estimate personal air pollution dose based on human energy expenditure rate data and air pollution concentration data acquired from wearable sensors. We believe that users with or without activity sensors, can all benefit from our application. A trial has been conducted to get a full day's data for one participant, from which we observed that inhalation dosage might be significantly higher while doing fitness activities, even under lower air pollution levels, though driving and walking do also contribute to the whole day's dosage. We also applied k-means clustering and several classification machine learning algorithms to find the association between location, activity mode and inhalation dosage. From the analysis of the results, we can summarize that (i) Dosage during sleeping, eating, working in a campus and doing general home activity is low; people will inhale more while working out, walking or driving outdoors. (ii) The performance of J48 classifier is the best, achieving nearly 94% accuracy within 0.03 seconds, while the mean absolute error is only 0.0471. In the future, we aim to release our application and gain more data to analyse the human daily dosage, making the classification result more convincing. Also, we will estimate the dosage based on the predictors instead of just classifying.

7. REFERENCES

[1] B. E. Ainsworth, W. L. Haskell, M. C. Whitt, M. L. Irwin, A. M. Swartz, S. J. Strath, W. L. O Brien, D. R. Bassett, K. H. Schmitz, P. O. Emplaincourt, et al. Compendium of physical activities: an update of activity codes and met intensities. *Medicine and science in sports and exercise*, 32:S498–S504, 2000.

[2] D. Arthur and S. Vassilvitskii. K-means++: The advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, 2007.

[3] J. Baker. A cluster analysis of long range air transport pathways and associated pollutant concentrations within the {UK}. *Atmospheric Environment*, 44:563 – 571, 2010.

[4] E. Bernmark, C. Wiktorin, M. Svartengren, M. LewnÃl', and S. ÃËberg. Bicycle messengers: energy expenditure and exposure to air pollution. *Ergonomics*, 49:1486–1495, 2006.

[5] C.-C. Chang and C.-J. Lin. Libsvm: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, 2:27:1–27:27, May 2011.

[6] W. W. Cohen. Fast effective rule induction. In *In Proceedings of the Twelfth International Conference on Machine Learning*. Morgan Kaufmann, 1995.

[7] T. Davison. *An iPhone application for visualising pollution maps*. UNSW, Australia, Tech. Rep., 2014.

[8] A. de Nazelle, D. A. Rodriguez, and D. Crawford-Brown. The built environment and health: Impacts of pedestrian-friendly designs on air pollution exposure. *Science of The Total Environment*, 407:2525 – 2535, 2009.

[9] A. de Nazelle, E. Seto, D. Donaire-Gonzalez, M. Mendez, J. Matamala, M. J. Nieuwenhuijsen, and M. Jerrett. Improving estimates of air pollution exposure through ubiquitous sensing technologies. *Environmental Pollution*, 176:92 – 99, 2013.

[10] S. Devarakonda, P. Sevusu, H. Liu, R. Liu, L. Iftode, and B. Nath. Real-time air quality monitoring through mobile sensing in metropolitan areas. In *Proceedings of the 2Nd ACM SIGKDD International Workshop on Urban Computing*, New York, NY, USA, 2013.

[11] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: An update. *SIGKDD Explor. Newsl.*, 11:10–18, Nov. 2009.

[12] D. Hasenfratz, O. Saukh, C. Walser, C. Hueglin, M. Fierz, and L. Thiele. Pushing the spatio-temporal resolution limit of urban air pollution maps. In *Pervasive Computing and Communications (PerCom), 2014 IEEE International Conference on*, March 2014.

[13] K. Hu, A. Rahman, V. Sivaraman, and P. Ray. Improving air pollution forecast with ubiquitous mobile sensor network. In *Ubiquitous and Future Networks (ICUFN), 2014 Sixth International Conf on*, July 2014.

[14] K. Hu, Y. Wang, A. Rahman, and V. Sivaraman. Personalising pollution exposure estimates using wearable activity sensors. In *IEEE Ninth International Conference on Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP)*, Singapore, Apr. 2014.

[15] G. H. John and P. Langley. Estimating continuous distributions in bayesian classifiers. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, 1995.

[16] T. Johnson. A guide to selected algorithms, distribution, and databases used in exposure models developed by the office of air quality planning and standards. In *U.S. Environmental Protection Agency*, North Carolina, 2002.

[17] S. Khedairia and M. T. Khadir. Impact of clustered meteorological parameters on air pollutants concentrations in the region of annaba, algeria. *Atmospheric Research*, 113:89 – 101, 2012.

[18] R. McConnell, T. Islam, K. Shankardass, M. Jerrett, F. Lurmann, F. Gilliland, J. Gauderman, E. Avol, N. KÃijnzli, L. Yao, J. Peters, and K. Berhane. Childhood incident asthma and traffic-related air pollution at home and school. *Environmental Health Perspectives*, 118:pp. 1021–1026, 2010.

[19] A. Morabia, P. N. Amstislavski, F. E. Mirer, T. M. Amstislavski, H. Eisl, M. S. Wolff, and S. B. Markowitz. Air pollution and activity during transportation by car, subway, and walking. *American Journal of Preventive Medicine*, 37:72 – 77, 2009.

[20] L. I. Panis, B. de Geus, G. Vandenbulcke, H. Willems, B. Degraeuwe, N. Bleux, V. Mishra, I. Thomas, and R. Meeusen. Exposure to particulate matter in traffic: A comparison of cyclists and car passengers. *Atmospheric Environment*, 44:2263 – 2270, 2010.

[21] J. Quinlan. C4.5: Programs for machine learning. In *Morgan Kaufmann*, San Mateo, 1993.

[22] R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, San Mateo, CA, 1993.

[23] A. Rahman, M. Murshed, and L. Dooley. Feature weighting methods for abstract features applicable to motion based video indexing. In *Information Technology: Coding and Computing, 2004. Proceedings. ITCC 2004. International Conference on*, April 2004.

[24] V. Sivaraman, J. Carrapetta, K. Hu, and B. Gallego Luxan. Hazewatch: A participatory sensor system for monitoring air pollution in sydney. In *IEEE SenseApp (co-located with IEEE LCN)*, Sydney, Australia, Oct. 2013.

[25] S. Steinle, S. Reis, and C. E. Sabel. Quantifying human exposure to air pollutionâĂŤmoving from static monitoring to spatio-temporally resolved personal exposure assessment. *Science of The Total Environment*, 443:184 – 193, 2013.

[26] Variable-Technologies. Node system. `http://www.variabletech.com/`.

[27] World-Health-Organization. Air Pollution. `http://www.who.int/topics/air_pollution/en/`.

[28] B. Yeganeh, M. S. P. Motlagh, Y. Rashidi, and H. Kamalan. Prediction of {CO} concentrations based on a hybrid partial least square and support vector machine model. *Atmospheric Environment*, 55:357 – 365, 2012.

[29] T.-C. Yu, C.-C. Lin, C.-C. Chen, W.-L. Lee, R.-G. Lee, C.-H. Tseng, and S.-P. Liu. Wireless sensor networks for indoor air quality monitoring. *Medical Engineering & Physics*, 35:231 – 235, 2013.