

Estimate Air Quality Based on Mobile Crowd Sensing and Big Data

Cheng Feng, Wendong Wang, Ye Tian, Xirong Que and Xiangyang Gong

State Key Lab of Networking and Switching Technology

Beijing University of Posts and Telecommunications, Beijing, China

Email: {fengcheng124, wdwang, yetian, rongqx, xygong}@bupt.edu.cn

Abstract—PM_{2.5} (particulate matter in the atmosphere with a diameter no more than 2.5 microns) in the air can cause great damage to human beings. It's a great challenge to offer a fine-grained and accurate PM_{2.5} monitoring service in urban areas as the required facilities are very expensive and huge. Since the PM_{2.5} has a significant scattering effect on visible light, the large-scale user-contributed image data collected by the mobile crowd sensing bring a new opportunity for understanding the urban PM_{2.5}. After the analysis of image data, we find that several image features are very discriminative in PM_{2.5} inference. In this paper, we propose a fine-grained PM_{2.5} estimation method based on the random forest model without any PM_{2.5} measurement devices. We evaluate our approach with experiments based on five data sources: the meteorological data, the traffic data, the records from the monitoring sites, the POIs and the photos collected. The experimental results show that we have a high evaluation accuracy on PM_{2.5} estimation (precision: 0.875, recall: 0.872), which outperforms existing methods (Logistic, Naive Bayes, Random Tree, and BP ANN).

I. INTRODUCTION

Since atmospheric pollutants may cause respiratory diseases such as lung cancer and severe environmental problems [1], urban air pollution is a critical issue in both developed and developing countries. It is necessary to give relatively accurate evaluation and analysis of air pollution in urban areas.

To communicate to the public the pollution levels of the air pollution, such as the concentration of NO₂, PM_{2.5} (particulate matter in the atmosphere with a diameter no more than 2.5 microns), and PM₁₀, the government agency has defined the Air Quality Index (AQI). To measure AQI values, the fixed and high cost air-quality monitoring stations are required. Up to 2016, there are only 38 air pollution monitoring stations in Beijing. Figure 1 shows the monitoring conditions of PM_{2.5} published by Beijing government [2]. Although more and more monitoring stations are established in recent years, these monitoring stations are far enough for fine-grained air pollution monitoring.

Previous studies on air pollution monitoring are mainly based on wireless sensor networks. In static sensor networks, sensors are usually placed on walls or street lamps. For example, CitySense [3] aims to provide an urban-scale wireless networking testbed for air pollution monitoring. In vehicle sensor networks, the sensor nodes are usually deployed on buses or taxis. For example, Volgyesi et al. [4] present the mobile air quality monitoring network system by a large



Fig. 1. PM_{2.5} monitoring stations in Beijing

number of car-mounted sensor nodes. As the gas sensors are usually low cost, light weight, and with fast response time, the programs above have a good performance on gas pollution monitoring. However, to monitor the PM_{2.5} accurately, the professional monitoring facilities are required, which are usually very expensive and huge.

Some studies have tried to provide fine-grained air quality estimation. For example, Zheng et al. [5] propose U-AIR to infer real-time and fine-grained air quality, which is based on a variety of data sources observed in the city and the air quality data reported by existing monitoring stations. Liu et al. [6] use the physics models [7] to get the extinction coefficient of the image and use the haze models [8] [9] to get the transmittance of the image to estimate the PM_{2.5}.

With the rapid development of mobile social network (MSN), MSN services (in particular for smartphones) are fast growing in recent years. The large-scale user-contributed data brings a new opportunity for understanding the air pollution in the city. Photos shared by users are valuable resources. However, after a large number of image analysis, we found that it's hard to use photos to estimate PM_{2.5} in some specific situations. For example, when there is a heavy fog, it will have a serious impact on the precision of estimation.

Fortunately, we find that if we add the image data (photos shared by users) to other data sources observed in the city and the air quality data reported by existing monitoring stations, the accuracy of air quality estimation can be significantly



Fig. 2. Interface of the APP

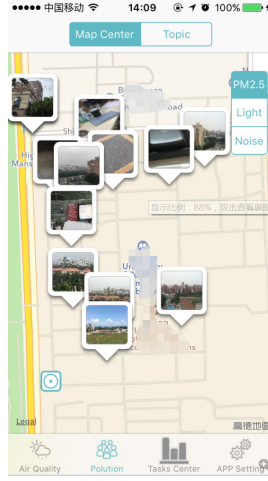


Fig. 3. Photo sharing

improved.

In this paper, we propose an air pollution monitoring platform based on mobile crowd sensing and computing (MCSC) [10] [11] [12]. We develop a smartphone APP (shown in Figure 2 and Figure 3) for the users to share their photos and also offer the users a fine-grained air pollution (e.g. PM2.5) monitoring service in real time without any additional device. According to the existing studies [5], the traffic flow, POI, and meteorology data have a strong correlation with air qualities. Thus, cloud servers of our platform also collect the historical and real-time data mentioned earlier from the public websites or APIs while the images are shared by the users.

In our platform, we infer the real-time and fine-grained PM2.5 throughout Beijing using the meteorology data, traffic flow, the POI data we observed, and the image data we collect from our APP. The algorithm based on random forest model (MCS-RF) is proposed to predict the urban area air quality (PM2.5). When new data arrives, we use online random forest [13] to change our model in real time. Meanwhile, we recalculate our model with all the data in an offline mode [14] after a fixed time.

II. OVERVIEW

A. Preliminaries

1) *Air Quality Index (AQI)*: AQI is a number used by government agencies to communicate to the public how polluted the air is currently. Different countries have different definitions on AQI. In China, the AQI is based on the levels of six atmospheric gasses, SO₂, NO₂, PM₁₀, PM_{2.5}, CO and O₃. Each gas has its Individual Air Quality Index (IAQI). In this paper, we use the IAQI of PM_{2.5} published by the China's government as shown in Fig. 4.

2) *Image data*: The image data are photos shared by users of our platform. All photos are labeled with GPS information and timestamps. Since the sky and the scene contribute different features to PM_{2.5} estimation, we encourage users to share

PM2.5(ug/m ³)	Air Quality Index	Air Quality Level	Color
0~35	0~50	Excellent	Green
36~75	51~100	Good	Yellow
76~115	101~150	Slight Pollution	Orange
116~150	151~200	Moderate Pollution	Red
151~250	201~300	Heavy Pollution	Purple
>250	>300	Severe Pollution	Maroon

Fig. 4. The definition of AQI in China



Fig. 5. Photos at location A

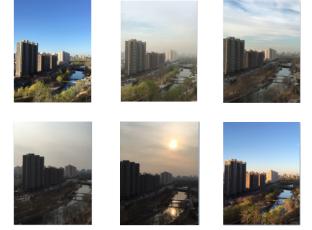


Fig. 6. Photos at location B

photos having the proper proportion between the sky and the scene as shown in Fig. 5 and Fig. 6.

3) *Traffic status*: Traffic status describes the traffic conditions on a certain road [15] as shown in Fig. 7. Different colors denote different levels of congestion. Green means the traffic is smooth, yellow means that there are several cars on the road, while red means heavy traffic jams.

4) *Meteorological data*: The meteorological data collected in our platform mainly contains temperature, humidity, barometer pressure, wind speed and weather. According to the meteorological data released by the government [16] as shown in Fig. 8, we divide the weather information into five categories: cloudy, foggy, rainy, sunny and snowy.

5) *Point of interest (POI)*: The geographic point of interest (POI) is used when referring to hotels, campsites, fuel stations or any other categories used in navigation systems. The category and density of POIs in a region indicate the land use and the function of the region as well as the traffic status in the region.

6) *Grid*: We divide the city into disjointed grids (e.g. 1km*1km) and assume the air quality in a grid is unified as shown in Fig. 9. Each grid g has a geographical position coordinate $g.l$, a timestamp $g.t$ and a PM_{2.5} label $g.PM2.5$

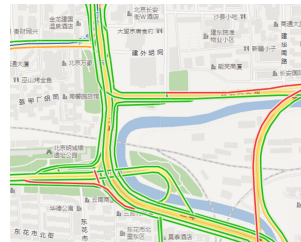


Fig. 7. Traffic status

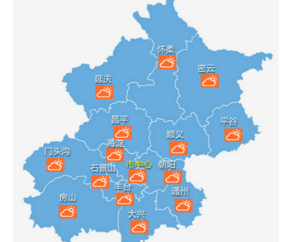


Fig. 8. Meteorological status

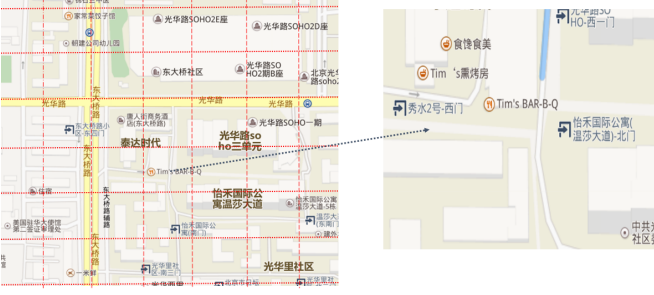


Fig. 9. Divide the city into disjointed grids

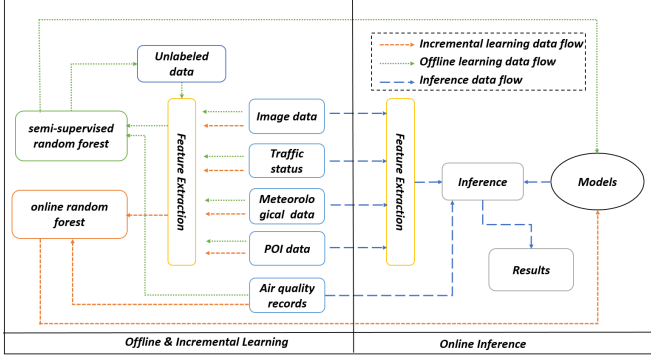


Fig. 10. Framework of the platform

to be inferred or already associated if having an air quality monitoring station located. We define $g.f = \{Fp, Fw, Ft, \dots\}$ (Section III) as the features extracted from the data we observed in the city.

B. Framework

Fig. 10 shows the framework of our platform, which consists of three major parts: offline learning, incremental learning and online inference. The framework generates three kinds of data flows: offline learning data flow, incremental learning data flow and inference data flow.

1) *Instance*: The instances used in this paper are based on grids. For each instance, we extract features from a variety of data based on its $g.l$ and $g.t$. The meteorological and traffic features can be calculated online from the public websites [16] [2], while as the POIs don't change frequently, we calculate the POI features for each grid in an offline mode. The image data are uploaded by the users, for each photo, we locate the grid based on its GPS information and then add the image features to the grid with the photo's timestamp. If the grid has an air quality monitor station located, we use the station's data to label the grid's $g.PM2.5$. We define $D_{(1,1)}$ as the straight-line distance between the center of the grid g_1 and g'_1 on the map, where g'_1 is a labeled grid. Since there are 12 PM2.5 records published by the PM2.5 monitoring stations as shown in Fig. 1, the instance of grid g_x can be denoted as

$$(g_x.l, g_x.t, g_x.f, D_{(x,1)}, g'_1.f, \dots, D_{(x,12)}, g'_{12}.f, PM2.5) \quad (1)$$

where $g_x.f$ are the extracted features of the grid g_x . We use a database to store the instances on our cloud servers.

2) *Offline learning data flow*: In the offline learning data flow (represented by broken green arrows), only the grids with an air quality monitor station located have true PM2.5 labels. As there are many places to infer, we choose a semi-supervised learning approach based on random forest [14]. It is unrealistic to recalculate all data immediately when the database updates, so we just restart the offline learning module to get a new model after a fixed time (e.g. one day).

3) *Incremental learning data flow*: Our framework is based on mobile crowd sensing, which means a large number of image data may be shared by users in real time. Our experimental results (detailed in Section 5) show that the photos can contribute a lot to the accuracy of the inference. In the incremental learning data flow (represented by the orange arrows), we choose the online random forest algorithm [13] to update the model with the labeled instances.

4) *Inference data flow*: In this flow (represented by blue arrows), we infer the PM2.5 for the unlabeled instances. Since the AQI records published by the monitoring sites updates every hour, we infer the PM2.5 for each grid each hour. Especially, when a photo is uploaded, an instance will be created based on the photo's GPS and timestamp, if the instance has a PM2.5 label, we will update the model with online random forest algorithm. Otherwise, we will infer the instance's PM2.5.

5) *Problem statement*: Given a collection of grids $G \doteq G1 \cup G2 \doteq \{g_1, g_2, g_3, \dots, g_n\}$, where $g.PM2.5 \in G1$ is unknown and $g'.PM2.5 \in G2$ is known, $|G1| \geq |G2|$, we collect images I , meteorological data M , traffic data T , POIs P and air quality monitor station reports R in G , we aim to infer $g.PM2.5 \in G1$ in real time.

III. FEATURE EXTRACTION

A. Image features (F_i)

1) *Image Preprocessing*: Our image analysis method based on mobile phone cameras, most mobile phones contain image processing functions which often apply a non-linear transformation that maps the observed image from the irradiance to the brightness. We use the radiometric calibration method to recover image [6].

Mobile phone photos usually contain two parts: the sky and the scene. These two parts are quite different in image features. S. Poduri et al. [17] use sky luminance to estimate air turbidity with mobile phones. It is very necessary to separate the sky area and the scene area with a low complexity algorithm. We find that, in most of the images of outdoor scenes, if we divide the image into nine parts (3*3), there will always be some parts which have little sky or scenery. We collect photos taken by mobile phones from different locations in Beijing and divide each image into nine parts. By calculating the image features of different parts, we find that, if the images are converted from RGB model to HSI model, there would be a significant difference in the variance of intensity between the sky part and the scene part. According to the above conclusions, when

we get a new image to estimate pm 2.5, it would be divided into nine parts. By calculating the variance of Intensity of each part, we choose the highest three parts as the SCPs (scenery parts) and the lowest part as the SKP (sky part).

2) *Spatial Contrast (F_{ig})*: Atmospheric transmission refers to how well light radiating from a scene is preserved when it reaches an observer. The atmospheric transmission model

$$I(x) = J(x)t(x) + A(1 - t(x)) \quad (2)$$

has been widely accepted nowadays, where $I(x)$ is the image irradiance, $J(x)$ is the scene radiance, A is the atmospheric light, and $t(x)$ is the atmospheric transmission. Atmospheric light extinction is inversely related to transmission through the following exponential equation [18]

$$t(x) = \exp^{-b_{ext}r(x)} \quad (3)$$

where b_{ext} is the extinction coefficient and $r(x)$ is the length of the visual pathway. According to Graves's [19] study,

$$|\nabla_x I(x)| = t(x) |\nabla_x J(x)| \quad (4)$$

transmission has the intuitive interpretation as the ratio of the observed contrast to the true contrast. So we define the F_{ig} as $F_{ig} = |\nabla_x I(x)|$.

3) *Dark channel (F_{id})*: The dark channel feature, recently proposed by He [10], has been widely used in haze removal. Based on the assumption that there is at least one color channel including pixels with very low or close to zero intensity in most of the non-sky blocks of the image. In this situation, only the SCPs will be used in this situation. The dark channel of an image is defined by

$$J_{dark}(x) = \min_{y \in \Omega(x)} \left\{ \min_{c \in \{r, g, b\}} J^c(y) \right\} \quad (5)$$

Where $\Omega(x)$ is a small block around the pixel x , J is the scene radiance and J^c is a color channel. From the equation, we can see that the dark channel value of a given pixel is the minimum intensity of the three color channels of the image block around it. A priori knowledge based on a large number of haze-free image shows [9], the dark channel of a haze-free image should be zero, which has the property of

$$J_{dark} \rightarrow 0 \quad (6)$$

By applying dark channel prior to Eq 2, the estimated transmission $t(x)$ can be obtained.

$$t(x) = 1 - \min_{y \in \Omega(x)} \left\{ \min_c \frac{I^c(y)}{A^c} \right\} \quad (7)$$

Where A^c is the global atmospheric light. In this situation, A^c is picked from the highest intensity of the image, and we estimate A^c with the brightest 0.1 percent of the pixels in the dark channel. According to the theoretical model proposed by H. Ozkaynak et al. [7], there is an exponential relationship between the $t(x)$ and the $PM2.5$

$$b_{ext} \approx \gamma PM2.5 \quad (8)$$

where b_{ext} is the extinction coefficient and γ is a constant usually set to 3.75 in the urban using the H. Ozkaynak et al.'s [7] model. As shown in Eq 3, we denote the estimated $t(x)$ as the F_{id} .

4) *HSI Color Difference (F_{ih}, F_{is}, F_{ii})*: According to Kim et al.'s [20] study, the sky's color difference in HSI color space has an exponential relation with the light extinction b_{ext} , which can be denoted by

$$b_{ext} = ae^{b\Delta D} \quad (9)$$

where a and b are coefficients in this model, ΔD is used to describe the difference in HSI color space. In this situation, we choose the image's SKP to complete the feature extraction. Since it's hard to obtain the influence coefficient of the three components (Hue, Saturation and Intensity) on the extinction coefficient b_{ext} , we use the difference of the three components in HIS color space as our features, which can be denoted as:

$$F_{ih} = \frac{1}{m * n} \sum_{y=1}^n \sum_{x=1}^m \sqrt{d_x(h)^2 + d_y(h)^2} \quad (10)$$

$$d_x(h) = I_h(x, y) - I_h(x + 1, y) \quad (11)$$

$$d_y(h) = I_h(x, y) - I_h(x, y + 1) \quad (12)$$

where I is the input image with $m * n$ points, and $I_h(x, y)$ is the h value of point (x, y) . In the same way, we can get the definition of F_{is} and F_{ii} as follows:

$$F_{is} = \frac{1}{m * n} \sum_{y=1}^n \sum_{x=1}^m \sqrt{d_x(s)^2 + d_y(s)^2} \quad (13)$$

$$F_{ii} = \frac{1}{m * n} \sum_{y=1}^n \sum_{x=1}^m \sqrt{d_x(i)^2 + d_y(i)^2} \quad (14)$$

5) *Correlation Analysis*: Fig. 11 shows the correlation matrix between the image features (F_i) and the $PM2.5$, using the image data we collected from May 2014 to March 2015. As the photos are taken from $PM2.5$ monitoring sites, all the image data are with true $PM2.5$ labels. Apparently, a high F_{id} usually means a low concentration of $PM2.5$ (the green points shown in Fig. 11). A high concentration of $PM2.5$ (the dark points) is always with a low F_{ii} . The green points always appear when the F_{id} and F_{ig} are all high, and the heavy pollution usually appears when the F_{is} and F_{ii} are all low. In short, these features are very discriminative in $PM2.5$ inference.

B. Meteorological features (F_m)

It's widely accepted that the concentration of air pollutants is influenced by meteorology. Air quality is strongly dependent on weather and therefore sensitive to climate change [21]. A high wind speed disperses the concentration of PM, while high humidity usually causes a high concentration [5]. Accordingly,

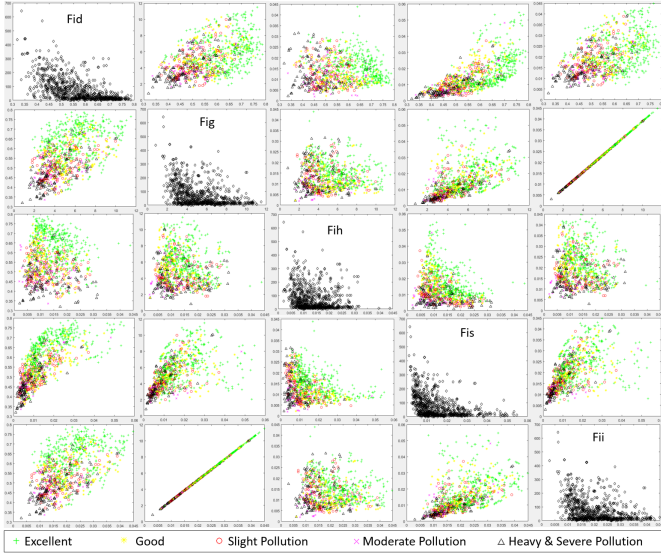


Fig. 11. Correlation matrix between Fi and PM2.5

we collect the meteorological data from the government's website [16] and identify five features: humidity (F_{wh}), temperature (F_{wt}), wind speed (F_{ws}), barometer pressure (F_{wb}), and weather (F_{ww}). The weather features have five classifications: sunny, cloudy, foggy, rainy and snowy. So the F_w for grid g can be defined as

$$F_w(g) = \{F_{wh}(g), F_{wt}(g), F_{ws}(g), F_{wb}(g), F_{ww}(g)\} \quad (15)$$

C. Traffic features (F_t)

It is widely believed that traffic flow is one of the major sources generating air pollutants [22]. The levels of PM increased during traffic rush hours and then decreased by the reduction of traffic [23]. Accordingly, we collect the traffic data from the Baidu map website [15]. The http request records of Baidu map and identify the traffic feature with the labeled conditions shown in Fig. 12. The pixels of different colours (green, yellow, red) can describe the traffic condition. Since all the http request records are in the same size, the number of the green, yellow, red pixels in a grid can be used to describe the traffic condition. We divide the map into $1km \times 1km$ grids, the traffic status in each grid is hourly-updated, and we keep all the grids' traffic status in our system. So the F_t for grid g can be defined as

$$F_t(g) = \{g.g(num), g.y(num), g.r(num)\} \quad (16)$$

where $g.g(num)$ is the green pixels in grid g , $g.y(num)$ is the yellow pixels in grid g and $g.r(num)$ is the red pixels in grid g . For each position to be inferred, the grid must be confirmed based on its GPS to get its traffic feature.

D. POI features (F_p)

The category of POIs indicates the land use and the function of the region. Some POI categories even have a direct relation with air pollution. For example, a factory may cause the increase of pollutant, a shopping mall may cause the



Fig. 12. The traffic information

TABLE I
CATEGORY OF POIS

$C1$	Transportation	$C6$	Park
$C2$	Entertainment	$C7$	Company
$C3$	Education	$C8$	Gas station
$C4$	Restaurant	$C9$	Shopping malls
$C5$	Factory	$C10$	Sports

aggregation of the people and the cars, which contributes to the air pollution, a park with many trees may contribute to a good air quality for the regions nearby. In this paper, we divided the POIs into ten categories shown in Table I, the POI feature F_p for grid g can be defined as

$$F_p(g) = g.(\{Num(C1), Num(C2), ..., Num(C10)\}) \quad (17)$$

where $Num(C1)$ is the number of the POIs $\in C1$.

IV. LEARNING AND INFERENCE

As shown in Fig.10, the learning process is divided into two parts, Offline Learning and Incremental Learning. As the data collected have a lot of features, Random Forest can be a good choice because they are inherently multi-class and very fast in both training and evaluation. We also test other models in the experiment part (detailed in Section 5). Since there are only a few air quality monitoring stations, most data are unlabeled, we use the Semi-Supervised Random Forest (SRF) [14] model to perform the off-line learning process. As the image data in our system are collected in real time, to improve the estimation's accuracy, the On-line Random Forest (ORF) [13] model is used to achieve the incremental learning. The two models are combined in our system as shown in Algorithm 1.

A. Random Forest

A Random Forest (RF) is a set of decision trees. Each tree in the forest is built and tested independently from other trees.

We denote the t^{th} tree of the ensemble as $f_t = f(x, \theta_t) : X \rightarrow Y$, where θ_t is a random vector capturing the various stochastic elements of the tree. The entire forest is denoted as $F = \{f_1 f_2 ... f_T\}$, where T is the number of trees in the forest. The estimated probability $p(PM2.5|x)$ for predicting class $PM2.5$ can be defined as

Algorithm 1 MCS-RF

Require: : Sequential sets of features: F
Require: : Some labeled grids G_l
Require: : Some unlabeled grids G_u
Require: : The recalculation time of the offline learning: T
Ensure: : The model produced by the offline learning: M_{off}
Ensure: : The model produced by the online learning: M_{on}
 $i \leftarrow CurrentTime$
while F is not empty **do**
 put F to the offline learning data sets;
 if $CurrentTime - i \geq T$ **then**
 $CurrentModel \leftarrow M_{off}$;
 $i \leftarrow CurrentTime$
 else
 $CurrentModel$ unchange;
 end if
 if $F \in G_l$ **then**
 $CurrentModel \leftarrow M_{on}$;
 else
 $CurrentModel$ unchange;
 end if
end while

$$p(PM2.5|x) = \frac{1}{T} \sum_{t=1}^T p_t(PM2.5|x) \quad (18)$$

where $p(PM2.5|x)$ is the estimated density of class labels of the leaf of the t^{th} tree. Then the final decision function of the forest can be defined as

$$C(x) = \underset{PM2.5 \in Y}{argmax} (PM2.5|x) \quad (19)$$

B. Offline learning

For many semi-supervised learning algorithms, the supervised loss function of unlabeled instances can be defined in the form of:

$$\sum_{(x,y) \in X_l} \lambda(y, h(x)) + \theta \sum_{x \in X_u} \lambda_u(h(x)) \quad (20)$$

where X_l is the labeled data, X_u is the unlabeled data, $(h.)$ is a binary classifier and $\lambda_u(.)$ encodes the regularizer based on the unlabeled samples. The regularization paradigm can be further subdivided into two main approaches: Manifold Assumption [24] [25] and Large Margin Approaches [26] [14].

Since we target applications with a large amount of data and manifold regularization leads to algorithms that are quadratic, i.e. $O(n^2)$, in terms of number of samples, in this work, we chose to use the Leistner et al.'s [14] approach. Given a margin maximizing loss function $\lambda(g_{PM2.5}(x))$, the local score for a decision node R_j is defined as

$$\ell(R_j) = \sum_{i=1}^T p_i^j \lambda(p_i^j - \frac{1}{T}) \quad (21)$$

the margin for the unlabeled data is defined as

$$m_u(x_u) = \max_{i \in PM2.5} g_i(x_u) \quad (22)$$

Based on the definition of the margin for unlabeled samples, the overall loss can be defined as

$$\ell(g) = \frac{1}{|X_l|} \sum_{(x, PM2.5) \in X_l} \lambda(g_{PM2.5}(x)) + \frac{\alpha}{|X_u|} \sum_{x \in X_u} \lambda(m_u(x)) \quad (23)$$

where α defines the contribution rate of the unlabeled samples, Leistner et al. [14] apply deterministic annealing to iteratively solve Eq 23.

C. Online learning

To perform online bagging, we use the N.Oza et al.'s [27] method where the sequential arrival of the data is modeled by a Poisson distribution. That means for each tree $T(x)$, it will update k times on each instance, where k is generated by $Poisson(\theta)$ and θ is usually defined as a constant number. To grow random trees on-the-fly, Saffari et al.'s [13] method can be a good choice, which grows extremely randomized trees by generating the test functions and thresholds randomly. The usual choices for quality measures are the entropy

$$L(R_j) = - \sum_{i=1}^K p_i^j \log(p_i^j) \quad (24)$$

or the Gini index

$$L(R_j) = \sum_{i=1}^K p_i^j (1 - (p_i^j)) \quad (25)$$

where p_i^j means label density of class i in node j , while K is the number of classes. Then the gain with respect to a test t can be measured as

$$\Delta L(R_j, t) = L(R_j) - \frac{|R_{jlt}|}{|R_j|} L(R_{jlt}) - \frac{|R_{jrt}|}{|R_j|} L(R_{jrt}) \quad (26)$$

where R_{jlt} and R_{jrt} are the left and right partitions made by the test t , $|R_j|$ is the number of instances in R_j . when splitting a node, the test with highest gain can produces the best splits of the data with respect to reduce the impurity of a node. Saffari et al. [13] perform non-recursive strategy for on-line learning, which means that a generated tree starts with only one root node. In our system, the online random forest is performed on the decision trees generated by the offline learning process. A node splits when $|R_j| > \alpha$ and $\Delta L(R_j, t) > \beta$, where α is the minimum number of samples a node has to collect before splitting, β is the minimum gain.

V. EVALUATION

A. Datasets

In the experiments, the data from May 2014 to March 2015 is collected, and the following datasets of Beijing are all available to the public except the image data.

1) *Meteorological data:* We collect fine-grained meteorological data, consisting of temperature, humidity, barometer pressure, wind strength and weather, which consists of five statuses: sunny, cloudy, foggy, rainy and snowy, published by meteorological authorities [16].

TABLE II
EVALUATION INDEX

Index	Description
Precision	For a classification, the correct number of examples of the total number of cases
Recall	Indicates the number of examples that are correctly identified.
F-measure	This value is a is the harmonic mean of precision and recall
TP rate	The recognition rate, the probability of recognizing an instance of this classification
FP rate	The error rate, the probability of identifying an instance of other classification as this classification
ROC area	The ROC curve is created by plotting the TP rate against FP rate at various threshold settings

2) *Traffic data*: We employ traffic features based on grids from Baidu maps. We collect the fine-grained traffic status(green, yellow and red) from the website [15] and calculate the traffic features for each grid every hour.

3) *Air quality records*: We collect real-valued AQI of six kinds of air pollutants, consisting of SO_2 , NO_2 , CO , O_3 , pm_{10} , $pm_{2.5}$, reported by air quality monitoring stations in Beijing every hour.

4) *POI*: We employ a POI database from Baidu Maps to extract F_p for each grid. As the POIs don't change frequently, we update our database every day.

5) *Image data*: The image data consists of photos taken by the participants with their smartphones, which range from May 2014 to March 2015 in Beijing. All the photos uploaded to our participatory sensing platform have accurate GPS positioning information. The photos near the PM2.5 stations (within one kilometer) will be labeled with the PM2.5 reported by the station.

6) *Evaluation Index* : The evaluation indexes we used in this paper are shown in the table II.

B. Baselines and Ground Truth

We compare our method with four baselines:

1) *logistic*: It is a classification model for building and using a logistic regression model with a ridge estimator.

2) *Naive Bayes*: It is a classification model for a Naive Bayes classifier using estimator classes [28].

3) *Random tree*: It is a classification model for constructing a tree that considers K randomly chosen attributes at each node. we perform no pruning in the experiment.

4) *BP ANN*: Back-propagation neural network (BP ANN) is a classifier that uses back-propagation to classify instances, which is used by U-Air [5].

Ground truth: We deliberately remove a station from a grid and infer its Pm2.5 from the other stations. The actual PM2.5 reported by the station is then used as the ground truth to measure the inference. The cross validation method is used to get the final result.

C. Results

1) *Results of models*: Fig. 13 shows the performance of random forest and the three aforementioned baselines. The

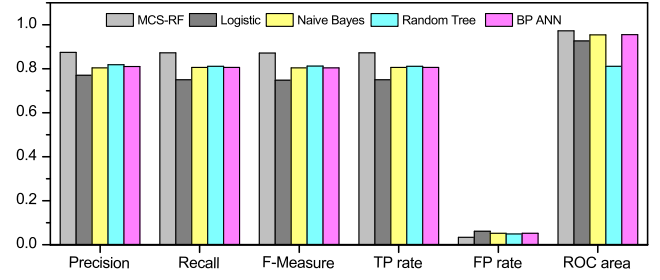


Fig. 13. Results of the models

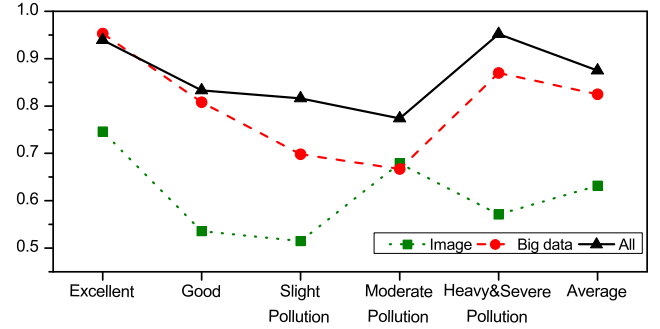


Fig. 14. Precision of the MCS-RF

results demonstrate the advantage of our method over logistic and Naive Bayes, and Random tree models based on our data sets in this situation.

2) *Results of Features*: In this area, zheng [5] has already proved that the meteorological data, traffic data, POI and the air quality records have a close relationship with the air quality. As shown in the figure 11, the image's features have a fuzzy relationship with the PM2.5. In this paper, the features are divided into two categories: the image features and the other features (Big data features), where the four models are employed to study the performance of features and their combinations. As shown in table III, adding the image feature sets into the models brings a significant precision and recall improvement on different models. The other evaluation indexes also have positive improvements except the ROC area in the Random Tree model.

3) *Results of the classifications with MCS-RF*: Fig. 14, Fig. 15 and Fig. 16 show the classification results of precision, recall and F-Measure based on MCS-RF. Adding the image feature sets into the model brings a significant precision improvement on the all classifications. Especially, the image features has a good performance on the classification of the Moderate Pollution, the overall precision of the Moderate Pollution experiences a sharp increase when we add the image features. The conclusion can be the same when it comes to the Recall and F-Measure. Although the performance of the image features is not very satisfactory without other features, adding the image features to the feature sets will always bring an improvement on all the classifications.

TABLE III
RESULTS RELATED TO MODELS

Model	MCS-RF			Logistic			Naive Bayes			Random Tree			BP ANN		
Datasets	Image	Big data	All	Image	Big data	All	Image	Big data	All	Image	Big data	All	Image	Big data	All
Precision	0.632	0.825	0.875	0.603	0.75	0.77	0.597	0.759	0.804	0.551	0.783	0.818	0.576	0.804	0.81
Recall	0.633	0.822	0.872	0.6	0.744	0.75	0.517	0.75	0.806	0.561	0.772	0.811	0.572	0.806	0.806
F-Measure	0.632	0.822	0.871	0.599	0.744	0.748	0.516	0.751	0.804	0.544	0.769	0.812	0.573	0.803	0.804
TP rate	0.633	0.822	0.872	0.6	0.744	0.75	0.517	0.75	0.806	0.561	0.772	0.811	0.572	0.806	0.806
FP rate	0.105	0.043	0.034	0.117	0.059	0.061	0.105	0.054	0.052	0.129	0.053	0.049	0.126	0.053	0.052
ROC area	0.867	0.954	0.972	0.828	0.914	0.927	0.808	0.935	0.954	0.716	0.87	0.811	0.799	0.922	0.955

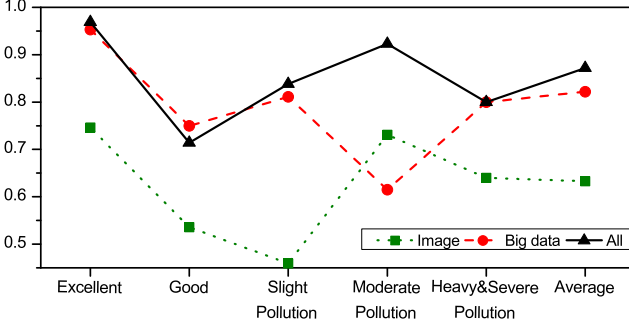


Fig. 15. Recall of the MCS-RF

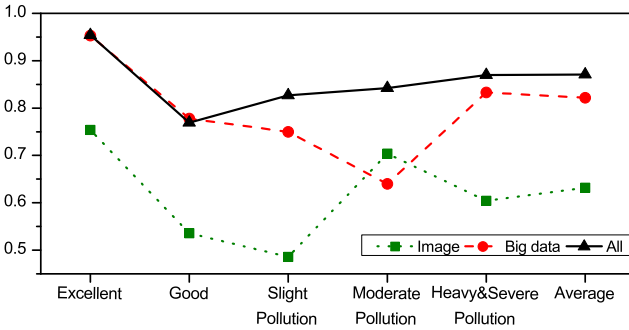


Fig. 16. F-Measure of the MCS-RF

VI. RELATED WORK

There is a lot of research on air quality monitoring based on WSN in the past few years. At present, there are three ways to achieve the air quality monitoring: Static Sensor Network (SSN), Vehicle Sensor Network (VSN) and Community Sensor Network (CSN). In SSN, the sensor nodes are typically mounted on the streetlight or traffic light poles, or walls [3] [29]. It can provide accurate and reliable data and can also guarantee the network connectivity. However, to achieve the fine-grained air quality monitoring, many sensors and a customized wireless network are required. What's more, the SSN may also cause resource wasting [30]. In VSN, the sensor nodes are typically carried by the public transportations like buses or taxis [4]. It can provide accurate and reliable data, and the sensor nodes have high mobility. However, to achieve fine-grained air quality monitoring the cost inefficiency on carriers could be very high. Uncontrolled or semi-controlled mobility can not be accepted in this way. Also, the Spatial-to-Temporal

resolution trade-off problem can't be ignored [31]. In CSN, the sensor nodes are carried by the public or professional users [32]. Since the sensors used are usually very cheap and easy to operate with low accuracy, the quality and quantity of the data can be very poor in some situation. For PM2.5, the sensors with acceptable accuracy are usually very high cost, large size and with heavy weight. It's hard to provide fine-grained PM2.5 monitoring service unless the PM2.5 sensors have a significant improvement.

There have been many studies on estimating the extinction coefficient and the transmittance of the scene based on the photos. He [9] uses the dark channel to remove the haze by calculating the transmittance with a single photo. Graves [19] uses the spatial contrast to estimate atmospheric light extinction with cameras. As the wavelength range of visible light in air is between 390 and 780nm, according to Mie scattering principle [33], the fine particulate matter with the diameter at the range of 0.38-0.78 (the major component of PM2.5) would obviously affect the extinction coefficient and the transmittance of the scene. However, the correlation between PM2.5 and extinction coefficient or transmittance is not intuitively clear.

The Mobile Crowd Sensing and Computing (MCSC) has been developing fast in recent years. The MCSC applications mainly focus on environment monitoring, transportation, traffic planning and urban dynamics sensing with smartphones. Maisonneuve's team [34] measure and map noise pollution with mobile smartphones' microphone. Eisenman [35] deploys BikeNet, an extensible mobile sensing system to measure the air pollution. Cui [36] schedules tasks fairly for large-scale mobile device systems with mobile cloudlets. Yang et al. [37] propose an incentive mechanism design for mobile phone sensing. F. Restuccia et al. [38] propose a trust-based framework for secure user incentivization in participatory sensing. Zheng et al. [5] achieves the fine-grained air pollution monitoring with the data they observed, which can also be achieved in our cloud servers.

VII. CONCLUSION AND FUTURE WORK

In this paper, we find that several image features are very discriminative in PM2.5 inference. However, it is difficult to get accurate PM2.5 concentration only with the images. when we add the image features to the other features (the meteorological features, the traffic features, and the POI features) we collected, we have a high evaluation accuracy on PM2.5 estimation (precision: 0.875, recall: 0.872) without any PM2.5

measurement devices. Our method outperforms other methods: Logistic, Naive Bayes, Random Tree and BP ANN.

In the future, to offer a better fine-grained air pollution monitoring service, it is essential to design the incentive mechanism to get the image data of some specific areas. Since the image features have strong effects on the evaluation results, the low-quality photos must be removed from our datasets. An algorithm based on image recognition need to be performed on our cloud servers to achieve the image filtering.

VIII. ACKNOWLEDGMENT

This work was partly supported by the National Natural Science Foundation of China (No.61370197, No.61402045 and No.61602051), Research and technology verification of address-driven network architecture (2015AA015601) and Research on architecture and key technology system of service-oriented software-defined network (2015AA016101).

REFERENCES

- [1] Nasa climate research. <http://climate.nasa.gov/causes/>. accessed on 17 April 2017.
- [2] Monitoring stations in beijing. <http://106.37.208.233:20035/>. accessed on 17 April 2017.
- [3] G. Murty, R. N. adn Mainland, I. Rose, A. R. Chowdhury, A. Gosain, J. Bers, and M. Welsh. Citysense: An urban-scale wireless sensor network and testbed. In *Technologies for Homeland Security, 2008 IEEE Conference on*, pages 583–588. IEEE, 2008.
- [4] P. Völgyesi, A. Nádas, X. Koutsoukos, and Á. Lédeczi. Air quality monitoring with sensormap. In *Proceedings of the 7th international conference on Information processing in sensor networks*, pages 529–530. IEEE Computer Society, 2008.
- [5] Y. Zheng, F. Liu, and H. P. Hsieh. U-air: when urban air quality inference meets big data. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1436–1444. ACM, 2013.
- [6] X. Liu, Z. Song, E. Ngai, J. Ma, and W. Wang. Pm2: 5 monitoring using images from smartphones in participatory sensing. In *2015 IEEE Conference on Information Communications Workshops (INFOCOM WKSHPS)*, pages 630–635. IEEE, 2015.
- [7] H. Ozkaynak, A. D. Schatz, G. D. Thurston, R. G. Isaacs, and R. B. Husar. Relationships between aerosol extinction coefficients derived from airport visual range observations and alternative measures of airborne particle mass. *Journal of the Air Pollution Control Association*, 35(11):1176–1185, 1985.
- [8] S. K. Nayar and S. G. Narasimhan. Vision in bad weather. In *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, volume 2, pages 820–827. IEEE, 1999.
- [9] K. He, J. Sun, and X. Tang. Single image haze removal using dark channel prior. *IEEE transactions on pattern analysis and machine intelligence*, 33(12):2341–2353, 2011.
- [10] B. Guo, X. Yu, Z. and Zhou, and D. Zhang. From participatory sensing to mobile crowd sensing. In *Pervasive Computing and Communications Workshops (PERCOM Workshops), 2014 IEEE International Conference on*, pages 593–598. IEEE, 2014.
- [11] F. Restuccia, S. K. Das, and J. Payton. Incentive mechanisms for participatory sensing: Survey and research challenges. *ACM Transactions on Sensor Networks (TOSN)*, 2016.
- [12] X. Zhang, Z. Yang, and W. Sun. Incentives for mobile crowd sensing: A survey. *IEEE Communications Surveys & Tutorials*, 2016.
- [13] A. Saffari, C. Leistner, J. Santner, M. Godec, and H. Bischof. On-line random forests. In *Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on*, pages 1393–1400. IEEE, 2009.
- [14] C. Leistner, A. Saffari, and H. Santner, J. and Bischof. Semi-supervised random forests. In *2009 IEEE 12th International Conference on Computer Vision*, pages 506–513. IEEE, 2009.
- [15] Baidu map. <http://map.baidu.com/>. accessed on 17 April 2017.
- [16] Meteorological data. <http://data.cma.cn/>. accessed on 17 April 2017.
- [17] S. Poduri, A. Nimkar, and G. S. Sukhatme. Visibility monitoring using mobile phones. *Annual Report: Center for Embedded Networked Sensing*, pages 125–127, 2010.
- [18] J. H. Seinfeld, S. N. Pandis, and K. Noone. *Atmospheric chemistry and physics: from air pollution to climate change*. John Wiley & Sons, 2016.
- [19] N. Graves and S. Newsam. Using visibility cameras to estimate atmospheric light extinction. In *Applications of Computer Vision (WACV), 2011 IEEE Workshop on*, pages 577–584. IEEE, 2011.
- [20] K. W. Kim and Y. J. Kim. Perceived visibility measurement using the hsi color difference method. *Journal of the Korean Physical Society*, 46(5):1243–1250, 2005.
- [21] D. J. Jacob and D. A. Winner. Effect of climate change on air quality. *Atmospheric environment*, 43(1):51–63, 2009.
- [22] S. Vardoulakis, B. E. Fisher, K. Pericleous, and N. Gonzalez-Flesca. Modelling air quality in street canyons: a review. *Atmospheric environment*, 37(2):155–182, 2003.
- [23] J. Cyrys, J. Heinrich, G. Hoek, M. Meliefste, K. and Lewn, U. Gehring, and H. E. Wichmann. Comparison between different traffic-related particle indicators: elemental carbon (ec), pm2. 5 mass, and absorbance. *Journal of Exposure Science and Environmental Epidemiology*, 13(2):134–143, 2003.
- [24] O. Chapelle, J. Weston, and B. Schölkopf. Cluster kernels for semi-supervised learning. In *Advances in neural information processing systems*, pages 585–592, 2002.
- [25] M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of machine learning research*, 7(Nov):2399–2434, 2006.
- [26] T. Joachims. Transductive inference for text classification using support vector machines. In *ICML*, volume 99, pages 200–209, 1999.
- [27] N. C. Oza and S. Russell. Experimental comparisons of online and batch versions of bagging and boosting. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 359–364. ACM, 2001.
- [28] G. H. John and P. Langley. Estimating continuous distributions in bayesian classifiers. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, pages 338–345. Morgan Kaufmann Publishers Inc., 1995.
- [29] J. H. Liu, Y. F. Chen, T. S. Lin, D. W. Lai, T. H. Wen, C. H. Sun, and J. A. Jiang. Developed urban air quality monitoring system based on wireless sensor networks. In *Sensing Technology (ICST), 2011 Fifth International Conference on*, pages 549–554. IEEE, 2011.
- [30] A. Dobre, S. J. Arnold, R. J. Smalley, J. W. D. Boddy, J. F. Barlow, A. S. Tomlin, and S. E. Belcher. Flow field measurements in the proximity of an urban intersection in london, uk. *Atmospheric Environment*, 39(26):4647–4657, 2005.
- [31] K. J. Wong, C. C. Chua, and Q. Li. Environmental monitoring using wireless vehicular sensor networks. In *Proceedings of the 5th International Conference on Wireless communications, networking and mobile computing*, pages 4986–4989. IEEE Press, 2009.
- [32] D. Hasenfrazt, O. Saukh, S. Sturzenegger, and L. Thiele. Participatory air pollution monitoring using smartphones. *Mobile Sensing*, pages 1–5, 2012.
- [33] M. Kerker. *The scattering of light and other electromagnetic radiation: physical chemistry: a series of monographs*, volume 16. Academic press, 2013.
- [34] N. Maisonneuve, M. Stevens, M. E. Niessen, and L. Steels. Noisetube: Measuring and mapping noise pollution with mobile phones. In *Information technologies in environmental engineering*, pages 215–228. Springer, 2009.
- [35] S. B. Eisenman, E. Miluzzo, N. D. Lane, R. A. Peterson, G. S. Ahn, and A. T. Campbell. Bikenet: A mobile sensing system for cyclist experience mapping. *ACM Transactions on Sensor Networks (TOSN)*, 6(1):6, 2009.
- [36] Y. Cui, J. Song, and K. Ren. Software defined cooperative offloading for mobile cloudlets. *Transactions on Networking*, 2017.
- [37] D. Yang, G. Xue, X. Fang, and J. Tang. Crowdsourcing to smartphones: incentive mechanism design for mobile phone sensing. In *Proceedings of the 18th annual international conference on Mobile computing and networking (Mobicom '12)*, 2012.
- [38] F. Restuccia and S. K. Das. Fides: A trust-based framework for secure user incentivization in participatory sensing. *Proceeding of IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks 2014*, 2014.