U-Air: When Urban Air Quality Inference Meets Big Data

Yu Zheng, Furui Liu, Hsun-Ping Hsieh
Microsoft Research Asia, Beijing China
{yuzheng, v-ful, v-hshsie}@microsoft.com

ABSTRACT

Information about urban air quality, e.g., the concentration of $PM_{2.5}$, is of great importance to protect human health and control air pollution. While there are limited air-quality-monitor-stations in a city, air quality varies in urban spaces non-linearly and depends on multiple factors, such as meteorology, traffic volume, and land uses. In this paper, we infer the real-time and fine-grained air quality information throughout a city, based on the (historical and real-time) air quality data reported by existing monitor stations and a variety of data sources we observed in the city, such as meteorology, traffic flow, human mobility, structure of road networks, and point of interests (POIs). We propose a semi-supervised learning approach based on a co-training framework that consists of two separated classifiers. One is a spatial classifier based on an artificial neural network (ANN), which takes spatially-related features (e.g., the density of POIs and length of highways) as input to model the spatial correlation between air qualities of different locations. The other is a temporal classifier based on a linear-chain conditional random field (CRF), involving temporally-related features (e.g., traffic and meteorology) to model the temporal dependency of air quality in a location. We evaluated our approach with extensive experiments based on five real data sources obtained in Beijing and Shanghai. The results show the advantages of our method over four categories of baselines, including linear/Gaussian interpolations, classical dispersion models, well-known classification models like decision tree and CRF, and ANN.

Categories and Subject Descriptors

H.2.8 [Database Management]: Database Applications - data mining, Spatial databases and GIS;

General Terms

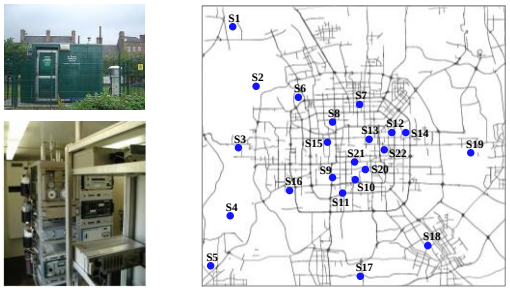
Algorithms, Management, Experimentation

Keywords

Air quality, city dynamics, human mobility, spatial trajectories.

1. INTRODUCTION

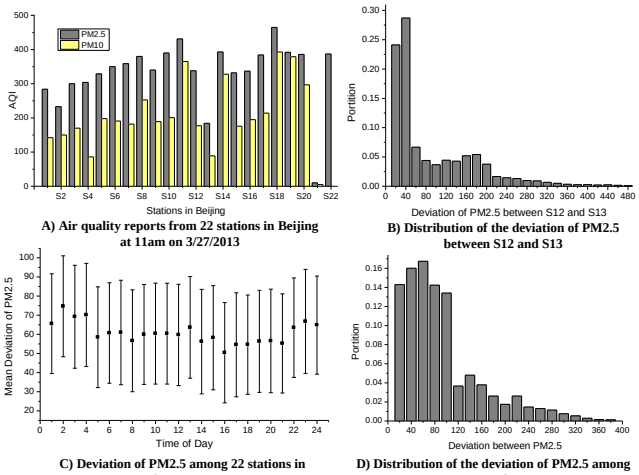
Real-time air quality information, such as the concentration of NO_2, $PM_{2.5}$, and PM_{10}, is of great importance to support air pollution control and protect humans from damage by air pollution. In reality, however, there are insufficient air quality measurement stations in a city due to the expensive cost of building and maintaining such a station. As demonstrated in Figure 1 A), an air quality monitor station usually needs a certain size of land, non-trivial money (about 200,000 USD for construction and 30,000 USD per year for maintenance), and human resources to regularly take care of it. As a result, for instance, Beijing only has 22 stations covering a 50×50km land(113km^2/per station), as illustrated in Figure 1 B).



A) Configuration of a station B) Air quality measurement stations in Beijing

Figure 1. Examples of air quality monitor stations

Unfortunately, urban air quality varies by locations non-linearly and depends on multiple factors, such as meteorology, traffic, land use, and urban structures. As depicted in Figure 2 A), the air quality indices (AQIs) reported by stations S_{12} and S_{13} are quite different at 11am on 3/27/2013, though they are geospatially close (about 3km away). As shown in Figure 2 B), the phenomenon is not a coincidence according to the distribution of the deviation between the $PM_{2.5}$ of the two stations reported at the same time of day (from Feb. 8 to May 27, 2013). Over 37 percent of the cases have a deviation greater than 100. Figure 2 C) further shows the mean deviation among the 22 stations in Beijing, changing over time of day. Figure 2 D) presents the distribution of the deviation among these 22 stations cross 3 months. All these results well demonstrate the skew of air quality in urban spaces.



A) Air quality reports from 22 stations in Beijing at 11am on 3/27/2013

B) Distribution of the deviation of PM2.5 between S12 and S13

C) Deviation of PM2.5 among 22 stations in Beijing (Feb. 8 to May 27, 2013)

D) Distribution of the deviation of PM2.5 among 22 stations in Beijing (Feb. 8 to May 27, 2013)

Figure 2. The difference between AQIs from different stations

In this paper, we infer the real-time and fine-grained air quality information throughout a city using (historical and real-time) air quality data reported by a limited number of existing monitor stations and a variety of data sets we observed in the city, such as meteorology, traffic flow, human mobility, structure of road networks, and POIs. Although environment scientists have been proposing models to approximate the relation between air quality and some factors like traffic and wind, these models are usually based on empirical assumptions and parameters that may not be

KDD'13, August 11–14, 2013, Chicago, Illinois, USA.

applicable to all urban environments [9] (detailed in the related work section). The methodology based on crowd and participatory sensing (e.g., using sensor-equipped mobile phones) could work for a very few kinds of gas like CO_2 but not applicable to aerosols and other pollutants, such as $PM_{2.5}$, PM_{10}, and NO_2. The devices for detecting the latter pollutants are not easily portable and usually need a relatively long sensing period (e.g., 1~2hours) before generating an accurate AQI.

Recently, big data reflecting city dynamics have become widely available [11][14], e.g., traffic flow, human mobility, and meteorology, enabling us to solve this challenging problem from a data perspective. According to existing studies [9], these data have a strong correlation with air qualities (detailed in Section 3). Using machine learning and data mining techniques, we build a network between air quality labels and features observed across multiple heterogeneous data sources. Figure 3 A) shows an example (Sept. 19, 2012 1pm in Beijing) of the results inferred by our method (U-Air), demonstrating the advantage beyond linear-interpolation, as depicted in Figure 3 B). To verify the validity of our method, we first *remove* two stations (S_1 and S_2) from Beijing (marked with two boxes) and predict the AQIs of the two with the rest of stations (denoted by gray points). The reports of the two stations are then employed as a ground truth (S_1=M, S_2=G) to evaluate the prediction (refer to Table 1 for the semantic meanings of the colors and AQI descriptors). Clearly, our method well reflects the ground truth, while linear interpolation generated incorrect results (S_1= S_2=U-S). The result also indicates another story. Besides providing accurate information of air quality, the research reported here can also suggest a location to setup a monitor station, where the inference of U-Air always differs from that of linear interpolation.

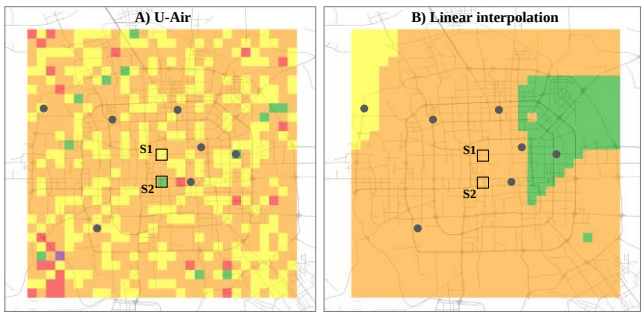


Figure 3. Results of PM_{10} generated by different methods

The challenges of our approach lie in three aspects. The first is to identify discriminative features from a variety of data sources. The second one is how to incorporate heterogeneous features into a data analytics model effectively. Equally treating these features does not work well. Third, the labeled data is insufficient though we have many observations represented by big data. While having many places to infer, only a few stations generate training data.

The contribution of this paper lies in the following three aspects:

- We propose a co-training-based semi-supervised learning approach, which leverages unlabeled data to improve the inference accuracy. Additionally, the approach consists of two classifiers respectively modeling the spatial and temporal factors that influence air qualities.
- We identify spatially-related (such as POIs, road networks, and distance to an existing station) and temporally-related features (e.g., humidity, temperature, and traffic flow), contributing to not only our application but also the general problem of air quality inference. Moreover, instead of treating these features equally, we feed them into the corresponding classifier in the co-training framework, therefore, leading to a high inference accuracy.
- We evaluated our approach using 5 data sources consisting of the POIs, road networks, meteorological data, and air quality records of Beijing and Shanghai, and the GPS trajectories generated by over 30,000 taxis in Beijing, justifying the advantages of our approach over 4 baselines.

2. OVERVIEW

2.1 Preliminary

***Definition 1**: Air quality index*. AQI is a number used by government agencies to communicate to the public how polluted the air is currently. As the AQI increases, an increasingly large percentage of the population is likely to experience increasingly severe adverse health effects. To compute the AQI requires an air pollutant concentration from a monitor or model. The function used to convert from air pollutant concentration to AQI varies by pollutants, and is different in different countries. Air quality index values are divided into ranges, and each range is assigned a descriptor and a color code. In this paper, we use the standard issued by United States Environmental Protection Agency, as shown in Table 1. The descriptor of each AQI level is regarded as the class to be inferred, i.e., $\mathcal{C}$={*G, M, U-S, U, V-U, H*}, and the color is employed in the following visualization figures.

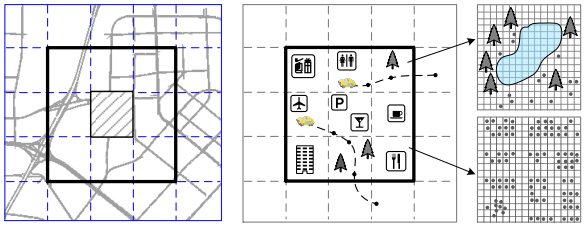
Table 1 AQI values, descriptors, and color codes

AQI	Values Levels of Health Concern	Colors
0-50	Good (G)	Green
51-100	Moderate (M)	Yellow
101-150	Unhealthy for sensitive groups (U-S)	Orange
151-200	Unhealthy (U)	Red
201-300	Very unhealthy (VU)	Purple
301-500	Hazardous (H)	Maroon

***Definition 2**: Trajectory*. A spatial *trajectory* τ is a sequence of time-ordered spatial points, $\tau: p_1 \rightarrow p_2 \rightarrow \cdots \rightarrow p_n$, where each point has a geospatial coordinate set and a timestamp, $p = (l, t)$.

***Definition 3**: POI*. A point of interest POI is a venue (like a school and shopping mall) in the physical world, having a name, address, coordinates, category, and other attributes.

***Definition 4**: Road Network*. A road network *RN* is comprised of a set of road segments $\{r\}$ connected among each other in a format of graph. Each road segment r is a directed edge having two terminal points, a list of intermediate points describing the segment, a length $r.len$, and a level $r.lev$ denoting its capacity.



A) Grid and its neighbors B) Data from the affecting region C) POI density

Figure 4. Illustration of grid, affecting region, POI, and trajectory

Definition 5: *Grid and Affecting Region*. We divide a city into disjointed grids (e.g., 1km×1km in the experiments) as illustrated in Figure 4 A), assuming the air quality in a grid g is uniform (while different grids may have different results). Each g has a geospatial coordinate $g.loc$ and a set of AQI labels $g.Q = \{q_1, q_2, \ldots, q_k\}$ to be inferred or already associated if having an air quality monitor station located. Here, k denotes the type of

pollutants, and $q_k \in \mathcal{C}$ (defined in Table 1) means the AQI label of the k-th type of pollutant, such as PM_{10}. We believe the air quality of a grid (filled by slashes in Figure 4 A)) would be influenced by the data (e.g., trajectories and POIs) observed in the *affecting region* $g.R$ that consists of the grid and its eight neighbor grids, as shown in Figure 4 B).

2.2 Framework

As shown in Figure 5, our framework consists of two major parts, offline learning and online inference, which generate three kinds of data flows: preprocessing, inference, and learning data flows.

Preprocessing data flow: In this flow (denoted by dotted black arrows), we receive spatial trajectories generated by vehicles (taxicabs in the experiments) and map each trajectory onto a road network using a map-matching algorithm [12]. The mapped data is then stored in a trajectory database for offline learning and also geo-indexed to improve the efficiency of online inference.

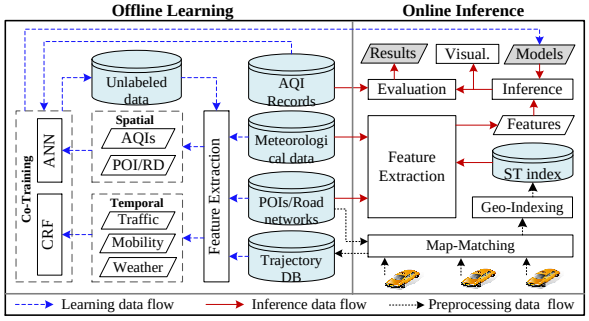


Figure 5. Framework of our system

Learning data flow: In this data flow (represented by broken blue arrows), we first extract features for each grid from a variety of data observed in its affecting region. In terms of spatio-temporal properties, these features can be categorized into two sets. One set is temporally-related (i.e., the value of the features vary with time), such as temperature, humidity, and average speed of vehicles, which are extracted from meteorological data and the spatial trajectories. The other feature set is only spatially-related, e.g., the density of POIs and the length of roads in a region, extracted from POI and road network databases. See Section 3 for details.

If an air quality monitor station is located in a grid, the grid is labeled by the AQIs reported from the station. The features extracted from the data observed in the affecting region of such a grid and the corresponding labels formulate a training set. As we only have a few air quality stations in a city while there are many places to infer, the data with labels are very few. To address this issue, we propose a semi-supervised learning approach based on co-training, where unlabeled data are used to improve the inference accuracy. Two separate classifiers are first trained respectively based on the labeled data using two separated feature sets. One is a temporal classifier (*TC*) based on a linear-chain conditional random field (CRF), which uses temporally-related features to estimate the temporal transformation of air quality in a location. The other is an artificial neural network (ANN)-based spatial classifier (*SC*) that uses spatially-related features to model the spatial correlations between air qualities of different locations. The AQIs reported by existing stations are also employed as an input in the *SC*. As different air pollutants (e.g., NO_2 and PM_{10}) are influenced by these factors differently, we build an individual model for each kind of pollutant, as detailed in Section 4.

Inference data flow: In this flow (denoted as the red solid arrows), we calculate the features for each grid based on the data observed in the grid's affecting region. While the spatially-related features like distribution of POIs can be calculated offline, the temporally-related features are computed online; e.g., the traffic-related features are extracted based on the spatio-temporal (ST)-index built in the preprocessing flow. For each grid, we respectively feed the spatially-related features into the *SC* and temporally-related features into *TC*, generating two probability scores. By multiplying the two scores, we can select the most possible class as a label. As monitor stations usually update the reports every hour, we conduct the inference every hour. Detailed in Section 4.

Problem statement

Given a collection of grids $G = G_1 \cup G_2 = \{g_1, g_2, \dots, g_n\}$, where $g.Q$ ($g \in G_1$) is known and $g'.Q$ is unobserved ($g' \in G_2$), $|G_1| \ll |G_2|$, a road network RN crossing G, POIs located in G, a trajectory dataset Tr passing G, and meteorological data of G, we aim to infer $g'.Q$, at periodic intervals, e.g., every 1 hour.

3. FEATURE EXTRACTION

3.1 Meteorological Features: F_m

The concentration of air pollutants is influenced by meteorology. Accordingly, we identify five features: temperature, humidity, barometer pressure, wind speed, and weather (such as cloudy, foggy, rainy, sunny, and snowy). Figure 6 shows the correlation matrix between the AQI of PM_{10} and the first four features, using the data we collected from August to Dec. 2012 in Beijing, where each row/column denotes one feature and a plot means the AQI label of a location. Apparently, a high wind speed disperses the concentration of PM_{10}, and high humidity usually causes a high concentration. A high pressure would result in a good AQI. The impact of temperature is not very clear, but, a good AQI is more likely when temperature is high and humidity is low, or when pressure is high and temperate is low. In short, these features are very discriminative in AQI inferences.

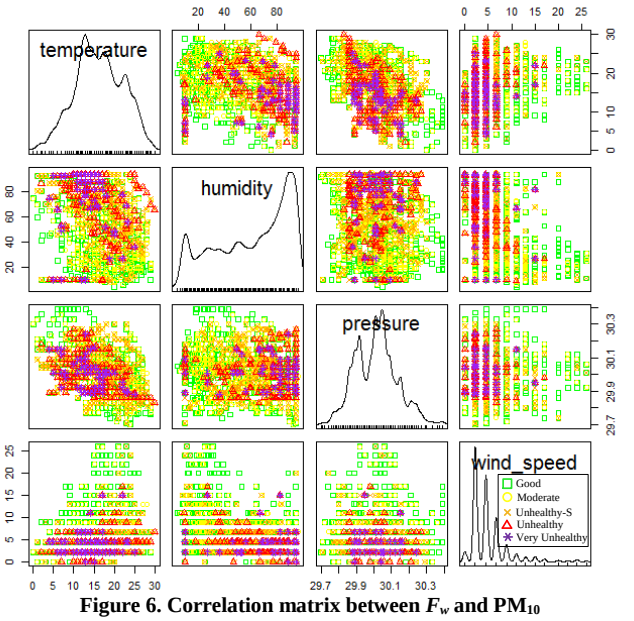


Figure 6. Correlation matrix between F_w and PM_{10}

3.2 Traffic-Related Features: F_t

It is widely believed that traffic flow is one of the major sources generating air pollutants that could damage air quality [9], though researchers are still exploring the specific correlation. Here, we identify the following three features for each grid. These features are calculated from the spatial trajectories of vehicles traversing the grid in the past hour:

1) Expectation of speeds: $E(V)$. Given a spatial trajectory generated by a vehicle, we retrieve the points that fall in the affecting region of each gird (let us say $p.l \in g.R$). We calculate the distance between any two consecutive points, then compute the speed of each vehicle at each point according to Equation 1. As the sampling rate of each GPS device is different, we calculate the expectation of speed w.r.t. time as Equation 2, which denotes the overall travel speed of vehicles in $g.R$.

$$p_i.v = \frac{Dist(p_i.l, p_{i+1}.l)}{|p_{i+1}.t - p_i.t|}, \quad (1)$$

$$E(v) = \frac{\sum_{p_i.l \in g.R} p_i.v \times |p_{i+1}.t - p_i.t|}{\sum_{p_i.l \in g.R} |p_{i+1}.t - p_i.t|}, \quad (2)$$

2) Standard deviation of speeds: $D(v)$. We calculate the feature according to Equation 3, which reflects how variably different vehicles were traveling in $g.R$ in the past hour. Similar to Equation 2, it is normalized based on time.

$$D(v) = \sqrt{\frac{\sum_{p_i.l \in g.R} [p_i.v - E(v)]^2 \times |p_{i+1}.t - p_i.t|}{\sum_{p_i.l \in g.R} |p_{i+1}.t - p_i.t|}}. \quad (3)$$

3) The distribution of speeds: $P(v)$. We employ a widely-used empirical setting to discretize the speed into three intervals (i.e., $0 \leq v < 20$, $20 \leq v < 40$, and $v \geq 40$), calculating the distribution of speeds over the three intervals in terms of Equation 4.

$$P(v_1 \leq v < v_2) = \frac{\sum_{p_j.l \in g.R \wedge v_1 \leq p_j.v < v_2} |p_{j+1}.t - p_j.t|}{\sum_{p_i.l \in g.R} |p_{i+1}.t - p_i.t|} \quad (4)$$

Figure 7 shows the correlation matrix between the aforementioned traffic features F_t and NO_2, where each row/column still denotes one feature and a plot denotes the AQI of a grid. Here, F_t is extracted from a GPS trajectory dataset generated by over 30,000 taxicabs. As taxis generate about 20% of traffic flow on road surfaces of Beijing [14], the dataset is big enough to represent the traffic patterns there. Additionally, GPS-equipped taxis can be regarded as mobile sensors probing the travel speed of each road. As a result, the features w.r.t. speeds are reliable [13].

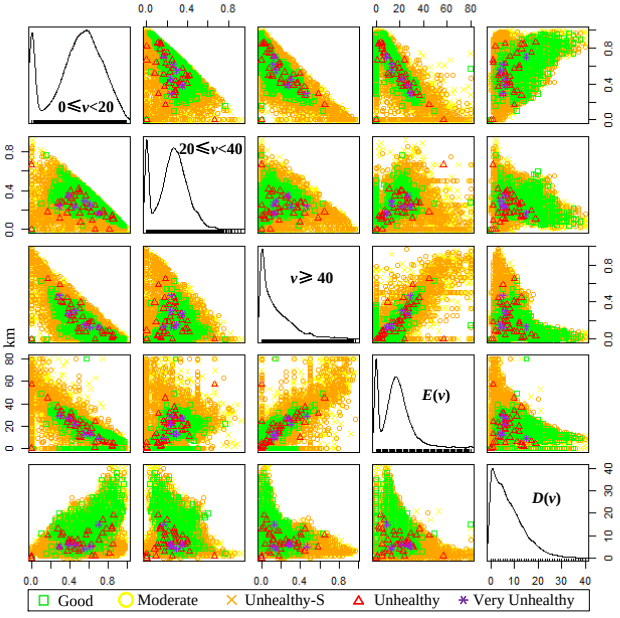


Figure 7. Correlation matrix between traffic features and NO_2

Clearly, the more vehicles traveling with a speed smaller than 20km/h, i.e., when $P(0 \leq v < 20)$ becomes larger, more instances of unhealthy and very unhealthy occurred. On the contrary, the larger $P(v > 40)$ is the better AQI would be (e.g., see the sub chart on the first row and the third column).This is very intuitive to understand given the fact that more air pollutants would be issued by a vehicle when traveling in a traffic jam, i.e., the gasoline would not be burned efficiently. A surprising discovery is that a bigger $D(v)$ indicates a better air quality while a smaller one has a very high probability of resulting in a worse AQI of NO_2, as depicted in the fifth column of Figure 7. It is actually very reasonable if we consider the speed limitation of different road segments. If there is no traffic jam, vehicles should travel with quite different speeds on different roads, e.g., vehicles traveling on highways (with a speed limitation of 120km/h) should move much faster than those on a local street (usually with a speed limitation of 40km/h). As a grid could contain road segments of different speed limitations, $D(v)$ tends to be large when the traffic condition in the grid is not bad. On the contrary, every vehicle has to move very slowly in a traffic jam, leading to a small $D(v)$. The results well matches the commonsense that traffic jams cause much heavier air pollution than normal traffic conditions.

3.3 Human Mobility Features: F_h

F_h is comprised of two features, denoting the number of people arriving at (f_a) and departing from (f_l) a grid's affecting region $g.R$ in the past hour. In practice, people themselves are not major air-pollutant-generators. Human mobility, however, implies useful information, such as land use of a location, traffic flow, and function of a region (like residential or business areas) [11], which could contribute to air quality inference. In the experiment, we extract the two features from the aforementioned taxi trajectories because the data tell the pickup and drop off points of each trip. The feature can actually be extracted from other data sources or a combination of multiple datasets, like mobile phone signal.

Figure 8 shows the correlation between AQIs and the human mobility features. Apparently, the concentration of PM_{10} in a grid g becomes denser when the number of people arriving at and departing from $g.R$ increases. When f_a and f_l becomes very small, however, there might be two results. One is a very good AQI; the other is very unhealthy. Both results actually make sense, as these places may have nature parks (good) or factories (unhealthy) with very few people traveling to.

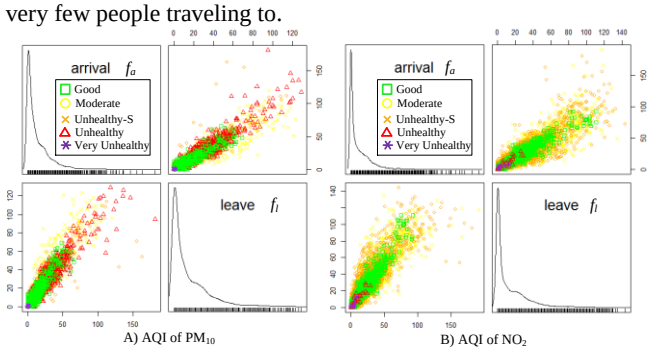


Figure 8. The correlation between AQIs and human mobility

While the traffic-related and human mobility features are calculated online, the feature extraction is very time-consuming. To address this issue, we build a ST-index between grids and the trajectories [5], as illustrated in Figure 9, where each grid is associated with two first-in-first-out lists respectively storing the taxi IDs traversing a grid and the pickup/drop-off points falling in the grid in the past hour. The two lists are sorted by arriving time and pickup/drop-off time respectively, and the trajectory data of a taxi is connected to the taxi ID by a hash table. Given an affecting region consisting of 9 grids (refer to Figure 4 as an example), we merge the taxi IDs falling in these grids by checking the sorted list.

We can then quickly retrieve the point data falling in the time interval from each trajectory by searching for the hash table.

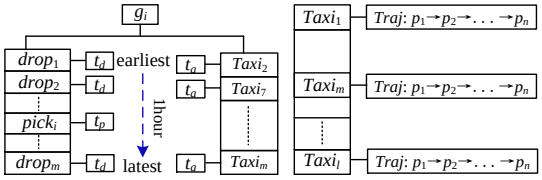


Figure 9. ST-index between grids and trajectories.

3.4 Road-network-related features: F_r

The structure of a road network has a strong correlation with its traffic patterns, therefore providing a good complement to traffic modeling. As demonstrated in Figure 4 A), we identify the following three features for each grid based on a road network database: 1) total length of highways f_h, 2) total length of other (low-level) road segments f_r, and 3) the number of intersections f_s in the grid's affecting region. Figure 10 presents the portion of instances with different AQI classes (NO_2) changing over f_h and f_r, by analyzing the data we collected in Beijing. The increase of road segments in a region significantly brings down the portion of *good* instances, enhancing the presence of *U-S and beyond* instances. We however did not see the phenomenon w.r.t. highways. We could say highways are relatively greener than other road segments in terms of generating air pollutants (as it does not usually contain traffic lights). This is also the reason why we need to separate f_h from f_r.

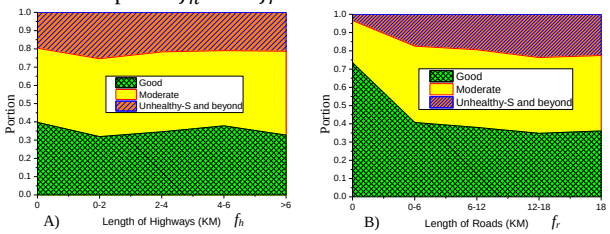


Figure 10. AQI of NO_2 changing over f_h and f_r

3.5 POI-related features: F_p

The category of POIs and their density in a region indicate the land use and the function of the region as well as the traffic patterns in the region, therefore contributing to the air quality inference of the region. Some POI category may even have direct causal relation with air quality. For example, if a region has some chemical factories, its air quality tends to be bad. A park, however, usually leads to a good air quality. Accordingly, we identify the following three features for each grid:

1) The number of POIs in some categories in $g.R$: f_n. We count the number of POIs belonging to the categories shown in Table 3.

Table 3. Category of POIs we studied

C_1: Vehicle Services (gas stations, repair)	C_7: Sports
C_2: Transportation spots	C_8: Parks
C_3: Factories	C_9: Culture & education
C_4: Decoration and furniture markets	C_{10}: Entertainment
C_5: Food and beverage	C_{11}:Companies
C_6: Shopping malls and Supermarkets	C_{12}:Hotels and real estates

2) The portion of vacant places in $g.R$: f_ρ. As illustrated in Figure 4 C), we further divide a grid into small cells, counting the number of cells without a POI located. In short, the bigger f_ρ the larger vacant places contained in a grid, therefore facilitating the dispersion of air pollutants; e.g., the upper subfigure in Figure 4 C) shows more vacant places than the bottom one due to the presence of a lake.

3) *The change in the number of POIs*: f_c. We compare the POI data of two consecutive quarters, calculating the change in the number of POIs in the following five categories (C_3, C_4, C_6, C_8, and C_{12}) in each grid's affecting region. The change implies the construction in which infrastructure was built or removed from a region. According to [9], construction is one of the major sources of air pollutants, such as PM_{10} and NO_2.

4. LEARNING AND INFERENCE

We propose the model based on the framework of co-training and the philosophy shown in Figure 11, where a circle denotes a location and a plane means the states of these locations at a timestamp. We can understand the philosophy of the model from the perspective of the state of air quality. First, air quality has temporal dependency on its current observations and that of its previous state. For example, the AQI of a location tends to be good if the AQI of the past hour is also good. Second, the air quality of a location is also reflected by its spatial neighbors. For instance, the AQI of a location is likely to be bad if the air quality of the places close to the location is bad. We can also understand the model from the perspective of the generation of air pollutants. The AQI of a location is determined by the air pollutants issued in the location and that propagated from other locations.

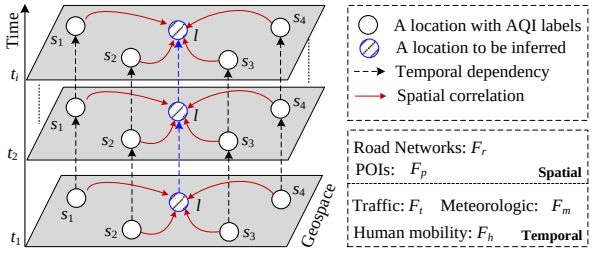


Figure 11. The philosophy of the inference model

4.1 Co-Training

Co-training is a semi-supervised learning technique that requires two *views* of the data. It assumes that each example is described by two different feature sets that provide different and complementary information about an instance. Ideally, the two feature sets of each instance are conditionally independent given the class, and the class of an instance can be accurately predicted from each view alone. Co-training can generate a better inference result as one of the classifiers correctly labels data that the other classifier previously misclassified [8].

Aligning with the co-training framework, we propose a spatial classifier (*SC*) to model the spatial correlation and a temporal classifier (*TC*) to model the temporal dependency of AQI. The two models are integrated into a co-training-based learning framework presented in Algorithm 1. As shown in line 3 and 4,

Algorithm 1: U-AIR Co-Training

Input: A set of features (F_m, F_t, F_h, F_r, F_p), some labeled grids G_1, and a set of unlabeled grids G_2, a threshold θ controlling the rounds
Output: The spatial classifier SC and temporal classifier TC.
1. $i \leftarrow 0$;
2. **Do**
3. $SC \leftarrow SC$.Learning (F_r, F_p, G_1);
4. $TC \leftarrow TC$.Learning (F_m, F_t, F_h, G_1);
5. Apply SC to each $g \in G_2$, for each class c_i, pick n_i grids that SC most confidently classifies as c_i, and add them to G_1.
6. Apply TC to each $g \in G_2$, for each class c_i, pick n_i grids that TC most confidently classifies as c_i, and add them to G_1.
7. i++;
8. **Until** G_2 is empty or $i > \theta$;
9. **Return** SC and TC;

we first train the two classifiers with two separated sets of features. The trained *SC* and *TC* are then used to infer unlabeled grids G_2 iteratively, adding the most confidently classified examples into the labeled dataset G_1 for the next round of training, until G_2 becomes empty or a certain rounds θ have been performed. When this algorithm ends, *SC* and *TC* are returned.

At the inference time, we apply SC and TC to the corresponding features separately, determining the AQI of a grid by the product of the two probability scores (P_{SC} and P_{TC}) generated by the two classifiers, defined in Equation 5.

$$c = \arg_{c_i \in \mathcal{C}} Max(P_{sc}^{c_i} \times P_{Tc}^{c_i}). \quad (5)$$

4.2 Temporal Classifier: TC

The temporal classifier infers the air quality of a grid given the temporally-related features consisting of F_m, F_t, and F_h of the grid, using a linear-chain CRF, which is a discriminative undirected probabilistic graphical model for parsing sequential data like natural language texts [4]. The advantage of CRFs over hidden Markov models is the relaxation of the independence assumptions between features. Additionally, CRFs avoid the label bias problem exhibited by maximum entropy Markov models.

Figure 12 shows the graphical structure $\boldsymbol{G}$ of the temporal classifier, which consists of two kinds of nodes $\boldsymbol{G} = (\boldsymbol{X}, \boldsymbol{Y})$. The gray nodes $\mathbf{Y} = \{Y_1, Y_2, \dots, Y_n\}$ represent hidden state variables to be inferred given the sequence of observations denoted by white nodes $\mathbf{X} = \{X_1, X_2, \dots, X_n\}$, $X_i = \{F_m, F_t, F_h, t\}$ (t is a timestamp by hour, e.g., 8am). The $Y_i \in \mathbf{Y}$ is structured to form a chain with an edge between each Y_{i-1} and Y_i, as well as having an AQI "label" belonging to $\mathcal{C}$. When conditioned on $\mathbf{X}$, the random variables Y_i obey the Markov property with respect to the graph $\boldsymbol{G}$:

$$P(Y_i|\mathbf{X}, Y_j, i \neq j) = P(Y_i|\mathbf{X}, Y_j, i \sim j),$$

where $i \sim j$ means i and j are neighbors in $\boldsymbol{G}$.

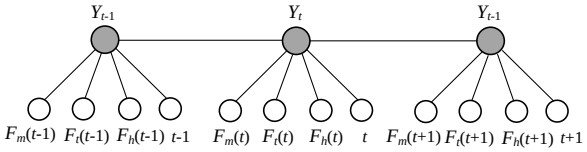


Figure 12. The graphic presentation of the temporal model

The probability of a particular label sequence $\boldsymbol{y}$ given observation sequence $\boldsymbol{x}$ is defined as a normalized product of potential functions as follows:

$$exp(\textstyle\sum_j \lambda_j\, t_j(y_{i-1}, y_i, x, i) + \sum_k \mu_k\, s_k(y_i, x, i)), \quad (6)$$

where $t_j(y_{i-1}, y_i, x, i)$ is a transition feature function of the entire observation sequence and the label at positions i and $i-1$; $s_k(y_i, x, i)$ is a state feature function of the label at position i and the observation sequence; λ_j and μ_k are parameters to be estimated from training data.

Writing $s_k(y_i, x, i) = s_k(y_{i-1}, y_i, x, i)$, we transfer Equation 6 to

$$P(y|x, \lambda) = \frac{1}{Z(x)} exp(\textstyle\sum_j \lambda_j\, f_j(y_{i-1}, y_i, x, i)), \quad (7)$$

where $Z(x)$ is a normalized factor [4]. This can be informally thought of as measurements on the input sequence that partially determine the likelihood of each possible value for Y_i. The model assigns each feature a numerical weight and combines them to determine the probability of a certain value for Y_i.

Given k sequences of the training data $\{(\boldsymbol{x}^{(k)}, \boldsymbol{y}^{(k)})\}$, learning the parameters λ is done by maximum likelihood learning $\mathrm{P}(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\lambda})$, which can be solved by gradient descent.

$$L(\boldsymbol{\lambda}) = \textstyle\sum_k \left[log \frac{1}{Z(x)} + \sum_j \lambda_j\, f_j(y_{i-1}, y_i, x, i) \right]. \quad (8)$$

4.3 Spatial Classifier: SC

The spatial classifier infers the AQI of a grid, using its own geospatial features and that of some grids having a monitor station. As depicted in Figure 13, the *SC* consists of two parts: input generation (in the left box) and an artificial neural network, where F_p^k, F_r^k, l^k, and c^k denotes the POI features, road network features, location, and the AQI label of grid k; x is the grid to be inferred; D_1 is a distance function between features (e.g., we use the Pearson correlation in the experiments) and D_2 calculates the geo-distance between the center of two grids, e.g.,

$$\Delta P_{kx} = Pearson\_Cor(F_p^k, F_p^x), \quad (9)$$

$$\Delta R_{kx} = Pearson\_Cor(F_r^k, F_r^x), \quad (10)$$

$$d_{kx} = Geo\_Distance(l^k, l^x). \quad (11)$$

Input generation: In this phase, we randomly choose n grids with labels, $\mathcal{G}_1 = (g_1, g_2, \dots, g_n)$, $\mathcal{G}_1 \in G_1$, to pair with the grid to be inferred (we found n=3 achieves the best inference accuracy in the experiments). The input of the ANN is then calculated according to Equation 9 to 11. To learn the impact of different scales of distance between grids, we perform this pairwise process m times to formulate a collection of inputs. The labeled grids involved in each round of input formulation should have at least e different grids from existing ones, formally defined as: $\mathbb{Q} = \{\mathcal{G}_1, \mathcal{G}_2 \dots, \mathcal{G}_m\}$, $\forall \mathcal{G}_i, \mathcal{G}_j \in \mathbb{Q}, |\mathcal{G}_i \cap \mathcal{G}_j| \leq e$, e.g., e=2 and n=3 means at least one out of the three grids is different from those used previously. Another reason for doing so is to vary the input. As the POI and road network features extracted from a grid are static, the input (ΔP_{nx}, d_{nx}, ΔR_{nx}) do not change over c^k if we always select the same three grids. Accordingly, the three inputs will be neglected by the ANN model in the training process.

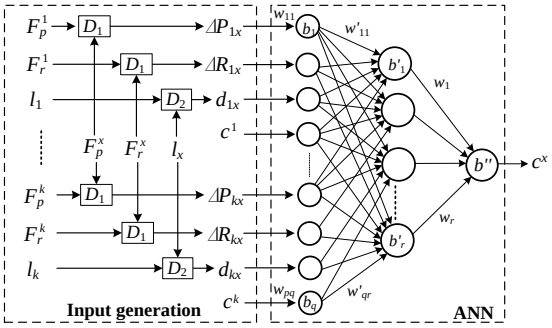


Figure 13. Structure of spatial classifier

Artificial neural network: Though many ANNs can be applied to our framework, we choose the widely-used Back-propagation (BP) neural network with one hidden layer in the experiments for its simplicity and generality. We set a linear function for the neurons (each of which accepts all the features) in the input layer and a sigmoid function $\varphi(x)$ for those in the hidden and output layers, formally defined as follows:

$$c^k = \varphi\left(\textstyle\sum_r w_r\, \varphi\left(\sum_q w'_{qr} \cdot \left(\sum_p f_p\, w_{pq} + b_q\right) + b'_n\right) + b''\right), \quad (12)$$

where f_p is a feature of input; b_m, b'_n, and b'' are the biases associated with the neuron in different layers; w_{pq}, w'_{qr}, and w_r denote the weight associated with the input of different layers.

In the inference process, we also pair a grid to be inferred with a certain sets of n labeled grids, generating a prediction of AQI label for each set. The frequency of each inferred label is then used as the probability score of the label, and the most frequent label will be selected as the final prediction result. The prediction of the *SC* can actually be regarded as a non-linear interpolation over geo-spaces, considering the road network and POIs of these

locations. This classifier is effective as road network and POI data are good supplementary of traffic data.

5. EXPERIMENTS

5.1 Datasets

In the evaluation we use the following five real datasets detailed in Table 4, where the first four sources are available in Beijing and Shanghai.

1) *Meteorological data*: We collect fine-grained meteorological data, consisting of weather, temperature, humidity, barometer pressure, wind strength, from a public website every hour.

2) *Air quality records*: We collect both real valued and labeled AQI of four kinds of air pollutants, consisting of SO_2, NO_2, $PM_{2.5}$, and PM_{10}, reported by ground-based air quality monitor stations in the four cities every hour. As a station may not have reports sometimes, we present the hours of effective records in Table 4.

3) *POIs*: We employ a POI database from Bing Maps to extract F_p for each city. The data of the first and third quarters of 2012 are used to identify the number of POIs changed in the five categories defined in Section 3.4.2. Figure 14 shows the POI distributions of Beijing and Shanghai.

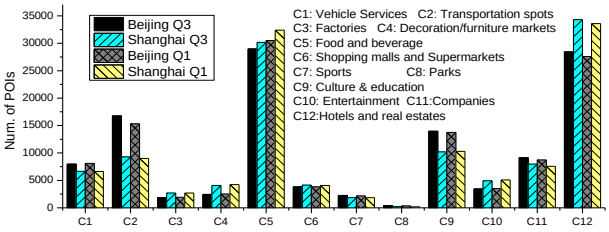


Figure 14. Number of POIs in different categories

4) *Road networks*: The road network data is also from Bing Maps.

5) *Taxi trajectories*: We use a GPS trajectory dataset generated by over 32,000 taxicabs in Beijing from August 21 to Nov. 30, 2012 to calculate F_t and F_h. GPS-equipped taxis can be regarded as mobile sensors probing the travel speed on roads, and the data also tell the pick-up and drop-off points of each trip. The total distance of the dataset is over 495 million kilometers, and the number of points reaches 1.45 billion. We found over 32 million occupied trips. As taxis generate about 20 percent of traffic flow on road surfaces of Beijing [14], the dataset is big enough to represent the traffic patterns there. Of course, the features can be extracted from other data sources or a combination of multiple datasets, like mobile phone signal.

Table 4. Details of the datasets

Data sources		Beijing	Shanghai
POIs	2012 Q1	271,634	321,529
	2012 Q3	272,109	317,829
Roads	#.Segments	162,246	171,191
	Highways	1,497km	1,963km
	Roads	18,525km	25,530km
	#. Intersections	49,981	70,293
AQIs	#. Stations	22	10
	Hours	23,300	8,588
	Time spans	8/24/2012-3/8/2013	1/19/2013-3/8/2013
Urban Sizes (grids)		50×50km (2500)	50×50km (2500)

5.2 Baselines and Ground Truth

We compare our method (U-Air) with five baselines:

1) *Linear interpolation (Linear)*: This is a distance-weighted interpolation algorithm using the AQI values reported by existing monitor stations, as shown in Equation 11,

$$\mathrm{g}_x.\mathrm{AQI} = \sum_i \frac{g_i.AQI \times \frac{1}{d_{xi}}}{\sum_i \frac{1}{d_{xi}}}, \quad (13)$$

where d_{xi} denotes the geo-distance between the location x and the i-th monitor station.

2) Another interpolation method is based on a Gaussian distribution (*Gaussian*) $X \sim N(0, \sigma)$, where σ is the average distance between any two existing air quality monitor stations in a city. Formally defined as

$$g_x.\mathrm{AQI} = \sum_i g_i.AQI \times f(d_{xi}), \quad (14)$$

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{\sigma^2}}. \quad (15)$$

We use continuous values in the interpolation and then discretize the aggregated result into AQI labels for evaluation.

3) *Classical Dispersion Model (Classical)*: This is a simple (also well-known) mathematical model that is typically applied to point source emitters, simulating them as point or line emission source propagating as a Gaussian plume. We use a widely-used tool CALPUFF [10] with default values for the parameters (as these parameters, e.g., vehicle emission rates, are difficult to obtain).

4) *Decision tree (DT)*: We choose this baseline to answer people's question why not just use a simple supervised learning model. In this baseline, we feed all the features equally into a decision tree.

5) *CRF-ALL and ANN-ALL*: Instead of dividing the features into two sets, the two baselines feed all features equally into the *SC* and *TC* respectively. We choose them as baselines to justify the features are used effectively in our approach.

Ground Truth: We deliberately *remove* a station from a grid and infer its air quality with the AQIs from other stations. The actual AQI reported by the station is then used as the ground truth to measure the inference. Each grid with a station is tested in this way every hour. For example, Beijing has 22 stations, generating 528 (22×24) test instances per day and 3,696 instances per week. In addition, we separate the training data from the test data by time, guaranteeing they have no overlap. Moreover, we apply the model trained in Beijing to the other cities, further verifying its effectiveness and adaptability to different cities.

5.3 Results

Evaluation on Features: We first justify the effectiveness of the features, using the data shown in Table 5, where a DT model is employed to study the performance of individual features and their combinations. Clearly, adding one feature set into the model brings a significant improvement on both precision and recall.

Table 5. Results related to features

	PM10		NO2	
Features	**Precision**	**Recall**	**Precision**	**Recall**
F_m	0.572	0.514	0.477	0.454
F_t	0.341	0.36	0.371	0.35
F_h	0.327	0.364	0.411	0.483
F_p+F_r	0.441	0.443	0.307	0.354
F_m+F_t	0.664	0.675	0.634	0.635
F_m+F_t+F_p+F_r	0.731	0.734	0.701	0.691
F_m+F_t+F_p+F_r+F_h	0.773	0.754	0.723	0.704

Overall Results: Figure 15 shows the performance of U-Air and the five aforementioned baselines, where U-Air outperforms other methods in terms of the mean precision and mean recall over time of day. The results demonstrate the advantage of our method over linear and Gaussian interpolations, and classical air pollutant dispersion models (though the latter may have a better performance with all the parameters accurately obtained, getting such parameters could be even difficult than building the model).

Additionally, simply using some supervised machine learning models (like DT and CRF) or ANN is less effective than the co-training-based approach.

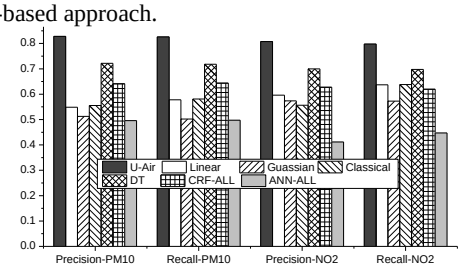


Figure 15. Overall results of different methods for PM_{10} and NO_2

Results of Co-Training: Figure 16 further reveals the co-training progress of our approach, where we add an instance into the training data if *SC* or *TC* predicts it as a class with a probability score over 0.85 (i.e., very confidently inferred). The unlabeled data gradually improves the inference performance, justifying the ability of the co-training framework in dealing with data sparsity.

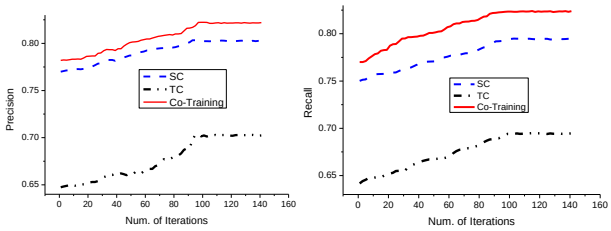


Figure 16. Learning progress of Co-training

Table 6 shows the confusion matrix of U-Air in inferring PM_{10} in Beijing (we do not show other pollutants given the limited spaces).

Table 6. Confusion matrix of U-Air on PM_{10}

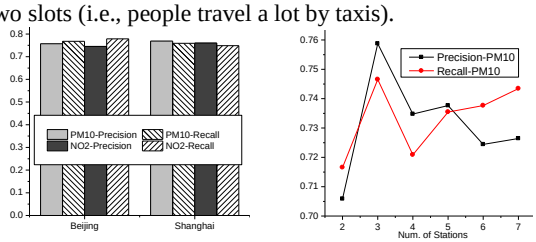
Ground Truth	Predictions					
	G	M	S	U		
G	3789	402	102	0	0.883	Recall
M	602	3614	204	0	0.818	
S	41	200	532	50	0.646	
U	0	22	70	219	0.704	
	0.855	0.853	0.586	0.814	0.828	
	Precision					

Results of SC: To further study the ability of our approach in differentiating between more AQI labels, we solely test the spatial classifier (no traffic-related features needed). Note that this is the result of SC rather than co-training. We use a half of the data for training and the rest for testing, ensuring both parts of data have a relatively balanced distribution over different AQI labels. Unbalanced data will result in impractically high accuracy in the test. We also apply the spatial classifier (trained based on Beijing data) to Shanghai data. As depicted in Figure 17 A), our SC has almost the same performance as that of Beijing, justifying its ability adapting to different urban environments. As shown in Figure 17 B), pairing a location with three stations in the SC generates a better result than using other number of stations (e.g., 2 or 4).

Table 7 Confusion matrix of the Spatial Classifier

Ground Truth	Predictions						
	G	M	U-S	U	V-U&H		
G	656	141	3	0	0	0.820	Recall
M	81	594	114	11	0	0.743	
U-S	1	90	278	183	23	0.483	
U	0	0	41	488	43	0.853	
V-U & H	0	0	0	2	190	0.989	
	0.889	0.720	0.638	0.713	0.742	0.751	
	Precision						

Results of TC: Figure 18 presents the performance of our TC respectively using temporally-related features and all features to infer PM_{10} in Beijing, showing two discoveries. First, feeding all features into the TC does not help, in most cases, becoming even worse than only using temporal features. Second, the performances of two times slots (around 8am and 6pm) are higher than others. The two time slots actually correspond to the morning and evening rush hours of Beijing, in which traffic flows would be the major cause of air pollutants. Another reason is we also have enough number of taxi trajectories representing the traffic flow in the two slots (i.e., people travel a lot by taxis).



A) Results of different cities B) Results w.r.t. number of stations

Figure 17. Study on the spatial classifier

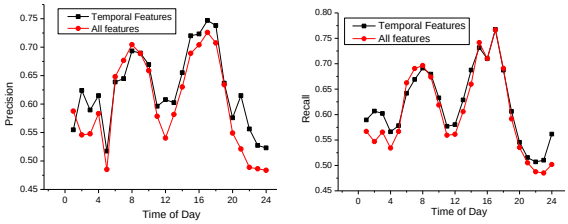


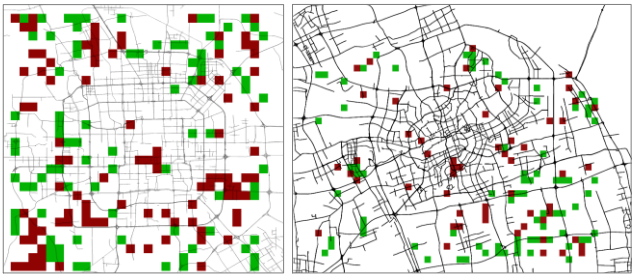
Figure 18. Study on TC using Beijing data (PM_{10})

Efficiency: Table 8 presents the online efficiency of our approach, which was tested on a 64-bit server with a Quad-Core 2.67G CPU and 16GB RAM. On average, we can infer the air quality of a grid in 131ms, generating the AQIs for entire Beijing in 5 minutes.

Table 8. Efficiency study

Procedures		Time(ms)	Procedures		Time(ms)
Feature extraction (per grid)	F_t & F_h	53.2	Inference (per grid)	SC	21.5
	F_p	28.8		TC	13.1
	F_r	14.4	Total		131

Visualization: We infer the AQI of each location in the urban areas of Beijing and Shanghai by using our approach, coming up with two visualizations shown in Figure 19 A) and B), where green and red grids respectively denote the top 100 locations that could have the best and worst AQIs in the two cities during the corresponding periods. The visualization can benefit air pollution analytics by identifying the locations always having a bad AQI.



A) Beijing B) Shanghai

Figure 19. Top 100 locations with the best and worst AQIs

6. RELATED WORK

6.1 Classical Bottom-up Emission Models

There are two major ways calculating the air quality of a location using the emission observed at ground surfaces, called "bottom-up" methods. One is interpolation using the reports from nearby air

quality monitor stations. The method is usually employed by public websites releasing AQIs. As air quality varies in locations non-linearly, the inference accuracy is quite low (see Figure 15).

The other is classical dispersion models, such as Gaussian Plume models, Operational Street Canyon models, and Computational Fluid Dynamics. These models are in most cases a function of meteorology, street geometry, receptor locations, traffic volumes, and emission factors (e.g., g/km per single vehicle), based on a number of empirical assumptions and parameters that might not be applicable to all urban environments [9]. For example, Gaussian Plume model requires vehicle emission rates (e.g., g/km per hour) as input and assumes that the concentration is dispersed in the vertical and horizontal directions in a Gaussian manner. Some models may even require the height, length, and orientation of a street canyon, the gaps between buildings, as well as the roughness coefficient of the urban surface. As these parameters are difficult to obtain precisely, the results generated by such kinds of models may not be very accurate either. Compared with these models, our approach does not need empirical assumptions and parameters. Therefore, it is easy to conduct and applicable to different city environments.

6.2 Satellite Remote Sensing

Satellite remote sensing of surface air quality has been studied intensively in past decades [7], which can be regarded as top-down methods. For example, [1] compared $PM_{2.5}$ inferred from the moderate resolution imaging spectroradiameter (MODIS) with surface $PM_{2.5}$ measurements in Canada and the United States. Likewise, [6] estimated surface NO_2 concentrations by applying local scaling factors from a global three-dimensional model to tropospheric NO_2 columns retrieved from the Ozone Monitoring Instrument onboard the Aura satellite. However, this category of methods is extremely influenced by clouds and would be sensitive to other factors, such as humidity, temperature, and location [1]. In addition, the results inferred from Satellite images only reflect the air quality of atmosphere rather than the ground air quality that people care more about.

6.3 Crowd Sensing

Crowdsourcing or participatory sensing [2][3] may be a potential solution solving this problem in the future, if every person can carry a gas-sensor-equipped smart phone to probe the air quality around them. While this approach is feasible for some gasses like CO_2, it is not practical for other air pollutants like $PM_{2.5}$ and NO_2 so far as the devices for sensing such kinds of air pollutants are not easily portable (refer to Figure 1). In addition, the devices need a long period of sensing time (e.g., 1 hour) before generating accurate results.

6.4 Urban Computing

A series of research on urban computing has been done recently, using big data to tackle the big challenges in big cities. For instance, Jing et al. inferred the functional regions in a city using human mobility data and POIs [11]. Zheng et al. detected the underlying problems in a city's transportation network using taxi trajectories [14]. Zhang et al. sense the urban refueling behavior based on GPS-equipped vehicles [15]. The research reported in this paper is also a step towards urban computing.

7. CONCLUSION

In this paper, from the perspective of big data, we infer the fine-granularity air quality in a city based on the AQIs reported by a few air quality monitor stations and four datasets (meteorological data, taxi trajectories, road networks, and POIs) observed in the city. We identify five sets of features (F_m, F_t, F_h, F_r, and F_p) based on the datasets and propose a co-training-based semi-supervised learning approach consisting of a spatial classifier and a temporal classifier. We first evaluated our co-training-based approach using the data obtained in Beijing, resulting in an overall (Precision=0.828, Recall=0.826) for PM_{10} and (Precision=0.808, Recall=0.798) for NO_2. The results outperform that of linear interpolation, a classical dispersion model, and some well-known supervised learning models like Decision Tree and CRF. Solely applying all features to the *SC* or *TC* is worse than our co-training-based approach. We applied the *SC* learnt from Beijing data to Shanghai, obtaining a result as good as that generated in Beijing (about 0.76). These results demonstrate our approach is applicable to different city environments and seasons.

The key experiences we learned from the research lies in three aspects. First, features should be selected carefully from the data and used properly in the inference models. Second, the design of *SC* and *TC* is helpful as they respectively model the temporal dependency of air quality in a location and the spatial correlation between locations. Third, the co-training-based framework does a good job of addressing the data sparsity problem by leveraging the unlabeled data and the mutual reinforcement relationship between the two feature sets (e.g., POIs and road networks are good complementary of traffic-related features). In the future, we would like to apply our approach to more cities and study the root causes of air pollution.

8. REFERENCES

[1] A. V. Donkelaar, R. V. Martin, and R. J. Park (2006), Estimating ground-level PM2.5 using aerosol optical depth determined from satellite remote sensing, J. Geophys. Res., 111, D21201.

[2] D. Hasenfratz, O. Saukh, S. Sturzenegger, and L. Thiele. Participatory Air Pollution Monitoring Using Smartphones. In the 2nd International Workshop on Mobile Sensing.

[3] Y. Jiang, K. Li, L. Tian, R. Piedrahita, X. Yun, O. Mansata, Q. Lv, R. P. Dick, M. Hannigan, and L. Shang. Maqs: A personalized mobile sensing system for indoor air quality. In Proc. of UbiComp 2011.

[4] J. Lafferty, A. McCallum, F. Pereira (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proc. of 18th International Conf. on Machine Learning.

[5] S. Ma, Y. Zheng, O. Wolfson. T-Share: A Large-Scale Dynamic Taxi Ridesharing Service. In Proc. of ICDE 2013.

[6] L. N. Lamsal, R. V. Martin, A. V. Donkelaar, M. Steinbacher, E. A. Celarier, E. Bucsela, E. J. Dunlea, and J. P. Pinto (2008), Ground-level nitrogen dioxide concentrations inferred from the satellite-borne Ozone Monitoring Instrument, J. Geophys. Res., 113, D1630.

[7] R. V. Martin. Satellite remote sensing of surface air quality, Atmospheric Environment (2008), doi:10.1016.

[8] K. Nigam, R. Ghani. Analyzing the Effectiveness and Applicability of Co-Training. In Proc. of CIKM 2000.

[9] S. Vardoulakis, B. E. A. Fisher, K. Pericleous, N. Gonzalez-Flesca. Modelling air quality in street canyons: a review. Atmospheric Environment 37 (2003) 155-182.

[10] J.S. Scire, D.G. Strimaitis and R.J. Yamartino, 2000b: User's Guide for the CALPUFF Dispersion Model, (Version 5.0), Earth Tech, Inc.

[11] J. Yuan, Y. Zheng, X. Xie. Discovering regions of different functions in a city using human mobility and POIs. In Proc. of KDD 2012.

[12] J. Yuan, Y. Zheng, C. Zhang, X. Xie, G. Sun. An Interactive-Voting based Map Matching Algorithm. In Proc. of MDM 2010.

[13] J. Yuan, Y. Zheng, X. Xie, G. Sun. Driving with Knowledge from the Physical World. In Proc. of KDD 2011.

[14] Y. Zheng, Y. Liu, J. Yuan, X. Xie. Urban Computing with Taxicabs. In Proc. of UbiComp 2011.

[15] F. Zhang, D. Wilkie, Y. Zheng, X. Xie. Sensing the Pulse of Urban Refueling Behavior. In Proc. of UbiComp 2013.