Inferring Air Quality for Station Location Recommendation Based on Urban Big Data

Hsun-Ping Hsieh
Graduate Institute of Networking and Multimedia,
National Taiwan University
d98944006@csie.ntu.edu.tw

Shou-De Lin
Dept. of Computer Science and Information Engineering,
National Taiwan University
sdlin@csie.ntu.edu.tw

Yu Zheng
Urban Computing Team
Microsoft Research
Beijing, China
yuzheng@microsoft.com

ABSTRACT
This paper tries to answer two questions. First, how to infer real-time air quality of any arbitrary location given environmental data and data from very sparse monitoring locations. Second, if one needs to establish few new monitor stations to improve the inference quality, how to determine the best locations for such purpose? The problems are challenging since for most of the locations (>99%) we do not have any air-quality data to train a model from. We design a semi-supervised inference model utilizing existing monitoring data together with heterogeneous city dynamics, including meteorology, human mobility, structure of road networks, and point of interests (POIs). We also propose an entropy-minimization model to suggest the best locations to establish new monitoring stations. We evaluate the proposed approach using Beijing air quality data, resulting in clear advantages over a series of state-of-the-art and commonly used methods.

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, sensor placement, location recommendation, monitoring station, semi-supervised inference

1. INTRODUCTION
In recent years, people are increasingly concerned with urban air qualities, such as the concentration of NO_2, $PM_{2.5}$, and PM_{10}. Government agency has defined the *Air Quality Index* (AQI) to communicate to the public the pollution levels of the air[1]. To measure AQI values, accurate air-quality monitoring stations are required. Unfortunately it is usually not feasible to establish such stations in many places. An air quality monitor station occupies a good amount of space, with non-trivial cost (about 200k USD for construction and 30k USD per year for maintenance) and labor efforts. Solutions based on crowd sourcing and participatory sensing (e.g., using sensor-equipped mobile phones) might not be reliable, as only a very limited number of gas like CO_2 are detectable by those equipment, and is not widely 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 portable and usually need a relatively long sensing period (e.g. 1~2 hours) before accurate measurement can be produced.

Given the fact that it is costly to establish an air-quality monitoring station, highly demanded is a model that can recommend a suitable location for building such stations. In this paper, we especially want to answer a practical question: Given a set of existing air monitoring stations, where to establish the next ones? This task is challenging in several aspects. Let's elaborate them by first investigate some plausible solutions:

(1) An immediate thought would be to establish the stations to maximize the coverage area. Such proposal makes more sense when the air-quality values are smooth. However, this might not be an ideal solution as research has shown that air-quality values are affected by many factors such as weather, traffic, and land usage, which leads to geographically non-smooth values, thus can hardly be captured through the interpolation-based approaches. For example, Figure 1 illustrates a real snapshot data of Beijing's air quality monitoring results. We can find that the AQI (i.e. green represents best, yellow is medium, and red is worst) shown on this map is not smooth. For instance, although stations S_{12} and S_{14} are close to each other, their $PM_{2.5}$ diverse a lot. We observe that over 35% of the monitored time, data of S_{12} and S_{14} have deviations higher than 80 from the period 8/24/2012 to 5/2/2013. It might be caused by the fact that S_{12} is located in a business area with dense buildings and heavy traffic, while S_{14} is located at a scenic spot with a lake nearby. Thus, from the coverage point of view, establishing two near-by stations at S_{12} and S_{14} is not a good choice. However, it might not be a bad idea since S_{14}'s values differ from those of S_{12} significantly.

(2) Another reasonable solution is to choose locations whose air-quality values are harder to be inferred based on data from the existing monitoring stations. To obtain the suitable unobserved locations are the ones with larger margin of inference error, we need not only (a) a certain technique to accurately infer the air quality values of the unobserved locations but also (b) the ground truth data of all the unobserved locations. Requirement (b) is not realistic since we cannot obtain the true air-quality values of a location without a monitoring station.

(3) A third proposal is to establish new stations at locations such that by doing so the inference capability of a given model can be boosted significantly. It seems to be a reasonable and practical idea since, after all, the establishment of monitoring stations is very sparse and we want to accurately infer the AQ values of the unobserved locations given the observed ones. However, it is hard to know 'observation in which locations can best boost the inference accuracy of other locations', since we do not really have any observation data about the candidate locations.

As the above proposals all have their own limitations, we feel that (3) brings the most benefits since it not only reveals the AQI of the

[1] In this paper, AQI is considered as a real value to indicate the quality of air, and focus on only $PM_{2.5}$ and PM_{10}.

KDD'15, August 10–13, 2015, Sydney, Australia.

new observation spots but also boost the inference accuracy of other unobserved locations, although it is very difficult to identify such spots given only monitoring data of few stations that are currently available (e.g. 22 in Beijing). To approximately achieve (3), we propose a two-stage framework. In the first stage, we try to create an AQI inference mechanism that not only can infer the AQI values of any arbitrary unobserved location but also reveal the confidence of its inference. Then in the 2nd stage we propose to establish new stations at the locations that can minimize the uncertainty of the inference model. Based on our proposal, by adding the observed values of these new stations, the inference model is less uncertain about the inferred values of the remaining locations without monitoring stations. The intuition is that although we cannot directly measure the improvement on inference accuracy due to lack of data, as long as we can minimize the uncertainty of a relatively accurate model, it is more likely to yield more accurate results. Our experiments confirm that minimizing the uncertainty of the model leads to a significant improvement on the inference quality.

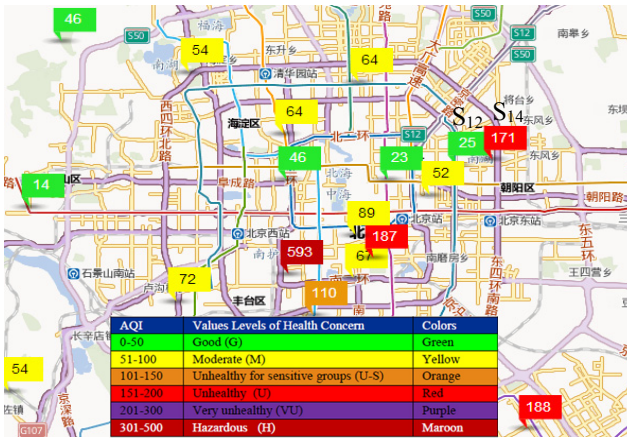

Figure 1. Our problem and application scenarios

Thus, we are required to solve two tasks. First, it is necessary to design an accurate inference model that can infer the AQI of the unobserved locations. Second, we need some mechanism to determine a set of locations that, assuming their AQI values are known, can significantly reduce the uncertainty of the inference model. Both tasks are nontrivial and each has its own challenges.

One typical solution for the first task is through interpolation. However, as have been discussed earlier it might not be an ideal solution since air quality values are not smooth at all. Another plausible solution is to exploit a supervised learning approach to build a regression model for inference. Unfortunately it works poorly for missing records, and for locations that have no historical values previously. Finally, it is known that air quality is affected by many factors such as traffic, land usage and weather. A model that can exploit such information is desirable. Based on the above observations, this paper proposes a semi-supervised learning framework to infer the air quality values of arbitrary unobserved locations in a city. The proposed framework assumes the observed locations are very sparse (e.g. covers less than 1% in Beijing), and takes the aforementioned factors into consideration to infer the unobserved AQI values.

Our 2nd task is to determine a set of k locations, which with their observed data, can reduce the uncertainty of the inference by the biggest margin. To handle this task, we first need a mechanism to determine the uncertainty of the model developed in the previous task. Then we need a method to predict how much uncertainty can be reduced given a new set of locations hat are previously unobserved. Finally, we need an efficient search mechanism to find these k locations that can maximize their effect. To achieve this goal, we design a greedy-based entropy-minimization (GEM) framework. The central idea of GEM lies in that we consider the entropy of the AQI distribution of every inferred location as the approximation of the model uncertainty for the location. Then an interactive process is designed to gradually determine a set of locations that jointly have better potential to reduce the entropy of the inferred locations. We show the overall framework in Figure 2.

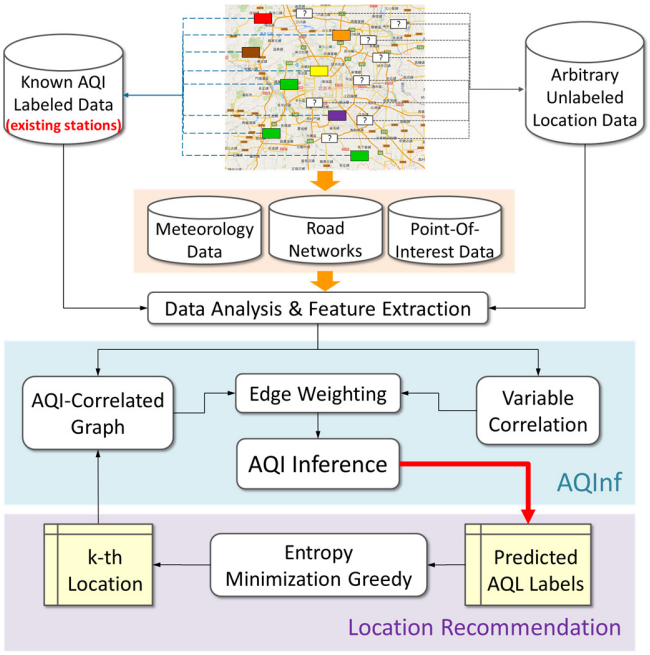

Figure 2. The proposed framework.

The last challenge we are facing lies in the difficulty of performing an evaluation on the proposed model. Due to limited amount of ground truth available, we divide them into three parts. The first part is used as observed data for inference, the second part is treated as the candidate stations (unobserved) to be recommended, while the third part is used to evaluate adding which candidates can best improve the inference quality. The results show that our model can significantly outperform the competitors in both inference and recommendation tasks.

2. DATA AND FEATURES

We utilize real datasets collected from Beijing air quality monitoring stations. The datasets consist of four parts, as elaborated in the following. The statistics are shown in Table 1.

(a) **Air Quality Records.** The data contains the real-valued AQI of two kinds of pollutants, $PM_{2.5}$ and PM_{10}, measured by ground-based air quality monitor stations every hour. Note that there are missing values in the data, while in our model they are treated as unobserved instance to infer.

(b) **Meteorological Data.** Previous study has shown that the concentration of air pollutants is influenced by meteorology [18]. For instance, high wind speed disperses the concentration of $PM_{2.5}$, high humidity usually results in high concentration of pollutants, and high pressure generally would result in a better AQI, etc. Accordingly, we identify five features, temperature, humidity, barometer pressure, wind speed, and weather condition (categorized as cloudy, foggy, rainy, sunny, and snowy). The list of fine-grained meteorological data is collected hourly

from a public website, http://aqicn.org/. For the locations without observation stations, we use Google Weather to capture the meteorological data given latitude and longitude.

(c) **Point-Of-Interests (POIs).** The category of POIs and their density in a region indicate the land usage and the function of the region, which has high correlation to the air quality of the region (e.g. poor air quality might be associated with locations with many factories). We extract 12 POI features using a POI database from Microsoft Bing Maps of Beijing (Table 2).

(d) **Road Networks.** The road network data is collected using Microsoft Bing Maps. It is known that air quality is strongly affected by the traffic condition. We exploit the structure of a road network since it has strong correlation with the real traffic condition. Three features are identified for each grid: (1) total length of highways, (2) total length of other (low-level) road segments, and (3) the number of intersections in the grid's affecting region.

Table 1. Statistics of the Beijing data.

Data Sources		Statistics
POI	2012 Q3	272,109
Road Network	# of road segments	162,246
	Highways length	1,497km
	Roads	18,525km
	# of intersections	49,981
AQI	# of stations	22
	# of hours	10416*22
	Time spans	8/24/2012 – 10/31/2013
Urban Size (grids)		1km (2500)
Meteorological Data	# of hours	10416*2500

Table 2. The list of types of POIs in this paper.

T1: Vehicle Services (gas stations, repair)	T7: Sports
T2: Transportation spots	T8: Parks
T3: Factories	T9: Culture and education
T4: Decoration and furniture markets	T10: Entertainment
T5: Food and beverage	T11: Companies
T6: Shopping malls and supermarkets	T12: Hotels and real estates

3. INFERRING ARBITRARY AQI VALUES

First, we divide geo-spatial area into disjointed grids, which becomes the basic unit or instance in our inference. Each grid, denoted by r, is a 1km*1km sub-area, with its own geographical coordination. Each grid is associated with an AQI value, of which some need to be inferred. Here we define a set of grids $R=\{r_1, r_2, ..., r_m\}$ in an area over a certain time intervals (in hours) $T=\{t_1, t_2, ..., t_n\}$. The AQI values of most grids are completely unknown while the historical AQI values of a small amount of grids can be obtained through existing monitoring stations (e.g. for Beijing city's data, only 0.88% of the locations are monitored). The meteorology, road network, and POI information of each grid are assumed to be available. The goal is to infer the *AQI distribution* $P(v(t_i))$ of any unobserved location v at any given time stamp t_i.

We design a semi-supervised learning algorithm to achieve such goal. The proposed algorithm consists of four stages. In the first stage a spatial-temporal graph, the *AQI Affinity Graph* (AG), is constructed to model the spatial-temporal correlation between grids. In the second stage we try to learn the weights of the edges, assuming they represent the correlations between nodes based on their features. The third stage emphasizes on inferring the AQI values for each location, which presumes those grids whose features are close to each other tend to share similar AQI values. In the final stage the feature weights are adjusted to minimize the uncertainty of the model on inferring the unobserved locations. Note that stages 3 and 4 are executed iteratively until convergence.

3.1 AQI Affinity Graph

The air quality values of different locations are correlated with each other in temporal and spatial perspectives. For example, the AQI of a location tends to be good if the AQI of the past few hours are also good; the AQI of a location is likely to be bad if the air quality of many of the neighboring places is bad. Inspired by above observations, we propose to create a graph to model such spatial-temporal correlations over locations. We first divide an urban area into disjointed grids (e.g. 1km*1km), and then construct an AQI *Affinity Graph* (AG) to model the correlation of different grids. The AG is designed to be a 3-dimensional weighted connected graph, in which only few nodes possess known AQI information (i.e., those places established with monitoring stations) while other nodes have no AQI record. The spatial correlation is reflected by both the geographical distance and the demographic spatial features. The temporal correlation is reflected by connecting nodes that represent different time stamps. We first describe how to construct the graph, and then show how the weights in the graph can be generated.

Definition: Affinity Graph. An affinity graph (See Figure 3) is a multi-layer weighted connected graph $G = \langle \mathcal{G}^1, \mathcal{G}^2, \dots, \mathcal{G}^n \rangle$, where each layer represents one time interval from $t_1, t_2, \dots, to\ t_n$, and $\mathcal{G}^i = \langle V, E, W^{t_i} \rangle$ is the graph at time interval t_i, where V is the set of all grids, E is the set of edges, and W^{t_i} represent edge weights at t_i. The first two dimensions in AG connect the geographical neighbors of grids, while the 3rd dimension connects nodes across temporal domain. Every single node in AG can be regarded as a random variable whose AQI distribution has to be inferred. The node set $V = U \cup \boldsymbol{\mathcal{V}}$ consists of a subset U of query grids without air quality monitoring stations, and the subset $\boldsymbol{\mathcal{V}}$ of grids with monitoring stations. We term nodes in $\boldsymbol{\mathcal{V}}$ as labeled ones and nodes in U as unlabeled ones. Each unlabeled node u is associated with an AQI distribution $P(u)$ to be inferred. The construction of an AG consists of four parts.

(a) **Connecting to Station Locations.** Because we aim to leverage few grids with existing stations to infer the temporal AQI values of arbitrary locations, we connect every unobserved node $u \in U$ to all the observed nodes $v \in \boldsymbol{\mathcal{V}}$ of the same time stamp, regardless of their geographical distance (i.e. the lines connecting white and black nodes in Figure 3). Note that since the observed nodes are very sparse, adding those connections does not affect the efficiency significantly.

(b) **Connecting to Near-by Locations.** Since the AQI values of near-by locations are naturally highly correlated, within each layer of graph, every node is connected to the neighboring nodes $w \in U$ within a given geographical radius r.

(c) **Connecting to Recent Layers.** Due to the fact that the AQI value of a location is highly correlated to its historical AQI values, we connect each node $u \in U$ of time stamp t_i to the previous z corresponding nodes of the same location. That is, the nodes of the same grid but with different time stamps: t_{i-1}, t_{i-2}, ..., and t_{i-z} are connected (i.e. the blue line in Figure 3).

(d) **Connecting to Similar Layers.** Since the environmental factors can repeat themselves within certain period (e.g. some phenomenon are observed every 24 hours while some are observed during a specific season of the year), it is also possible that the AQI value of a node correlates with that further away from the current time stamp. Our idea is to connect a node in the current layer to the corresponding nodes of certain past layers with the most similar environmental features. The similarity between layers is computed based on the features. See the red lines in Figure 3 as an example.

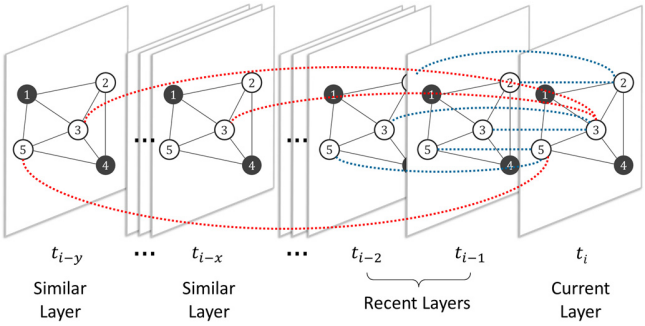

Figure 3. An example of the *affinity graph*. There are five location nodes, in which two contains measurement stations (in black) and three needs to be inferred (in white). We construct i layer graphs capturing temporal correlation, while the connections are identical for each layer. The temporal correlations are modeled by connections across layers such as the red and blue lines.

Next, it is required to learn the correlation between nodes as edge weights in the AQI affinity graph. The concept is that if two nodes have higher feature affinity, their AQI values should have higher correlation. The heterogeneous features are exploited to characterize the affinity between nodes. Various features might have different degree of effect on the correlation. Therefore, we propose to learn such effect separately from data. Finally, we combine affinity functions of all the features through a weighted sum, while the weights can further be adjusted later on to reduce the inference uncertainty.

Affinity Function. Given a particular type of feature $\boldsymbol{f_k}$, its *affinity* value $a_{\boldsymbol{f_k}}(u,v)$ between nodes u and v, $(u,v) \in E$ can be derived from the affinity function $AF_{\boldsymbol{f_k}}(\Delta \boldsymbol{f}_k(u,v))$, where $\Delta \boldsymbol{f}_k = \|\boldsymbol{f}_k(u) - \boldsymbol{f}_k(v)\|$. $AF_{\boldsymbol{f_k}}$ is a *linear* function $a \cdot \Delta f_k(u,v) + b$ to model the correlation between feature difference and AQI similarity, where the parameters a and b are learned using from Maximum Likelihood Estimation [1].

Combined Affinity Function. Given a set of features $F = \{f_1, f_2, \dots, f_m\}$, the combined affinity $a(u,v)$ between nodes u and v, $(u,v) \in E$can be derived based on the weighted sum of AF:

$$a(u,v) = \exp\left(-\sum_{k=1}^{m} \pi_k^2 \times AF_{\boldsymbol{f_k}}(\Delta \boldsymbol{f}_k(u,v))\right) \quad (1)$$

where π_k is the weight of feature f_k. Note that the parameters a and π_k seem to have the same linear effect on the model, they are not combined as a single parameter since each is learnt with a different purpose; a is learnt to capture the correlation with the AQI and, as will be described later, π_k is learnt to minimize the uncertainty of the model.

3.2 Affinity-based AQI Inference

Based on the affinity graph and the affinity function learned from data, we propose the *affinity-based AQI inference* (AQInf) *model*, which is a graph-based semi-supervised learning solution. The fundamental idea is three-fold. First, the observed AQI on labeled nodes $v \in \boldsymbol{\mathcal{V}}$ are utilized to infer the AQI distributions $P(u)$ of unlabeled nodes $u \in U$. Second, we assume that nodes with similar features should have similar AQI distributions. This relationship is modeled by edge weights through the combined affinity function. Third, since the AQI values for the unobserved locations are not available and observed data are sparse, it is less practical to tune our model parameters to minimize the inference error. Instead we propose to tune the parameters to minimize the model uncertainty. By putting these three ideas together, we seek for an optimal set of edge weights W such that a) after inference, the unlabeled nodes shall possess similar AQI distributions with its close neighbors and b) the learned label distribution should possess small entropy to minimize the uncertainty of inference. Figure 4 describes our iterative learning framework.

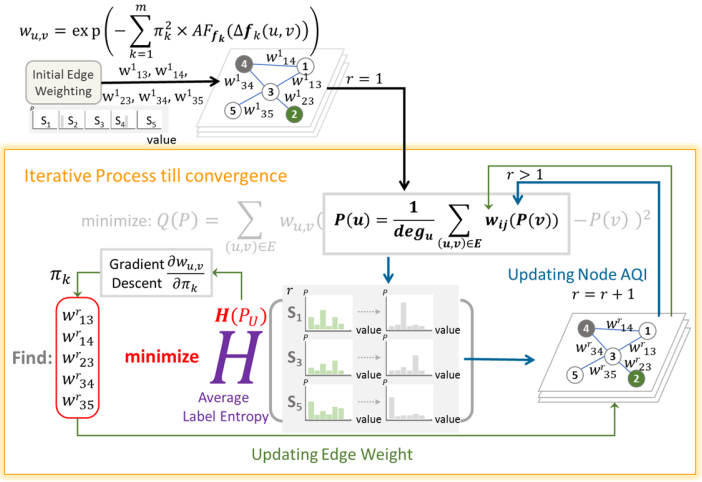

Figure 4. The process of inferring AQI values.

To realize a), we propose to optimize a loss function on the affinity graph to enforce the label distributions to be propagated to nodes with higher edge weights:

$$Q(P) = \sum_{(u,v)\in E} w_{u,v} \cdot (P(u) - P(v))^2 \quad (2)$$

using the quadratic loss function. We exploit the *symmetric Kullback-Leibler* (KL) *Divergence* to measure the difference between two AQI distributions, i.e.,

$$P(u) - P(v) = D_{KL}(P(u)||P(v)) + D_{KL}(P(v)||P(u)) \quad (3)$$

where $D_{KL}(P(u)||P(v)) = \int_{x=0}^{q_{max}} P(u)[x] \ln \frac{P(u)[x]}{P(v)[x]} dx$

where q_{max} is the maximum AQI value among locations. Putting these together, our goal is to find the AQI distributions for unlabeled nodes such that $Q(P)$ is minimized, expressed by:

$$P = argmin_{P|\boldsymbol{\mathcal{V}}}\, Q(P) \quad (4)$$

It is not hard to show that such minimization is *harmonic* [21], which refers to $\Delta P = 0$ for unlabeled nodes U, while $\Delta P = P(v)$ for labeled nodes $v \in \boldsymbol{\mathcal{V}}$. Δ is known as the *combinatorial graph Laplacian* matrix defined as $\Delta = D - W$, where the matrix D is a diagonal matrix whose diagonal elements are given by $D_{ii} = \sum_j W_{ij}$ and $W = [w_{ij}]$ is the weight matrix of affinity graph. The underlying rationale is that minimizing $Q(P)$ drives P to leverage the AQI distributions of labeled nodes $\boldsymbol{\mathcal{V}}$ and propagates smoothly on unlabeled nodes U based on the weight matrix W. The harmonic property of the function P derives the solution to assign the AQI distribution of each unlabeled node using the weighted average of its neighboring nodes via:

$$P(u)[x] = \frac{1}{deg_u} \sum_{(u,v)\in E} w_{u,v} \cdot P(v)[x], \quad x = 0, 1, 2, \dots, q_{max} \quad (5)$$

which corresponds to the smooth propagation of P in affinity graph, where deg_u is the degree of node u. We can further obtain the harmonic solution $\Delta P = 0$ in terms of matrix operations for unlabeled nodes subject to the AQI distributions of labeled nodes, expressed by:

$$P_U = (D_{UU} - W_{UU})^{-1} W_{U\boldsymbol{\mathcal{V}}} P_{\boldsymbol{\mathcal{V}}} = -\Delta_{UU}^{-1} \Delta_{U\boldsymbol{\mathcal{V}}} P_{\boldsymbol{\mathcal{V}}} \quad (6)$$

where P_U is the AQI distributions of unlabeled nodes, $P = [P_{\boldsymbol{\mathcal{V}}}; P_U]$, and both the weight matrix W and the combinatorial graph Laplacian matrix Δ can be split into labeled and unlabeled parts, given by

$$W = \begin{bmatrix} W_{\boldsymbol{\mathcal{V}}\boldsymbol{\mathcal{V}}} & W_{\boldsymbol{\mathcal{V}}U} \\ W_{U\boldsymbol{\mathcal{V}}} & W_{UU} \end{bmatrix} \text{ and } \Delta = \begin{bmatrix} \Delta_{\boldsymbol{\mathcal{V}}\boldsymbol{\mathcal{V}}} & \Delta_{\boldsymbol{\mathcal{V}}U} \\ \Delta_{U\boldsymbol{\mathcal{V}}} & \Delta_{UU} \end{bmatrix}.$$

The results obtained from the above equation is the *soft* labeling of AQI values because the derived $P(u)$ for unlabeled node $u \in U$ is a probability distribution of AQI values, and thus does not provide exact AQI values for unlabeled nodes. To have a *hard* labeling of AQI for unlabeled node $u \in U$, we find the AQI value q^* with the highest quantized probability from its AQI distribution and assign q^* as its final predicted AQI value, given by

$$q^*(u) = argmax_x P(u)[x], \;\; x = 0, 1, 2, \dots, q_{max} \quad (7)$$

Note that since an unlabeled location $u \in U$ has n node instances $u(t_1)$, $u(t_2)$, ..., $u(t_n)$ in the affinity graph over time intervals t_1, t_2, ..., t_n, and the edge weights (i.e., affinity) vary based on the features of locations within/across time intervals, the predicted AQI, $q^*(u(t_i))$, shall be different given different time stamp.

So far we have elaborated the central idea of the affinity-based AQI inference (AQInf) model, i.e., minimizing $Q(P)$ based on the given edge weights $w_{u,v}$. Recall that in Section 3.1, edge weights between nodes are determined by the affinity values between locations, which are obtained through the weighted sum over the location affinity functions of features f_k with feature weights $\{\pi_k\}$. That says, $\{\pi_k\}$ first influences the affinity between locations, i.e. edge weights, and consequently the weights take effect on the inference of the unlabeled nodes. Thus, learning a suitable set $\{\pi_k\}$ becomes the key to the success of the inference.

The intuitive approach is to adjust $\{\pi_k\}$ to maximize the likelihood of labeled nodes using validation data. However, this idea is problematic because the observed data is very sparse, and thus doing so would likely overfit the model to the validation data. Here we propose an objective for learning π_k, which is to minimize the entropy of the inferred AQI distribution of unlabeled nodes. It is intuitive since the inference model would become useless if the inferred distribution has high entropy (i.e. unpredictable values).

The average AQI distribution entropy $H(P_U)$ for unlabeled nodes U can be defined as:

$$H(P_U) = \frac{1}{|U|}\left(-\sum_{u \in \mathbb{V}} \int_x \begin{pmatrix} P(u)[x]\log(P(u)[x]) + \\ (1-P(u))[x]\log((1-P(u))[x]) \end{pmatrix} \mathrm{d}x \right) \quad (8)$$

where $|U|$ is the number of unlabeled nodes in the affinity graph. For brevity, we denote $\int_x (P(u)[x]\log(P(u)[x]) + (1 - P(u))[x]\log((1-P(u))[x]))\mathrm{d}x$ as $P(u)\log\big(P(u)\big) + (1 - P(u))\log\big(1 - P(u)\big)$. We want to minimize the uncertainty of the model, which is equivalent to minimizing $H(P_U)$.

To derive $\{\pi_k\}$, the minimization of $H(P_U)$ is embedded into the AQInf model using $w_{u,v} = \exp\big(-\sum_{k=1}^{m} \pi_k^2 \times AF_{\boldsymbol{f_k}}(\Delta \boldsymbol{f}_k(u,v))\big)$ and $P_U = -\Delta_{UU}^{-1}\Delta_{U\boldsymbol{\mathcal{V}}} P_{\boldsymbol{\mathcal{V}}}$. We exploit the technique of gradient descent on π_k to obtain an updated set of feature weights $w_{u,v}$ that minimizes $H(P_U)$. The gradient can be derived by computing $\frac{\partial H(P_U)}{\partial \pi_k}$, given by

$$\frac{\partial H(P_U)}{\partial \pi_k} = \frac{1}{|U|} \sum_{u \in U} \log \frac{1 - P(u)}{P(u)} \frac{\partial P(u)}{\partial \pi_k} \quad (9)$$

Using $w_{u,v} = \exp\big(-\sum_{k=1}^{m} \pi_k^2 \times AF_{\boldsymbol{f_k}}(\Delta \boldsymbol{f}_k(u,v))\big)$, $P(u)[x] = \frac{1}{deg_u}\sum_{(u,v)\in E} w_{u,v} \cdot P(v)[x]$, and $P_U = -\Delta_{UU}^{-1}\Delta_{U\boldsymbol{\mathcal{V}}} P_{\boldsymbol{\mathcal{V}}}$, together with the chain rule of differentiation, we obtain the final gradient as:

$$\frac{\partial w_{u,v}}{\partial \pi_k} = 2 w_{u,v} \cdot AF_{\boldsymbol{f_k}}\big(\Delta \boldsymbol{f}_k(u,v)\big) \cdot \pi_k \quad (10)$$

Such integration produces a *mutually reinforced* inference flow, the learned feature weights π_k update the AQI distributions $P(u)$ of unlabeled nodes $u \in U$, and $P(u)$ determines the average AQI distribution entropy $H(P_U)$ to be minimized in the next iteration. We develop the inference model to follow this mutual reinforcement mechanism, in which each change of feature weights π_k triggers an update of edge weights $w_{u,v}$ that further generates new AQI distribution $P(u)$ based on the affinity graph, and proceeds iteratively till convergence. The pseudocode of AQInf model is described in Algorithm 1.

Algorithm 1: Affinity-based AQI Inference (AQInf)

Input: (a) a set of locations $\boldsymbol{\mathcal{V}}$ with existing measurement stations; (b) a set of query locations U without stations; and (c) the time interval t_i of interest.

Output: the AQI value $q(u)$, where $u \in U$ and $t_i \in T$.

$V \leftarrow \boldsymbol{\mathcal{V}} \cup U$.
$\boldsymbol{f_k}(v) \leftarrow$ extracting feature f_k, where $k = 1, 2, \dots, m$, $v \in V$
Construct affinity graph AG from V and $\boldsymbol{f_k}(v)$, $v \in V$.
Initialize feature weight $\pi_k = 1$, where $k = 1, 2, \dots, m$.
$w_{u,v} \leftarrow \exp\big(-\sum_{k=1}^{m} \pi_k^2 \times AF_{\boldsymbol{f_k}}(\Delta \boldsymbol{f}_k(u,v))\big)$.
$H(P_U) \leftarrow \frac{1}{|U|}\big(-\sum_{u\in U} P(u)\log\big(P(u)\big) + (1-P(u))\log\big(1-P(u)\big)\big)$
$\Delta h \leftarrow H(P_U)$.
while $\Delta h > \epsilon$ **do**:
  $\pi_k \leftarrow \pi_k - 2 w_{u,v} \cdot AF_{\boldsymbol{f_k}}(\Delta \boldsymbol{f}_k(u,v)) \cdot \pi_k$
  $w_{u,v} \leftarrow \exp\big(-\sum_{k=1}^{m} \pi_k^2 \times AF_{\boldsymbol{f_k}}(\Delta \boldsymbol{f}_k(u,v))\big)$
  $P_U = (D_{UU} - W_{UU})^{-1} W_{U\boldsymbol{\mathcal{V}}} P_{\boldsymbol{\mathcal{V}}}$
  $H'(P_U) \leftarrow \frac{1}{|U|}\big(-\sum_{u\in U} P(u)\log\big(P(u)\big) + (1-P(u))\log\big(1-P(u)\big)\big)$
  $\Delta h \leftarrow |H(P_U) - H'(P_U)|$
  $H(P_U) \leftarrow H'(P_U)$
end
$q(u) = argmax_x P(u)[x], \;\; x = 0, 1, 2, \dots, q_{max}$
return: $q(u)$.

4. BUILDING MEASUREMENT STATIONS

With the proposed *AQInf* model, now we can talk about how to use it to recommend locations for building new measurement stations. The ultimate goal is to recommend k locations such that the establishment of new stations can lead to the best improvement of AQI inference of other locations. Unfortunately, such goal cannot directly be achieved since we do not know the exact AQI values for locations without monitoring stations. Instead, we focus on recommending locations that have better potential to reduce the uncertainty of the AQI inference model. Recall previously in our AQInf model, the distribution of each unobserved location is inferred. Thus, we can model the uncertainty of the model as *the sum of the entropies of all unobserved nodes* in Equation 8

4.1 Greedy-based Entropy Minimization

Our goal is to select k locations that by using their AQI values together with those of the known sites, we can construct an update AQInf model whose uncertainty on the remaining locations are minimized. Let us first assume that once the k locations are picked,

their AQI values can be known. Then the overall task becomes an optimization problem. That is, for each of the C_k^N possible combination, we can infer the corresponding uncertainty for each AQInf model, and finally choose the minimum one. Unfortunately, searching among the C_k^N combinations is intractable. One might resort to a greedy method to choose one location first, obtain the updated AQInf model, and then use this updated model to choose the 2nd location, and so on so forth. However, it is not clear how the AQInf model can be updated in the first place because we do not have the observed values for the selected locations. Furthermore, the first selected locations are the ones with high entropy (more uncertain), thus their true AQI values can hardly be inferred.

Alternatively, we might simply pick the top-k locations with the highest entropy as the recommended outcome. The concern for this proposal is that some of these uncertain regions might be highly correlated. That says, once we obtain the values for some locations, the rest might not as unpredictable as before. Our experiments confirm such hypothesis.

We propose a method called greedy-based entropy minimization (GEM) that aims at ranking locations based on their capability to reduce uncertainty. Instead of focusing on the high-entropy locations, we start from the low-entropy ones first. GEM is performed with following steps, which corresponds to the five steps specified in Figure 5.

1. Given the obtained AQInf model, first identify the location X_0 with the *lowest* entropy, meaning that our model is very confident about its inferred AQI value. Rank X_0 the last candidate to be recommended.
2. Choose the most likely value inferred from the original AQInf of X_0 as its 'pseudo observed AQI value'. The mark X_0 as labeled.
3. Use X_0's pseudo AQI value together with the original observed data to build a new influence model, $AQInf_1$.
4. Identify another location X_1 with the lowest entropy in $AQInf_1$, assign it as the second-to-last candidate in the recommendation list.
5. Repeat 1~4 to iteratively rank the locations to be recommended from last to first.

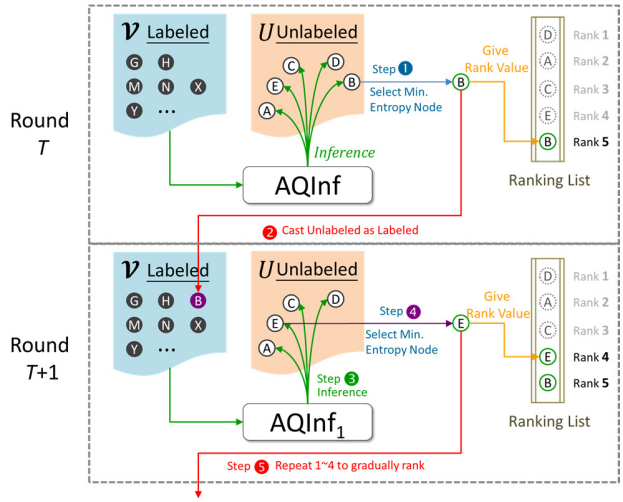

Figure 5. An illustration of GEM at time interval $t_i \in T$.

Note that for each specific time stamp t_i, we perform GEM to obtain a ranked list of unobserved nodes. Since the unpredictability of nodes can vary with time, eventually we would like to average the ranking over certain period of time to obtain the overall ranking. Note that this step is important to boost the performance of GEM. It is because conceptually the nodes with lower rank in GEM are those not very correlated with the ones in the higher rank. We find that for different time stamp, the nodes with the highest entropy might differ. Thus the high ranked nodes might not always be the same in each ranking list across time. This implies that if there are some nodes that are consistently ranked very low (i.e. those with lowest average rank), they are very likely to be independent with many other nodes, and thus should be picked. The pseudocode of the proposed method is described in Algorithm 2.

Algorithm 2: Greedy-based Entropy Minimization (GEM)

Input: (a) a set of locations $\boldsymbol{\mathcal{V}}$ with existing measurement stations; (b) a set of candidate locations C without stations; (c) the time stamps $T = t_1, t_2, \ldots, t_m$; and (d) the number of locations k to be selected for new stations.

Output: the set S of k recommended locations ($|S| = k$).

for each $t_i \in T$ **do:**
  $U \leftarrow C$.
  for $num \leftarrow |C|$ to 1 **do:**
    $H(P_U) \leftarrow AQInf(\boldsymbol{\mathcal{V}}, U, t_i)$.
    $u^* \leftarrow argmin_u H(P(u))$. // select the min entropy one
    $rank(u^*, t_i) \leftarrow num$. // give the rank value reversely
    $\boldsymbol{\mathcal{V}} \leftarrow \boldsymbol{\mathcal{V}} \cup \{u^*\}$. // turn unlabeled to labeled
    $U \leftarrow U \setminus \{u^*\}$. // exclude the turned candidate
  end
end
for each $u \in C$ **do:**
  $rank(u) = (\sum_{t_i \in T} rank(u, t_i))/|T|$.
end
$sort(rank(u \in C))$ in a *descending* order.
for $selected = 0$ to $k - 1$ **do:**
  $S = S \cup rank(C)[selected]$.
end
return: S.

5. EXPERIMENTS

In this section, we introduce two evaluations to verify the performance of our proposed method using the air quality data introduced in Section 2.1, and compare the results with several state-of-art and baseline methods. In the following experiments, we repeat each of them 1000 times to obtain the average results

5.1 The Effectiveness of AQInf

Settings In this experiment, we would like to investigate whether our designed AQInf can accurately infer AQI. In the experiment, the Beijing area is decomposed into 50*50 grids, in which 22 of the grids have the monitoring stations (i.e., have AQI values) while the AQI values of the other 2478 grids are unknown. The experiment period spans from 8/24/2012 to 10/31/2013, containing 10416 time stamps (each represents an hour). Note that not all 10416 Since we only have ground truth for locations with monitoring stations, we propose to conduct cross validation by randomly choosing 15 of the 22 grids to be the labeled data, resulting in 15*10416 labeled instances. The model is evaluated based on the inference accuracy of the leftover 7*10416 instances whose ground truth measurement is available. All the experiments are repeated 1000 times and the average results are reported. For each repetition the training and testing sets are randomly selected.

To evaluate the usefulness of the proposed features, we build 3 slightly different models with incremental feature sets (1) geographical distances features plus three recent and three similar time

layers as features (denoted by D+T3), (2) features in (1) plus the meteorology data (denoted by D+T3+M), (3) features in (2) plus Roadnet and POI features (denoted by ALL). We use Root Mean Square Error (RMSE) to measure the difference between the predicted AQI and ground-truth AQI.

Competitors. Competitors can be divided into three categories. Three interpolation-based methods, two learning-based methods, and two semi-supervised learning methods as the state-of-the-art. All the features proposed in this paper are exploited in the last two categories of models:

- **Spatial *k*NN.** *Spatial k-Nearest Neighbors* (*k*NN) simply generates the AQI value by averaging the values of the labeled data from *k* spatially closest stations.
- **Inverse Distance Weighting (IDW).** It is a well-known interpolation method. The idea of IDW is to assign values of unknown data points using the weighted averages of the values available from the known data points. The weights are proportional to the inverse of the geographical distances.
- **Ordinary Kriging (OK).** The Ordinary Kriging is one of the most used spatial point interpolation. The estimations are weighted averaged input point values, similar to IDW. The main difference with IDW is that it utilizes semi-variogram to express the spatial variation, and minimizes the error of predicted values which are estimated by spatial distribution of the predicted values.
- **ANN.** We choose *Artificial Neural Network* (ANN) with back propagation technique as another baseline. The constructed ANN contains one hidden layer. The ANN method simply treats all historical labeled data from all stations as the training data to build a model.
- **SVR.** The version of SVM for regression is chosen to predict the AQI value. Similar to ANN, SVR utilizes stations' historical data for training and then infer the values of unlabeled ones.
- **Co-training.** The co-training model proposed by U-Air [18] is the state-of-the-art method for air quality inference problem. The co-training model consists of two separated classifiers. One is a spatial classifier based on artificial neural network, which employs the spatially-related features (e.g. POIs) to model the spatial correlation. The other is a temporal classifier based on a linear-chain conditional random field (CRF), which utilizes temporally-related features (e.g. meteorology) to model the temporal dependency.
- **RBF-SSL.** We employ a state-of-the-art graph-based semi-supervised learning (SSL) with a radial basis function (RBF) classifier [21] to predict the AQI labels. All the features proposed in this paper are exploited, and the weights in RBF are estimated by $w_{ij} = \exp(-\sum_{d=1}^{m}(x_{id} - x_{jd})^2/{\sigma_d}^2)$.

Error of Air Quality Inference. The experimental results on inferring PM_{10} and $PM_{2.5}$ are shown in Figure 6, which show that in general our method outperforms all the competitors significantly, especially when all the features are considered. The results also show that each additional feature set brings certain level of improvement (3~ 7.9 for PM_{10} and 2.3 ~ 8.5 for $PM_{2.5}$).

Robustness of Air Quality Inference. In Figure 7, we test how sensitive the models are to the number of observed neighbors. When more spatial neighbors of the labeled instance are removed, the proposed model always outperforms other competitors, up to 12.9~114.8 for PM_{10} and 14.4~106.5 for $PM_{2.5}$ in average. We can further find that the increases of RMSE of the AQInf are relatively slow, comparing to the competitors. Such results suggest that our proposed model is the most survivable when there is only very rare labeled data available. The interpolation-based methods are suffered from labeled data the most.

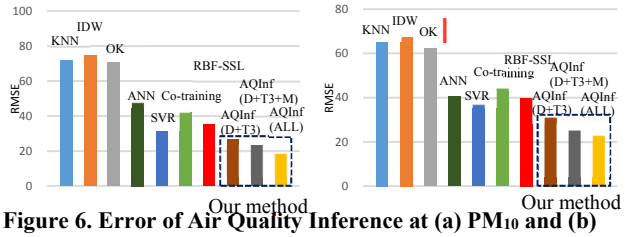

Figure 6. Error of Air Quality Inference at (a) PM_{10} and (b) $PM_{2.5}$ for different algorithms.

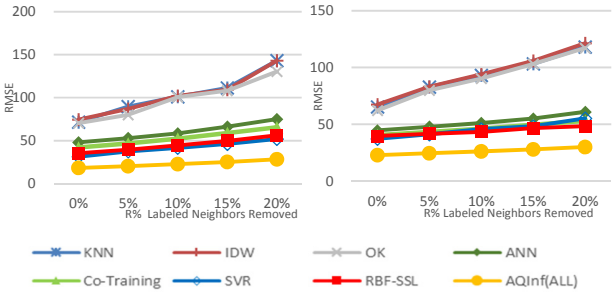

Figure 7. Error of Air Quality Inference at (a) PM_{10} and (b) $PM_{2.5}$ for different algorithms when R% labeled neighbors removed.

5.2 The Effectiveness of GEM

Settings. In this experiment, we would like to verify whether our model recommends locations that bring the most improvement in terms of inference accuracy. Since there are total 2,478 unobserved grids, assuming we want to recommend k=5 locations to establish monitoring stations, we will need to evaluate C_5^{2478} combinations to find the optimal solution. However, we need to point out that there are no ground truths for those locations without monitoring stations established. Since we only have the ground truth AQI for locations with stations, we propose to conduct cross validation by randomly choosing 5 locations among 22 grids to be the labeled data, i.e., pretending there are only 5 locations already with stations, and then we reserve 10 locations to be the candidate locations to build new stations. The rest 7 locations are used for evaluation of prediction quality. To elaborate, we first train the AQInf model using the data of 5 training locations, and utilize the model to infer the air quality of the 7 evaluation locations to obtain the inference error (i.e. RMSE). Now assume we would like to pick k locations from the 10 candidate locations to build new monitoring stations. There will be C_k^{10} combinations to be considered. To generate the ground truth, for each candidate combination, we turn them from unlabeled into labeled ones to form the new labeled set whose size is *5+k*. Then we can use the expanded labeled set to build an AQInf model and infer the AQI of the remaining 7 locations again to obtain the updated RMSE. The combination with highest improved RMSE is best choice to construct the new stations because it can bring the most accuracy improvement of AQI. Based on this strategy, we generate and rank the RMSE improvements for all C_k^{10} combinations, which we call the *ground-truth ranking*. The goal of the experiment then becomes checking whether the combination returned by our method is indeed a combination that ranks high in the *ground-truth ranking*.

Evaluation Metrics. We designed two evaluation metrics, *top-rank ratio TRR* that reflects how the returned combination by each

algorithm ranks in the *ground-truth ranking* and *RMSE-improvement* which shows how much inference accuracy can be improved after establishing the stations in the recommended locations. The TRR of a combination C is determined as, $TRR(C) = \frac{Rank(C)}{C_k^{10}}$, with value range from $1/C_k^{10}$ to 1 (the lower the better).

Competitors. Since to our knowledge, there is currently no study on solving the proposed problem, here we devise several methods as the competitors in our experiments.

- **Distance-based greedy.** This method greedily chooses the farthest location to the existing stations as the target location to build new station. The central idea of this method is to evenly distribute stations in geographical space.
- **Temporal feature-dissimilarity greedy.** This method greedily selects the locations with the most dissimilar meteorology features with the existing stations as the candidates. We exploit the Pearson Correlation to measure the similarity between the candidate location and existing stations considering normalized meteorology features such as temperature, humidly, pressure and wind speed.
- **Spatial feature-dissimilarity greedy.** Same as the previous methods, but replaces temporal features with spatial features.
- **Hybrid feature-dissimilarity greedy.** Similar to the previous methods but considers both spatial and temporal features together. It assumes equal weights for spatial and temporal features.
- **Entropy-based search.** This method simply ranks locations using entropy of the inferred AQI distribution at each time interval $t_i \in T$. Nodes with the smallest k ranking values averaged over intervals $t_i \in T$ will be selected as the final recommended locations. The entropy-based search cannot be performed greedily.
- **Low edge weight search.** For each time stamp, we sort the locations using their average edge weight to the neighbors. Conceivably higher weights means better correlation with others. Eventually we choose the k locations with the smallest total weights, averaging over a span of time T.

Note that since our model creates a ranking of locations for each time stamp, and then average the ranking over all time stamps to obtain the final ranking, the same procedure is applied to all of our competitors.

Evaluating GEM with Top-rank ratio. In this experiment, we compare the TRR of each competitor. We vary the number of recommended locations k and report the TRR value in Figure 8. Note that we only report results for k=1~9 all candidates are selected. The left chart is for PM_{10} and the right one is $PM_{2.5}$. The quality of entropy-based and low edge search decreases sharply when the number of combinations becomes large. It is probably because they do not take into account previously recommended locations while selecting the next location. In summary, we show that our method can steadily outperform other competitors with TRR value less than 0.2.

Evaluating GEM with RMSE-improvement. We also report how much RMSE improvement each model can obtain, with varying parameter k, in Figure 9. Theoretically, RMSE should improve more when k increases, as more locations are added into the training data. We find that our method generally brings much better improvement than others for PM_{10} and $PM_{2.5}$. Note that here we also display the optimal improvement in RMSE (the result of the top-1 combination in the ground truth list) curve as black line in Figure 9. The results show that our model's performance is not too far away from the optimal results.

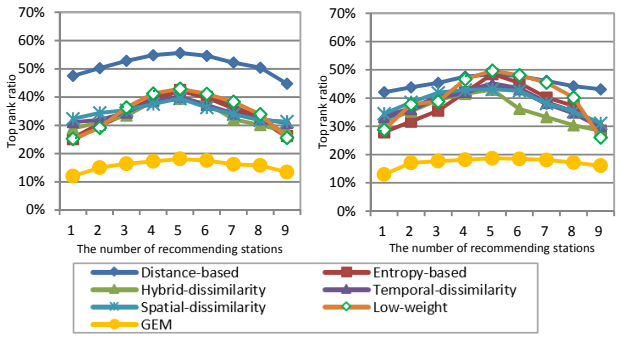

Figure 8. The TRR results for PM_{10} and $PM_{2.5}$ with varying number of recommended locations.

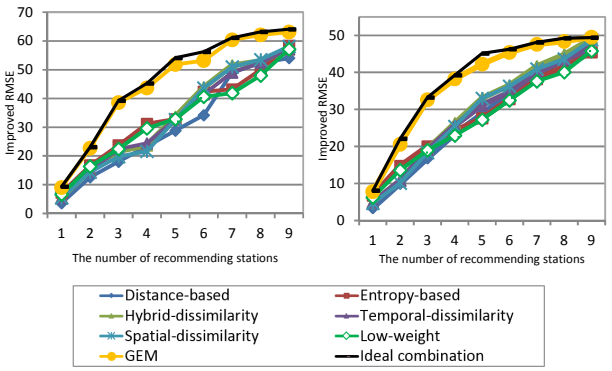

Figure 9. The improved RMSE results for PM_{10} and $PM_{2.5}$ with varying number of recommended locations.

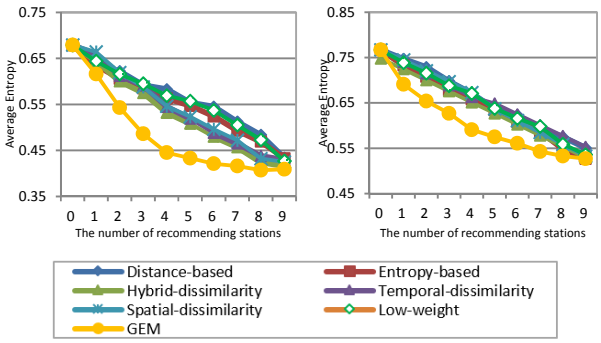

Figure 10. The average entropy for PM_{10} and $PM_{2.5}$ with varying number of recommended locations.

Entropy variation. In Figure 10 we further report the model uncertainty in entropy $H(P_U)$ as the number of recommended location k increases. We can observe that the average entropy in the affinity graph can effectively be reduced using our method for both PM_{10} and $PM_{2.5}$. This result together with Figure 9 verifies that entropy is highly correlated with improved RMSE.

Evaluating GEM with different number of training stations. In Figure 11, instead of varying k, we vary the number of labeled nodes x and set k as 5. That is, the x-axis represents the total number of observed stations. There are 15 sites available to pick 5. When x becomes large, all models saturated to the same outcome since there are fewer combinations available to choose from. It shows that better quality can be obtained with more training stations available.

Evaluating GEM with different time spans. In previous experiments, the ranking at each time stamp is averaged to find the final ranking. Here we evaluate how the length of the time span affects the outcome. We vary the span from 0.5 to 14 months and report the corresponding TRR in Figure 12. Note that the spatial-dissimilarity and distance-based greedy methods are not influenced by time span change since they only consider static spatial features. The results show that in general when we average a longer time span, the performance improves. With 6 months of data, the TRR can be lower than 0.2.

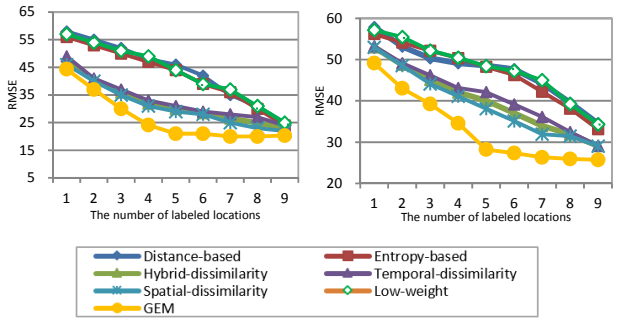

Figure 11. The RMSE of recommending 5 locations for PM_{10} and $PM_{2.5}$ with varying number of labeled locations.

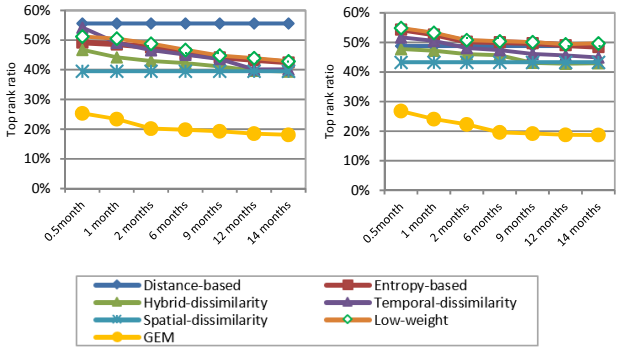

Figure 12. The TRR results for PM_{10} and $PM_{2.5}$ when varying the length of time span N_T.

6. RELATED WORK

The research is a step towards urban computing [20], which aims to use big data to tackle big challenges in cities. The research is also the part of urban air project [18, 19].

6.1 Inferring Unobserved Sensor Values

We first introduce four kinds of models to infer sensor values:

Emission Models. Two types of techniques exist under this category. Arguably the most popular is *interpolation*. Air quality values of a location can be derived through interpolating records from its nearby air quality monitoring stations [3]. Two commonly used interpolation models are Inverse Distance Weighting (IDW) [2] and Ordinary Kriging (OK) [14]. However, the applicability of this type of model is questionable when trying to infer non-smooth values, as shown in our experiments. The dispersion model represents another type of knowledge-driven techniques, which considers the air quality as a function of street geometry, meteorology, traffic condition, and other emission factors (e.g. g/km per road segment or per vehicle) [5]. Representative dispersion models include Gaussian Plum, Operational Street Canyon, and Computational Fluid Dynamics. However, the dispersion model requires detailed knowledge about the parameters of the spatial deployments (e.g. the height, orientation, and gaps between buildings and the roughness coefficient of different urban surfaces), which could hinder its generality and applicability.

Satellite Remote Sensing. Satellite remote sensing [8] is a top-down approach to derive the air quality of urban surface, which has been used for many years. One of the most representative studies is conducted by A. V. Donkelaar et al. [3] who propose to estimate $PM_{2.5}$ using the modern resolution imaging spectroradiameter. However, this method suffers from following limitations. First, the imaging technique considers only the atmosphere effect but not the human factors such as traffic and land usage. Second, its outcome is quite sensitive to the weather condition, thus might not be applicable to every city.

Crowdsourcing. The crowdsourcing approach provides an alternative approach to track the air quality [5, 17] using wearable gas sensors. The basic idea of crowdsourcing is to solicit the contribution of probed air quality from a large group of people. Crowdsourcing-based methods, however, suffers several limitations. First, the current wearable sensors are only capable of probing certain gas like CO_2 but not $PM_{2.5}$ or NO_2, and the cost is still pretty high (e.g. 300 USD). Second, the portable sensors usually need a much longer sensing time (up to 1 hour) for accurate sensing.

Machine Learning methods. Recently, machine-learning techniques have been proposed as alternatives to model air quality. There have been several learning-based techniques for air quality inference based on ANN [7, 13] and SVMs [12], such as U-Air [18]. However, our experiment results show that given sparse data, none of the above methods is capable of producing reasonably good results. The most similar model to what we proposed might be U-Air, but there are still several main differences. First, U-Air adopts existing classifiers to create accurate AQI classification based on features while this work aims to make the unlabeled nodes whose neighboring nodes with similar features tend to have similar AQI values. Our experiments confirm that our feature modeling method is more effective than U-Air's for spatial-temporal case. Second, U-Air exploits CRF to handle the temporal dependency for the past few hours. However in our model we can associate a node with not only recent layers but also far-away layers with similar conditions. Finally, the U-Air model does not explicitly minimize and output the uncertainty of the model, thus it is not apparent how such approach can be incorporated into a station-recommendation model.

6.2 Sensor Deployment Strategies

Another relevant thread of research focuses on selecting proper locations to place sensors such that the benefits can be maximized. This section discusses some relevant works in this direction. Generally, existing works can be categorized based on two factors:

1. Does it focus on deploying sensors from scratch, or it focuses on adding sensors incrementally? Our work belongs to the later as it assumes some sensors have already been deployed and the goal is to recommend locations for new ones.
2. Does it exploit observed sensor data during the process of deployment? In our solution the observed data is used.

Deploying from scratch without observed data. Krause et al. [10] proposes a deployment model with the goal to detect water contaminations as early as possible. Krause et al. [11] propose to find a set of locations such that the wireless sensors can best predict some future events, such as road speeds on a highway. Erdos et al. [6] aim to deploy sensors in an information delivery network to optimize the detection of duplicate data contents. The above works formulate the task as optimization problems and propose approximation solutions to deal with them.

Deploying from scratch using observed data. Pourali et al. [16] utilize Bayesian brief network to find a set of functional locations such that the placement of sensors can best monitor a complex power systems. Du et al. [4] aim to find a set of locations for sensor deployment to best measure the surface wind distribution over a large urban reservoir. They solve this problem by finding locations with the largest mutual information with others based on some heuristics. Their solutions cannot be directly applied to our problem since they do not consider incremental deployment.

Incremental deployment using observed data. This type of works is the most relevant to our task, though we have only identified one work in this category. Karamshuk et al. [9] aim to find a set of locations such that the placement of new retail stores can bring maximum number of customers. They formulate the task as a learning-to-rank problem based on geographical and human mobility features. However, we find it hard to adopt their methods to our problem because of several reasons. First, they aim to find locations that have higher potential to attract more customers while our objective is to find locations that can boost the model's AQI inferring accuracy. Second, they only focus on predicting the overall popularity of a location, while we want to infer the AQI values at different time steps.

To summarize, although there are existing studies proposed to find locations to deploy sensors under different scenarios, to the best of our knowledge, the goal we are achieving given sparse data is unique and challenging. The models we propose, however, might be applicable to solve the problems they are facing.

7. CONCLUSIONS

This paper proposes a model to recommend the most proper locations in which building new air quality monitoring stations can lead to the largest accuracy improvement on air quality inference. A framework that jointly infers air quality and recommends new locations is developed. We believe the proposed framework is general enough to be applied to the inference and deployment of other kinds of sensors. For example, it can be applied to the monitoring of traffic flows and noise at arbitrary locations and time. Several reasons lead to the success of the proposed model. First, the Affinity Graph seamlessly integrates spatial and temporal correlations together. Second, the weights are learned to not only capture the correlation between features and AQI but also to minimize the uncertainty of the model. Finally, the proposed entropy-minimization greedy tries to identify a set of nodes that are uncorrelated with the more confident (i.e. low entropy) ones most of the time as the recommended locations for deployment. It is much more effective than myopically minimize entropy or other heuristics. In the future, we will focus on improving the efficiency of this model through parallelization. Moreover, we will seek for more applications of our model, in particular in the area of traffic monitoring and surveillance in urban areas.

8. REFERENCES

[1] T. H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. 2009. Introduction to Algorithms (3rd ed.), MIT Press.

[2] S. Donald. A two-dimensional interpolation function for irregularly-spaced data. In Proc. of the National Conference. pp. 517–524. 1968.

[3] 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.

[4] W. Du, Z. Xing, M. Li, B. He, L. H. C. Chua, and H. Miao. Optimal sensor placement and measurement of wind for water quality studies in urban reservoirs. In Proc. of IEEE International Symposium on Information Processing in Sensor Networks ISPN, 2014.

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

[6] D. Erdös, V. Ishakian, A. Lapets, E. Terzi, and A. Bestavros. The filter-placement problem and its application to minimizing information multiplicity. In Proc. VLDB 2012.

[7] J. Hooyberghs, C. Mensink, G. Dumont, F. Fierens, O. Brasseur (2005). Aneural network forecast for daily average PM10 concentrations in Belgium. Atmospheric Environment 39 (2005) 3279-3289.

[8] 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.

[9] D. Karamshuk, A. Noulas, S. Scellato, V. Nicosia, and C. Mascolo. Geo-spotting: mining online location-based services for optimal retail store placement. In Proc. of KDD 2013.

[10] A. Krause, J. Leskovec, C. Guestrin, J. VanBriesen, and C. Faloutsos. Efficient Sensor Placement Optimization for Securing Large Water Distribution Networks. Journal of Water Resources Planning and Management, 134(6), 2008.

[11] A. Krause, R. Rajagopal, A. Gupta, and C. Guestrin. Simultaneous Optimization of Sensor Placements and Balanced Schedules. IEEE Transactions on Automatic Control, 2011.

[12] Wei-Zen Lu, Wen-Jian Wang (2005). Potential assessment of the "support vector machine" method in forecasting ambient air pollutant trends. Chemosphere 59: 693-701.

[13] H. Niska, T. Hiltunen, A.Karppinen, J. Ruuskanen, and M. Kolehmainen (2004). Evolving the neural network model for forecasting air pollution time series. Engineering Applications of Artificial Intelligence 17, 159-167.

[14] M. A. Oliver and R. Webster. Kriging: a method of interpolation for geographical information system. INT. J. Geographical Information Systems, VOL. 4, No. 3, 313-332, 1990.

[15] P. Perez, R. Palacios and A. Castillo (2004). Carbon Monoxide Concentration Forecasting in Santiage, Chile. Journal of the air and waste management association 54:908-913. ISSN 1047-3289. 2004.

[16] M. Pourali and A. Mosleh. A Functional Sensor Placement Optimization Method for Power Systems Health Monitoring, IEEE Transactions on Industrial Applications, 49(4), 2013.

[17] 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.

[18] Y. Zheng, F. Liu, H- P. Hsieh, U-Air: When Urban Air Quality Inference Meets Big Data. In Proc. of KDD 2013.

[19] Y. Zheng, L. Capra, O. Wolfson, H. Yang. Urban Computing: concepts, methodologies, and applications. ACM Transaction on Intelligent Systems and Technology (ACM TIST). 5(3), 2014.

[20] Y. Zheng, X. Chen, Q. Jin, Y. Chen, X. Qu, X. Liu, E. Chang, W-Y. Ma, Y. Rui, W. Sun. A Cloud-Based Knowledge Discovery System for Monitoring Fine-Grained Air Quality. MSR-TR-2014-40.

[21] X. Zhu, Z. Ghahramani and J. Lafferty. Semi-Supervised Learning Using Gaussian Fields and Harmonic Functions. ICML 2003.