AIREX: Neural Network-based Approach for Air Quality Inference in Unmonitored Cities

Anonymous authors
Anonymous affiliation
Anonymous@email

Abstract

Urban air pollution is a major environmental problem affecting human health and quality of life. Monitoring stations have been established to continuously obtain air quality information, but they do not cover all areas. Thus, there are numerous methods for spatially fine-grained air quality inference. Since existing methods aim to infer air quality of locations only in monitored cities, they do not assume inferring air quality in *unmonitored* cities.

In this paper, we first study the air quality inference in unmonitored cities. To accurately infer air quality in unmonitored cities, we propose a neural network-based approach *AIREX*. The novelty of AIREX is employing a mixture-of-experts approach, which is a machine learning technique based on the divide-and-conquer principle, to learn correlations of air quality between multiple cities. To further boost the performance, it employs attention mechanisms to compute impacts of air quality inference from the monitored cities to the locations in the unmonitored city. We show, through experiments on a real-world air quality dataset, that AIREX achieves higher accuracy than state-of-the-art methods.

1 Introduction

Urban air pollution poses a severe and global problem. The fine-grained assessment of urban air quality is crucial for both the governments and citizens to establish means to improve human health and quality of life. Monitoring stations have been established in numerous cities to continuously obtain air quality information. However, due to high construction and management costs, monitoring stations are sparsely installed and concentrated only in areas of higher importance, such as cities with large populations. As a result, it is crucial to infer air quality in areas without monitoring stations.

The development of neural network techniques has accelerated a neural network-based approach for inferring spatially fine-grained air quality cite [Zheng *et al.*, 2013; Cheng *et al.*, 2018]. This approach leverages available external data related to the air quality, such as point-of-interest, traffic, and meteorology, to capture features of locations. Existing neural network-based methods evaluated air quality of target locations only in monitored cities (i.e., cities with monitoring stations). However, since not all cities have monitoring stations, it is necessary to infer air quality in *unmonitored* cities.

In this paper, we study a new problem, *air quality inference in unmonitored cities*, to globally solve the urban air pollution problem. A straightforward approach for the problem is the use of existing models that are trained by air quality data of cities in the vicinity of the target unmonitored city [Chang and Hanna, 2004]. However, even the state-of-the-art method ADAIN [Cheng *et al.*, 2018] deteriorates the inference accuracy in unmonitored cities, even when using air quality data of numerous monitored cities as training data (see Table 1 in experiments section).

Therefore, we need a new neural network architecture in this problem. For developing a neural network architecture, we face two challenges: (1) how to design a neural network architecture to capture the correlations of air quality between monitored and unmonitored cities and (2) how to train models without available air quality data of the unmonitored cities. For the first challenge, since features of cities differ, architectures must capture their differences and reflect them in the inference of air quality. It is difficult to select optimal cities and compute the correlations between cities before model training. We require architectures that automatically capture the correlations between monitored and unmonitored cities. For the second challenge, since we do not have air quality data of the unmonitored city, architectures must be trained only by using air quality data of monitored cities and external data. This indicate that we cannot directly learn the correlations between monitored and unmonitored cities. We require architectures that can be effectively trained in an unsupervised manner and their training methods.

We propose a novel neural network-based architecture *AIREX* to solve the above two challenges. AIREX employs a mixture-of-experts approach [Jacobs *et al.*, 1991; Masoudnia and Ebrahimpour, 2014; Guo and Barzilay, 2018], which is a machine learning technique based on the divide-and-conquer principle. The mixture-of-experts approach uses multiple models (called *experts*) and aggregates outputs of experts for deriving the final output. AIREX infers air quality in unmonitored cities by aggregating air quality assessed from individual monitored cities. In order to further boost the performance, it employs attention mechanisms [Bahdanau *et al.*, 2014] for computing weights of influences from monitored cities to unmonitored cities. In addition, there are the correlations between monitoring stations and the target locations [Cheng *et al.*, 2018]. Thus, AIREX captures the importance of both monitored cities and monitoring stations individually by employing two attentions, namely, city-based and station-based attentions. AIREX effectively combines the mixture-of-experts and attention mechanisms for accurately inferring air quality in unmonitored cities.

For training AIREX, we develop a training method using a meta-training approach [Guo and Barzilay, 2018], which is suitable for training of the mixture-of-experts approach in an unsupervised manner. In our training method, we regard one of the monitored cities as an unmonitored city at the training phase so that AIREX can be learned in an unsupervised manner. We use multi-task learning [Caruana, 1997] for training both the whole AIREX and experts with capturing the difference among cities. This training method enables to learn the correlations between monitored and unmonitored cities without air quality data of unmonitored cities.

Our contributions presented in this study are as follows:

- We address a novel problem that infers air quality information in unmonitored cities by using the air quality data obtained from other cities. We show that state-of-the-art methods are not suitable for this problem.
- We propose AIREX that can accurately infer air quality information in unmonitored cities. This employs the mixture-of-experts approach and attention mechanism to capture the correlations of air quality between monitored and unmonitored cities.
- Through experiments with 20 cities in China, we show that AIREX achieves higher accuracy than the-state-of-the-art method.

2 Problem Formulation

We describe the notations and definitions used in the formulation of the problem that we solve in this study.

There are two types of cities, namely, target and source cities, that denote unmonitored and monitored cities, respectively. Each city c has its representative specific location l_c (e.g., the center of c). We assume that we have a single target city c_{tgt} and a set C_{src} of source cities. We denote the set of monitoring stations by S and each monitoring station $s \in S$ has its location l_s, which periodically monitors a quantity of air pollutants, such as PM$_{2.5}$, over the time domain $T = \langle t_1, t_2, \ldots, t_{|T|} \rangle$. Source city $c_k \in C_{src}$ has a set of monitoring stations $S_k \subseteq S$. We denote $s_{k,i}$ as monitoring station $s_i \in S_k$. We define air pollutant data as follows:

Definition 1 (Air pollutant data)**.** Air pollutant data D^A consists of quantities of air pollutants monitored by stations, and they are time-dependent.

Cities have characteristics that affect air quality. To infer air quality, we use three external data that were frequently employed in prior studies [Xu and Zhu, 2016; Zheng *et al.*, 2013; Cheng *et al.*, 2018]; Point-of-interest (PoI), road network, and meteorological information.

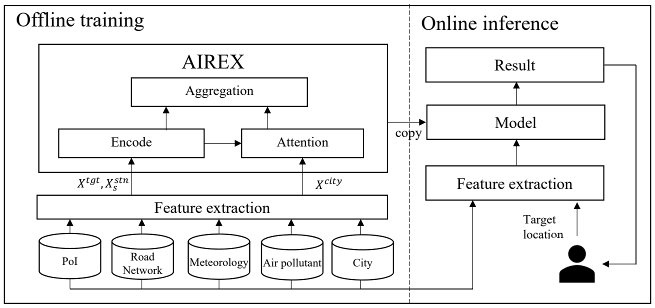

Figure 1: Our framework for air quality inference

Definition 2 (PoI data)**.** PoI data D^P consist of PoI information p, which is a triple of an identifier, specific location l_p, and category v_p (e.g., factory).

Definition 3 (Road network data)**.** A road network D^R consists of road segments r. Each road segment includes coordinates of the start and end points, and road category v_r (e.g., highway).

Definition 4 (Meteorology data)**.** Meteorology data D^M consist of distinct-level meteorological information. Meteorological information includes meteorological measurements, such as weather and temperature. The meteorology data are time-dependent data.

In this study, we aim to infer spatially fine-grained air quality in the target unmonitored city.

Problem statement. Given target city c_{tgt}, target location l_{tgt} in c_{tgt}, a set C_{src} of source cities, a set of monitoring stations in C_{src}, air pollutant data D^A, PoI data D^P, road network data D^R, and meteorology data D^M, we aim to infer air quality of l_{tgt} over time domain T.

We focus on regression for evaluating quantities of air pollutants in this paper, but our models can be used for classification for evaluating the air quality index [Cheng *et al.*, 2018].

3 Proposal

We present our neural network-based architecture AIREX and training method after describing our framework and feature extraction.

3.1 Framework and Design Policy

Figure 1 illustrates our framework. This framework consists of offline training and online inference. In the offline training, we build our inference model after extracting features, and in the online inference, we infer the air quality of the given target location by using the built model.

We describe a design policy of the offline training. Air quality of the target location is assessed by data related to target location, monitoring station, and cities. Thus, our framework extracts features of target location, monitoring stations, and cities, from data sources. We leverage these features to capture the correlations of air quality between the target and source cities, and the target location and monitoring stations.

We design our inference architecture AIREX for automatically capturing the correlations and being trained in an unsupervised manner. For this purpose, AIREX is based on the mixture-of-experts approach [Guo and Barzilay, 2018] and attention mechanism [Bahdanau *et al.*, 2014]. The mixture-of-experts approach compute the final output by aggregating the output of multiple models (i.e., *experts*). In AIREX, each expert is a model for inferring air quality by using data of source city. Each source city and monitoring station does not equally contribute the air quality inference in the target city, and thus we use the attention mechanism to compute the importance of cities and monitoring stations. AIREX can accurately infer the air quality in the target city by elegant combination of the mixture-of-experts approach and attention mechanism. Furthermore, AIREX can be trained in an unsupervised manner by using the meta-training approach [Guo and Barzilay, 2018] and multi-task learning [Caruana, 1997]. We describe the training method later.

AIREX consists of three main components: encoding, attention, and aggregation. First, in the encode, it encodes raw input features to obtain latent features for capturing interactions between inferred values and raw input. Then, in the attention, AIREX computes the importance of source cities and monitoring stations for inferring air quality of the target city. Finally, in the aggregation, it computes output of experts for each source city by aggregating the transformed features and importance of monitoring stations, and then compute the final output by aggregating the outputs of experts and importance of cities.

3.2 Feature extraction

We introduce our features for assessing air quality at l_{tgt}. We extract the three features, namely, the target location feature $\mathbf{X}^{tgt}$, monitoring station feature $\mathbf{X}_s^{stn}$, and city feature $\mathbf{X}^{city}$. These features comprise (1) PoI factor, (2) road network factor, (3) meteorological factor, (4) air pollutant factor, (5) station location factor, and (6) city location factor. We describe our three features after explaining how to extract each factor from the data.

The PoI, road network, and meteorological factors are associated with location l (e.g., locations of monitoring stations and the target location). l has its own factors that are extracted from the data within affecting region $\mathcal{L}(l)$. We set $\mathcal{L}(l)$ as a circle whose center and radius are l and 1 km, respectively.

PoI factor X_l^P: X_l^P includes the numbers of PoIs, which represents the characteristics of locations, such as the numbers of factories and public parks. We consider a set Υ_P of PoI categories and count the number of PoIs belonging to each PoI category. Let $X_l^P = \{x_\upsilon^P(l)\}_{\upsilon \in \Upsilon_P}$ denote the PoI factor for l. We compute x_υ^P as follows:

$$x_\upsilon^P(l) = |\{p \in D^P | l_p \subset \mathcal{L}(l) \wedge \upsilon_p = \upsilon\}|. \tag{1}$$

Road network factor X_l^R: X_l^R includes the numbers of road segments, which affects local air quality, as vehicles are one of the sources of air pollutants. We consider a set Υ_R of road categories and count the number of road segments belonging to each road category. Let $X_l^R = \{x_\upsilon^R(l)\}_{\upsilon \in \Upsilon_R}$ denote the road network features extracted for l. We define $\overline{r}$ as arbitrary points between the start and end of road segment r. We compute x_υ^R as follows:

$$x_\upsilon^R(l) = |\{r \in D^R | \overline{r} \subset \mathcal{L}(l) \wedge \upsilon_r = \upsilon\}|. \tag{2}$$

Meteorological factors X_l^M: X_l^M is the sequence of meteorological measurements of l, such as weather and temperature, which influences the concentrations and flows of air pollutants. The meteorological measurements have two types of values; categorical values (e.g., weather and wind direction) and numerical values (e.g., temperature and wind speed). For categorical and numerical values, we adopt one-hot encoding and raw values, respectively. We denote the meteorological factor at time step t as X_l^{Mt}.

These factors have demonstrated their usefulness in previous studies [Xu and Zhu, 2016; Cheng *et al.*, 2018]. We normalize numerical values in factors by dividing the largest values among each factor.

The monitoring and station location factors are associated with station s, and the city location factor is associated with city c.

Monitoring factor X_s^A: Quantities of air pollutants monitored by station s represent the most important information for inferring air quality. X_s^A is the sequence of air pollutant quantities in D^A of station s. We denote the monitoring factor at time step t as X_s^{At}.

Station location and city location factors X_c^C and X_s^S: The distance and direction from a location to another location are likewise important factors to measure the influence of their respective air quality levels. X_s^S (resp. X_c^C) is the relative position that depecits the distance and angle from station s (resp. source city c) to the target location l_{tgt} (resp. target city c_{tgt}).

Our features combine the above factors. The target location feature $\mathbf{X}^{tgt}$, monitoring station feature $\mathbf{X}_s^{stn}$, and city feature $\mathbf{X}^{city}$ are given as follows:

$$\begin{aligned}
\mathbf{X}^{tgt} &= X_{l_{tgt}}^P \cup X_{l_{tgt}}^R \cup X_{l_{tgt}}^M, \\
\mathbf{X}_s^{stn} &= X_{l_s}^P \cup X_{l_s}^R \cup X_{l_s}^M \cup X_s^A \cup X_s^S, \text{ and} \\
\mathbf{X}^{city} &= \cup_{c \in C_{src}} \{X_c^C\}.
\end{aligned}$$

Here, since the air quality changes time by time, it is preferable that all factors are time-dependent. Due to limited data sources, it is necessary to support both time-independent and time-dependent data.

3.3 Inference architecture

We introduce our inference architecture AIREX. Figure 2 shows components of AIREX. AIREX has three input types: $\mathbf{X}^{tgt}$, $\mathbf{X}_s^{stn}$ for $\forall s \in S$, and $\mathbf{X}^{city}$, and it contains five layers: encode, station-based attention, city-based attention, experts, and mixture layers. We describe each layer in the following.

Encode layer: The encode layer transforms $\mathbf{X}^{tgt}$ and $\mathbf{X}_s^{stn}$. Each feature includes time-independent (e.g., PoI) and time-dependent (e.g., meteorology) data. We transform time-dependent and time-independent factors by LSTM and FC, respectively [Cheng *et al.*, 2018]. We use different models for $\mathbf{X}^{tgt}$ and $\mathbf{X}_s^{stn}$ because they include different factors; however, we use the same LSTM and FC for all monitoring stations to increase generalization ability.

We first explain models for time-depending factors in $\mathbf{X}_s^{stn}$. $X_{l_s}^M$ and $X_{l_s}^A$ at time step t are transformed into $\boldsymbol{h}_s^t$

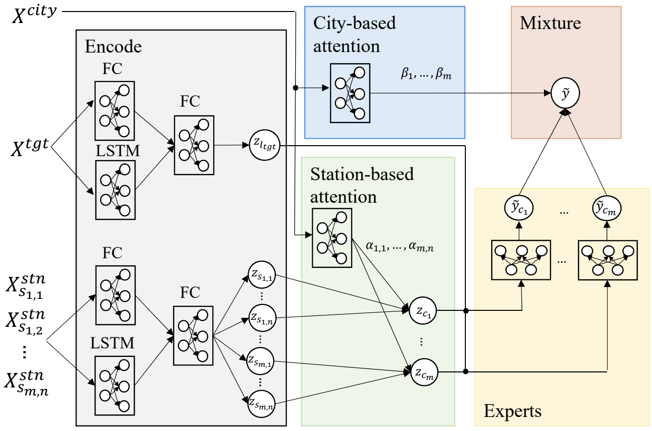

Figure 2: Neural network structure of AIREX

as follows:

$$\begin{aligned}
\mathbf{i}^t_s &= \sigma(\mathbf{W}_{ix}(X^{Mt}_{l_s} \oplus X^{At}_{l_s}) + \mathbf{W}_{ih}\mathbf{h}^{t-1}_s + \mathbf{W}_{ic} \odot \mathbf{c}^{t-1}_s + \mathbf{b}_i)\\
\mathbf{f}^t_s &= \sigma(\mathbf{W}_{fx}(X^{Mt}_{l_s} \oplus X^{At}_{l_s}) + \mathbf{W}_{fh}\mathbf{h}^{t-1}_s + \mathbf{W}_{fc} \odot \mathbf{c}^{t-1}_s + \mathbf{b}_f)\\
\mathbf{c}^t_s &= \mathbf{f}^t_s \odot \mathbf{c}^{t-1}_s + \mathbf{i}^t_s \odot tanh(\mathbf{W}_{cx}(X^{Mt}_{l_s} \oplus X^{At}_{l_s}) + \mathbf{W}_{ch}\mathbf{h}^{t-1}_s + \mathbf{b}_c)\\
\mathbf{o}^t_s &= \sigma(\mathbf{W}_{ox}(X^{Mt}_{l_s} \oplus X^{At}_{l_s}) + \mathbf{W}_{oh}\mathbf{h}^{t-1}_s + \mathbf{W}_{oc} \odot \mathbf{c}^{t}_s + \mathbf{b}_o)\\
\mathbf{h}^t_s &= \mathbf{o}^t_s \odot tanh(\mathbf{c}^t_s)
\end{aligned}$$

where, ***W*** is weight matrix, ***b*** is bias vector, and $\odot$ indicates Hadamard product. ***i***, ***f***, ***o***, ***c***, and ***h*** are input gate, forget gate, output gate, memory cell, and final states of hidden layer, respectively.

Next, we describe models for time-independent factors. $X^P_{l_s}$, $X^R_{l_s}$, and $X^S_{l_s}$ in $\mathbf{X}^{stn}_s$ are translated into embedding $\mathbf{z}^{(n)}_s$ as follows:

$$\mathbf{z}^{(i)}_s = \begin{cases} ReLU(\mathbf{W}^{(i)}_s(X^P_{l_s} \oplus X^R_{l_s} \oplus X^S_{l_s}) + \mathbf{b}^{(i)}_s), i = 1\\ ReLU(\mathbf{W}^{(i)}_s\mathbf{z}^{ni1}_s + \mathbf{b}^{(i)}_s), 1 < i \leq L \end{cases}$$

where L denotes the number of hidden layers.

$\mathbf{X}^{tgt}$ is transformed in the same way as $\mathbf{X}^{stn}_s$. The difference is the input factors.

Finally, the transformed features generated by the LSTM and FC are concatenated to input another FC to obtain the features $\mathbf{z}^{(n')}_*$ as follows:

$$\mathbf{z}^{(i')}_* = \begin{cases} ReLU(\mathbf{W}^{(i')}_{*'}(\mathbf{z}^L_* \oplus \mathbf{h}^t_*) + \mathbf{b}^{(i')}_{*'}), i' = L + 1\\ ReLU(\mathbf{W}^{(i')}_{*'}\mathbf{z}^{i'-1}_* + \mathbf{b}^{(i')}_{*'}), i' \in [L + 2, L + L'] \end{cases}$$

where $*$ indicates either l_{tgt} or s and L' denotes the number of hidden layers.

City-based Attention layer: Not all source cites contribute equally to inference in the target city. AIREX automatically captures the importance of different city data by employing the attention mechanism. The city-based attention layer computes *city-attention factor* which represents the weights of influences of source cities to air quality in the target city. The city-attention factor β_{c_k} of source city c_k is computed as follows:

$$\begin{aligned}
\mathbf{z}^{(L+L')}_{\oplus k} &= \mathbf{z}^{(L+L')}_{s_{k,1}} \oplus \cdots \oplus \mathbf{z}^{(L+L')}_{s_{k,n}}\\
\beta'_{c_k} &= \mathbf{w}^{\intercal}_{\beta} ReLU(\mathbf{W}_{\beta}(\mathbf{z}^{(L+L')}_{l_{tgt}} \oplus \mathbf{z}^{(L+L')}_{\oplus k} \oplus X^C_{c_k}) + \mathbf{b}_{\beta}) + b_{\beta}\\
\beta_{c_k} &= \frac{exp(\beta'_{c_k})}{\Sigma_{c \in C_{src}} exp(\beta'_c)}
\end{aligned}$$

Station-based Attention layer: Each monitoring station has a different impact to the target location, as distances and angles between each monitoring station and target location are different as well as similarity of their features. In the station-based attention layer, we compute *station-affect factor*, which is a weight of influence of monitoring stations on the air quality of the target location. The station-affect factor $\alpha_{k,i}$ for stations s_i in source city c_k is calculated by the following equation:

$$\begin{aligned}
\alpha'_{k,i} &= \mathbf{w}^{\intercal}_{\alpha} ReLU(\mathbf{W}_{\alpha}(\mathbf{z}^{(L+L')}_{l_{tgt}} \oplus \mathbf{z}^{(L+L')}_{s_{k,i}}) + \mathbf{b}_{\alpha}) + b_{\alpha}\\
\alpha_{k,i} &= \frac{exp(\alpha'_{k,i})}{\Sigma_{s_i \in S_k} exp(\alpha'_{k,i})}
\end{aligned}$$

We then compute embedding $\mathbf{z}_{c_k}$ of source city with station affect-factors as follows:

$$\mathbf{z}_{c_k} = \sum_{s_i \in S_k} \alpha_{k,i}\mathbf{z}^{(L+L')}_{s_{k,i}}.$$

$\mathbf{z}_{c_k}$ represents how much is influence air quality of source city c_k to the target location.

Experts layer: The experts layer computes an inferred value on each source city. Inferred value $\tilde{y}_{c_k}$ of c_k is computed by the following equation:

$$\tilde{y}_{c_k} = \mathbf{w}^{\intercal}_k ReLU(\mathbf{W}_k(\mathbf{z}^{(L+L')}_{l_{tgt}} \oplus \mathbf{z}_{c_k})) + \mathbf{b}_k) + b_k.$$

This equation represents an expert model. We use this simple model for all cities to eliminate the the impact of performance of experts to the final output in this paper.

Mixture layer: We obtain the inferred value by summing outputs of experts weighted by city attention factors as follows:

$$\tilde{y} = \sum_{c_k \in C_{src}} \beta_{c_k}\tilde{y}_{c_k}.$$

3.4 Training method

One of the major challenges of our study is the training of AIREX because we cannot directly train our model due to missing air quality data of the target city. We develop a training method in an unsupervised manner. We describe our approach and loss function in the training phase.

Overall idea: We employ a meta-training approach [Guo and Barzilay, 2018], which supports to learn the differences between individual features and cities in an unsupervised setting. Given a set of source cities, the meat-training approach regards a single source city as a *temporal* target city, and then trains models using the pair of temporal target and other source cities. The temporal target and other source cities are referred to as the *meta-target* c_t and *meta-sources* $c_i \in C_s$, respectively. We obtain $|C_{src}|$ training pairs of meta-target and meta-sources.

We use a multi-task learning method with a shared encoder. We design loss functions for accurately inferring air quality and capturing the difference between source and target cities.

Loss functions: The main objective of our training is that the final outputs are closer to the actual value. Since we have multiple experts, we additionally train them. It is not sufficient to evaluate the difference between outputs and true values because we must capture the correlations between the source and target cities. Since we do not have air quality data in the target city, we must indirectly learn the correlations. For this purpose, we use a loss for minimizing the difference between the transformed features of cities. We note that the true values in training phase are air quality of the meta-targets instead of the actual target location.

The loss $\mathcal{L}_f$ is the main loss function for evaluating the inference accuracy. $\mathcal{L}_f$ is computed by the mean squared error (MSE) between the final output $\tilde{y}$ and true value y as follows:

$$\mathcal{L}_f = \frac{1}{|\mathcal{T}|} \sum_{x \in \mathcal{T}} (\tilde{y}(x) - y(x))^2$$

where $\mathcal{T}$ denotes the set of training pairs.

The loss $\mathcal{L}_e$ is one for evaluating the inference accuracy of an individual expert. $\mathcal{L}_e$ is computed by MSE between $\tilde{y}_{c_k}$ for source city c_k of outputs of experts and y.

$$\mathcal{L}_e = \frac{1}{|C_s|} \sum_{c_i \in C_s} \left(\frac{1}{|\mathcal{T}|} \sum_{x \in \mathcal{T}} (\tilde{y}_{c_i}(x) - y(x))^2 \right).$$

The loss $\mathcal{L}_a$ is for evaluating the difference of cities. It is computed based on maximum mean discrepancy (MMD) [Gretton *et al.*, 2012] as the adversary to minimize the divergence between the marginal distribution of target and source cities. MMD is known as effective distance metric measures for evaluating the discrepancy between two distributions explicitly in a non-parametric manner.

$$\begin{aligned}
\mathcal{L}_a &= MMD^2(\mathbf{z}_{\cup c_1} \cup \cdots \cup \mathbf{z}_{\cup c_{|C_s|}}, \mathbf{z}_{\cup c_t}), \\
\mathbf{z}_{\cup c_i} &= \cup_{s \in S_i} \mathbf{z}_s^{(L+L')}, \\
MMD(\mathcal{X}, \mathcal{X}') &= \left\| \frac{1}{|\mathcal{X}|} \sum_{\mathbf{x} \in \mathcal{X}} \phi(\mathbf{x}) - \frac{1}{|\mathcal{X}'|} \sum_{\mathbf{x}' \in \mathcal{X}'} \phi(\mathbf{x}') \right\|_{\mathcal{H}},
\end{aligned}$$

In MMD computation, $\mathcal{H}$ indicates the reproducing kernel Hilbert space (RKHS) and ϕ is the mapping function to RKHS space. In our method, we compute the MMD score by the kernel method [Bousmalis *et al.*, 2016]. The kernel method computes the MMD score as follows:

$$\begin{aligned}
MMD(\mathcal{X}, \mathcal{X}') &= \frac{1}{|\mathcal{X}|(|\mathcal{X}|-1)} \sum_{\mathbf{x}, \mathbf{x}' \in \mathcal{X}, \mathbf{x} \neq \mathbf{x}'} \mathcal{K}(\mathbf{x}, \mathbf{x}') \\
&+ \frac{1}{|\mathcal{X}'|(|\mathcal{X}'|-1)} \sum_{\mathbf{x}, \mathbf{x}' \in \mathcal{X}', \mathbf{x} \neq \mathbf{x}'} \mathcal{K}(\mathbf{x}, \mathbf{x}') \\
&- \frac{2}{|\mathcal{X}||\mathcal{X}'|} \sum_{\mathbf{x} \in \mathcal{X}} \sum_{\mathbf{x}' \in \mathcal{X}'} \mathcal{K}(\mathbf{x}, \mathbf{x}')
\end{aligned}$$

where, $\mathcal{K}$ indicates a Gaussian karnel function: $\mathcal{K}(\mathbf{x}, \mathbf{x}') = exp(-\frac{1}{2\sigma^2}||\mathbf{x} - \mathbf{x}'||^2)$.

We further use regularization of β to avoid overfitting. The regularization computes the entropy of β and the sum of them.

$$\mathcal{R} = \sum_{c_i \in C_s} \beta_{c_i} \log \beta_{c_i} \tag{3}$$

The total loss function to be minimized in our training phase is defined as follows:

$$\mathcal{L} = \lambda \cdot \mathcal{L}_f + (1 - \lambda) \cdot \mathcal{L}_e + \gamma \cdot \mathcal{L}_a + \zeta \cdot \mathcal{R}$$

where λ, γ, and ζ are hyper parameters.

4 Experiments

In this section, we evaluate the inference accuracy of AIREX compared with the state-of-the-art methods. We aim to validate that AIREX can accurately infer air quality in unmonitored cities and other methods cannot[1].

4.1 Experimental settings

Dataset: We use data of 20 cities in China spanning four months from June 1st 2014/6/1 to September 30th 2014. We collect air quality data, road network, PoI, and meteorology data as follows. Air quality data is provided as open data by Microsoft[2]. We focus on inferring $\mathrm{PM}_{2.5}$. We collect PoI data from Foursquare[3] and categorize them into ten categories according to the official categories provided by Foursquare. For road network data, we use OpenStreetMap[4], and roads are categorised into three types; highway, trunk, and other. For meteorology data, we use weather, temperature, air pressure, humidity, wind speed, and wind direction, which is also provided by Microsoft. Air quality and meteorology data are sampled every hour.

Evaluation: We select four cities as target cities; Beijing, Tianjin, Shinzhen, and Guangzhou. Beijing and Tianjin are cities in the northern area of China, whereas Shinzhen and Guangzhou are in the south. We randomly select five monitoring stations from each city for training and test data. The ratio of training and test data is $|C_{src}|$ to one.

As evaluation metrics of inference accuracy, we use the root mean squared error (RMSE) for $\mathrm{PM}_{2.5}$ and accuracy for the air quality index derived by RMSE, which are standard metrics [Xu and Zhu, 2016; Cheng *et al.*, 2018]. We run three times for training by changing monitoring stations. Due to space limitations, we only report the average RMSE and accuracy of the air quality index has the same trends of RMSE.

Compared methods and hyper parameters: We compare AIREX with three approaches: (a) k nearest neighbors (KNN): This method selects the k monitoring stations closest to the target location, and compute the average air pollutant quantities from these stations as result. We set k to be three in our experiments. (b) Feedforward neural networks (FNN): This method uses a simple neural network model, whose inputs are $\mathbf{X}^{tgt}$ and $\mathbf{X}_s^{stn}$ for all stations. In our experiments, the model consists of three layers with 200 units. For sequential features, we only use their values at the same time step of the inferred air quality. (c) ADAIN: This method represents the state-of-the-art neural network model for inferring air quality [Cheng *et al.*, 2018]. We use two cases of source cities: ADAIN5 and ADAIN19, whose source cities are the five cities closest to the target city and all source cities, respectively.

In parameter settings of AIREX and ADAIN, we follow the setting in experiments of ADAIN [Cheng *et al.*, 2018]. We construct a single basic FC layer ($L = 1$) with 100 neurons and two LSTM layers with 300 memory cells per layer. We

[1] Please see a supplementary file for detail implementation, data statistics, and additional results.

[2] www.microsoft.com/en-us/research/project/urban-computing/

[3] developer.foursquare.com

[4] www.openstreetmap.org/

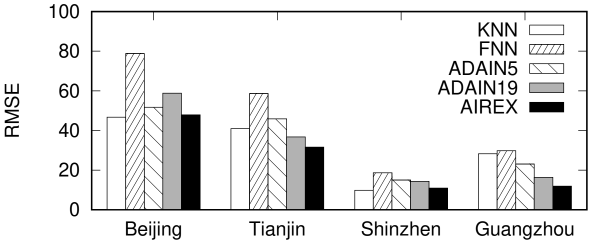

Figure 3: An overview of accuracy

then build two layers of the high-level FC network ($L' = 2$) with 200 neurons per layer. The time-dependent data is input in 24 time steps (i.e., one day). The number of epochs, the batch size, learning rate are selected from [100, 200, 300], [32, 64, 128, 256, 512], and [0.005, 0.01], respectively, by grid search. In our model, λ, γ, and ζ in AIREX are 0.5, 1.0, and 1.0, respectively. Further detail is provided in our codes.

4.2 Experimental results

Figure 3 shows the inference accuracy for each method. AIREX achieves the best accuracy in Tianjin and Guanzhou and the second best in Beijing and Shinzhen. Since AIREX learns the difference between target and source cities, it can accurately infer air quality without air quality data in the target city. KNN achieves the best accuracy in Beijing and Shinzhen, as these are monitoring stations that very close to the target location, whereas KNN fails the accurate inference when there are no monitoring stations close to the target location like Tianjin and Guangzhou. ADAIN and FNN do not perform well in all target cities. In particular, although ADAIN is the state-of-the-art method for inferring air quality, it does not perform well when the source and target cities are different.

We further investigate the difference between AIREX and ADAIN, as ADAIN may perform well if we use optimal source cities. Table 1 shows the accuracy of AIREX and ADAIN in Beijing and Guangzhou as target cities (see appendix for Tianjin and Shenzhen). In ADAIN, we use each city as the source city in addition to ADAIN5 and ADAIN19. In Beijing, ADAIN accurately infers the air quality when its source city is Beijing (i.e., target and source cities are the same). However, the accuracy of ADAIN significantly decreases when ADAIN uses different cities even when the source cities are close to Beijing. In Guangzhou, AIREX achieves better performance than ADAIN even when ADAIN uses Guangzhou as the source city. This result indicates that the use of multiple cities increases the inference accuracy if we can capture the correlations of air quality between cities. From these results, we can confirm that our mixture-of-experts approach combined with attention mechanisms performs well for accurately inferring the air quality in unmonitored cities without selecting source cities.

Table 1: AIREX vs ADAIN in different source cities. City names indicate the result obtained by ADAIN, where training data is the city. A distance of zero kilometers indicates that the target and source cities are the same.

Method		Beijing		Guangzhou	
		RMSE	Dist. [km]	RMSE	Dist. [km]
AIREX		47.88	—	11.90	—
ADAIN	5 NN cities	51.70	—	23.05	—
	19 cities	58.83	—	16.34	—
	Beijing	30.49	0	78.15	1883.5
	Langfang	48.58	47.1	61.57	1847.9
	Tianjin	52.05	113.8	43.01	1807.9
	Baoding	68.46	140.3	52.01	1758.1
	Tangshan	60.53	154.9	54.85	1887.7
	Zhangjiakou	81.79	160.9	24.0	1961.8
	Chengde	74.67	176.0	27.18	2024.9
	Cangzhou	59.57	181.5	47.22	1716.5
	Hengshui	66.90	248.8	52.68	1635.9
	Shijiazhuang	55.80	263.8	87.56	1657.7
	Qinhuangdao	69.13	273.0	43.79	1956.8
	Zibo	69.62	372.1	41.98	1585.2
	Shantou	85.63	1,835.3	25.95	350.6
	Huizhou	84.25	1,871.4	22.30	118.0
	Guangzhou	76.62	1,883.5	19.70	0
	Dongguan	80.07	1,888.8	20.29	51.4
	Foshan	76.56	1,897.4	21.87	18.9
	Shenzhen	86.14	1,937.7	23.64	104.1
	Jiangmen	82.32	1,946.5	22.75	63.8
	Hong Kong	86.99	1,953.3	26.46	118.8

5 Related Work

We review neural network-based approaches for spatially fine-grained air quality inference. Numerous methods have been proposed [Shad *et al.*, 2009; Hasenfratz *et al.*, 2014; Xu and Zhu, 2016], such that employ linear regression, matrix factorization and neural networks. For example, Zheng et al. [Zheng *et al.*, 2013] proposed U-air, which is a neural network-based classifier model that captures both spatial and temporal correlations. Hu et al. [Hu *et al.*, 2018] proposed an architecture that employs deep reinforcement learning for optimizing air quality sensing systems. Zhong et al. [Zhong *et al.*, 2020] proposed AirRL, which consists of station selector that distills monitoring stations using reinforcement learning. Cheng et al. [Cheng *et al.*, 2018] proposed ADAIN, which employs the attention mechanism to assign weights to station-oriented features. We used ADAIN to design encode and station-based attention layers. To the best of our knowledge, we first employ a mixture-of-experts approach for air quality inference.

None of them addresses the problem of air quality inference in unmonitored cities. In contrast to these studies, our method assigns weights to each city automatically, without selecting monitoring stations.

6 Conclusion

We addressed a new problem that infers air quality information in unmonitored cities. For the problem, we proposed AIREX, which can accurately infer air quality in unmonitored cities. Experimental studies using real data showed that AIREX outperforms the state-of-the-art methods.

As our future works, we address air quality inference in different countries, in particular, countries across sea, and support environments that each city has different data sources.

References

[Bahdanau *et al.*, 2014] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[Bousmalis *et al.*, 2016] Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. Domain separation networks. In *Proceedings of the NIPS*, pages 343–351, 2016.

[Caruana, 1997] Rich Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997.

[Chang and Hanna, 2004] Joseph C Chang and Steven R Hanna. Air quality model performance evaluation. *Meteorology and Atmospheric Physics*, 87(1-3):167–196, 2004.

[Cheng *et al.*, 2018] Weiyu Cheng, Yanyan Shen, Yanmin Zhu, and Linpeng Huang. A neural attention model for urban air quality inference: Learning the weights of monitoring stations. In *Proceedings of the AAAI*, 2018.

[Gretton *et al.*, 2012] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012.

[Guo and Barzilay, 2018] Jiang Guo and Regina Barzilay. Multi-source domain adaptation with mixture of experts. In *Proceedings of the ACL EMNLP*, pages 4694–4703, 2018.

[Hasenfratz *et al.*, 2014] David Hasenfratz, Olga Saukh, Christoph Walser, Christoph Hueglin, Martin Fierz, and Lothar Thiele. Pushing the spatio-temporal resolution limit of urban air pollution maps. In *Proceedings of the IEEE PerCom*, pages 69–77, 2014.

[Hu *et al.*, 2018] Zhiwen Hu, Zixuan Bai, Kaigui Bian, Tao Wang, and Lingyang Song. Real-time fine-grained air quality sensing networks in smart city: Design, implementation and optimization. *arXiv preprint arXiv:1810.08514*, 2018.

[Jacobs *et al.*, 1991] Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991.

[Masoudnia and Ebrahimpour, 2014] Saeed Masoudnia and Reza Ebrahimpour. Mixture of experts: a literature survey. *Artificial Intelligence Review*, 42(2):275–293, 2014.

[Shad *et al.*, 2009] Rouzbeh Shad, Mohammad Saadi Mesgari, and Arefeh Shad. Predicting air pollution using fuzzy genetic linear membership kriging in GIS. *The ELSEVIER Computers, Environment and Urban Systems*, 33(6):471–481, 2009.

[Xu and Zhu, 2016] Yanan Xu and Yanmin Zhu. When remote sensing data meet ubiquitous urban data: Fine-grained air quality inference. In *Proceedings of the IEEE Big Data*, pages 1252–1261, 2016.

[Zheng *et al.*, 2013] Yu Zheng, Furui Liu, and Hsun-Ping Hsieh. U-Air: When urban air quality inference meets big data. In *Proceedings of the ACM SIGKDD*, pages 1436–1444, 2013.

[Zhong *et al.*, 2020] Huiqiang Zhong, Cunxiang Yin, Xiaohui Wu, Jinchang Luo, and JiaWei He. Airrl: A reinforcement learning approach to urban air quality inference. *arXiv preprint arXiv:2003.12205*, 2020.