Incorporating Spatio-Temporal Smoothness for Air Quality Inference

Xiangyu Zhao[2], Tong Xu[1], Yanjie Fu[3], Enhong Chen[1], Hao Guo[1]
[1]Anhui Province Key Lab of Big Data Analysis and Application, University of Science and Technology of China
[2] Data Science and Engineering Lab, Michigan State University
[3] Department of Computer Science, Missouri University of Science and Technology
zhaoxi35@msu.edu, {tongxu, cheneh}@ustc.edu.cn, fuyan@mst.edu, guoh916@mail.ustc.edu.cn,

***Abstract*—It is well recognized that air quality inference is of great importance for environmental protection. However, due to the limited monitoring stations and various impact factors, e.g., meteorology, traffic volume and human mobility, inference of air quality index (AQI) could be a difficult task. Recently, with the development of new ways for collecting and integrating urban, mobile, and public service data, there is a potential to leverage spatial relatedness and temporal dependencies for better AQI estimation. To that end, in this paper, we exploit a novel spatio-temporal multi-task learning strategy and develop an enhanced framework for AQI inference. Specifically, both time dependence within a single monitoring station, and spatial relatedness across all the stations will be captured, and then well trained with effective optimization to support AQI inference tasks. As air-quality related features from cross-domain data have been extracted and quantified, comprehensive experiments based on real-world datasets validate the effectiveness of our proposed framework with significant margin compared with several state-of-the-art baselines, which support the hypothesis that our spatio-temporal multi-task learning framework could better predict and interpret AQI fluctuation.**

I. Introduction

Urban air quality is closely related to the health and lives of residents, which raises an urgent need for air quality control. However, though a large amount of efforts have been made, it is still challenging for meteorological departments to infer fine-grained and high-accuracy **A**ir **Q**uality **I**ndex (AQI). On the one hand, high cost of building and maintaining leads to insufficient monitoring station and imbalanced distribution, e.g., in Figure 1, only dozens of stations exist in a huge city like Shanghai, which results in extremely sparse monitoring records. On the other hand, complicated impact factors, e.g., historical air quality, meteorology, traffic volume, and human mobility, may cause significant challenges in summarizing AQI fluctuation rules.

Generally, prior arts may help to estimate the air quality index via statistical inference, but the performances are usually limited due to the small-scale course-grained data sources. Meanwhile, traditional spatial interpolation methods mainly exploit the spatial autocorrelation of AQI in terms of geographical distance, but they may fail with only considering the static factors, but completely ignoring their dynamics. Indeed, AQI is influenced by sophisticated and dynamic factors, which not just vary *spatially*, but also *temporally*. Therefore, it highly necessitatesa new inference framework with fine-grained spatio-temporal resolution, to overcome the challenge of dynamic spatio-temporal coupling.

Fig. 1. Examples of air quality monitoring stations in Shanghai city.

Recently, with the development of new techniques for collecting and integrating AQI-related data, such as air quality data, meteorology data, taxi trajectories data, and human mobility data near monitoring stations during a period, there is a potential to not only describe the AQI fluctuation rules, but also reveal their latent correlations, to be specific, the **temporal dependencies** and **spatial relatedness** between monitoring stations. Along these lines, in this paper, we attempt to infer AQI values with high spatio-temporal resolution through a data-driven perspective. Specially, we design our framework based on the assumptions as follows. First, current AQI values usually depend upon the previous status, which could be summarized as *"intra-station temporal dependencies"*. Second, two stations who are spatially close, or similar in profiles, usually hold correlated AQI value, which could be summarized as *"inter-station spatial relatedness"*. By jointly measuring and modeling these two basic assumptions, we design several constraints in regularization term, in which, particular, a new metric is incorporated to summarize the inter-station correlations by considering both geographic distance and profile similarities. Then, the comprehensive lost functions will be well optimized to support AQI inference tasks.

Finally, as air-quality related features from cross-domain data have been extracted and quantified, we conduct comprehensive experiments on real-world datasets to validate the effectiveness of our proposed framework, which outperforms several state-of-the-art baselines with significant margin. These

2374-8486/17 $31.00 © 2017 IEEE
DOI 10.1109/ICDM.2017.158

results not only prove the potential of our spatio-temporal multi-task learning framework to better interpret and predict AQI fluctuation, but also reveal some rules for measuring inter-station correlation, which could benefit the public service for station locating and building.

II. Air Quality Inference

In this section, we introduce our Spatio-Temporal real-time Feature-based Multi-Task Regression (stfMTR) framework for air quality inference. Our framework can perform 1) **temporal prediction**: to capture time dependence within a single monitoring station and provide forecasts for air quality in the near future as well as 2) **spatial interpolation**: to explore the spatial relatedness across all the stations and provide real-time inference for air quality in any given location.

A. Data Representations and Symbols

Let $\mathbf{S} \in R^N$ denote the air quality monitoring stations, where N is the number of stations. Suppose there are totally K time points, i.e., $\mathbf{T} \in R^K$. Let $\mathbf{X} \in R^{N\times K\times M}$ encode the feature tensor, where M is the dimension of features, and $\mathbf{X}_{:,k} \in R^{N\times M}$ is the feature matrix of all stations at time point t_k. Correspondingly, we use $\mathbf{W} \in R^{N\times K\times M}$ as the weight tensor of the model, and $\mathbf{W}_{:,k} \in R^{N\times M}$ is the weight matrix of all stations at time point t_k. For station s_n at time point t_k, $\mathbf{X}_{n,k} \in R^M$ is the feature vector, and $\mathbf{W}_{n,k} \in R^M$ is the related weight vector.

Let matrix $\hat{\mathbf{Y}} \in R^{N\times K}$ encode the ground truth matrix of AQIs, e.g., $\hat{\mathbf{Y}}_{n,k}$ is the AQI of station s_n in time point t_k. Let matrix $\mathbf{Y} \in R^{N\times K}$ encode the target matrix. For different task, $\mathbf{Y}$ has different elements. For **temporal prediction**, if we want to predict the air quality for h hour later, in the model *training stage*, we set $Y_{n,k} = \hat{Y}_{n,k+h}$. For **spatial interpolation**, $\mathbf{Y}$ is the matrix of real-time AQIs, i.e., $Y_{n,k} = \hat{Y}_{n,k}$. Specifically, $\mathbf{Y}_{:,k} \in R^N$ is the target vector for time point t_k. Thus, we can integrate two tasks into one framework.

B. Distance-based Spatial Smoothness for AQIs Inference

For inferring air quality of multiple stations at time point t_k, $\mathbf{X}_{:,k} \odot \mathbf{W}_{:,k} = \{\mathbf{X}_{1,k}\mathbf{W}_{1,k}^{\mathrm{T}}, \ldots, \mathbf{X}_{N,k}\mathbf{W}_{N,k}^{\mathrm{T}}\} \in R^N$ should be a good approximation of $\mathbf{Y}_{:,k}$, i.e., the air quality is the linear weighted sum of individual feature. We formulate the loss function as a least-squares loss function. At time point t_k, we minimize the loss function to learn $\mathbf{W}_{:,k}$ as:

$$\min_{\mathbf{W}_{:,k}} L_k = \|\mathbf{Y}_{:,k} - \mathbf{X}_{:,k} \odot \mathbf{W}_{:,k}\|_2^2 + \gamma\|\mathbf{W}_{:,k}\|_F^2, \tag{1}$$

where we add a penalty term $\|\mathbf{W}_{:,k}\|_F^2$ to regularize the magnitude of $\mathbf{W}_{:,k}$, and $\|\cdot\|_F$ is the *Frobenius norm*. As mentioned in [1] traditional multi-task regression models usually adopt a graph regularization term to enforce spatial smoothness as follows:

$$\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} D_{ij}\|\mathbf{W}_{i,k} - \mathbf{W}_{j,k}\|_2^2, \tag{2}$$

where D_{ij} is the spatial proximity between station s_i and station s_j. One frequently-used choice of D is a power law exponential function $D_{ij} = d(i,j)^{-\mathcal{K}}$, where $d(i,j)$ is the *Vincenty distance* between station s_i and station s_j. This spatial penalty automatically encodes Tobler's first law of geography and imposes a soft constraint that spatially close stations tend to have similar air quality.

C. Real-Time Feature-based Smoothness for AQIs Inference

However, this traditional spatial penalty is simply based on static spatial distance, which cannot reveal the accurate relatedness among stations. In fact, the air quality reported by two stations may be quite different though they are spatially close, e.g., station 2 and 10 in Figure 1, because of the difference of function of region, traffic pattern and human mobility pattern. Therefore we introduce a feature-based penalty term to capture the real-time proximity across stations as:

$$\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} F_{ij}^k\|\mathbf{W}_{i,k} - \mathbf{W}_{j,k}\|_2^2, \tag{3}$$

where F_{ij}^k is the real-time feature proximity between station s_i and station s_j at time point t_k. We borrow the classic *Cosine similarity* to measure this feature-based similarity F as:

$$F_{ij}^k = cosine(\mathbf{X}_{i,k}, \mathbf{X}_{j,k}) = \frac{\mathbf{X}_{i,k}\cdot\mathbf{X}_{j,k}}{\|\mathbf{X}_{i,k}\|\|\mathbf{X}_{j,k}\|}, \tag{4}$$

where $\mathbf{X}_{n,k} \in R^M$ is the feature vector for station s_n at time point t_k. This feature-based penalty softly constrain that stations having similar real-time features tend to have similar air quality. Thus, the objective loss function at time point t_k can be restated as:

$$\begin{aligned}\min_{\mathbf{W}_{:,k}} L_k = &\|\mathbf{Y}_{:,k} - \mathbf{X}_{:,k} \odot \mathbf{W}_{:,k}\|_2^2 + \gamma\|\mathbf{W}_{:,k}\|_F^2 \\ &+ \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\alpha D_{ij} + \beta F_{ij}^k\right)\|\mathbf{W}_{i,k} - \mathbf{W}_{j,k}\|_2^2,\end{aligned} \tag{5}$$

where γ, α and β are regularization parameters. Thus, the loss function Equation (5) captures both static spatial proximity and real-time feature proximity among air quality monitoring stations.

D. Temporal Smoothness for AQIs Inference

In reality, aside from inter-station spatial relatedness, the air quality in the near future have dependency on the present air quality. Thus, we introduce temporal smoothness into the air quality inference framework from multi-task perspective. Specifically, we divide the entire time domain into K discrete points ordered by time. For a single station s_n, the feature matrix $\mathbf{X}_{n,:} \in R^{K\times M}$ is divided into K separate feature vectors $\{\mathbf{X}_{n,1}, \mathbf{X}_{n,2}, \ldots, \mathbf{X}_{n,K}\}$. Correspondingly, we separate weight matrix $\mathbf{W}_{n,:} \in R^{K\times M}$ into K separate weight vectors, i.e., $\mathbf{W}_{n,:} = \{\mathbf{W}_{n,1}, \mathbf{W}_{n,2}, \ldots, \mathbf{W}_{n,K}\}$.

We now propose a temporal regularization to reveal inherent temporal smoothness of AQIs within each single station. Intuitively, considering the smooth evolution of AQIs, the weight on each feature should also change smoothly. We use a discrete weight sequence over time to capture the temporal dynamics of our framework, and we add the following temporal regularization term to the loss function, which is a variant of the fused lasso:

$$\sum_{n=1}^{N}\left(\sum_{k=2}^{K}\|\mathbf{W}_{n,k} - \mathbf{W}_{n,k-1}\|_2^2\right), \tag{6}$$

where $\mathbf{W}_{n,k} = \{W^1_{n,k}, W^2_{n,k}, \ldots, W^M_{n,k}\} \in R^M$ is the weight vector for station s_n at time point t_k. Scalar $W^m_{n,k}$ is the m^{th} element of $\mathbf{W}_{n,k}$. We leverage l_2-norm because l_2-norm treats each feature equally and shrinks all the quantities with the same magnitude, but not l_1-norm which defeats our purpose of tracking temporal dynamics of weights.

E. Optimization Task

To capture both spatial smoothness and temporal smoothness into our air quality inference framework, we integrate Equation (5) and Equation (6), then the overall objective function can be restated as:

$$\begin{aligned}\min_{\mathbf{W}} L &= \sum_{k=1}^{K} L_k + \lambda \sum_{n=1}^{N} \left(\sum_{k=2}^{K} \|\mathbf{W}_{n,k} - \mathbf{W}_{n,k-1}\|_2^2 \right) \\ &= \sum_{k=1}^{K} \Bigg(\|\mathbf{Y}_{:,k} - \mathbf{X}_{:,k} \odot \mathbf{W}_{:,k}\|_2^2 + \gamma \|\mathbf{W}_{:,k}\|_F^2 \\ &\quad + \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \left(\alpha D_{ij} + \beta F^k_{ij} \right) \|\mathbf{W}_{i,k} - \mathbf{W}_{j,k}\|_2^2 \Bigg) \\ &\quad + \lambda \sum_{n=1}^{N} \left(\sum_{k=2}^{K} \|\mathbf{W}_{n,k} - \mathbf{W}_{n,k-1}\|_2^2 \right),\end{aligned} \tag{7}$$

where γ, α, β and λ are regularization parameters.

Then, we solve the optimization problem of the loss function in Equation (7). Considering the differentiability of our formulation, we approach weight tensor $\mathbf{W}$ by deriving the gradient of L with respect to vector $\mathbf{W}^m_{n,:}$. Specifically, in each iteration, we update one vector of $\mathbf{W}$, say $\mathbf{W}^m_{n,:}$, while keep other vectors fixed at current values to solve subproblem:

$$\begin{aligned}\min_{\mathbf{W}^m_{n,:}} \sum_{k=1}^{K} \Bigg[& \left| Y_{n,k} - \mathbf{X}^{(-m)}_{n,k} \odot \mathbf{W}^{(-m)}_{n,k} - X^m_{n,k} W^m_{n,k} \right|^2 \\ & - \left(2\alpha \sum_{n' \neq n} D_{nn'} W^m_{n',k} + 2\beta \sum_{n' \neq n} F^k_{nn'} W^m_{n',k} \right) W^m_{n,k} \\ & + \left(\alpha \sum_{n' \neq n} D_{nn'} + \beta \sum_{n' \neq n} F^k_{nn'} + \gamma \right) {W^m_{n,k}}^2 \Bigg] \\ & + \lambda \sum_{k=2}^{K} \left| W^m_{n,k} - W^m_{n,k-1} \right|^2,\end{aligned} \tag{8}$$

where $\mathbf{X}^{(-m)}_{n,k}$ is $\mathbf{X}_{n,k}$ with the m^{th} element removed, and $\mathbf{W}^{(-m)}_{n,k}$ is $\mathbf{W}_{n,k}$ with the m^{th} element removed correspondingly. We divide and rearrange Equation (8) according to each time point t_k as follows:

$$L(W^m_{n,k}) = A^m_{n,k} {W^m_{n,k}}^2 + 2B^m_{n,k} W^m_{n,k} + C^m_{n,k} + E^m_{n,k} \tag{9}$$

where we define $A^m_{n,k}$, $B^m_{n,k}$ and $C^m_{n,k}$ as follows:

$$A^m_{n,k} = {X^m_{n,k}}^2 + \alpha \sum_{n' \neq n} D_{nn'} + \beta \sum_{n' \neq n} F^k_{nn'} + \gamma, \tag{10}$$

$$\begin{aligned}B^m_{n,k} = & \alpha \sum_{n' \neq n} D_{nn'} W^m_{n',k} + \beta \sum_{n' \neq n} F^k_{nn'} W^m_{n',k} \\ & + \left(Y_{n,k} - \mathbf{X}^{(-m)}_{n,k} \odot \mathbf{W}^{(-m)}_{n,k} \right) X^m_{n,k},\end{aligned} \tag{11}$$

$$C^m_{n,k} = \left(Y_{n,k} - \mathbf{X}^{(-m)}_{n,k} \odot \mathbf{W}^{(-m)}_{n,k} \right)^2. \tag{12}$$

$E^m_{n,k}$ varies along with time point, i.e., no prior status of temporal relatedness works when $k = 1$, so we have $E^m_{n,k} = \lambda (W^m_{n,k+1} - W^m_{n,k})^2$. When $k \in \{2, 3, \ldots, K-1\}$, $E^m_{n,k} = \lambda (W^m_{n,k} - W^m_{n,k-1})^2 + \lambda (W^m_{n,k+1} - W^m_{n,k})^2$. When $k = K$, $E^m_{n,k} = \lambda (W^m_{n,k} - W^m_{n,k-1})^2$. Thus, we can calculate the gradient of $L(W^m_{n,k})$ as follows:

$$\frac{\partial L(W^m_{n,k})}{\partial W^m_{n,k}} = \begin{cases} 2\left(A^m_{n,k} + \lambda\right) W^m_{n,k} - 2\left(B^m_{n,k} + \lambda W^m_{n,k+1}\right), \\ \qquad\qquad if\, k = 1 \\ 2\left(A^m_{n,k} + 2\lambda\right) W^m_{n,k} \\ \quad -2\left(B^m_{n,k} + \lambda W^m_{n,k-1} + \lambda W^m_{n,k+1}\right), \\ \qquad\qquad if\, 1 < k < K \\ 2\left(A^m_{n,k} + \lambda\right) W^m_{n,k} - 2\left(B^m_{n,k} + \lambda W^m_{n,k-1}\right). \\ \qquad\qquad if\, k = K \end{cases} \tag{13}$$

Thus in each iteration solving subproblem Equation (8), we use gradient-based methods with the gradient calculated in Equation (13) to update the current $W^m_{n,k}$, while all $W^{m'}_{n',k'} (n' \neq n, k' \neq k, m' \neq m)$ are fixed at the current values. The time complexity of each optimization iteration is $\#iter * O(N + M^2)$, where $\#iter$ is the number of iterations for the gradient descent procedure.

F. Inference Task

Based on the above preliminaries, now we can formally present the inference task of our spatio-temporal multi-task learning framework.

For **temporal prediction**, as mentioned in Section II-A, if we want to predict the air quality for h hour later, in the model *training stage*, we set $Y_{n,k} = \hat{Y}_{n,k+h}$, then we get corresponding weight vector $\mathbf{W}_{n,k}$ at time point t_k. Because features have temporal dependency within in single station, i.e., features tends not to change sharply in near future, we use $\mathbf{W}_{n,k}$ as the weight vector for time t_{k+1}, i.e., at time t_{k+1}, we have feature vector $\mathbf{X}_{n,k+1}$, then we can predict the AQI for h hours later, i.e., t_{k+1+h}, as $Y_{n,k+1+h} = \mathbf{X}_{n,k+1} \mathbf{W}^{\mathrm{T}}_{n,k}$.

For **spatial interpolation**, as the test location has no historical AQIs, we use feature extracted from other four data sets to train our model, i.e., meteorological, traffic, POIs and human mobility. In test stage, we estimate the weight vector as $\mathbf{W}_{test,k} = \frac{\sum_{i=1}^{N} \left(\alpha D_{test,i} + \beta F^k_{test,i}\right) \mathbf{W}_{i,k}}{\sum_{i=1}^{N} \left(\alpha D_{test,i} + \beta F^k_{test,i}\right)}$, i.e., locations which are spatially close or have similar real-time features tend to have similar weight vectors. Then we infer the AQI of test location as $Y_{test,k} = \mathbf{X}_{test,k} \mathbf{W}^{\mathrm{T}}_{test,k}$.

III. Experiments

In this section, we conduct extensive experiments with multiple cross-domain datasets to evaluate the air quality inferring performance of our Spatio-Temporal real-time Feature-based Multi-Task Regression (stfMTR) model.

A. Experimental Settings

1) Data and Features: We evaluate our method with five datasets collected from April 1 to April 30, 2015 in Shanghai

City, China. Here we could only collect a short-term data set due to the limited overlap of multi-source data.

- **Air quality data**: Air quality has temporal dependency within in single station, i.e., current AQI is correlated with historical AQIs. We collect hourly historical AQIs reported by 10 air quality monitor stations in Shanghai.
- **Meteorological data**: Meteorology influences the concentration of air pollutants. We collect hourly fine-grained meteorological data, consisting of weather, temperature, wind speed, wind direction and precipitation.
- **Taxi trajectories**: It is widely believed that traffic flow is a major source generating air pollutants. We use a GPS trajectory dataset generated by taxicabs to calculate number of taxis appear in the target region, and the expectation, standard deviation and distribution of speeds.
- **Point of interests (POIs)**: The density of POIs indicates the land use rate and function of region, which could contribute to air quality inference. We crawl 12 categories of POIs, i.e., crossroad, services of auto, restaurant, shopping, hotel, school, sports leisure, scenic spot, housing, transportation, finance, company.
- **Human mobility**: Human mobility implies useful information, such as land use of a location, traffic flow and function of a region. We extract pick-up points, drop-off points, occupancy rate of taxis[2] from the aforementioned taxi trajectories and swiping data of smart card, which are frequently-used indices to value the human mobility.

2) Ground Truth and Metrics: For **temporal prediction**, the ground truth is obtained from test station's later reports, we evaluate the predictive performance with respect to its reports in next 1 and 3 hours. For **spatial interpolation**, we use leave-one-out validation, i.e., we first remove one station (testing set) from Shanghai and infer the AQIs of this station with the remaining 9 stations (training set). The reports of this station are then employed as a ground truth to evaluate the interpolation. The performance is evaluated in terms of the avergae root-mean-square-error (RMSE) of all stations as $RMSE = \frac{1}{N}\sum_{n=1}^{N}\sqrt{\frac{1}{K}\sum_{k=1}^{K}(y_k - \hat{y_k})^2}$.

B. Baselines

1) Baselines of Temporal Prediction: To validate the temporal prediction of our stfMTR model, we compare it with the following five baselines:

- **ARIMA**: Auto-Regression-Integrated-Moving-Average is well-known for predicting time series data, which makes predictions solely based on historical data.
- **VAR**: Vector Auto-Regressive is a multi variate forecasting technique accounting for cross correlation and temporal correlation.
- **LASSO**: Lasso tries to minimize the objective function $\frac{1}{2}\|\mathbf{Y}_{:,k} - \mathbf{X}_{:,k} \odot \mathbf{W}_{:,k}\|_2^2 + \gamma\|\mathbf{W}_{:,k}\|_1$ and encodes the sparsity over all weights in $\mathbf{W}_{:,k}$.
- **stMTL** [1]: Spatio-Temporal Multi-Task Learning enhances static spatial smoothness regression framework by learning the temporal dynamics of features in an non-parametric manner.
- **stMTMV** [3]: Spatio-Temporal Multi-Task Multi-View learning contains: 1) intra-station view, which combines local spatio-temporal information within each station; 2) inter-station view, which performs co-predictions through spatial correlations among stations.

2) Baselines of spatial interpolation: To validate the spatial interpolation of our stfMTR model, we compare it with the following five baselines:

- **Average**: We use the remaining 9 stations' average air quality to interpolate the test station' air quality.
- **IDW+**: IDW is a Inverse-Distance-Weighting linear interpolation algorithm [4] using AQIs of existing monitoring stations. We enhance it to capture both spatial proximity and feature similarity, named IDW+.
- **CoKrigring** [5]: CoKrigring is a multi-variate extension of kriging to interpolate data in a multi-variate scenario, which could capture both spatial and feature proximity.
- **ANN**: Artificial Neural Network treats historical features from all stations as the training data to build interpolation model. The ANN contains one hidden layer.
- **SFST** [6]: Shape-Function-based Spatio-Temporal method presents AQI interpolation by accounting for the existence of a temporal pattern on spatial measures of time-evolving geophysical fields.

TABLE I
OVERALL PERFORMANCE (RMSE) OF EACH APPROACH.

Temporal	1 hour	3 hour	**Spatial**	real-time
ARIMA	30.225	45.787	Average	46.563
VAR	28.756	42.907	IDW+	39.016
LASSO	25.387	38.653	CoKriging	35.291
stMTL	18.176	30.009	ANN	29.667
stMTMV	13.989	24.239	SFST	25.290
stfMTR	**12.595**	**20.562**	stfMTR	**22.633**

C. Experiment Results

1) Comparison of Overall Performance: First of all, we show the overall inference performance of our approach comparing with different baselines and the results are shown in Table I. The performance is evaluated in terms of the average RMSE of 10 stations. It's worth noting that we select the best performing hyperparameters (through grid-search), i.e., $\gamma, \alpha, \beta, \lambda$, for stfMTR in Table I. Correspondingly, we do parameter-tuning for baselines to make comparison fairer.

For the **temporal prediction**, we have following observations: 1) All the techniques perform better in short-term prediction (1 hour v.s. 3 hours), which shows that prediction of the near time is easier than prediction of distant future. 2) stMTL, stMTMV and stfMTR outperform the other three single-task learning methods, which demonstrates that the air monitoring stations are correlated. 3) stMTL performs worse than stfMTR, since stMTL only uses air quality data, while

stfMTR can incorporate multi-source heterogeneous information. 4) stfMTR performs better than stMTMV, because stfMTR can capture the time-dependence within each station and real-time feature relatedness among all the stations.

For the **spatial interpolation**, we have folwing observations: 1) Our stfMTR model performs better in temporal prediction than spatial interpolation, because spatial interpolation lacks historical air quality data. 2) CoKriging outperforms IDW+, because air quality interpolation is a non-linear system and Kriging could capture the distribution of the stations. 3) ANN outperforms the former two methods, as ANN incorporates multi-source heterogeneous information. 4) stfMTR performs better than SFST, because stfMTR captures the feature relatedness among all the stations, and stfMTR is more adept at interpolation in fine-grained time scale.

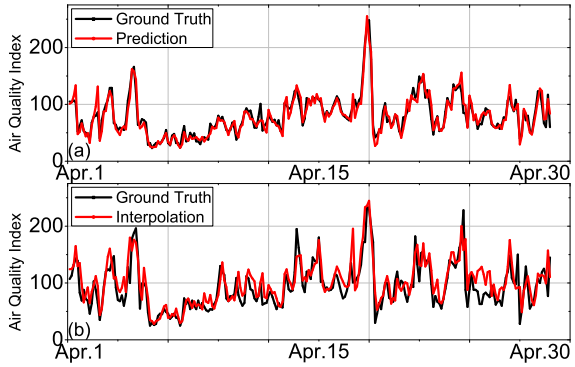

Fig. 2. Inferences of stfMTR against the ground truths. (a)Temporal prediction over next one hour. (b)Real-time spatial interpolation.

2) Air Quality Curve Fitting: Figure 2 depicts the temporal prediction results of our method over the next one hour in Figure (a) and the spatial interpolation in Figure (b) against the ground truth of two selected stations from April 1 to April 30, 2015, where the black curves are the ground truth and the red curves are the inferences. In general, our model is more accurate in temporal prediction, because spatial interpolation lacks historical air quality data.

3) Evaluation on Model Components: To evaluate each component of our stfMTR model, we compared it with four different variants of stfMTR:

- stfMTR$-t$: This derivation is to evaluate the performance of temporal smoothness, so we set parameters of spatial smoothness as 0, i.e., $\alpha = 0$ and $\beta = 0$.
- stfMTR$-s$: In this derivation, we evaluate the performance of spatial smoothness, so we derive it by setting parameter of temporal smoothness as 0, i.e., $\lambda = 0$.
- stfMTR$-dis$: This derivation is to evaluate the performance of spatial distance proximity smoothness, we can derive it by setting $\beta = 0$.
- stfMTR$-fea$: In this derivation, we evaluate the performance of real-time feature similarity smoothness, we can derive it by setting $\alpha = 0$.

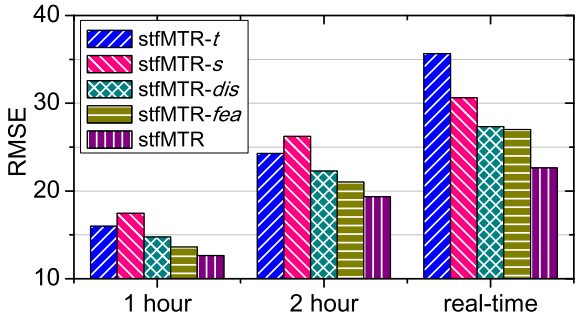

Fig. 3. Performance comparison on model components.

The experimental results are demonstrated in Figure 3. From this figure, we have following observations: 1) stfMTR$-t$ outperforms stfMTR$-s$ in temporal prediction, while stfMTR$-s$ outperforms stfMTR$-t$ in spatial interpolation, which verifies that the spatial and temporal smoothness parts have advantages in their related tasks respectively. 2) stfMTR$-fea$ performs better than stfMTR$-dis$ because air quality reported by two spatially close stations are usually quite different, which motivates us to introduce the real-time feature proximity. 3) stfMTR outperforming stfMTR$-t$ indicates that the air monitoring stations are not independent. Introducing the relatedness among stations can improve inference performances.

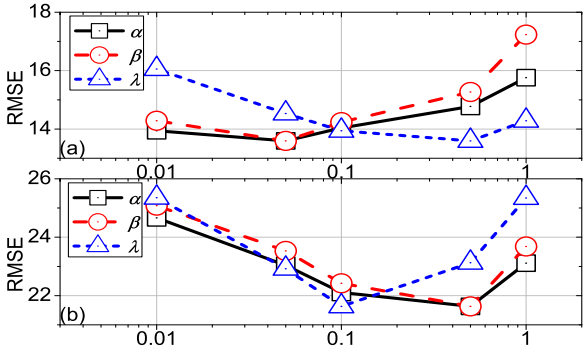

Fig. 4. Parameter sensitiveness. (a) Temporal prediction performance with different parameters. (b) Spatial interpolation performance with different parameters.

4) Parametric Sensitivity Analysis: We mainly concern three parameters in our approach, i.e., α that controls spatial proximity smoothness, β that controls real-time feature smoothness and λ that controls temporal smoothness. Specifically, we observe the RMSE of the stfMTR model by adjusting one parameter, while keeping other parameters fixed at best performance values. Note that the best performing results in Table I are the average best performances of 10 stations (hyperparameters are selected through grid-search), while the analysis in Figure 4 is based on the hyperparameters of two selected stations, since each station has its own best performing hyperparameters respectively.

In Figure 4, (a), (b) separately show the parametric sensitivities in temporal prediction/spatial interpolation task. Figure 4(a) shows that performance achieves the peak when $\alpha = \beta = 0.05$, or when $\lambda = 0.5$. In other words, temporal prediction is mainly depended on the historical air quality data within single station. On the contrary, for spatial interpolation in Figure 4(b), performance achieves the peak when $\alpha = \beta = 0.5$ or $\lambda = 0.1$, which means that the test station's air quality is mainly affected by the other stations' current features.

IV. Related Work

In this section, we briefly introduce the related works of our study, which can be mainly grouped into three categories.

The first category is about air quality model. Numerous efforts have been made on understanding air quality prediction model. [7] used data-driven models to predict fine-grained AQI. [8] introduced a nonparametric method for the continuous online prediction. The other topic closely related to this category is air quality interpolation. [9] inferred the AQI by a co-training learning approach. [6] investigated interpolation methods through the temporal pattern on spatial measures.

The second category related to this paper is cross-domain data mining. For social event analysis, authors in [10], [11] analyzed how cross-domain knowledge influences users' decision making process of social event participation. For complex network analysis, Zhao et. al [12], [13] attempted to find effective multiple spreaders in complex networks by generalizing the idea of the coloring problem in graph theory. For urban computing, Zhu et. al [14] measured real estate liquidity by examining multiple factors in a holistic manner. For recommendation, [15] proposed a probabilistic latent factor model by jointly considering the social correlation, geographical influence and users preference.

The third category related to this paper is application of multi-task learning. For water safety, [3] forecast the water quality of a station over the next few hours. For traffic, [1] proposed an extension to static trajectory regression framework. However, the model of [1] has three drawbacks in the following aspects: 1) the model only fed a variety of feature about a task; 2) only static spatial proximity between tasks depended on distance was involved in this model; and 3) samples in the same time slot share the same weights. Unlike prior literature, we jointly take into account the three drawbacks and propose an enhanced framework.

V. Conclusion

In this paper, we propose a novel Spatio-Temporal real-time Feature-based Multi-Task Regression (stfMTR) learning framework for better AQI inference, in which the intra-station time dependences and the inter-station spatial relatedness could be both captured. Specially, a new metric is incorporated in our model to summarize the inter-station correlations by considering both geographic distance and profile similarities. Then, an integrated framework with well optimization is designed to support AQI inference tasks in recent future and in any given place. Extensive experiments have validated the effectiveness of our framework with significant margin compared with several state-of-the-art baselines, which prove the potential of spatio-temporal smoothness in AQI inference. In the future, we plan to extend our multi-task learning framework by exploring more complex relatedness, and exploit more applications of our framework in urban areas instead of only air quality inference.

Acknowledgment

This research was partially supported by grants from the National Natural Science Foundation of China (Grant No. U1605251 and No. 61703386), the Anhui Provincial Natural Science Foundation (Grant No. 1708085QF140), and the Fundamental Research Funds for the Central Universities (Grant No. WK2150110006). Also, this research was partially supported by University of Missouri Research Board (UMRB) via the proposal number: 4991.

References

[1] J. Zheng and L. M.-S. Ni, "Time-dependent trajectory regression on road networks via multi-task learning," in *Proceedings of the 27th AAAI Conference on Artificial Intelligence, AAAI 2013, Bellevue, Washington, USA*, 2013, p. 1048.

[2] T. Xu, H. Zhu, X. Zhao, Q. Liu, H. Zhong, E. Chen, and H. Xiong, "Taxi driving behavior analysis in latent vehicle-to-vehicle networks: A social influence perspective," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2016, pp. 1285–1294.

[3] Y. Liu, Y. Zheng, Y. Liang, S. Liu, and D. S. Rosenblum, "Urban water quality prediction based on multi-task multi-view learning," *Urban water*, 2016.

[4] A. Appice, A. Ciampi, F. Fumarola, and D. Malerba, "Missing sensor data interpolation," in *Data Mining Techniques in Sensor Networks*. Springer, 2014, pp. 49–71.

[5] B. Gräler, E. Pebesma, and G. Heuvelink, "Spatio-temporal interpolation using gstat," *R Journal*, vol. 8, no. 1, pp. 204–218, 2016.

[6] L. Li, X. Zhang, J. B. Holt, J. Tian, and R. Piltner, "Spatiotemporal interpolation methods for air pollution exposure." in *SARA*, 2011.

[7] Y. Zheng, X. Yi, M. Li, R. Li, Z. Shan, E. Chang, and T. Li, "Forecasting fine-grained air quality based on big data," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015, pp. 2267–2276.

[8] V. C. Guizilini and F. T. Ramos, "A nonparametric online model for air quality prediction." in *AAAI*, 2015, pp. 651–657.

[9] Y. Zheng, F. Liu, and H.-P. Hsieh, "U-air: When urban air quality inference meets big data," in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2013, pp. 1436–1444.

[10] X. Zhao, T. Xu, Q. Liu, and H. Guo, "Exploring the choice under conflict for social event participation," in *International Conference on Database Systems for Advanced Applications*. Springer, 2016, pp. 396–411.

[11] T. Xu, H. Zhong, H. Zhu, H. Xiong, E. Chen, and G. Liu, "Exploring the impact of dynamic mutual influence on social event participation," in *Proceedings of the 2015 SIAM International Conference on Data Mining*. SIAM, 2015, pp. 262–270.

[12] X.-Y. Zhao, B. Huang, M. Tang, H.-F. Zhang, and D.-B. Chen, "Identifying effective multiple spreaders by coloring complex networks," *EPL (Europhysics Letters)*, vol. 108, no. 6, p. 68005, 2015.

[13] B. Huang, X.-Y. Zhao, K. Qi, M. Tang *et al.*, "Coloring the complex networks and its application for immunization strategy," *Acta Physica Sinica*, vol. 62, p. 218902, 2013.

[14] H. Zhu, H. Xiong, F. Tang, Q. Liu, Y. Ge, E. Chen, and Y. Fu, "Days on market: Measuring liquidity in real estate markets." in *KDD*, 2016, pp. 393–402.

[15] H. Guo, X. Li, M. He, X. Zhao, G. Liu, and G. Xu, "Cosolorec: Joint factor model with content, social, location for heterogeneous point-of-interest recommendation," in *International Conference on Knowledge Science, Engineering and Management*. Springer, 2016, pp. 613–627.