Inferring Air Pollution by Sniffing Social Media

Shike Mei, Han Li, Jing Fan, Xiaojin Zhu and Charles R. Dyer
Department of Computer Sciences, University of Wisconsin-Madison, Madison, WI, USA 53706
{mei, hanli, fanj, jerryzhu, dyer}@cs.wisc.edu

***Abstract*—The first step to deal with the significant issue of air pollution in China and elsewhere in the world is to monitor it. While more physical monitoring stations are built, current coverage is limited to large cities with most other places under-monitored. In this paper we propose a complementary approach to monitor Air Quality Index (AQI): using machine learning models to estimate AQI from social media posts. We propose a series of progressively more sophisticated machine learning models, culminating in a Markov Random Field model that utilizes the text content in social media as well as the spatiotemporal correlation among cities and days. Our extensive experiments on Sina Weibo data from 108 cities during a one-month period demonstrate the accurate AQI prediction performance of our approach.**

I. Introduction

Air pollution is a significant issue in China and elsewhere around the world. For example, in 2013 Beijing had 58 days when the Air Quality Index (AQI) was higher than 200 or "heavy pollution."[1] In December 2013 the east and central regions of China, which have more than 600 million people, experienced heavy pollution for more than two weeks. Air pollution is harmful to people's health, causing "eye irritation, lung and throat irritation, lung cancer and problems with babies at birth"[2].

To better deal with the problems of air pollution, the first step is to monitor air quality. From January 1 to November 1, 2013, the coverage of physical monitoring stations has increased from 74 cities to 108 cities in China. Also, the Chinese government has started to include PM2.5 (a major and dangerous air pollutant) into AQI monitoring[3].

The cost of establishing and maintaining physical monitoring stations limits their deployment currently to large and medium cities only. As a result, AQI monitoring in many regions such as small cities and rural towns is still lacking. To help people in these regions obtain air quality information, we consider the following question: can we estimate AQI without physical monitoring by using other, already available, information sources?

In this paper we estimate AQI using social media data as the information source. Social media is a rich and timely information source about air pollution in China. The most popular social media site in China, Sina Weibo, has about 100 million messages posted every day from all over the country[4]. Our key observation is that high AQI (poor air quality) in a region causes more Weibo posts from that region to discuss air pollution. For instance, here are some random Weibo posts (in Chinese) that contain the word "mai" (霾, haze):

起风了希望把雾霾吹走
雾霾中的太阳没有了光芒
雾霾污染指数450整个世界都模糊
雾霾一天比一天严重了雾鲁木齐啊
雾霾天气收到口罩是一件很幸福的事情
雾霾好大门窗紧闭都依然感到嗓子不舒服
今天又是雾霾能见度较低请驾驶员减速慢行
落霞与孤鹜齐飞秋水共长天一色可惜有雾霾
南京今天雾霾红色警报全城中小学幼儿园停课

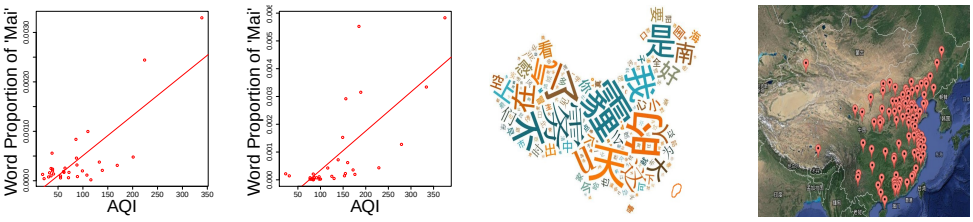
In fact, the word "mai" is positively correlated with AQI as shown in Figures 1(a,b). Figure 1(c) further visualizes the frequency of some Chinese characters that co-occur with "mai" in Weibo posts.

Our main machine learning model is a Markov Random Field that exploits this and other correlations. This paper demonstrates that our method can accurately estimate AQI from publicly-available Weibo posts, thereby offering an inexpensive way to obtain information about air quality in diverse areas in China, not limited to cities with AQI monitoring stations.

We also point out the main limitations of our approach upfront. First, our model does not *forecast* future AQI but rather estimates current AQI from near-realtime population reactions in social media. Second, our model is subject to the availability of social media posts and therefore does not apply to remote regions with extremely low social media user populations. Still, this work provides complementary value to existing AQI monitoring approaches, and can be an integral part in the overall solution to the air pollution problem.

The present paper is related to a line of recent work that attempts to gather air pollution information based on sources other than monitor stations. Honicky *et al.* [3] suggested collecting air pollution information by sensors attached to mobile phones. Poduri *et al.* [4] estimated the extent of airborne particulate matter by commodity cameras that are commonly used in mobile phones. Aoki *et al.* [1] used vehicles to monitor air quality by deploying mobile air quality sensing platforms on street sweeping trucks in San Francisco. Recently, Zheng *et al.* [6] and Chen *et al.* [2] estimated the air quality in big cities by fusing monitor stations data with meteorological and traffic data. Our computational model is distinct and builds upon the recent work by Xu *et al.* [5], which monitored spatial-temporal signals from social media. However, we focus on predicting air pollution from the text content in social media, whereas they only counted the number of wildlife roadkill event occurrences in social media.

[1] http://www.cnemc.cn
[2] http://www.cdc.gov/air/particulate_matter.html
[3] The PM2.5 information in China is reported at http://www.cnemc.cn
[4] According to http://en.wikipedia.org/wiki/Sina_Weibo



(a) Beijing (b) Shanghai (c) word cloud (d) 108 monitoring cities

Fig. 1: (a,b) Daily proportion of the word "mai" (haze) in Weibo posts vs. AQI in Beijing and Shanghai between November 13 and December 12, 2013. The correlation coefficient ρ are 0.799 and 0.709 for Beijing and Shanghai, respectively. (c) Word cloud showing the most frequent Chinese characters that occur in Weibo posts containing "mai," produced using tagxedo.com. (d) The 108 cities with air quality monitor stations in China.

II. Collection and Processing of Data

A. Data Collection

We collected posts to Sina Weibo from 108 cities (i.e., all cities with publicly available AQI data) in China during the time period November 18 to December 18, 2013. The location of the 108 cities is shown in Figure 1(a). The collection was done by calling the "nearby photos" API of Sina Weibo[5]. There are constraints: the API returns at most 200 posts; the posts' GPS coordinates should lie within a 10-km radius circle defined by the center's coordinates; and the time when the posts were posted should be in a specific one hour long period. We specified the center of the circles at each city's geographic center found by Google Maps. Due to the limitation of the API, we collected at most 200 posts per hour per city. On average, we obtained about 1,380 posts in each spatiotemporal bin (city and day).

We collected AQI information for these 108 cities every hour from the Ministry of Environmental Protection of China [6]. These hourly data were averaged to produce a daily AQI value for each city. Table I shows the data distribution.

TABLE I: Distribution of spatiotemporal bins according to AQI

AQI	$[0, 100)$	$[100, 200)$	$[200, 300)$	≥ 300	All
Number of bins	1422	1313	397	185	3317
Proportion of bins	42.87%	39.58%	11.96%	5.57%	100.00%

B. Text Processing

From each Weibo post we extracted the text information contained in the posts. First, we used the open-source Chinese segmentation software ANSJ[7] to segment the Chinese text in each post. Each output segment in each post was regarded as a word. We filtered out all the stopwords using the stopword list at http://nlp.csai.tsinghua.edu.cn/thulac/. We created a vocabulary by removing all word types appearing less than 10 times in the whole dataset. Our vocabulary contained 100,000 word types. We aggregated all the posts in one (city, day) bin as one "document." Finally, we represented each document as a bag-of-words vector to be defined below.

[5]http://open.weibo.com/wiki/2/place/nearby/photos
[6]http://113.108.142.147:20035/emcpublish/
[7]https://github.com/ansjsun/

C. Evaluation

We first introduce our notation for the data. For spatiotemporal bin (s, t), the bag-of-words vector representing the pooled Weibo posts in that bin is denoted $\mathbf{x}_{s,t}$, and the daily average AQI is denoted $y_{s,t}$. For evaluation, we divided the cities as training cities S_{train} and test cities S_{test}. All bins (s, t) with $s \in S_{train}$ form the training set. The other bins form the test set. Our goal is to estimate $y_{s,t}$ in the test set given the AQI and Weibo content in the training set, and the Weibo content in the test set.

In the rest of the paper we introduce a series of machine learning methods for estimating AQI. The differences between these methods will shed light on the merits of different features and spatiotemporal correlations in predicting air pollution. To compare different machine learning methods, we used mean square error (MSE) between the predicted AQI $\hat{y}_{s,t}^{test}$ and the actual AQI $y_{s,t}^{test}$ to evaluate the performance: $MSE = \frac{1}{\#TestDataPoints} \sum_{s,t} (\hat{y}_{s,t}^{test} - y_{s,t}^{test})^2$.

III. AQI Prediction Based on "Mai"

A. The Models

To begin, we consider very simple regression models based on only one feature: the proportion of the word "mai" in each bin. This proportion is one element of $\mathbf{x}_{s,t}$ and we denote it as $x_{s,t}^{mai}$. This assumption captures the intuition that poor air quality (i.e., high $y_{s,t}$ values) will lead to more complaints about "mai" in social media, and is justified by Figures 1(a,b).

We consider two machine learning models: ridge regression and support vector regression. Ridge regression learns the slope β_1 and the offset β_0 by solving the following optimization problem on the training set, $\text{argmin}_{\beta_0,\beta_1} \frac{1}{2} \sum_{s,t} (y_{s,t} - \beta_1 x_{s,t}^{mai} - \beta_0)^2 + \frac{C}{2}(\beta_0^2 + \beta_1^2)$, where C is the regularization parameter to trade off between training loss and model complexity. Letting $\mathbf{z}_{s,t} = [1, x_{s,t}^{mai}]$ and $\boldsymbol{\beta} = [\beta_0, \beta_1]$, then the problem has the standard notation:

$$\text{argmin}_{\boldsymbol{\beta}} \frac{1}{2} \sum_{s,t} (y_{s,t} - \boldsymbol{\beta}^T \mathbf{z}_{s,t})^2 + \frac{C}{2} \|\boldsymbol{\beta}\|_2^2. \quad (1)$$

Another commonly used linear regression model is support vector regression (SVR). It solves the following constrained

TABLE II: Prediction performance with only one feature, "mai"

AQI	$[0, 100]$	$(100, 200]$	$(200, 300]$	$(300, 1000)$	All
MSE for Ridge Regression	4739 ± 74	15494 ± 167	38562 ± 677	65048 ± 1601	17032 ± 470
MSE for SVR	4718 ± 23	16511 ± 130	40791 ± 511	68485 ± 1880	17914 ± 430

optimization problem,

$$\begin{aligned} \text{argmin}_{\boldsymbol{\beta}} \quad & \frac{1}{2}\|\boldsymbol{\beta}\|_2^2 + C_{SVR}\sum_{s,t}\xi_{s,t} \\ s.t. \quad & |y_{s,t} - \boldsymbol{\beta}^T \mathbf{z}_{s,t}| \leq \epsilon + \xi_{s,t};\ \xi_{s,t} \geq 0 \end{aligned} \tag{2}$$

where C_{SVR} is a regularization parameter and ϵ is the tolerance of error.

B. Experimental Results

We randomly divided the 108 cities into a training set with 80 cities and a test set with 27 cities. The parameter C, C_{SVR} and ϵ were tuned using 5-fold cross-validation (CV) and the best values were $C = 10^6$, $C_{SVR} = 10^5$ and $\epsilon = 10^{-2}$. The average test MSEs of five runs on random train/test splits are shown in Table II. the overall prediction MSE is large (a MSE around 17000 translates to roughly off by 130 in AQI value). Therefore, prediction based on only the most intuitive feature is not enough. We need to utilize more information, as we explain in the next sections.

IV. AQI Prediction Using Full BOW Features

A. The Models

We generalize the above methods as regression based on all Weibo bag-of-words features (with dimensionality 100,000). The same ridge regression and SVR as in Eq (1) and Eq (2) are used except for higher dimensional parameter vectors. We re-tuned the parameters with cross-validation.

B. Experimental Results

The experiment settings were exactly the same as in Section III-B. The best parameter settings tuned by 5-fold CV were $C = 10^3$, $C_{SVR} = 10^3$ and $\epsilon = 10^{-2}$.

Table IV shows the results. The prediction performance using the full BOW vector is much better than prediction based on only "mai". The overall MSE reduces to about 3500, which roughly translates to AQI error of less than 60. For spatiotemporal bins with AQI<200 ("light pollution" [8]), the prediction has MSE less than 2300 (AQI error less than 50). For bins with AQI>300 ("severely polluted"), the AQI prediction is off by no more than 150 on average. So it usually does not predict a severely polluted day as good air quality (with $AQI \leq 100$). As before, the MSEs of ridge regression and SVR are similar. Therefore, the problem is not very sensitive to the specific form of training loss (hinge loss in SVR or L2 norm in ridge regression).

We are also interested in the words with the largest absolute value of weights (i.e., top words). First, we removed the top words which are city names because they are not easily interpretable. Then, the words with the largest positive weights and the smallest negative weights learned in ridge regression are shown[9] in Table III. The words with the largest positive weights are all strongly indicative of poor air quality. The words with the smallest negative weights are indicative of good weather and perhaps cold fronts which sweep air pollution away. Therefore, our regression utilizes the Weibo content strongly related with air quality to estimate the AQI accurately.

TABLE III: Features with extreme weights

Chinese word	English translation	weight
霾	haze	12496
污染	pollution	8865
室内	indoor	5562
严重	heavy	5501
停留	stay	5396
指数	index	5214
...	...	...
阳光	sunshine	-4181
晴	sunny	-5087
冷	cold	-5715

Even though these prediction models are much improved, they only perform estimation on each spatiotemporal bin in isolation. We also want to know if the performance could be improved by considering the spatial correlation of the cities. After all, air pollution occurs in large pockets that often span several nearby cities. This will be investigated next.

TABLE IV: Prediction performance with the full BOW vector

AQI	$[0, 100]$	$(100, 200]$	$(200, 300]$	$(300, 1000)$	All
MSE for Ridge Regression	2231 ± 97	1101 ± 25	6990 ± 281	22904 ± 929	3469 ± 121
MSE for SVR	1705 ± 58	1291 ± 47	8275 ± 313	25772 ± 868	3598 ± 141

TABLE V: Prediction performance with KNN

AQI	$[0, 100]$	$(100, 200]$	$(200, 300]$	$(300, 1000)$	All
MSE for KNN	1336 ± 108	1910 ± 104	4396 ± 204	12607 ± 620	2646 ± 75

TABLE VI: Prediction performance with MRF

AQI	$[0, 100]$	$(100, 200]$	$(200, 300]$	$(300, 1000)$	All
MSE for MRF	1534 ± 96	1150 ± 77	3878 ± 231	11710 ± 782	2312 ± 105

V. AQI Prediction by KNN

A. The Model

To exploit the spatial correlation among cities, we start with a method that uses only nearby cities' AQI information: k-nearest-neighbor (KNN) method. In KNN, the AQI of a test data point $y_{s,t}^{test}$ is predicted by the average of AQIs in the nearest K training cities (denoted as $\text{KNN}^{train}(s)$) in the same day t. That is, $\hat{y}_{s,t}^{test} = \frac{\sum_{s' \in \text{KNN}^{train}(s)} y_{s',t}^{train}}{K}$. The distance between cities is the straight-line distance. Note that this KNN does *not* use any social media information.

B. Experimental Results

We first tuned the number of nearest neighbors K, by 5-fold CV. The CV MSE for different values of K from 1 to 64

[8] The definition of AQI levels can be found in http://en.wikipedia.org/wiki/Air_quality_index

[9] We report that similar phenomenon is observed for SVR.

and the MSE is the smallest when $K = 3$. This K potentially suggests the characteristic size of a pollution pocket. With $K = 3$, the test set MSE is shown in Table V. The KNN prediction is surprisingly good, considering that it does not utilize any Weibo content. This is an important observation. It suggests that there can be synergy between spatial correlation and text content, which we explore in the next section.

We point out that the earlier linear regression models are still valuable despite their slightly inferior MSE performance. KNN cannot predict anything without knowing the nearby cities' AQI, whereas our linear regression models can predict AQI based on a completely separate information source: Weibo text. When there are no nearby cities with AQI information, or should the AQI information become unavailable in the future for any reason, the linear regression models still work.

VI. Markov Random Field for AQI Estimation

A. The Model

The KNN results show the strength of spatial correlation and the regression model shows the power of Weibo content. Combining them together may further improve performance. Therefore, we now consider both in a Markov Random Field (MRF) model to model the correlation between AQI $y_{s,t}$ and social media information $\mathbf{x}_{s,t}$. As in linear regression, we assume that $y_{s,t}$ is related to the dot product between some weight $\boldsymbol{\beta}$ and $\mathbf{x}_{s,t}$, plus Gaussian noise ϵ with zero mean and σ^2 variance $y_{s,t} \approx \boldsymbol{\beta}^\top \mathbf{x}_{s,t} + \epsilon$ where $\epsilon \sim \mathcal{N}(0, \sigma^2)$. We also assume that the weight $\boldsymbol{\beta}$ is drawn from a Gaussian distribution with $\mathbf{0}$ mean and covariance $\sigma^2_{\boldsymbol{\beta}}\mathbf{I}$: $\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \sigma^2_{\boldsymbol{\beta}}\mathbf{I})$. This leads to a log potential term in the MRF:

$$\phi(\mathbf{y}, \mathbf{X}, \boldsymbol{\beta}) \triangleq \sum_{s,t} \frac{(y_{s,t} - \boldsymbol{\beta}^T \mathbf{x}_{s,t})^2}{2\sigma^2} + \frac{\|\boldsymbol{\beta}\|_2^2}{2\sigma^2_{\boldsymbol{\beta}}}. \quad (3)$$

To take into account the spatial correlation between cities and the temporal correlation within the same city, we define the potential term between two spatiotemporal bins $\phi(y_{s,t}, y_{s',t'})$. For AQI values on the same day t, nearby cities should have similar AQI. Therefore, we define KNN(s) as the set of K nearest neighbors of city s measured by geographical distance. And we define two cities s, s' as similar, denoted as $s \sim s'$, when $s \in$ KNN(s') or $s' \in$ KNN(s). Then we define $\phi(y_{s,t}, y_{s',t})$ as $\phi(y_{s,t}, y_{s',t}) = \frac{1}{2}\alpha_S I(s' \sim s)(y_{s,t} - y_{s',t})^2$, where $I()$ is an indicator function and α_S controls the strength of spatial correlation (tuned using cross validation).

The AQI in the same city s on two adjacent days may also have similar values. We denote two days t and t' as neighbors $t \sim t'$ when $|t - t'| = 1$. We model this by defining $\phi(y_{s,t}, y_{s,t'})$ as $\phi(y_{s,t}, y_{s,t'}) = \frac{1}{2}\alpha_T I(t' \sim t)(y_{s,t} - y_{s,t'})^2$. α_T controls the temporal correlation. It is also tuned using cross validation. In summary, the potentials $\phi(y_{s,t}, y_{s',t'})$ defined as

$$\phi(y_{s,t}, y_{s',t'}) = \begin{cases} \frac{1}{2}\alpha_S(y_{s,t} - y_{s',t'})^2 & \text{if } t = t', s' \sim s \\ \frac{1}{2}\alpha_T(y_{s,t} - y_{s',t'})^2 & \text{if } s = s', t' \sim t \\ 0 & \text{otherwise} \end{cases}$$

form an undirected graph among the y's.

Summing up the potential functions in Eqs. (3) and (4), we get an MRF model that accounts for both spatiotemporal correlations and social media. The joint probability is

$$p(\mathbf{y}, \mathbf{X}, \boldsymbol{\beta} | \sigma_\beta, \sigma, \alpha_S, \alpha_T) \propto \exp\left(-(\phi(\mathbf{y}, \mathbf{X}, \boldsymbol{\beta}) + \sum_{s,t}\sum_{s',t'} \phi(y_{s,t}, y_{s',t'}))\right). \quad (4)$$

When $\alpha_S = \alpha_T = 0$, the MRF only considers the correlation between $y_{s,t}$ and $\mathbf{x}_{s,t}$. When $\sigma = \infty$, the MRF degenerates to model only the spatiotemporal correlation. Therefore, our MRF model combines both information.

B. Inference with the MRF

To perform inference using the MRF, all Weibo content $\mathbf{X}$ is observable, the AQI values are divided into a training set $\mathbf{y}^{train}$ (observable) and a test set $\mathbf{y}^{test}$ (hidden), and the goal is to compute the MLE of $p(\boldsymbol{\beta}, \mathbf{y}^{test} | \mathbf{X}, \mathbf{y}^{train}, \sigma_W, \sigma, \alpha_S, \alpha_T)$. According to Eq (4), that is

$$\{\hat{\boldsymbol{\beta}}, \hat{\mathbf{y}}^{test}\} = \text{argmin}_{\boldsymbol{\beta}, \mathbf{y}^{test}} \sum_{s,t} \frac{(y_{s,t} - \boldsymbol{\beta}^T \mathbf{x}_{s,t})^2}{2\sigma^2} + \frac{\|\boldsymbol{\beta}\|_2^2}{2\sigma^2_{\boldsymbol{\beta}}} + \sum_{s,t}\sum_{s',t'} \phi(y_{s,t}, y_{s',t'})). \quad (5)$$

We can scale the terms so that $\sigma = 1$ without changing the solution. Then the problem becomes $\text{argmin}_{\boldsymbol{\beta}, \mathbf{y}^{test}} \sum_{s,t} \frac{(y_{s,t} - \boldsymbol{\beta}^T \mathbf{x}_{s,t})^2}{2} + \frac{\|\boldsymbol{\beta}\|_2^2}{2\sigma^2_{\boldsymbol{\beta}}} + \sum_{s,t}\sum_{s',t'} \phi(y_{s,t}, y_{s',t'}))$. This optimization problem is nonconvex. Our strategy is to alternately optimize $\boldsymbol{\beta}$ and $\mathbf{y}^{test}$ while keeping the other fixed, as follows. Given $\mathbf{y}^{test}$, the optimization problem for $\boldsymbol{\beta}$ is $\text{argmin}_{\boldsymbol{\beta}} \sum_{s,t} \frac{(y_{s,t} - \boldsymbol{\beta}^T \mathbf{x}_{s,t})^2}{2} + \frac{\|\boldsymbol{\beta}\|_2^2}{2\sigma^2_{\boldsymbol{\beta}}}$. This is a ridge regression problem where $C = \frac{1}{\sigma^2_{\boldsymbol{\beta}}}$ is the regularization parameter that trades-off the predictive error and the complexity of the model. This problem can be solved in closed form as

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X} + C\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}. \quad (6)$$

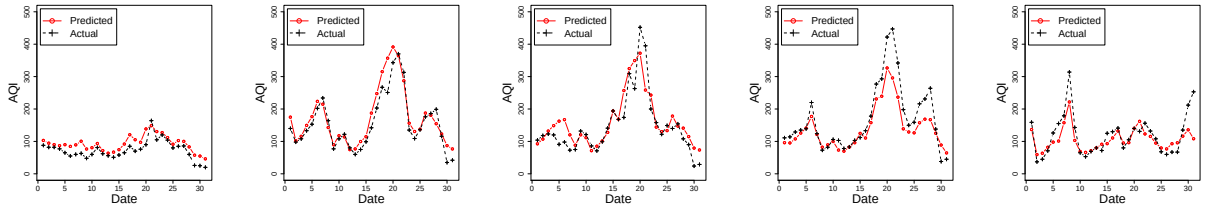
Optimizing y^{test} given $\boldsymbol{\beta}$ can be formulated as $\text{argmax}_{\mathbf{y}^{test}} \sum_{s,t} \frac{(y^{test}_{s,t} - \boldsymbol{\beta}^T \mathbf{x}_{s,t})^2}{2} + \sum_{s \sim s'}\sum_t \frac{1}{2}\alpha_S(y^{test}_{s,t} - y^{train}_{s',t})^2 + \frac{1}{2}\sum_{s,t} \alpha_T(y^{test}_{s,t} - y^{test}_{s,t+1})^2$. Letting the gradient zero, we obtain the system of equations for $\mathbf{y}^{test}$: $(I + A)\mathbf{y}^{test} = \mathbf{b}$, where $\mathbf{b}$ is a vector with $b_{s,t} = \frac{\alpha_S \sum_{s' \sim s, s' \in train} y^{train}_{s',t} + \boldsymbol{\beta}^T \mathbf{x}_{s,t}}{\alpha_S \sum_{s' \sim s, s' \in train} 1 + 1}$, and A is a matrix with the element at row (s,t) and column (s',t') defined as

$$a_{s,t,s',t'} = \begin{cases} -\alpha_S & \text{if } s \sim s', t = t' \\ \sum_{\tilde{s} \sim s} \alpha_S + \sum_{\tilde{t} \sim t} \alpha_T & \text{if } s = s', t = t' \\ -\alpha_T & \text{if } s = s', t \sim t' \end{cases}.$$

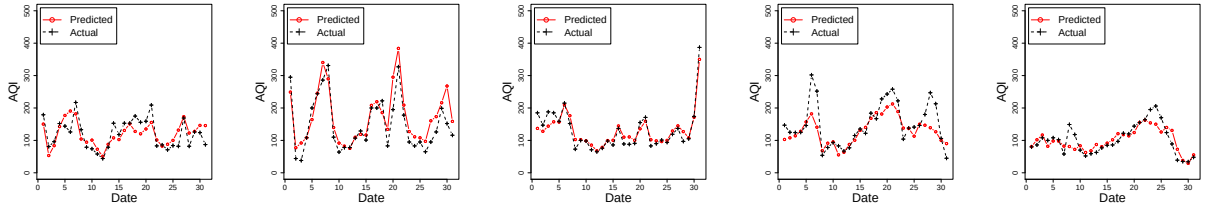
The optimal $\mathbf{y}^{test}$ is given by

$$\mathbf{y}^{test} = (I + A)^{-1}\mathbf{b}. \quad (7)$$

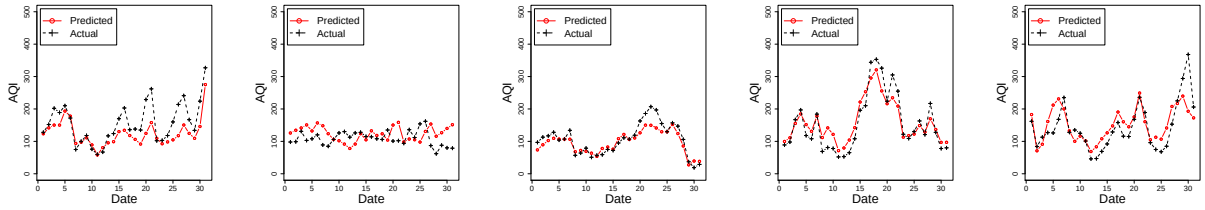
In summary, the inference algorithm is in Algorithm 1.



(a) Fuzhou (b) Hangzhou (c) Ningbo (d) Jinhua (e) Shenyang

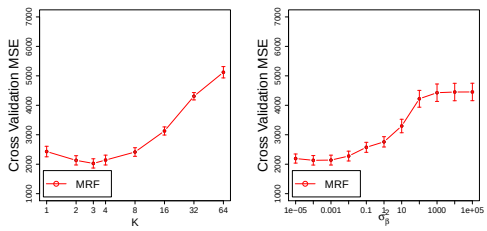


(f) Taiyuan (g) Tianjin (h) Xian (i) Zhuzhou (j) Beihai



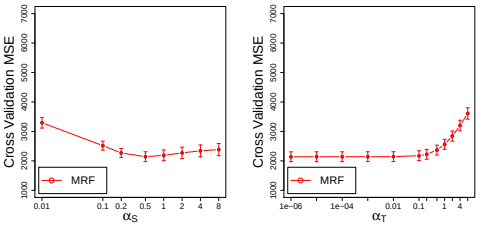
(k) Xiangyang (l) Xining (m) Shaoguan (n) Yangzhou (o) Zhengzhou

Fig. 3: Predicted AQI and actual AQI for 15 test cities from November 18 to December 18, 2013



(a) CV for K

(b) CV for σ_β^2



(c) CV for α_S

(d) CV for α_T

Fig. 2: Average MSE of folds in CV for four parameters

C. Cross Validation

In our MRF model we have five parameters: the number of iterations T, the number of neighbors K, the variance of the Gaussian prior $\sigma_{\boldsymbol{\beta}}^2$, the strength of spatial correlation α_S, and the strength of temporal correlation α_T. For computational efficiency, we simply set $T = 1$. That is, we only do one iteretion in the above algorithm. The other four parameters are tuned by 5-fold CV. The tuning curve of each parameter with the other parameters fixed at their best values is in Figures 2(a)–2(d). The best parameters are $K = 3$, $\sigma_{\boldsymbol{\beta}}^2 = 10^{-3}$, $\alpha_S = 0.5$ and $\alpha_T = 0.0$. Note that $\alpha_T = 0.0$ means that exploiting temporal correlation does not help improve the prediction performance. This may due to that the temporal unit (a whole day) is large compared to the time scale of typical AQI fluctuations. For exact AQI prediction, the fluctuations between days are too large (see Figure 3).

Algorithm 1 Inference for MRF

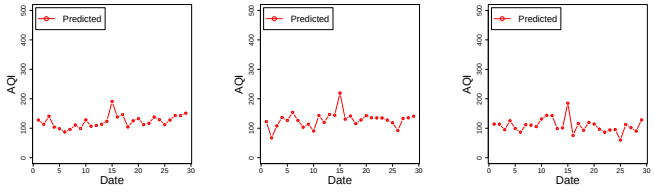
Require: $\mathbf{X}$, $\mathbf{y}^{train}$, σ_β, σ, α_S, α_T and T
  $\hat{\mathbf{y}}^{test} \Leftarrow \mathbf{0}$
  for $i = 1$ to T **do**
    update $\hat{\boldsymbol{\beta}}$ using Eq (6)
    update $\hat{\mathbf{y}}^{test}$ using Eq (7)
  end for
  return $\hat{\boldsymbol{\beta}}, \hat{\mathbf{y}}^{test}$

D. Experimental Results on the Test Set

We show the MRF's MSE on the test set in Table VI. The overall MSE is 2312, which is the best among the machine learning models we considered. Therefore, combining Weibo content and spatiotemporal correlation together improves performance. For the (city, day) bins with no heavy pollution (AQI$<$ 200), the predictions do not deviate more than 40 (on average). Therefore, our method rarely make false heavy air pollution predictions on good air quality days. For the bins with severe pollution (AQI$>$ 300), our predictions of AQI deviate no more than 110 (on average). So we seldom make false negative predictions. The errors for high AQI days are larger than the errors for lower AQI days because the training

data with high AQI (days with bad air quality) is fewer than the data with low AQI (days with good air quality).

To visualize our MRF model's predictions, we randomly selected 15 different cities from the test sets of five runs. We show the prediction AQI curves and the actual AQI curves from 11/18/2013 to 12/18/2013 in Figure 3. Our predicted AQI curves are close to the actual curves. Occasionally, on severely polluted days (AQI$>$ 300) the predictions are lower than the actual values. However, even in this case our prediction is still useful in that it predicts all of them as heavily polluted days (AQI$>$ 200). We also point out the similarity between that the actual AQI curves in three nearby cities Hangzhou, Ningbo, and Jinhua (see Figure 3(b), Figure 3(c) and Figure 3(d)), a fact that we exploited using spatial correlation potentials in our MRF. However, another nearby city, Yangzhou, in Figure 3(n) does not has similar AQI with Hangzhou. This shows that small cities' AQI information cannot always be predicted by their nearby big cities. The AQI can be influenced by many factors, not only geographical distance.



(a) Luoyang (b) Zhuji (c) Guilin

(d) Dunhuang (e) Lijiang (f) Dali

Fig. 4: Predicted AQI for several cities with no official AQI information.

E. Out-of-Sample Predictions on Cities Without AQI Monitoring Stations

To demonstrate that our method can help predict the air quality in places without AQI monitoring stations, we used our MRF model to predict the AQI for 26 additional cities that currently lack official air quality monitoring. The Weibo data collection and processing procedure was identical to that in Section II except that the study period of this dataset was from 01/17/2014 to 02/14/2014, which is after and does not overlap with the 108-city study.

We show the predicted AQI curves for several such cities in Figure 4. Because there is no official public AQI information on these 26 cities, we cannot judge our model prediction by comparing it with the true AQI. However, we are able to give some indirect evidence to justify our predictions. First, Figures 4(a)–4(c) all have a peak AQI value near the middle of the study period (the 15th day or 16th day). We note that the 15th day in the study period is Chinese new year's eve. Traditionally, people on this day celebrate with fireworks, which usually emit a lot of air pollutants. We hypothesize that the peak AQI may be attributed to these fireworks emissions. The Chinese Ministry of Environmental Protection proclaimed heavy pollution on that day because of fireworks in almost all major cities, which supports our hypothesis[10]. Also, the predicted AQI for Dunhuang increased during the 25th–29th days in the study period (see Figure 4(d)). According to news reports, Dunhuang had a dust storm in 2014 during that period, which probably contributed to the air pollution[11]. Finally, the air quality in Lijiang (see Figure 4(e)) looks much better than other cities. This is consistent with the impression that Lijiang is a tourist destination with good air quality.

VII. Discussion

We presented several estimators for air quality based on Weibo text content and spatiotemporal correlation between cities. Our methods complement physical AQI monitoring by monitoring stations. They may be particularly attractive for regions without monitoring stations. Our information source is inexpensively crawled from social media.

Our MRF model can easily combine other information sources, including the topography of the areas and the weather. Particularly interesting are the photos users post to social media. For future work, we may exploit the different behaviors of people from large cities and small cities on the social networks. Also, different cultures in different regions in China may be considered. We are also interested in predicting AQI, which depends heavily on human activities and weather. Weather can always be predicted. Therefore, understanding the pattern of human activities by social media may help predicting AQI.

Acknowledgment

This work was supported in part by National Science Foundation grant IIS-1148012.

References

[1] Paul M Aoki, RJ Honicky, Alan Mainwaring, Chris Myers, Eric Paulos, Sushmita Subramanian, and Allison Woodruff. A vehicle for research: using street sweepers to explore the landscape of environmental community action. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 375–384. ACM, 2009.

[2] Jiaoyan Chen, Huajun Chen, Guozhou Zheng, Jeff Z Pan, Honghan Wu, and Ningyu Zhang. Big smog meets web science: Smog disaster analysis based on social media and device data on the web. In *Proceedings of the 23th International World Wide Web Conference*, 2014.

[3] Richard Honicky, Eric A Brewer, Eric Paulos, and Richard White. N-smarts: networked suite of mobile atmospheric real-time sensors. In *Proceedings of the Second ACM SIGCOMM Workshop on Networked Systems for Developing Regions*, pages 25–30. ACM, 2008.

[4] Sameera Poduri, Anoop Nimkar, and Gaurav S Sukhatme. Visibility monitoring using mobile phones. *Annual Report: Center for Embedded Networked Sensing*, pages 125–127, 2010.

[5] Jun-Ming Xu, Aniruddha Bhargava, Robert Nowak, and Xiaojin Zhu. Socioscope: Spatio-temporal signal recovery from social media. In *Machine Learning and Knowledge Discovery in Databases (ECML/PKDD)*, pages 644–659. Springer, 2012.

[6] Yu Zheng, Furui Liu, and Hsun-Ping Hsieh. U-air: When urban air quality inference meets big data. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1436–1444. ACM, 2013.

[10] http://www.mep.gov.cn/gkml/hbb/qt/201401/t20140131_267406.htm

[11] http://gansu.gansudaily.com.cn/system/2014/03/20/014933184.shtml