Inferring urban air quality based on social media

Article *in* Computers Environment and Urban Systems · July 2017

DOI: 10.1016/j.compenvurbsys.2017.07.002

CITATIONS
21

READS
451

6 authors, including:

Yandong Wang
Wuhan University
48 PUBLICATIONS **546** CITATIONS
SEE PROFILE

Xiaokang Fu
Wuhan University
20 PUBLICATIONS **188** CITATIONS
SEE PROFILE

Ming-Hsiang Tsou
San Diego State University
126 PUBLICATIONS **2,668** CITATIONS
SEE PROFILE

Xinyue ye
Texas A&M University
318 PUBLICATIONS **4,778** CITATIONS
SEE PROFILE

Computers, Environment and Urban Systems 66 (2017) 110–116

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

journal homepage: www.elsevier.com/locate/ceus

Inferring urban air quality based on social media

Yan-dong Wang [a,b,*], Xiao-kang Fu [a], Wei Jiang [a], Teng Wang [a], Ming-Hsiang Tsou [c], Xin-yue Ye [d]

[a] State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China
[b] Collaborative Innovation Center for Geospatial Information Technology, Wuhan 430079, China
[c] Department of Geography, San Diego State University, San Diego, CA 92101, United States
[d] Computational Social Science Lab, Department of Geography, Kent State University, Kent, OH 44240, United States

ARTICLE INFO

Article history:
Received 27 November 2016
Received in revised form 30 May 2017
Accepted 10 July 2017
Available online xxxx

Keywords:
Air quality
Ensemble model
Feature extraction
Social media
Sina Weibo

ABSTRACT

Outdoor air pollution is a serious environmental problem in many developing countries; obtaining timely and accurate information about urban air quality is a first step toward air pollution control. Many developing countries however, do not have any monitoring stations and therefore the means to measure air quality. We address this problem by using social media to collect urban air quality information and propose a method for inferring urban air quality in Chinese cities based on China's largest social media platform, Sina Weibo combined with other meteorological data. Our method includes a data crawler to locate and acquire air-quality associated historical Weibo data, a procedure for extracting indicators from these Weibo and factors from meteorological data, a model to infer air quality index (AQI) of a city based on the extracted Weibo indicators supported by meteorological factors. We implemented the proposed method in case studies at Beijing, Shanghai, and Wuhan, China. The results show that based the Weibo indicators and meteorological factors we extracted, this method can infer the air quality conditions of a city within narrow margins of error. The method presented in this article can aid air quality assessment in cities with few or even no air quality monitoring stations.

1. Introduction

Air pollution control has become an urgent task in developing countries with rapid industrialization and urbanization like China facing increasingly serious air pollution problems (Huang et al., 2014). Air quality monitoring, allows decision makers and planners to understand urban pollution problems, and is the first step toward air pollution control and mitigation (Zheng, Liu, & Hsieh, 2013). Ideally, urban air quality is measured by a network of air quality monitoring stations. Most cities in the world, however, do not have any monitoring stations at all and thus lack any means to measure air quality. As of May 2015, only 16,700 monitoring stations have been built in 65 countries according to http://aqicn.org/sources.

Fortunately, there are other ways to get air quality information based on data various collected in cities without monitoring stations. These methods have been intensively studied over the past decade. For example, Air quality information can be inferred from remote sensing data (Martin, 2008; Van Donkelaar et al., 2010), from a crowdsensing based air quality monitoring system (Dutta, Chowdhury, Roy, Middya, & Gazi, 2017), by utilizing surveillance cameras as sensors (Zhang, Ma, Fu, Liu, & Zhang, 2016), or by using online photos (Li, Huang, & Luo, 2015). Researchers also are considering the use of data fusion technologies to combine a variety of these data, such as the use of remote sensing data combined with meteorological data (Xu & Zhu, 2016). Moumtzidou et al. (2016) present an open platform, which collects multimodal environmental data related to air quality from several sources including official open sources, social media, and citizens. All these efforts will help provide improved monitoring information that to support effective air pollution abatement and mitigation measures.

Social media, which disseminates popular ideas, opinions, and feelings, can be regarded as a social sensor of geographical phenomena in the era of big data (Sakaki, Okazaki, & Matsuo, 2010). Social media sites such as Facebook, Twitter, and Flickr, produce a large amount of user-generated data daily. Those data contain much useful information, attracting the attention of many researchers hoping to apply social media data to solve intractable problems across many domains. Researchers have applied social media data to understand dynamic urban processes (Ferrari, Rosi, Mamei, & Zambonelli, 2011) and to uncover human activity patterns (Caverlee, Cheng, Sui, & Kamath, 2013; Liu, Sui, Kang, & Gao, 2014; Zhi et al., 2016). Social media has been used to predict disease diffusion (Achrekar, Gandhe, Lazarus, Yu, & Liu, 2011; Nagel et al., 2013; Salathé & Khandelwal, 2011), to predict election outcomes (Tsou et al., 2013; Tumasjan, Sprenger, Sandner, & Welpe, 2010), and for detecting disasters (De Longueville, Smith, & Luraschi, 2009; Sakaki et al., 2010; Yates & Paquette, 2011).

Subjective human perceptions of air quality recorded in social media; such as smells in the air, might reflect objective air quality, and

* Corresponding author at: No.129 Luoyu Road, Wuhan 430079, China.
E-mail address: ydwang@whu.edu.cn (Y. Wang).

http://dx.doi.org/10.1016/j.compenvurbsys.2017.07.002
0198-9715/

thus, provide a new way to obtain urban air quality information (Smid, Mast, Tromp, Winterboer, & Evers, 2011). Researchers point out the possibility of using social media for air quality assessment. Du et al. (2016) mined pollutants data from social media and structured sources by using association rules, clustering, and classification to discover knowledge on air quality from a health standpoint, and found that knowledge discovered by data mining is useful in predictive analysis. Ni, Huang, and Du (2017) proposed a correlation analysis model for PM2.5 combining physical and social media data. An Autoregressive Integrated Moving Average Time Series model was applied to predict PM2.5 in a short-term time series. Tao, Kokas, Zhang, Cohan, and Wallach (2016) identify the key terms whose frequency was most correlated with PM2.5 levels and used these terms to construct an "Air Discussion Index" (ADI) for inferring daily PM2.5 levels based on the content of social media messages. They found a strong correlation between the ADI and measured PM2.5 in big cities.

There are several social media applications specifically for monitoring urban air quality in the public interest. Wang, Paul, and Dredze (2015) found that the volume of air pollution messages on Sina Weibo, a twitter-like microblogging site widely used in China could be used to infer the annual particle pollution levels. Jiang, Wang, Tsou, and Fu (2015) find that the filtered social media messages are strongly correlated to the AQI and can be used to monitor the daily air quality dynamics to some extent. Mei et al. (2014) applied a Markov Random Field model, to monitor the daily air quality index (AQI) by exploring the relationship between the terms in messages and AQI. Their monitoring method can predict low AQI values with the error less than 40 (on average) and high AQI values with the error less than 110 (on average). Nevertheless, the average error margin of their predictions is relatively large.

The objective of this article is to present a novel way to support air quality assessment in cities with few or no air quality monitoring stations. In this article we describe our method for inferring urban air quality based on social media and meteorological data; and discuss our validation case study experiments. We designed web crawlers to collect historical Weibo data on the topic of air quality. After filtering this data, we calculated two Weibo indicators reflecting subjective good and bad air quality information, based on regularities in Weibo message posts. We identified seven meteorological factors; including temperature, humidity, barometer pressure, wind speed, and weather conditions that affect urban air quality. Our Weibo indicators for good and bad air quality were integrated with meteorological factors using a Gradient Tree Boosting (GTB), a machine learning method, to infer the AQI of a city. Three cities were included in case study experiments: Beijing, Shanghai, and Wuhan, China. Our case studies show that air quality as inferred by our method follows the same trend as the AQI measurements from monitoring stations with narrow margins of error. These results demonstrate that the proposed method is useful for urban air quality assessment.

This article is organized as follows: Section 2 describes the study area and data. Section 3 describes the method. Section 4 contains experimental results. The work and outcomes are discussed in Section 5, and summarized in Section 6.

2. Study area and data

2.1. Study area

We chose three Chinese cities— Beijing, Shanghai, and Wuhan—as our study areas. In recent years, many cities in China have experienced periods of severe air pollution, most notably, the two big cities, Beijing and Shanghai (Kaiman, 2013; Wong, 2013). In addition, Wuhan is the largest city in central China was included as a research area. These three cities are located in central, coastal, and northern China, thus constituting a cross section of major metropolitan regions in China and suitable as our study areas.

2.2. Data

We used three types of data, Sina Weibo messages, meteorological information, and AQI measurements. Sina Weibo messages were collected by a web crawler and Sina Weibo application programming interface (API). The crawler simulated advanced Sina Weibo search requests, collecting Weibo messages containing specific keywords. To obtain detailed Weibo message information from the collected messages, we used the Sina Weibo API and specified the parameters to include Weibo identifying information (ID) extracted from the crawler responses.

Meteorological data were obtained from the online Weather Underground data services (http://www.wunderground.com/history). These include hourly meteorological data such as temperature, humidity, barometer pressure, wind speed, wind direction, and weather conditions, such as rain and haze. The AQI data came from the Ministry of Environmental Protection of China. The AQI data were the daily pollution measurement announcements, monitored by air quality stations. Sina Weibo and meteorological data were used to infer urban air quality; the AQI measured by air quality stations was used as the ground truth data.

Sina Weibo is currently the main micro-blogging service platform in China. As of December 2012, the number of registered Sina Weibo users exceeded 500 million, accounting for 37% of the total population JOSH, 2013). According to the Study Report on China's Social Application User Behaviour in 2015, 47.5% of the Sina Weibo users will micro-blog every day, and microblogging has become an important social media in their lives (China Internet Network Information Center, 2016). As of September 30, 2015, the number of active daily users (DAU) reached 100 million. Among these active Weibo users, the gender ratio is relatively balanced. Young people between the ages of 17 and 33 years old are main Sina Weibo user group, accounting for 83% of all the active Weibo users. Users in the economically developed areas such as Pearl River Delta, Yangtze River Delta and Beijing and cities with large populations accounted for a large proportion (Sina & Weibo Data Center, 2016).

We collected a sample of Weibos using the keywords: "空气污染" (air pollution), "雾霾"(smog), and"口罩"(mask), yielding 537,660 Weibos for Beijing from March to August 2014, 8584 Weibos for Wuhan during January 2014, and 19,127 Weibos for Shanghai during January 2014. Our data set included 565,371 messages in total.

The microblogging behaviours of Weibo users regarding air pollution is directly relevant on the credibility and generalizability of our work and data pre-processing. Our previous work explored how closely microblogging behaviours of Weibo users corresponded to air pollution condition in (Jiang et al., 2015), finding that the original individual messages, which created by social media users expressing their own personal opinions, are highly correlated to the AQI, while retweets were not strongly related to the AQI. In this study, only geotagged original individual messages containing information with location coordinates from mobile phones and other mobile terminal positioning devices, which might reflect the true feelings of users about their surroundings, were used in this study. In this way the data were filtered of retweeted messages and noises such as Weibos that were posted by special applications including "未通过审核应用" (Unapproved Application) and "皮皮时光机" (Pipi Time Machine). These were developed for special purposes

Table 1
Characteristics of Sina Weibo data used in our work.

	Total messages	Total users in messages	Total geotagged messages	Total users in geotagged messages
Beijing	537,660	388,083	73,494	33,089
Shanghai	19,127	15,989	881	837
Wuhan	8584	7393	819	777

such as marketing and produce large amounts of advertising or irrelevant Weibos. Sina Weibo data as described in Table 1.

3. Methodology

3.1. Constructing indicators from Weibo messages

We constructed two indicators for Weibo user perceptions of air quality using messages containing words related to good and bad air quality and normalizing them for further analysis in our model. The Weibo text messages were manually divided into two groups: group one included messages about bad air quality, and group two include messages about good air quality.

We defined the amount of group one Weibo messages as $W^T_{BAD}(t)$, and the amount of group two Weibo messages as $W^T_{GOOD}(t)$, the process for calculating the two indicators, f_{BAD} and f_{GOOD}, is as follows:

$$f_{BAD}(t) = \frac{W^T_{BAD}(t)}{U^t C(h(t))} \quad (1)$$

$$f_{GOOD}(t) = \frac{W^T_{GOOD}(t)}{U^t C(h(t))} \quad (2)$$

where U^T is the number of active Weibo users in a city estimated by the non-themed Weibo messages in that city. The C (h (t)) was the probability of Weibo messages posted at time t on a weekday.

Since Weibo messages posting activity varies temporally and spatially; weekly patterns and the number of active Weibo users in a city were used to normalize the quantity of group one Weibo messages.

Fig. 1(A) and 1(B) shows the relationships between AQI and the two Weibo air quality measures, f_{BAD}, and f_{GOOD} extracted from Beijing data sets. In this figure, the Weibo measure, f_{GOOD}, is larger when the AQI is small and f_{BAD}, is larger when the AQI is big, and is correlated to AQI.

3.2. Constructing factors from meteorological data

The diffusion of air pollutants is influenced by meteorological conditions at a location. We identified seven conditions that contribute to air pollution diffusion: temperature, humidity, barometer pressure, wind speed, wind direction, and weather such as rain and haze. We calculated daily mean values of the temperature, humidity, barometer pressure, and wind speed factors. In addition, we also calculated the rain and haze weather factors as the times they appeared in a day. We calculated the wind direction as the variance of the wind distribution of a day according to following formula:

$$f_{VANE}(t) = \frac{1}{N}\sum_{i=1}^{N}\left(d_i(t) - \bar{d}(t)\right)^2 \quad (3)$$

where d was calculated as the times the wind blew in a certain direction during a day. For this calculation, we defined 16 wind directions by dividing 360 degrees into 16 parts for a total of 16 wind directions.

Fig. 1(C) to 1(I) show the relationships between AQI and different meteorological factors. The relationships between humidity, pressure, temperature, and wind speed factors and AQI are weakly associated and similar to the results found in (Zheng et al., 2013). However, the scatter plot in 1(F) suggests that haze is correlated to AQI. Rainy weather correlated with AQI negatively as pollutants are absorbed by heavy rains. Wind direction also shows a negative correlation with AQI. When the wind is concentrated in one direction, it can easily to blow away air pollutants in that direction. Moreover, the influence of wind speed, wind direction, and rain on AQI often last until the next day.

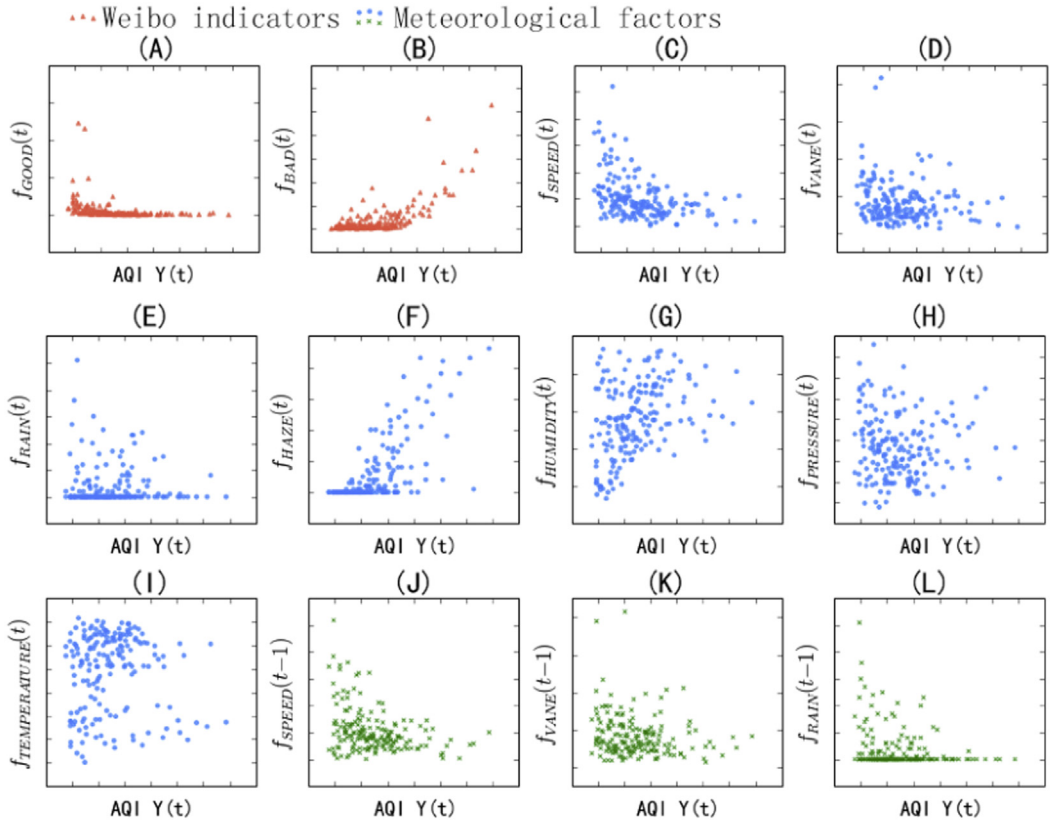

Fig. 1. The relationship between the extracted Weibo indicators or meteorological factors and AQI. The x-axis indicates AQI (denoted as Y (t)), the y-axis indicates the Weibo indicators or meteorological factors.

3.3. Inference model

The objective of our study is to extract air quality information from social media and meteorological data. We used a regression model, F, to infer the air quality information from a variety of data sources. The F model infers urban daily air quality based on the Weibo indicators $\boldsymbol{F}_W$ and meteorological factors $\boldsymbol{F}_q$ we extracted. In this article, an ensemble model, GTB, was deployed to handle the extracted indicators and factors.

GTB builds an additive model in a forward stage-wise fashion (Eq. (4)):

$$F_m(x) = F_{m-1}(x) + \nu\gamma_m h_m(x) \tag{4}$$

$$\mathbf{x} = \left(F_W(t), F_q(t), F_q(t-1)\right) \tag{5}$$

Here, $h_m(\boldsymbol{x})$ are decision trees that are usually called weak learners in the context of boosting. The vector $\boldsymbol{x}$ (Eq. (5)) consists of all of the Weibo indicators and meteorological factors at time t and some of meteorological factors (i.e., wind speed, wind direction, and rain) at time t-1. In the first stage during the iteration process, the initial value of the model was set as the average value of target AQI. The γ_m and ν values represent the lengths of steps in each stage and learning rates in all stages.

At each stage, the decision tree $h_m(\boldsymbol{x})$ is chosen to minimize the loss function L given the current model F_{m-1} and its fit $F_{m-1}(\boldsymbol{x}_i)$.

$$F_m(x) = F_{m-1}(x) + \nu \underset{h}{\operatorname{argmin}} \sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) - h(x)) \tag{6}$$

The loss function is chosen as the least squares function in this article, then $h_m(\boldsymbol{x})$ was solved as a minimization problem for the steepest descent, in which the direction was the negative gradient of the loss function.

$$F_m(x) = F_{m-1}(x) + \nu\gamma_m \sum_{i=1}^{n} \nabla_F L(y_i, F_{m-1}(x_i)) \tag{7}$$

where the step length γ_m is chosen using line search:

$$\gamma_m = \arg\min_{\gamma} \sum_{i=1}^{n} L\left(y_i, F_{m-1}(x_i) - \gamma \frac{\partial L(y_i, F_{m-1}(x_i))}{\partial F_{m-1}(x_i)}\right) \tag{8}$$

The hyperparameters for the model (Model iterations M, Tree depth S, and learning rate ν) were optimized with a grid search method, which is simply an exhaustive search method using a manually specified subset of the hyperparameter space of the learning algorithm, k-fold cross-validation was used to estimate the generalization performance in each parameter search.

4. Results

We implemented our method in case studies of Beijing, Shanghai, and Wuhan, three major cities in China. The Weibo indicators F_W and meteorological factors F_q we extracted were input GTB as explanatory variables to infer urban air quality. The daily AQI data monitored by air quality stations and announced by the Ministry of Environmental Protection of China were used as ground truth data. The hyperparameters of the model (the learning rate v, the number of iterations M, and the depth of the decision tree S) were optimized through the grid search and 5-fold cross-validation method on Beijing dataset. When the hyperparameters were optimized, the model was trained with 70% of the Beijing dataset, randomly shuffled and tested with the rest of the Beijing dataset. We also used the model trained on the Beijing dataset to infer the AQI in Shanghai and Wuhan, based on the Weibo indicators F_W and the meteorological factors F_q we extracted from Shanghai and Wuhan datasets. The inferred AQI were compared to the AQI obtained from the Ministry of Environmental Protection of China (ground truth data) to show the validity of our method in other China major cities.

4.1. Training and testing on Beijing data set

As three important hyperparameters in GTB, the learning rate ν, the number of iterations M, and the depth of the decision tree S must be determined and optimized. A detailed discussion of hyperparameter optimization is found in (Hastie, Friedman, & Tibshirani, 2001; Ridgeway, 2005); their research demonstrates that a learning rate ν strongly interacts with the number of iterations M. Smaller values of learning rate ν require larger numbers of iterations M to maintain a constant training error. Empirical evidence suggests that setting the learning rate ν to a small constant value (e.g. learning rate $<= 0.1$) and choosing the numbers of iterations M by early stopping reduces training error. Typically, the values for the depth of decision tree S works well when set to values between four and eight. We empirically, set the subset of hyperparameter space to M (200, 300,400,700), ν (0.005, 0.01, 0.05), S (2, 3, 4, …, 12); values consistent with Jiang et al. (2015). The hyperparameters of the model (Model iterations M, Tree depth S, and learning rate ν) were optimized as 400, 0.01, and 7, respectively through the grid search method. With that hyperparameters set, the mean coefficient of determination was calculated to be 0.84, and the mean RMSE was calculated to be 19 through the 5-fold cross-validation. These outcomes showed that the model has good data interpretation capability with small error margins.

When the hyperparameters were optimized, the model was trained with 70% of the Beijing dataset, randomly shuffled and tested with the rest of the dataset. Fig. 2 is a scatter plot of AQI observed by monitoring stations and AQI inferred by our model. The points are distributed diagonally, illustrating a strong correlation between the AQI inferred from the model and Observed AQI at monitoring stations.

4.2. Air quality inference results on shanghai and Wuhan

In order to illustrate the usefulness of the proposed method, we used it to infer the air quality of Shanghai and Wuhan in January 2014. This inferred air quality was compared to the observed AQI at monitoring stations. A comparison of the observed and inferred AQI for Shanghai is shown in Fig. 3; the RMSE between observed and inferred AQI was 28.07. The AQI varied widely over time, from about 50 to 250. These inferred values show the same overall trend as the observed AQI values.

The comparison of the observed and inferred AQI for Shanghai is shown in Fig. 4; the RMSE between observed and inferred AQI values

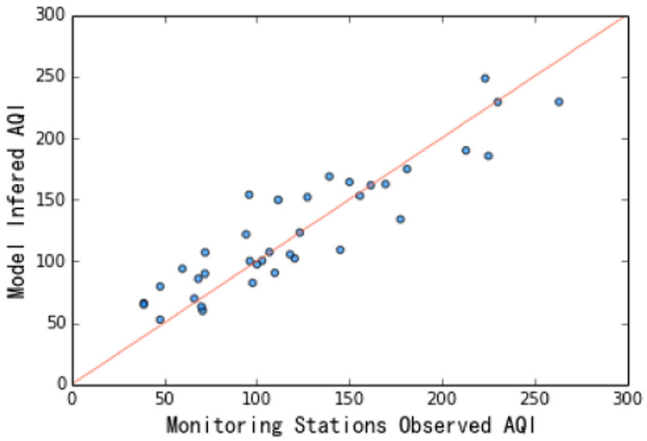

Fig. 2. The comparison scatter plot of observed and inferred AQI.

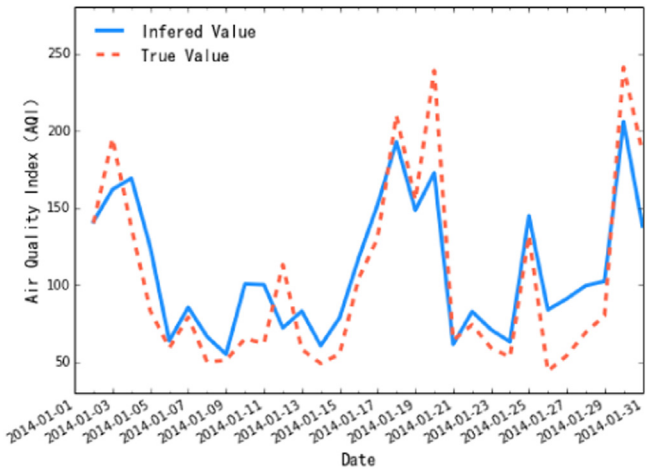

Fig. 3. The comparison of observed and inferred AQI of Shanghai in January 2014.

was 47.44. Although there is a great deal of error in the modelled AQI values on January 4 and January 29, nevertheless, the inferred values still show the same overall trend as the observed AQI values.

When comparing the model applications for these two cities as illustrated in Fig. 3 and Fig. 4, it is apparent that the model fit better in Shanghai than in Wuhan. The higher error in the inferred AQI values for January 4 in Wuhan can be understood by examining the Weibo messages. Fig. 5 shows the word cloud for this day. The most frequently appearing word was "雾霾"(smog), which suggests that the air quality on January 4 was bad. The word cloud also contained "流星雨"(meteor showers), "流量"(flow), "帐篷"(tents), and other words. Upon further examination we found that the Quadrantids meteor shower reached its peak at about 1 am (Molau et al., 2014). There were fog and smog at that time, in contrast to clear skies at midnight. So stargazers watching the meteor shower were likely to have posted more messages about smog on Weibos.

The differences between the applications of our model in the two cities possibly were related to variation in the meteorological data from them as well as differences in AQI. We compared the relationship between meteorological data and AQI in Wuhan with Beijing to understand the results of our model. As shown in Fig. 6, there was a negative correlation between rainy weather feature and AQI in Beijing dataset but none in Wuhan dataset. On January 29, the AQI in Wuhan reached 340 while there was a great deal of rain the day before, January 28. This did not correspond to the knowledge model trained with the Beijing dataset, resulting in greater error.

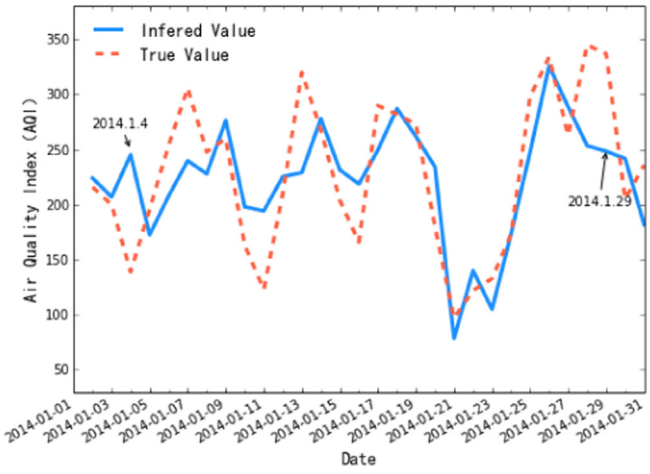

Fig. 4. The comparison of observed and inferred AQI of Wuhan in January 2014.

Fig. 5. The world cloud of Weibo text of Wuhan in January 4, 2014.

5. Discussion

The development of research on methods for extracting relevant information from social media feeds is still in its infancy. Our study contributes to this emerging body of research by exploring new ways to pre-process and extract meaningful indicators from social media feeds, and combining easily obtained data such as meteorological data to enhance the robustness and accuracy of an inference model. In our research, GTB is inference model particularly well suited to mix different types of data. The methods presented in this paper will provide a way to support air quality assessment in cities without air quality monitoring stations. Governments should consider these types of information sources concerning air quality to be more responsive and better serve the public interest. Environmental researchers can also deploy this method as an ancillary source of data in their studies.

5.1. Extracting relevant indicators from social media feeds

We followed a general pre-processing approach based on fetching keywords in Weibos, removing retweeted and special app posted Weibos. In addition, only Weibos that containing location coordinate information were retained. The Weibos that contained location coordinate information had less noise. We also considered the textual content of Weibo messages, the numbers of active Weibo users, and the weekly patterns of Weibo postings to fully exploit the information contained in Weibo data. In our study, we used a manual classification method to divide Weibo text messages into two categories. Further research is required to reduce the labour and time costs of manual classification.

5.2. Air quality inferring from multi-source data

Data from multiple sources were integrated in one model. In this study, we combined easily obtained meteorological data to enhance the robustness and accuracy of the inference model. Our study showed that temperature, humidity, barometer pressure, wind speed, wind direction, and weather factors contributed to more accurate AQI estimation results. We analysed the changes over time in relationship between meteorological factors and air quality and found that only three meteorological factors had an obvious relationship to air quality.

5.3. Generalization ability of inferring model

GTB is an ensemble method that combines the predictions of several base decision trees in order to improve applicability and robustness over a single tree (Friedman, 2001; Friedman, 2002). We used a GTB to model the relationship between extracted indicators and AQI in a city. The model can mix different types of data and has good robustness.

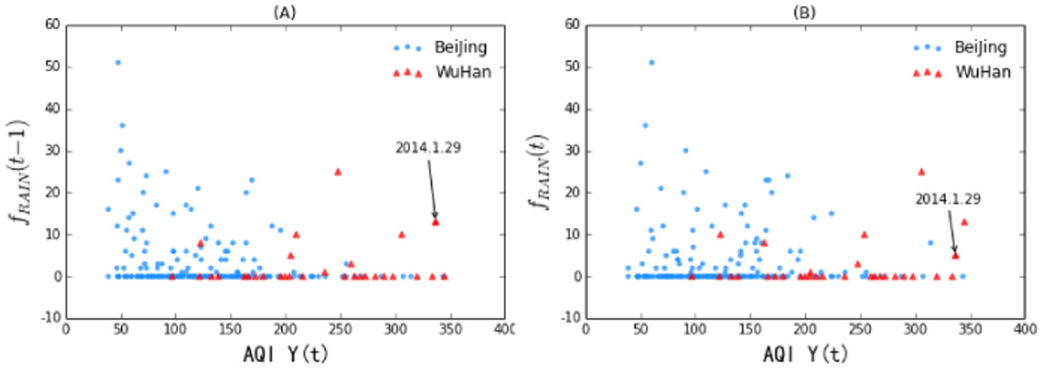

Fig. 6. Different rain meteorological factor and AQI relationship between Beijing and Wuhan.

Experiment results showed that our model fitted Shanghai but not Wuhan. One of the main reasons for this discrepancy was that there were different relationships between meteorological factors and AQI in different cities.

6. Conclusions

Social media can be used as a supporting information source for detecting dynamic real world environmental changes. In addition, social media is free, easy to obtain and might be a real-time human sensor source of air quality information. As more and more people participate in the use of social media, they will generate huge volumes of social media messages, reflecting their feelings and perceptions about their surroundings. These messages, if used properly, can be helpful for environmental monitoring and protection. In our research, we inferred air quality based on social media messages and meteorology data. In case studies of Shanghai and Wuhan, we found that social media could indeed be used as an information source to monitor air quality of a city. Weibo message texts, the number of Weibo active users in a city, and weekly patterns of Weibo postings can help to make use of the information that contained in Weibo data. We found that easily accessible data related to air quality, such as meteorological data, can be successfully integrated with social media to infer urban air quality. Our model, trained in one city, could also be applied in another city to support air quality monitoring.

Human sensors (derived from social media) are similar to actual environmental sensors, and provide only a narrow slice of air quality information. These data are affected by the number and type of samples. In our case, we illustrate that results obtained from social media are nevertheless, useful for detecting the change of air quality in major Chinese cities. In the future work, more cities will be included to further verify the effectiveness of our method and explore how different demographic compositions of Weibo users found in different cities influence the credibility and generalizability of results obtained from these data. In this work, we only use three keywords, "空气污染"(air pollution), "雾霾"(haze), and"口罩"(mask), to collect Sina Weibo data, to illustrate the usage of social media and meteorological data to assist in urban air quality assessment. More key words, including "PM2.5", "AQI", "particulate matter" will be included in our future work.

Moreover, further research is needed to improve this model. Special events relating to the environment in Weibo such as meteor showers need to be taken into account in this model. Differences in local environments need to be better represented in the data used to construct the relationships between the meteorological factors and air quality for different cities. This would increase the applicability of the model. Finally, the Weibo messages posting may be influenced by the publicly available AQI information.

References

Achrekar, H., Gandhe, A., Lazarus, R., Yu, S. -H., & Liu, B. (2011). Predicting flu trends using twitter data. *Proc. of the 2011 IEEE conference on computer communications workshops (INFOCOM WKSHPS), Shanghai* (pp. 702–707).

Caverlee, J., Cheng, Z., Sui, D. Z., & Kamath, K. Y. (2013). Towards geo-social intelligence: Mining, analyzing, and leveraging geospatial footprints in social media. *IEEE Data Engineering Bulletin., 36*(3), 33–41.

China Internet Network Information Center (2016). The Study report on China 's social application user behavior in 2015. http://www.cnnic.cn/hlwfzyj/hlwxzbg/sqbg/201604/P020160722551429454480.pdf/ (Accessed 17.05.03).

De Longueville, B., Smith, R. S., & Luraschi, G. (2009). Omg, from here, i can see the flames!: A use case of mining location based social networks to acquire spatio-temporal data on forest fires. *Proc. of the 2009 ACM international workshop on location based social networks, New York* (pp. 73–80).

Du, X., Emebo, O., Varde, A., Tandon, N., Chowdhury, S. N., & Weikum, G. (2016). Air quality assessment from social media and structured data: Pollutants and health impacts in urban planning. *Proc. of the 2016 IEEE 32nd international conference on data engineering workshops (ICDEW), Helsinki* (pp. 54–59). http://dx.doi.org/10.1109/ICDEW.2016.7495616.

Dutta, J., Chowdhury, C., Roy, S., Middya, A. I., & Gazi, F. (2017). Towards Smart City: Sensing air quality in city based on opportunistic crowd-sensing. *Proc. of the 18th international conference on distributed computing and networking, New York. 42.* (pp. 1–42). http://dx.doi.org/10.1145/3007748.3018286 (6).

Ferrari, L., Rosi, A., Mamei, M., & Zambonelli, F. (2011). Extracting urban patterns from location-based social networks. *Proc. of the 3rd ACM SIGSPATIAL international workshop on location-based social networks, New York* (pp. 9–16).

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics, 29*(5), 1189–1232.

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics and Data Analysis, 38*(4), 367–378.

Hastie, T., Friedman, J., & Tibshirani, R. (2001). *The elements of statistical learning* (2th ed.). New York: Springer New York. http://dx.doi.org/10.1007/978-0-387-21606-5 (Chapter 10).

Huang, R. J., Zhang, Y., Bozzetti, C., Ho, K. F., Cao, J. J., Han, Y., ... Zotter, P. (2014). High secondary aerosol contribution to particulate pollution during haze events in China. *Nature, 514*(7521), 218–222.

Jiang, W., Wang, Y., Tsou, M. H., & Fu, X. (2015). Using social media to detect outdoor air pollution and monitor air quality index (AQI): A geo-targeted spatiotemporal analysis framework with Sina Weibo (Chinese twitter). *PloS One, 10*(10), e0141185.

JOSH, O. N. G. (2013). China's Sina Weibo grew 73% in 2012, passing 500 million registered accounts. https://thenextweb.com/asia/2013/02/21/chinas-sina-weibo-grew-73-in-2012-passing-500-million-registered-account/ (Accessed 17.05.03).

Kaiman, J. (2013). Chinese struggle through "airpocalypse" smog. https://www.theguardian.com/world/2013/feb/16/chinese-struggle-through-airpocalypse-smog/ (Accessed 17.05.05).

Li, Y., Huang, J., & Luo, J. (2015). Using user generated online photos to estimate and monitor air pollution in major cities. *Proc. of the 7th international conference on internet multimedia computing and service, New York. 79.* (pp. 1–79). http://dx.doi.org/10.1145/2808492.2808564 (5).

Liu, Y., Sui, Z., Kang, C., & Gao, Y. (2014). Uncovering patterns of inter-urban trip and spatial interaction from social media check-in data. *PloS One, 9*(1), e86026.

Martin, R. V. (2008). Satellite remote sensing of surface air quality. *Atmospheric Environment, 42*(34), 7823–7843.

Mei, S., Li, H., Fan, J., Zhu, X., & Dyer, C. R. (2014). Inferring air pollution by sniffing social media. *Proc. of the 2014 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM), Beijing* (pp. 534–539).

Molau, S., Kac, J., Crivello, S., Stomeo, E., Barentsen, G., & Goncalves, R. (2014). Results of the IMO video meteor network-January 2014. *WGN, Journal of the International Meteor Organization, 42*(2), 83–86.

Moumtzidou, A., Papadopoulos, S., Vrochidis, S., Kompatsiaris, I., Kourtidis, K., Hloupis, G., ... Keratidis, C. (2016). Towards air quality estimation using collected multimodal environmental data. *In collective online platforms for financial and environmental*

awareness (pp. 147–156). Cham: Springer. http://dx.doi.org/10.1007/978-3-319-50237-3_7.

Nagel, A. C., Tsou, M. H., Spitzberg, B. H., An, L., Gawron, J. M., Gupta, D. K., ... Sawyer, M. H. (2013). The complex relationship of realspace events and messages in cyberspace: Case study of influenza and pertussis using tweets. *Journal of Medical Internet Research*, *15*(10), e237.

Ni, X. Y., Huang, H., & Du, W. P. (2017). Relevance analysis and short-term prediction of PM2.5 concentrations in Beijing based on multi-source data. *Atmospheric Environment*, *150*, 146–161. http://dx.doi.org/10.1016/j.atmosenv.2016.11.054.

Ridgeway, G. (2005). *Generalized boosted models: A guide to the gbm package*.http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.151.4024.

Sakaki, T., Okazaki, M., & Matsuo, Y. (2010). Earthquake shakes twitter users: Real-time event detection by social sensors. *Proc. of the 19th international conference on world wide web, New York* (pp. 851–860).

Salathé, M., & Khandelwal, S. (2011). Assessing vaccination sentiments with online social media: Implications for infectious disease dynamics and control. *PLoS Computational Biology*, *7*(10), e1002199.

Sina & Weibo Data Center (2016). The Report of Weibo Users' Development in 2015. http://data.weibo.com/report/reportDetail?id=333/ (Accessed 17.05.03).

Smid, H., Mast, P., Tromp, M., Winterboer, A., & Evers, V. (2011). Canary in a coal mine: Monitoring air quality and detecting environmental incidents by harvesting twitter. *Proc. of the ACM CHI'11 extended abstracts on human factors in computing systems, New York* (pp. 1855–1860).

Tao, Z., Kokas, A., Zhang, R., Cohan, D. S., & Wallach, D. (2016). Inferring atmospheric particulate matter concentrations from Chinese social media data. *PloS One*, *11*(9), e0161389. http://dx.doi.org/10.1371/journal.pone.0161389.

Tsou, M. H., Yang, J. A., Lusher, D., Han, S., Spitzberg, B., Gawron, J. M., ... An, L. (2013). Mapping social activities and concepts with social media (Twitter) and web search engines (Yahoo and Bing): A case study in 2012 US presidential election. *Cartography and Geographic Information Science*, *40*(4), 337–348.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment. *Proc. of the international AAAI conference on weblogs and social media, Washington, D.C. 10.* (pp. 178–185).

Van Donkelaar, A., Martin, R. V., Brauer, M., Kahn, R., Levy, R., Verduzco, C., & Villeneuve, P. J. (2010). Global estimates of ambient fine particulate matter concentrations from satellite-based aerosol optical depth: Development and application. *Environmental Health Perspectives*, *118*(6), 847–855.

Wang, S., Paul, M. J., & Dredze, M. (2015). Social media as a sensor of air quality and public response in china. *Journal of Medical Internet Research*, *17*(3), e22.

Wong, H. (2013). *Will be remembered as the year that deadly, suffocating smog consumed China. 2013.* (https://qz.com/159105/2013-will-be-remembered-as-the-year-that-deadly-suffocating-smog-consumed-china/Accessed 17.04.03).

Xu, Y., & Zhu, Y. (2016). When remote sensing data meet ubiquitous urban data: Fine-grained air quality inference. *Proc. of the 2016 IEEE international conference on big data (big data), Washington, DC* (pp. 1252–1261). http://dx.doi.org/10.1109/BigData.2016.7840729.

Yates, D., & Paquette, S. (2011). Emergency knowledge management and social media technologies: A case study of the 2010 Haitian earthquake. *International Journal of Information Management*, *31*(1), 6–13.

Zhang, Z., Ma, H., Fu, H., Liu, L., & Zhang, C. (2016). Outdoor air quality level inference via surveillance cameras. *Mobile Information Systems*, *2016*, e9825820. http://dx.doi.org/10.1155/2016/9825820.

Zheng, Y., Liu, F., & Hsieh, H. P. (2013). U-air: When urban air quality inference meets big data. *Proc. of the 19th ACM SIGKDD international conference on knowledge discovery and data mining, New York* (pp. 1436–1444).

Zhi, Y., Li, H., Wang, D., Deng, M., Wang, S., Gao, J., ... Liu, Y. (2016). Latent spatio-temporal activity structures: A new approach to inferring intra-urban functional regions via social media check-in data. *Geo-spatial Information Science*, 1–12.