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

{mei, hanli, fanj, jerryzhu, dyer}@cs.wisc.edu

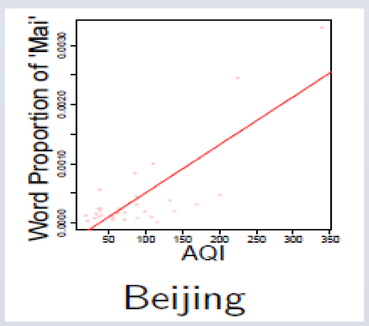
MOTIVATION

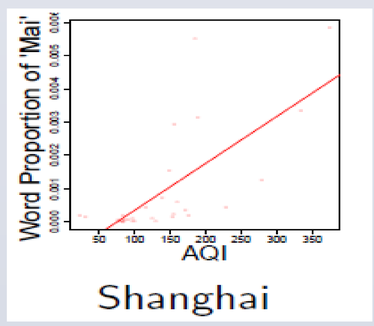
Air pollution is currently a big issue in China and elsewhere.

To deal with the air pollution, we first need to monitor it. However, physical monitoring stations are limited to large cities.

Cities without monitoring stations may also suffer air pollution

Can we use social media as another source to estimate Air Quality Index (AQI)?





We propose WeiboAQI, a complementary approach to monitoring AQI from social media posts by machine learning models.

DATA

Weibo Posts:

- All 108 cities in China with monitoring stations
- Time period from November 18 to December 18, 2013
- On average, we obtained about 1,380 posts in each (city and day) bin

AQI Information:

- Collect AQI information for these 108 cities every hour
- The daily AQI of each city is defined as the average of the AQI in the day and the city

Preprocessing

- Segment the Chinese text in each post
- Filter out all the stopwords and words with count <10
- Aggregate all the posts in one (city, day) bin as one document
- Represent each document as a bag-of-words vector

For spatiotemporal bin (s,t), $x_{s,t}$ is the bag-of-words vector of the pooled Weibo posts, for city s and day t.

$y_{s,t}$ is the daily average AQI.

For evaluation, we divided the cities as training cities S_{train} and test cities S_{test}.

Mean square error (MSE) between the estimated AQI $\widehat{y_{s,t}^{test}}$ and the actual AQI $y_{s,t}^{test}$ used to evaluate the performance:

$$MSE = \frac{1}{\#\,TestDataPoints} \sum_{s,t} (\hat{y}_{s,t}^{test} - y_{s,t}^{test})^2.$$

MACHINE LEARNING MODELS

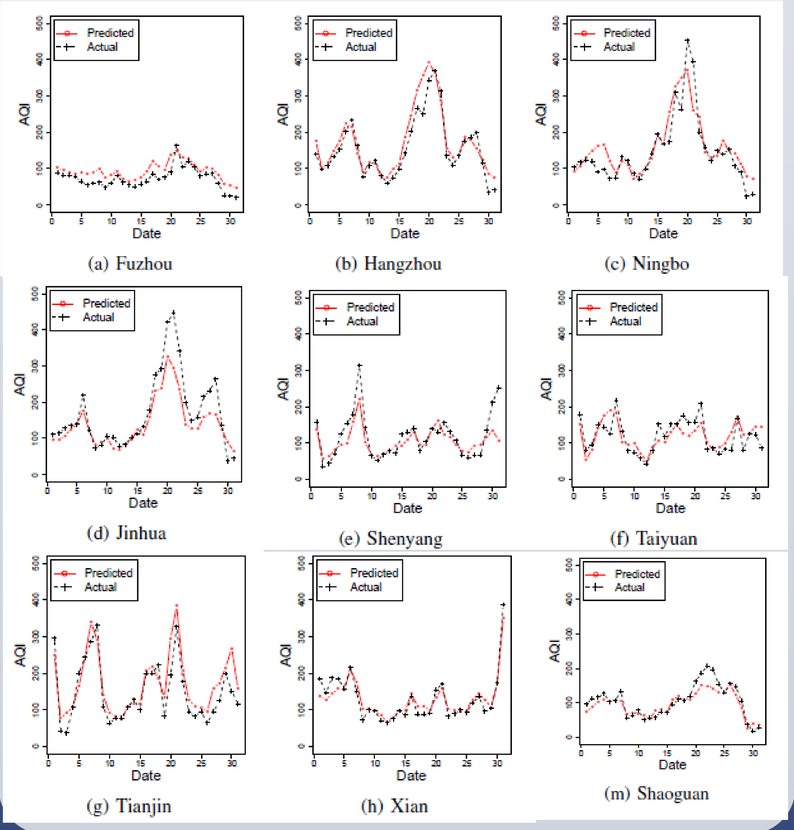
1. Linear regression model on Weibo bag-of-words features.
2. K nearest neighbor to predict the AQI of a city by average of nearest (geographically) K cities.
3. Combining linear regression model, spatiotemporal correlation in Markov random field model.

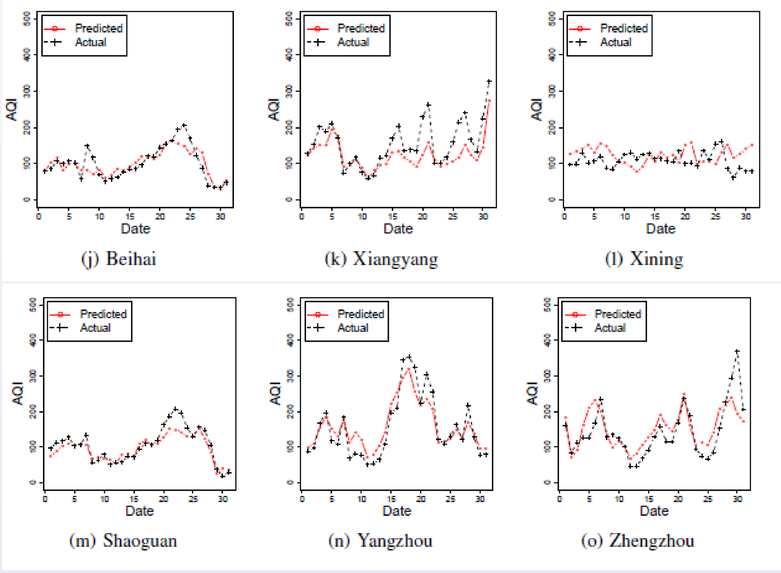
RESULTS

MSE of all three models

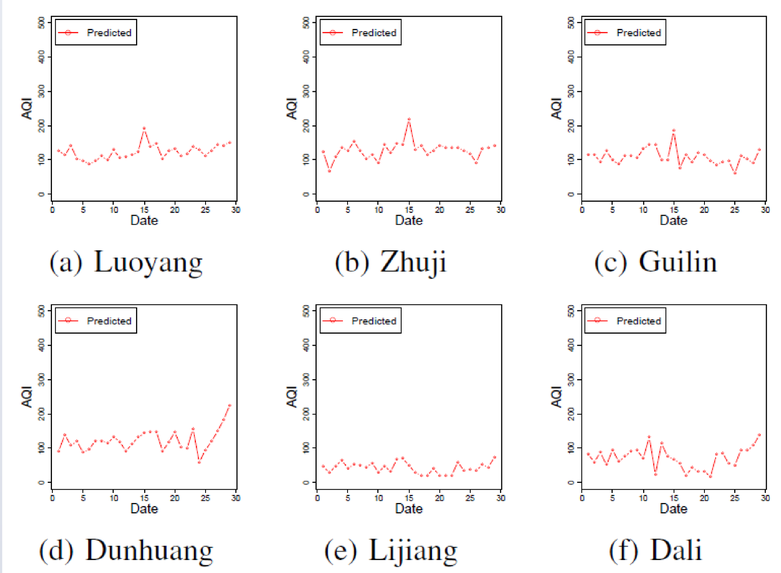
AQI	[0, 100]	(100, 200]	(200, 300]	(300, 1000)	All
MSE for Linear Regression	2231 ± 97	1101 ± 25	6990 ± 281	22904 ± 929	3469 ± 121
MSE for KNN	1336 ± 108	1910 ± 104	4396 ± 204	12607 ± 620	2646 ± 75
MSE for MRF	1534 ± 96	1150 ± 77	3878 ± 231	11710 ± 782	2312 ± 105

Predicted and actual AQI in test cities.





Predicted AQI in cities without AQI monitoring stations.



We are able to give some indirect evidence to justify our predictions:

- Figures (a-c) all have a peak AQI value near the middle of the study period (Chinese New Year). Heavy pollution is because of fireworks.
- The estimated AQI for Dunhuang increased during the 25th and 29th days in the study period. There is a dust storm during that period.
- The air quality in Lijiang (a famous tourist destination) looks much better than other cities.

CONCLUSION

- We estimate AQI based on social media by machine learning methods.
- It is a complement physical AQI monitoring stations for regions without stations
- Future work: forecast AQI
- Check our paper at http://pages.cs.wisc.edu/~jerryzhu/pub/airPollution.pdf