Large-scale and Automatic $\mathrm{PM}_{2.5}$ Data Monitoring: A Spatial Temporal Modeling Approach

This Version: June 28, 2020

Abstract

KEY WORDS:

1. INTRODUCTION

Air pollution is a crucial environmental issue globally. According to World Health Organization, in 2006, outdoor air pollution in both cities and rural areas caused 4.2 million premature deaths worldwide. Apart from being a major threat to public health, air pollution incurs huge amounts of economic loss. The Organisation for Economic Co-operation and Development (OECD) estimates that outdoor air pollution is costing its member countries 1.57 trillion USD per year in terms of prevention from premature deaths (OECD, 2014). With the rapid industrialization and intense energy consumption, China is troubled with severe air pollution, which leads to major environmental health problem as well as immense economic costs. According to China Ecological Environment Bulletin, 64.2% of prefecture-level cities in China failed to meet the urban air quality standards in 2018. As a result, it is estimated that 17% of nation's annual deaths are caused by heavy air pollution every year and that the costs of ambient outdoor air pollution mortalities mount to more than 8% of domestic GDP (Rohde and Muller, 2015; Roy and Braathen, 2017).

Particulate matter (PM) is considered as a major source of air pollution among all the pollutants. The notorious $PM_{2.5}$, which refers to small particulate matter of 2.5 microns or less in diameter, is believed to be detrimental in the following three major aspects to the general public. Firstly, exposure to fine particulate air pollution contributes to the increase of both natural-cause mortality and disease mortality (Pope III et al., 2002; Hoek et al., 2013; Beelen et al., 2014). Researches show that long exposure to high level of $PM_{2.5}$ causes increased risk for cardiovascular and respiratory disease, and cancers (Pope et al., 2006; World Health Origanization, 2018). Secondly, particulate matter can scatter and absorb sunlight. On one hand, it leads to visibility deterioration and further causes haze. On the other hand, accumulated particulate matter can obstruct photosynthesis, resulting in reduction on agricultural production (Gu et al., 2018; Zhou et al., 2018). Thirdly, $PM_{2.5}$ pollution also contributes to local climate changes. Studies show that particulate matter may alter aerosol and cloud properties and enhance heat island phenomenon in urban areas (Wang et al., 2015; Cao et al., 2016). With all negative impacts mentioned above, $PM_{2.5}$ pollution can result in huge amounts of direct or indirect economic loss (Qi et al., 2017; Xie et al., 2016). As a result, monitoring and controlling concentration level of $PM_{2.5}$ becomes a problem of fundamental importance.

To deal with the high-level $PM_{2.5}$ pollution, the Chinese government carried out a series of forceful measures and regulations. As early as 2013, State Council of China released the Action Plan for the Control of Air Pollution, which is a guideline document in controlling air pollution. Under the guidance of this plan, the Chinese government first set up 1436 monitoring stations equipped with remote quality control system in all prefecture-level cities (Kou, 2016). Furthermore, contribution to the control of $PM_{2.5}$ pollution was later incorporated into the performance evaluation system for the Chinese officials. Decrease in the level of pollution directly affects the amount of fiscal funds of local governments as well as the promotions of officials (Xinhua Net, 2014). In addition, since 2015, real-time $PM_{2.5}$ data of all stations are made public on the internet. China has built the largest air quality monitoring network in developing countries (Kou, 2016). With all measures mentioned above, the Chinese government has made great progress in air quality monitoring.

Particularly, the performance evaluation system for the Chinese government officials based on $PM_{2.5}$ pollution urges local government to take actions and therefore effectively controls the pollution level. However, such evaluation system at the same time allures some officials to risk themselves by faking the pollution data for political achievements. The country has at least witnessed two serious $PM_{2.5}$ data fraud cases so far. The first fraud case of automatic monitoring data of ambient air happened in Xi'an, 2016. The station staff deliberately blocked the detection devices hence interfered with the monitoring system. All seven defendants were sentenced to at least one year and three months imprisonment (Xinhua Net, 2017). Another data fraud case happened in Linfen, 2018, in which staff maneuvered the data reporting system (Legal Daily, 2018). Propelled by the continuously appeared data fraud cases, the Chinese government is determined to suppress $PM_{2.5}$ data fraud with complete and forceful laws and regulations, more frequent inspection as well as intensified punishment (Ministry of Ecology and Environment of the People's Republic of China, 2018). However, most of the inspection and monitoring are still human-powered, and such actions could be inefficient and costly. In this sense, we see the necessity of automation in data monitoring and detection of abnormal data behaviors, which will largely facilitate the supervision and improve the efficiency of the government.

We believe that the automatic monitoring is feasible for the following three reasons. Firstly, there exists temporal dependence for the $PM_{2.5}$ data. The concentration level of $PM_{2.5}$ at one particular time point should be correlated with its lagged values (Ağaç et al., 2015; Batterman et al., 2016). Secondly, the $PM_{2.5}$ data show certain spatial correlation. The concentration level of $PM_{2.5}$ in one particular monitoring station should be highly correlated to that of its neighbors (Wang et al., 2016). Thirdly, certain relationships can be found among different air pollutants. Particularly, there are 6 air pollutants recorded by the monitoring stations, which are $PM_{2.5}$, PM_{10}, SO_2, NO_2, CO and O_3. It is demonstrated that the other 5 pollutants are helpful in predicting $PM_{2.5}$ (Wang and Niu, 2009). To summarize, if we can take advantages of the spatial dependence, temporal dependence, and the relationships among pollutants, the modeling of $PM_{2.5}$ is feasible. This can surely help the automatic monitoring of $PM_{2.5}$ and further specifies outliers.

To this end, we propose a spatial temporal modeling approach. This approach contains three regression models, which can detect spatial, temporal, and content outliers respectively. For the spatial analysis, we fit a linear regression model relating the concentration level of $PM_{2.5}$ in one particular monitoring station to that of its neighbors. For the temporal analysis, an AR(1) model is built for the concentration level of $PM_{2.5}$, which is strongly suggested by the sample partial autocorrelation function (PACF). For the content analysis, concentration level of other 5 pollutants are chosen as predictors to fit a multiple linear regression model. Besides, some meteorological variables including atmospheric pressure, temperature, dew point, combined wind direction and wind speed are controlled in the above three models. By doing so, various dependence relationships can be effectively used. This helps us to construct a confidence interval for the normal $PM_{2.5}$ value. Accordingly, any value stays outside this interval should be suspectable for data quality issue and further inspection should be conducted.

The rest of the article is organized as follows. In Section 2, abundant descriptive analyses are conducted to get preliminary understanding of the air pollution data. Spatial and temporal correlations can be detected as discussed above. From Sections 3 to 5, spatial, temporal, and content analysis are conducted respectively. To be more specific, various regression models are built to fit $PM_{2.5}$ and further find out the outliers. In Section 6, we integrate the results and a labeling system for outliers is developed for automatic monitoring. Concluding remarks and some discussions are shown in Section 7.

2. DATA DESCRIPTION

The data adopted in this work are collected from 6 national controlled monitoring stations in the city of Shijiazhuang. As the capital city of Hebei and major industrial city in Northern China, Shijiazhuang is reported to be one of the most polluted cities nationwide(Learish, 2018). In this sense, the monitoring of $PM_{2.5}$ in Shijiazhuang is of great importance. Figure 1 exhibits the distribution of the 6 monitoring stations, namely, GaoXinQu (GXQ), RenMinHuiTang (RMHT), ShiJiGongYuan (SJGY), XiNanGaoJiao (XNGJ), XiBeiShuiYuan (XBSY), and ZhiGongYiYuan (ZGYY). Station XBSY locates in the north of the city, away from the other stations.

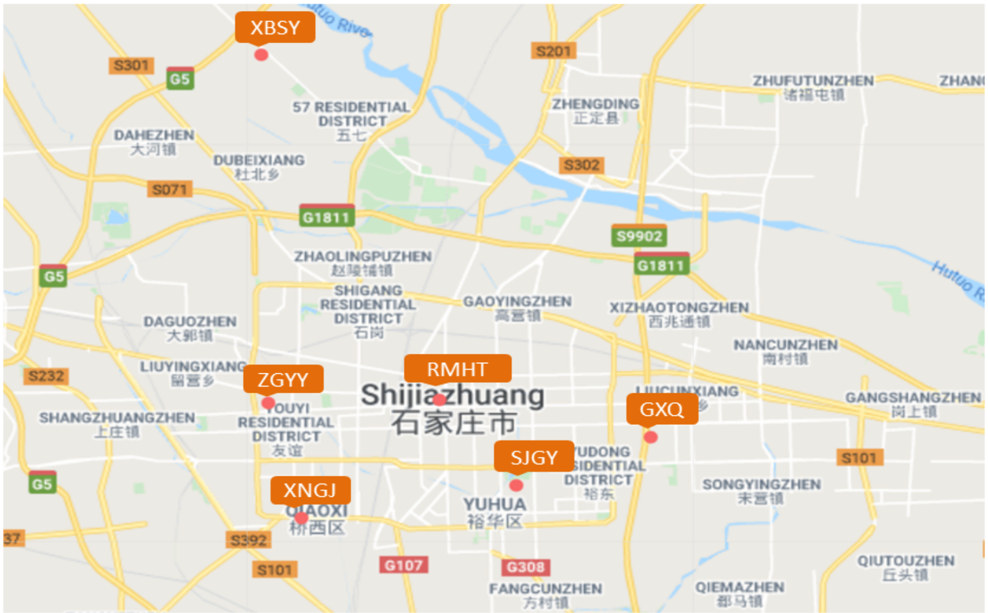


Figure 1: Geographic distributions of the 6 stations in Shijiazhuang. The XBSY station locates relatively far away from the other 5 stations.

The data are collected hourly from 2016.03.01 to 2017.02.28, including concentrations of 6 air pollutants (i.e., $PM_{2.5}$, PM_{10}, SO_2, NO_2, CO and O_3) and 5 meteorological variables (i.e., Atmospheric pressure, Temperature, Dew point, Combined wind direction and Wind speed).

It should be noted that there are missing data in the dataset, which can be caused by having observations lower than detecting threshold or other errors in the data collection. Detailed explanations of the variables as well as their missing rate are shown in Table 1. SO_2 has the highest missing rate due to its relatively low concentrations in magnitude.

Table 1: Detailed imformation and missing rate of variables observed by stations in Shijiazhuang. The missing rate is defined as the ratio of the number of missing observations over total possible number of observations for each variable. Take $PM_{2.5}$ as an example, the number of missing observations in $PM_{2.5}$ of all the 6 stations is 1290, and the number of possible observations is $N = 6 \times 365 \times 24 = 52,560$.

	Variable	Meaning	Unit	Missing Rate(%)
Air pollutants	$PM_{2.5}$	An air pollutant with a diameter of 2.5 micrometers or less	ug/m^3	2.45
	PM_{10}	An air pollutant with a diameter between 2.5 and 10 micrometers	ug/m^3	2.43
	SO_2	A gas produced by volcanoes and in various industrial processes	ug/m^3	5.40
	NO_2	An acutely toxic reddish-brown gas with a sharp smell	ug/m^3	1.30
	CO	A gas at room temperature produced by something organic burning	ug/m^3	1.32
	O_3	A pale blue gas with a distinctively pungent smell	ug/m^3	1.21
Meteorological variables	atmospheric pressure	The weight of the column of air above a designated area	hPa	1.95
	temperature	A measure of the hotness or coldness of the air around us	$^{\circ}C$	1.95
	dew point	The temperature to which air must be cooled for saturation to occur	$^{\circ}C$	1.95
	combined wind direction	The direction from which the wind blows	NE NW SE SW CV	1.55
	wind speed	The rate of motion of air	m/s	1.55

Since the main target of our work is to automatically monitor the $PM_{2.5}$ data, we then plot the histogram of $PM_{2.5}$ concentrations of all the 6 stations in Figure 2. The histogram is highly right skewed, with mean 111.27 and median 66.

——————————————————————-¿

Temporal Characteristics of $PM_{2.5}$

Firstly, by season. The boxplot shows the $PM_{2.5}$ of six stations by seasons. The right one is a time series diagram based on the average $PM_{2.5}$ of all stations per hour.

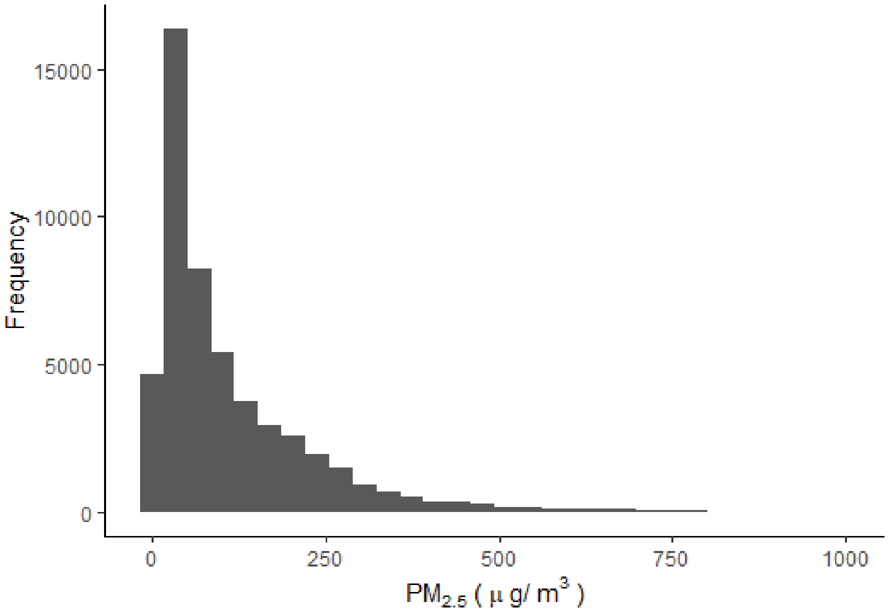
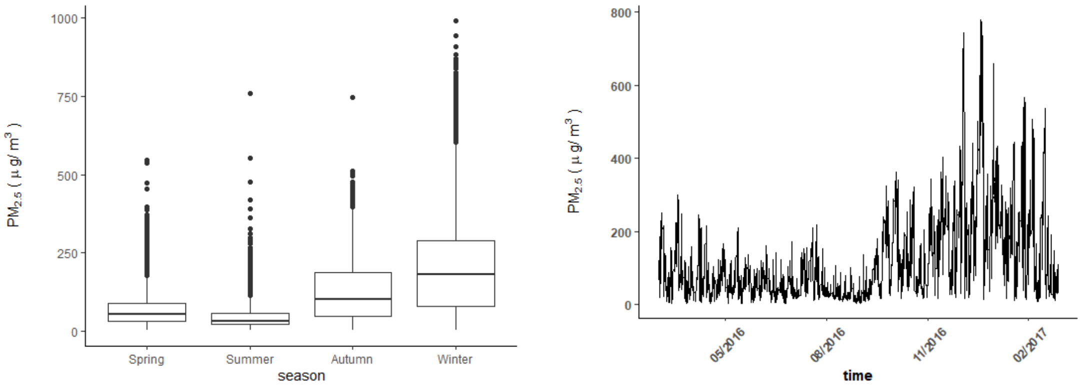


Figure 2: The histogram of PM$_{2.5}$ concentrations. Hourly data of 6 stations from 2016.03.01 to 2017.02.28 are combined to show.



(a) boxplot

(b) time series diagram

Figure 3: Boxplot of PM$_{2.5}$ concentration of six stations by seasons and the time series diagram based on the average PM$_{2.5}$ of all stations per hour.

Secondly, by hour. The scatter plot is employed to depict the relation of concentrations of the target station between time t and time t-1 (last hour) to illustrate the temporal dependence.

Spatial Characteristics of PM$_{2.5}$

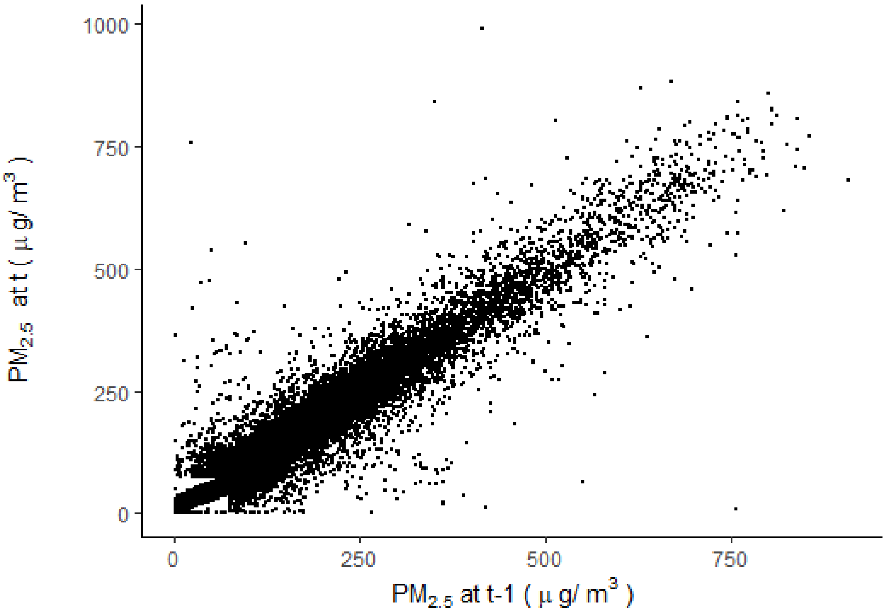


Figure 4: The scatter plot of $PM_{2.5}$ between adjacent times.

The average $PM_{2.5}$ of the 6 stations separately and seasonal average value of PM2.5 for the 6 stations is shown in the table.

Table 2: Seasonal average value of $PM_{2.5}$ for the 6 stations.

	GXQ	**RMHT**	**SJGY**	**XBSY**	**XNGJ**	**ZGYY**
Spring	67.70	73.88	72.23	65.55	62.58	63.14
Summer	44.56	48.83	50.64	49.63	48.04	43.79
Autumn	126.27	127.47	119.90	125.93	122.96	122.35
Winter	204.98	198.74	218.17	221.77	212.80	202.08
The whole year	110.05	111.83	113.28	114.97	110.60	107.01

Accordingly, the heatmap shows seasonal average value of $PM_{2.5}$ of different stationss.

The scatter plot is to depict the relation between $PM_{2.5}$ concentrations of the target station at time t and the weighted sum of concentrations of the rest of the stations at time t-1 (last hour) to illustrate the spatial dependence. The weights are proportional to the inverse of distance to the target station and sum up to 1. Figure 6 suggests that

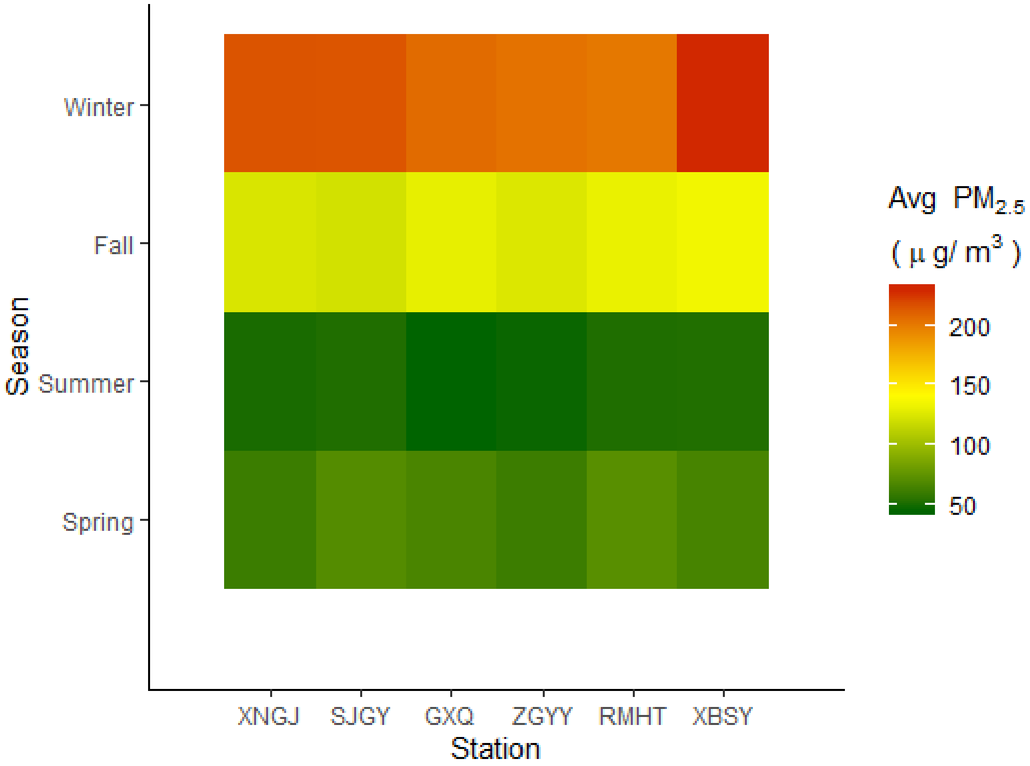


Figure 5: The heatmap of the seasonal average $PM_{2.5}$ of all the 6 stations.

the $PM_{2.5}$ concentrations have spatial dependence.

——————————————————————————¿

The Meteorological Variables

Lastly, we illustrate the distribution of meteorological variables. We bin the meteorological variables by quartiles and visualize the distributions with boxplots. Taking winter data as an example, Figure 7 shows that there is a strongly positive correlation between $PM_{2.5}$ concentrations and dew point while the correlations between $PM_{2.5}$ concentrations and atmospheric pressure, wind speed or temperature are slightly negative. $PM_{2.5}$ concentrations vary among different combined wind direction. The average of concentrations under northwest wind direction(NW) and calm wind(CV) are the two highest. The correlation coefficients between $PM_{2.5}$ concentrations and dew point, atmospheric pressure, wind speed or temperature are 0.56, -0.30, -0.25, -0.17. Therefore,

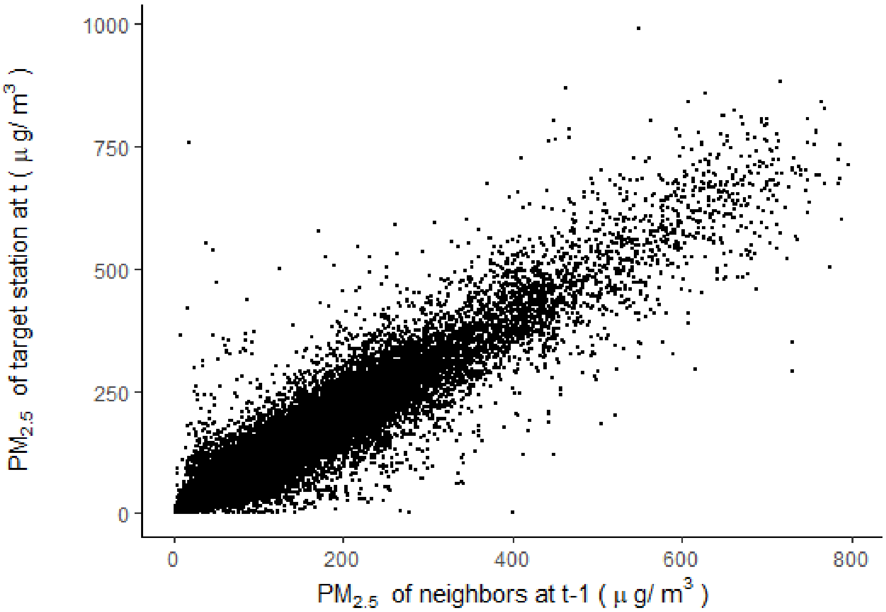
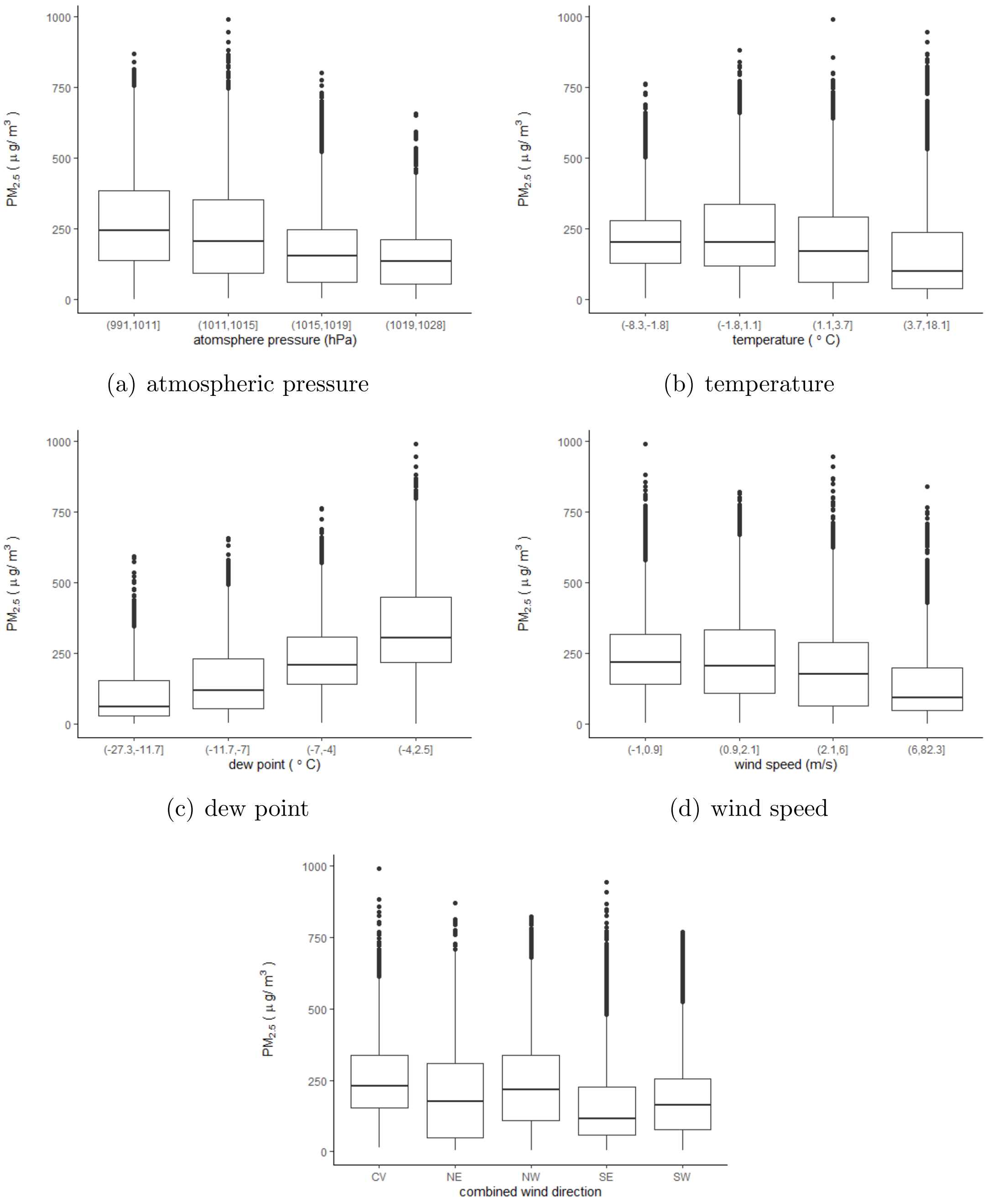


Figure 6: The scatter plot between PM$_{2.5}$ concentrations of the target station at time t and the weighted sum of concentrations of the rest of the stations at time t-1.

we include these meteorological variables into the regression models so as to eliminate the impact brought about by these variables.



(a) atmospheric pressure

(b) temperature

(c) dew point

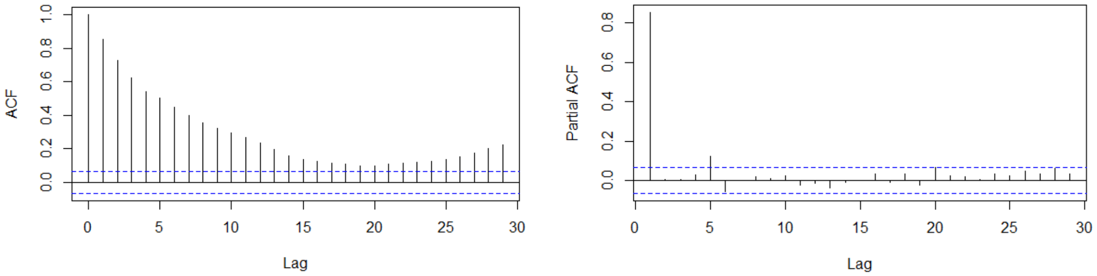
(d) wind speed

(e) combined wind direction

Figure 7: Boxplots of PM$_{2.5}$ concentrations and some meteorological variables. Winter data is chosen for instance.

(Professor Pan asked us to put it in the model part. So it's here for the time being)

We further calculate ACF and PACF based on the time series of $PM_{2.5}$ concentrations in Figure 8 to illustrate the temporal dependence. ACF shows slow decay while Partial ACF is truncated with a large value at lag 1, which suggests to fit an AR(1) model.



(a) acf

(b) pacf

Figure 8: ACF and PACF of $PM_{2.5}$ concentrations time series. A piece of continuous data without missing(GaoXinQu 2017.07.24-2017.08.31) is chosen for instance.

References

Ağaç, K., Koçak, K., and Deniz, A. (2015). Simulation And Forecasting of Daily Pm10 Concentrations Using Autoregressive Models In Kagithane Creek Valley, Istanbul. In *EGU General Assembly Conference Abstracts*, EGU General Assembly Conference Abstracts, page 5737.

Batterman, S., Xu, L., Chen, F., Chen, F., and Zhong, X. (2016). Characteristics of pm2.5 concentrations across beijing during 2013-2015. *Atmospheric Environment*, 145:104–114.

Beelen, R., Raaschounielsen, O., and Stafoggia, M. (2014). Effects of long-term exposure to air pollution on natural-cause mortality: an analysis of 22 european cohorts within the multicentre escape project. *The Lancet*, 383(9919):785–795.

Cao, C., Lee, X., Liu, S., Schultz, N., Xiao, W., Zhang, M., and Zhao, L. (2016). Urban heat islands in china enhanced by haze pollution. *Nature communications*, 7:12509.

Gu, X., Wang, L., and Zhuang, W. (2018). Reduction of wheat photosynthesis by fine particulate (pm2.5) pollution over the north china plain. *International Journal of Environmental Health Research*, 28(6):635–641.

Hoek, G., Krishnan, R. M., and Beelen, R. (2013). Long-term air pollution exposure and cardio-respiratory mortality: A review. *Environmental Health*, 12(1):43–43.

Kou, J. (2016). Air quality data is ”reliable” (green focus). `http://society.people.com.cn/n1/2016/1210/c1008-28939097.html`.

Learish, J. (2018). The most polluted cities in the world, ranked. `https://www.cbsnews.com/pictures/the-most-polluted-cities-in-the-world-ranked/30.html`.

Legal Daily (2018). The second case of environmental data fraud in china adjudicated. http://www.legaldaily.com.cn/index/content/2018-06/24/content_7576562.htm?node=20908.

Ministry of Ecology and Environment of the People's Republic of China (2018). Action plan for ecological environment monitoring quality supervision and inspection (2018-2020). http://society.people.com.cn/n1/2016/1210/c1008-28939097.html.

Ministry of Ecology and Environment of the People's Republic of China (2019). 2018 china ecological environment bulletin. http://www.mee.gov.cn/hjzl/sthjzk/zghjzkgb/201905/P020190619587632630618.pdf.

OECD (2014). The cost of air pollution. http://www.oecd.org/env/the-cost-of-air-pollution-9789264210448-en.htm.

Pope, Arden, C., and Dockery, D. W. (2006). Health effects of fine particulate air pollution: Lines that connect. *Journal of the Air & Waste Management Association*, 56(10):709–742.

Pope III, C. A., Burnett, R. T., Thun, M. J., Calle, E. E., Krewski, D., Ito, K., and Thurston, G. D. (2002). Lung Cancer, Cardiopulmonary Mortality, and Long-term Exposure to Fine Particulate Air Pollution. *JAMA*, 287(9):1132–1141.

Qi, Z., Chen, T., Chen, J., and Qi, X. (2017). Ambient pm 2.5 in china: its negative impacts and possible countermeasures. *Journal of the Air and Waste Management Association*, 68.

Rohde, R. and Muller, R. (2015). Air pollution in china: Mapping of concentrations and sources. *PloS one*, 10:e0135749.

Roy, R. and Braathen, N. A. (2017). The rising cost of ambient air pollution thus far in the 21st century. *OECD Environment Working Papers*, (124).

The State Council of China (2013). The action plan for the control of air pollution. `http://www.gov.cn/zwgk/2013-09/12/content_2486773.htm`.

Wang, C., Jiang, H., and et al (2016). A key study on spatial source distribution of pm2.5 based on the airflow trajectory model. *International Journal of Remote Sensing*, 37(24):5864–5883.

Wang, F., Guo, J., Huang, J., Min, M., Chen, T., Liu, H., Minjun, D., and Li, X. (2015). Multi-sensor quantification of aerosol-induced variability in warm clouds over eastern china. *Atmospheric Environment*, 113.

Wang, W. and Niu, Z. (2009). Var model of pm2. 5, weather and traffic in los angeles-long beach area. In *2009 International Conference on Environmental Science and Information Application Technology*, volume 3, pages 66–69. IEEE.

World Health Origanization (2018). Ambient (outdoor) air pollution. `https://www.who.int/en/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health`.

Xie, Y., Dai, H., Dong, H., Hanaoka, T., and Masui, T. (2016). Economic impacts from pm2.5 pollution-related health effects in china: A provincial-level analysis. *Environmental science and technology*, 50.

Xinhua Net (2014). China includes pm2.5 in the official assessment: ”reducing haze” becomes a new trend of political achievements. `http://www.xinhuanet.com/world/2014-05/30/c_126788602.htm`.

Xinhua Net (2017). Ministry of environmental protection responded to the environmental data fraud case in xi'an: A mechanism will be established to prevent and punish data fraud. `http://www.xinhuanet.com/politics/2017-06/17/c_1121162461.htm`.

Zhou, L., Chen, X., and Tian, X. (2018). The impact of fine particulate matter (pm 2.5) on china's agricultural production from 2001 to 2010. *Journal of Cleaner Production*, 131-141.