UNDERSTANDING MISSING DATA IN REAL-TIME POLLUTION MONITORING SYSTEM IN CHINA

A Thesis

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Master of Science

by

Congyan Han

August 2019

ABSTRACT

Using both remote sensing data on air pollution and publicly reported hourly $PM_{2.5}$ data from ground-level monitoring stations, this paper examines whether the quality of the publicly reported $PM_{2.5}$ is affected by selective reporting whereby high-level hourly pollution readings are dropped in the reported data. Our analysis shows that the contemporaneous level of air pollution measured by the Aerosol Optical Depth (AOD) has a negative relationship with the frequency of data missing. This relationship is weaker in dirty cities measured by the average AOD during the sample period and is reversed in very dirty cities.

Key Words: air pollution, real-time monitoring, missing value, satellite data, China

BIOGRAPHICAL SKETCH

Congyan Han is a Master student in the Dyson School of Applied Economics and Management at Cornell University. Email: ch884@cornell.edu.

This document is dedicated to all Cornell graduate students.

ACKNOWLEDGEMENTS

I would like to thank Prof. Shanjun Li for his tremendous support for my study and research. He helps me establish a framework for how to conduct a research, and motivates my interest in doing research and pursuing further study as a PhD student. I also wish to thank Prof. Ivan Rudik for his insightful comments and patience, which encourages me to keep trying and moving forward. Also, I am very grateful to Eric Zou, Lin Yang, Yuanning Liang and Jing Qian for their help and detailed suggestions. Finally, I want to express my gratitude to my parents who give me continuous support and encouragement during years of study.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

The quality of air quality data has been an issue that attracts both the public's and the government's attention in China[1]. In March 2016, it was reported[2] that the monitors of two monitoring stations in Xi'an were blocked by gauze to avoid high readings. The six monitoring stations in Linfen[3] were accused of manipulating air quality data by blocking or spraying water towards the monitors during April 2017 to March 2018. A recent scandal[4] in January 2018 was that the building of Environmental Protection Bureau in Shizhuishan was frozen when the officials sprayed water towards this building to improve the air quality around the monitors there.

In general, three ways to have better air quality data are commonly used, (a) strategy response to intermittent monitoring by behaving differently when being monitored and not being monitored(Zou (2018)), (b) falsifying data by not reporting the true concentration, improving the air quality just around the monitors or blocking the monitors, and (c) discarding data by not reporting the concentration or shutting down the monitors. This paper is aimed to check whether the third one exists, considering the time period covered by this data set and the corresponding policy background in China. The reason why this paper is focused on this case will be further explained in Chapter 2.

Air pollution data with low quality has many negative effects. First, it has a rather bad impact on the government's credibility, especially when there are some

[1]China to probe accuracy of its air pollution data. Some provincial governments have been manipulating figures to meet national standards, says minister.
https://www.downtoearth.org.in/news/china-to-probe-accuracy-of-its-air-pollution-data-49303

[2]http://news.cctv.com/2017/06/22/ARTIgJAAaCTZMXDUqRvkeWgs170622.shtml

[3]http://news.sina.com.cn/sf/news/ajjj/2018-08-06/doc-ihhhczfc4604136.shtml

[4]http://www.sohu.com/a/217980601_681337

other countries or institutes reporting air quality data at the same time[5] to the public with differences. Also, air quality data is an important factor taken into consideration by people when they are making the decisions about whether to go outside and whether they should use masks to protect themselves((Ghanem and Zhang (2014)).), which means wrong information will lead to potential loss in terms of health and social welfare, since air pollution has huge acute and chronic negative effect on health both in the short and long term, even at very low exposure(Brunekreef and Holgate (2002), Kampa and Castanas (2008), Chen et al. (2013b)). And many studies have already shown that in some cities in China, air pollution does lead to higher non-surgery outpatient visits and mortality(M.D. et al. (1995), Xu et al. (2000), Rohde and Muller (2015)). This kind of health damage will then lead to economic cost(Kan and Chen (2004)). Also, it makes the researches or policies based on it come to incorrect conclusions and applications.

There are several studies on quality of air pollution data. Based on intermittent monitoring data of air quality in United States, Zou (2018) shows that strategic responses exist and the widely used once-every-six-day monitoring schedule for outdoor particle pollution causes significant deterioration in air quality on unmonitored days compared to monitored days. And when it comes to air quality data in China, Ghanem and Zhang (2014) provide empirical evidence for data manipulation by testing the discontinuity around the cut-off for Blue-Sky Days, using self-reported PM_{10} data by Chinese cities over the period 2001-2010 as a proxy for API[6], using in-

[5]China Has No Good Answer to the U.S. Embassy Pollution-Monitoring. Lashing out at the U.S. only highlights the Chinese leadership's inability to clean up the country's air and further erodes their credibility with the public.
https://www.theatlantic.com/international/archive/2012/06/china-has-no-good-answer-to-the-us-embassy-pollution-monitoring/258447/

[6]Air Pollution Index. It is an indicator from 0 to 500, with 6 levels, instead of the pollution concentration. The higher the AQI value, the greater the level of air pollution. It converts the concentrations of PM_{10}, SO_2 and NO_2 into a single index by choosing the maximum of the indexes transformed from three pollutants.

visibility as a proxy for true air quality, with weather variables being controlled. Chen et al. (2013a) apply officially reported API data from 37 large cities in China during 2000-2009 and two proxies for air pollution(visibility data from China Meteorological Administration(CMA) and Aerosol Optical Depth(AOD) data from National Aeronautics and Space Administration(NASA)) and find the discontinuity at the threshold of Blue-Sky Days as well. In addition, they show that with higher pressure to achieve the target of exemplary city policy[7], the higher possibility a city that is about to win the award reports API or PM_{10} right below the threshold.

This paper investigates the quality of air pollution data in China in a different way, testing the patterns of missing values of air pollution data after 2012 in China. The studies above are focused on the discontinuity of API data in China before 2012. But the studies based on the time period after 2012 are rare, as well as the studies based on the patterns of missing values in air quality data. Exploration into data quality regarding missing values after 2012 is very necessary and meaningful, not just a study repeating a similar topic. This paper focuses on a different time period with many changes in environmental policy, monitoring, air quality standards. As a result, there are some differences. API used by previous studies is decided by the pollutant that has the highest index[8], and during the period they cover, it is mostly decided by PM_{10}. That is why they use PM_{10}[9]. However, AQI replaced API in 2012 as the main indicator for air quality. Except for the three indexes included in API, three more indexes, $PM_{2.5}$, O_3 and CO are added. AQI is to choose the maximum of the indexes of these six pollutant, which are calculated by piece-wise linear trans-

[7]The central government of China decides whether a city is "The National Environmental Protection Exemplary City" depending on four indicators, social economy, environmental quality, environmental construction and environmental management.

[8]$API = MAX(I_{SO_2}, I_{NO_2}, I_{PM_{10}})$

[9]PM_{10} is used instead of AQI because that they are going to check the discontinuity. But the index of pollutants that API depends on is not a linear transformation of pollution concentration. Ghanem and Zhang (2014) shows mathematically the calculation of AQI leads to discontinuity and can not be used for the test directly.

formation of the concentration. $PM_{2.5}$ mostly occurs as the pollutant deciding it. In addition, since the U.S. Embassy in Beijing reported $PM_{2.5}$ data to the public around 2009, it has became the pollutant that is the most eye-catching in China. Many studies focused on $PM_{2.5}$ are conducted, such as health effect, economic cost and source analysis. The government also tends to emphasize more on $PM_{2.5}$ than other pollutants. For example, in Thirteenth Five-Year Plan(covers 2016-2020)[10], $PM_{2.5}$ is highlighted and said to be reduced by 23.6% in 2015 compared to 2013 in the 74 cities under the first wave of monitoring. The general situation and several economic zones are also mentioned. This plan is a programmatic document in China and provides the direction for the government. The evaluations of many cities also emphasize a lot on this particular pollutant[11]. All these indicate $PM_{2.5}$ would be a representative pollutant for this time period. So in my study, it will be applied for analysis. Second, the new Ambient Air Quality Standards which came into effect at 2012 establishes a new national air quality monitoring system. It is said that the data will be uploaded automatically and remotely by the system without the potential interference by local officials, which means the data I use is not self-reported anymore and manipulations in terms of strategy response and falsifying data should have been eliminated theoretically. Based on this assumption, there will be differences of the incentives and measures for data manipulation between the time periods before 2012 and after 2012. Hence, it will be meaningful to conduct research based on air quality data after 2012. Moreover, if patterns of missing values do exist, only testing the accuracy of data is not enough because the data set could have already been biased. Also, to know what kind of manipulation exists could help policy makers and regulators to better manage the air quality monitoring and data reporting process. As a result, this paper is to check whether the quality of air pollution data is affected by selectively reporting

[10] $PM_{2.5}$ haven't been added into the air quality evaluation system when Twelfth Five-Year Plan came into effect.

[11] http://roll.sohu.com/20160224/n438349922.shtml

whereby discarding high readings.

CHAPTER 2

BACKGROUND

This part introduces the new air quality regulation in China and how environment is correlated to officials' promotion. By this, it is easier to understand why discarding data is feasible and why the officials may have the incentive for it.

2.1 New Ambient Air Quality Standards and Data Veracity

Economic growth and urbanization in China cause large pollution emission and many cities in China have been faced with severe air quality issues, not limited to major cities, with widespread source of pollution sources(Chan and Yao (2008), Rohde and Muller (2015)). As MEE[2012]NO.11[1] points out, pollution by NO_x, $VOCs$, O_3 and $PM_{2.5}$ is aggravating. The problem of PM_{10} and TSP pollution has not been fully solved. To protect and improve living environment, ecological environment and health, Ministry of Ecology and Environment(MEE) published Ambient Air Quality Standards (GB 3095-2012[2]) on Feb. 29, 2012, since when the new system of national air quality monitoring began to be constructed and came into use. The pollutants disclosed by this system include $SO_2, NO_2, PM_{10}, PM_{2.5}, O_3$ and CO and AQI. Implementation of the monitoring system with new standards is to improve environmental protection, environmental quality evaluation, monitoring and warning system, and government credibility. The policy led to the installation of real-time air pollution monitors across the country since 2012, which were built in 3 waves(2012, 2013 and

[1]MEE [2012]NO.11 is a document to notify that GB 3095-2012 is going into effect.

[2]GB 3095-2012 is a revised version of GB 3095-1996 and GB 9137-88. They are all air quality standards. Generally speaking, GB 3095-2012 includes something new, like the new pollutants added into the evaluation system. Also, it is stricter than the former ones, such as the cut-offs of number of missing values that make the data valid.

2015)[3]. During the exchange meeting in Beijing in January 4, 2015, it was announced[4] that the whole plan for the new national air quality standards was completed. And started with January 1, 2015, 1436 national monitoring sites in prefecture-level cities and the higher ones would come into use under the new standards and disclose the data of the 6 pollutants.

The new standards not only build up a new monitoring system, but also help to ensure the veracity of data disclosed to the public by constructing a platform for quality control. Seamless supervision on air quality monitoring data is realized by point-to-point transfer between city monitoring stations and remote online quality control platform. The data from these monitoring stations is automatically processed, reported and disclosed, getting rid of manual intervention. The 1436 national monitoring sites are internet-connected and disclose real-time data immediately to the city stations, provincial stations and China National Environmental Monitoring Centre(CNEMC). This theoretically allows no chance for strategy response and the data to be falsified. However, there are still some ways for the local officials to affect data quality, by discarding high pollution data. In fact, to prevent missing of air quality data, GB3095-2012 also sets cut-offs for numbers of missing values, above which will make the pollution concentration data invalid. For example, for $PM_{2.5}$, to ensure the validity, there must have at least 324 daily average data every year, 27 daily average data every month (25 in February) and 20 hourly average data every day. But the limit still allows chances to discard data and the fact is that there are many missing values in this air quality data set. This is why this paper is focused on the potential

[3]State Council authorized a Three Steps implementation plan for the new air quality standards. In MEE [2012] NO.11, it is stated that the plan is divided into 3 waves. The new standards should be applied in Beijing-Tianjin-Hebei Urban Agglomeration, Yangtze River Delta, Pearl River Delta, municipalities and provincial capitals by 2012(Wave 1), in 113 National Environmental Protection Key Cities and National Environmental Protection Exemplary Cities by 2013(Wave 2), in all prefecture-level cities and the higher administrative regions by 2015(Wave 3) and across the country at Jan. 1, 2016.

[4]http://finance.chinanews.com/ny/2015/01-04/6932330.shtml

patterns of these missing values.

2.2 Performance Evaluation of Officials

Performance evaluation is closely related to the promotions of officials. An important part in this evaluation is about economic development. The likelihood of promotion increases with the officials' economic performance(Li and Zhou (2005)). Local governments are provided with incentives to promote the economic prosperity(Montinola et al. (1995)). When facing potentially conflicting task, such as economic growth and environmental protection, the less measurable task will be ignored so that the environmental protection won't work well(Xu (2011)).

However, as the environmental issues have got more and more attention from the public and then the government, officials are also given the incentives to protect the environment. To encourage air pollution abatement, air quality has been included in the local officials' performance assessment. Chen et al. (2013a) carefully check the incentives and exist of gaming of air pollution data and indicate that the central personnel control over the local government is effective. According to the document, Decision of the State Council on Implementing the Scientific Outlook on Development and Strengthening Environmental Protection (Guofa [2005] No.39), environmental improvement is added into performance evaluations of officials in the way of 'Chengkao'[5]. During Twelfth Five-Year Plan(covers 2011-2015), air quality account for 15% of a city's environmental assessment. This indicator consists of the ratio of days with API $\leq$ 100, PM_{10}, SO_2 and NO_2[6]. GDP is a very essential index to evaluate the performance of local officials, but the concept of Green GDP has been

[5]http://websearch.mee.gov.cn/was5/web/search?
[6]http://www.mee.gov.cn/gkml/hbb/bgth/201111/W020111116343313075391.pdf

developed in order to push officials to work more on environmental protection. Technical specifications of Green GDP was finished in 2015 and applied in 7 pilot cities[7]. In addition, in Thirteenth Five-Year Plan[8], $PM_{2.5}$ is added as an important indicator for the environmental evaluation. These measures take environmental protection into the assessment system, pressuring the local officials and giving them the incentives to understate air pollution data.

Based on the policy, monitoring technology and performance evaluation, it is clear that feasibility and incentives to affect the data quality by discarding some high readings are satisfied.

[7]http://finance.people.com.cn/n/2015/0811/c1004-27441095.html

[8]Notice of the State Council on Printing and Distributing the "Thirteenth Five-Year Plan" Ecological Environmental Protection Plan

CHAPTER 3

DATA DESCRIPTION

3.1 Air Pollution

The air pollution data I use is from China National Environmental Monitoring Centre. This data set covers 1605 monitoring stations in 369 cities from May 14, 2014 to Dec 31, 2017, with hourly pollution concentration of 6 pollutants (15 indicators[1]), $PM_{2.5}$, PM_{10}, SO_2, NO_2, O_3 and CO. Beginning from 2008, $PM_{2.5}$ attracted more and more attention from the public with the disclosure of the daily concentration by U.S. Embassy in Beijing. It then was firstly included into the air quality standards in 2012. After Chinese Government replaced API with AQI, $PM_{2.5}$ also became the dominant pollutant instead of PM_{10}. As a result, to better assess the patterns of missing values, I choose $PM_{2.5}$ for the analysis. I count the number of missing values of $PM_{2.5}$ by hour-and-station level, and aggregate them to the day-and-city level. The summary statistics are reported in Table A.1[2]. The summary statistics of missing ratio by city level[3](Table A.2), season[4](Table A.3), location(Table A.4) and pollution level(Table A.5) are also provided. Two types of calculations of missing ratio are shown and used in the robustness check. One is calculated directly by the observations without $PM_{2.5}$. Another one includes the hours we don't have observations in this data set. In main part of this paper, I use the second missing ratio for analysis. It is impor-

[1] $PM_{2.5}$ hourly average, $PM_{2.5}$ 24hr average, O_3 hourly average, O_3 8hr average, O_3 24hr average, etc. AQI is one of the 15 indicators, which is an index calculated based on those 6 pollutants.

[2] Some of the maximums of Missing Ratio is 1, which means the data for that day at that city are all missing. After further looking into the data, it is found that city "Zhuji" accounts for most of the cases that the Missing Ratio equals to 1. These may be due to technical issues or some specific reasons. In the following analysis, I sometimes keep 95% quantile of the Missing Ratio for the analysis.

[3] Smaller number means larger cities. For example, Tier 1 means the largest cities, like Shanghai, Beijing, etc.

[4] Define March, April and May as Spring, June, July and August as Summer, September, October and November as Fall, December, January and February as Winter.

tant to note that these tables are just for summarizing the data, not for a strict causal inference.

3.2 Aerosol Optical Depth(AOD)

For the analysis of the potential patterns, proxy for air pollution may be needed. Chen et al. (2013a) use visibility and AOD data to show that the discontinuity of API or PM_{10} is driven by gaming instead of adopting real measures to improve the air quality when it comes closely to the threshold of Blue-Sky Days. Zou (2018) applies AOD data to compare pollution levels on off-days and on-days under the intermittent monitoring. Actually, remote data is widely applied in many fields. Donaldson and Storeygard (2016) demonstrate three main advantages of satellite data and conduct a comprehensive review of applications in Economics.They suggest that ground-based air pollution monitoring stations are not that widespread and the data may be affected by government manipulation. Many studies using satellite data are mentioned, like measuring air pollution caused by forest fires in Indonesia, testing the effect of air quality on infant mortality and evaluating potential causes of air pollution. Sullivan and Krupnick (2018) use satellite data to fill the gaps in the air quality monitoring network and estimate how many people live in areas with high but undetected pollution. New opportunities also will be available regarding application of satellite data into air quality data manipulation.

This paper uses the satellite data, MERRA-2 AOD[5](Summary statistics reported

[5]The Modern-Era Retrospective analysis for Research and Applications, Version 2 (MERRA-2) provides data beginning in 1980. It was introduced to replace the original MERRA dataset because of the advances made in the assimilation system that enable assimilation of modern hyperspectral radiance and microwave observations, along with GPS-Radio Occultation datasets. Spatial resolution remains about the same (about 50 km in the latitudinal direction) as in MERRA.

in A.6) as a proxy for air pollution to further check the potential relations between the missing and the pollution level. AOD is a measurement of the extinction of the solar beam by dust and haze, based on the Moderate Resolution Imaging Spectroradiometer (MODIS) on NASA's Terra satellite. AOD tells us how much direct sunlight is prevented from reaching the ground by these aerosol particles[6].

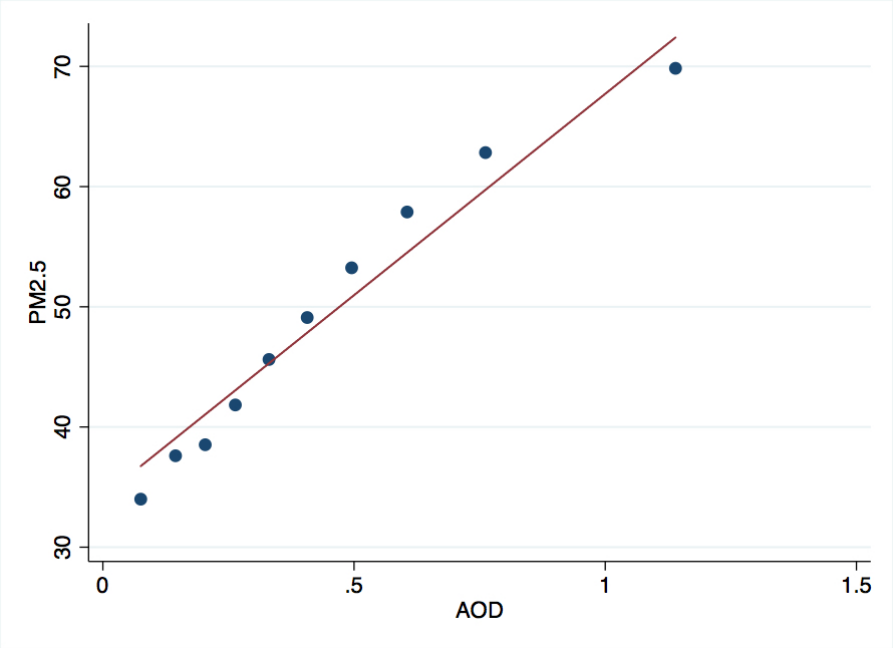
However, a widespread application of AOD doesn't mean it is perfect. Chu et al. (2002) demonstrate that The MODIS aerosol retrievals cover approximately 70% of the land surface. However, there are some cases we have no AOD data, like the high brightness, snow/ice covered regions — too bright in the visible wavelength to derive aerosol optical depth. Cloud cover more than 10% will also make the data unavailable(Chen et al. (2013a)). AOD provides data at at the satellite crossing time, which is about 10:30 am and 1:30pm local time(Zou (2018); Chen et al. (2013a)). In addition, AOD measures all particulate matter in the atmosphere, instead of air quality close to the ground.

Despite this fact, there are researches indicating the validity of applying AOD to predict air quality. van Donkelaar et al. (2010) indicate that with a chemical transport model, AOD could be used to estimate long-term $PM_{2.5}$ concentration. Many studies look into the relationship between AOD and air pollutants develop empirical models to make the prediction. Wang and Christopher (2003) show that the MODIS AOT (Aerosol Optical Thickness) has a good positive correlation with $PM_{2.5}$ mass (linear correlation coefficient, R = 0.7). They derive an empirical relationship between the MODIS AOT and 24hr mean $PM_{2.5}$ mass and conclude that the satellitederived AOT is a useful tool for air quality studies over large spatial domains to track and monitor aerosols. Koelemeijer et al. (2006) demonstrate that different meteorological conditions, such as cloud, humidity, make the difference and develop a relation between

[6]https://www.esrl.noaa.gov/gmd/grad/surfrad/aod/

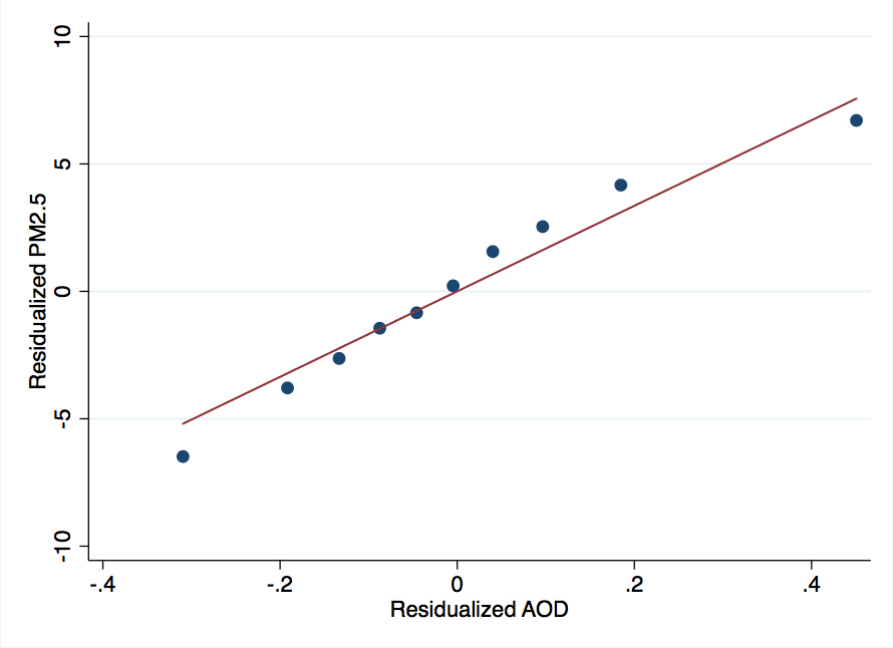
AOT and PM with local meteorological information taken into consideration.

The relation between AOD and PM is not that strict in this paper since what I need is the relative relation. When I look into whether there are more missing values when air pollution is higher, the trend of AOD matters, rather than the true relation between AOD and PM, as long as the positive relation between these two exists(Wang and Christopher (2003); Koelemeijer et al. (2006)). In this data set, the relation between $PM_{2.5}$ and AOD is shown in Figure 3.1 and 3.2, indicating the positive linear correlation. Based on this and the studies, this paper use AOD as a proxy for air pollution.



Note: Binscatter with n=10.

Figure 3.1: $PM_{2.5}$ and AOD(City-by-Day)



Note: Binscatter with n=10. Both are residualized with temperature, visibility, wind speed, precipitation, pressure, month_city FE, date FE and province_year FE.

Figure 3.2: Residualized $PM_{2.5}$ and AOD(City-by-Day)

3.3 Weather

In this study, I use weather data to correct for meteorological conditions. Koelemeijer et al. (2006) show that monthly average AOT and PM values show clear anti-correlation with rainfall. Ghanem and Zhang (2014) identify the conditions under which the manipulation is most likely to appear, using panel matching approach. It demonstrates that manipulation occurs under certain weather conditions but not others, and shows that higher levels of visibility and low wind speed are the two important factors. In addition, it is intuitively understandable that wind speed has a large effect on the visibility because if wind speed is high, the pollutants will be dissipated and visibility will be better. This paper uses wind speed (knots), visibility (miles), temperature(Fahrenheit), pressure(millibars) and precipitation amount (inches) data,

from Global Surface Summary of the Day(GSOD) data by National Oceanic and Atmospheric Administration(NOAA).

3.4 Summary Statistics

Table 3.1 is for data summary statistics. For more details about missing ratio and AOD, please refer to Tables in Appendix A.

Table 3.1: Summary Statistics

Variable	Obs	Mean	Std. Dev.	Min	Max
Missing Ratio	374,436	0.092	0.136	0	1
AOD	374,436	0.426	0.320	0.009	6.852
Temperature	374,436	58.286	19.959	-37.5	108.8
Pressure	339,687	945.597	95.181	576.7	1049.4
Visibility	373,572	8.933	5.334	0	18.6
Wind Speed	371,595	4.853	2.600	0	48.3
Precipitation	370,593	0.134	0.418	0	12.64

CHAPTER 4

PATTERNS OF MISSING VALUES

In this chapter, plots are used to summarize the general trend and variations, instead of showing rigorous causal correlations.

4.1 Threshold for Valid Data

To ensure the completeness of the air quality data and prevent discarding high readings on purpose, Ambient Air Quality Standards have set some cut-offs for the number of missing values to make the data valid. The new Ambient Air Quality Standards change the cut-off for hourly $PM_{2.5}$ from 18 to 20, which means for the data this paper uses, at least 20 hours of concentration should be reported every day. Otherwise, the data for that station at that day will be invalid, which if sums to exceed the cut-off for daily data, will lead to invalidity of all the data for the month or the year. I plot the distribution of number of missing values by hour-by-station level. We may expect discontinuity at the point of four if we assume manipulation exists. There is no sudden decrease around the threshold being found in Figure B.1. The figures for clean cities(Sanya, Kunming), mega cities(Beijing, Shanghai), dirty cities(Shijiazhuang, Zhengzhou) and cities that are reported to falsify data(Linfen, Xi'an, Shizuishan) can be found Figure B.2. However, this is not strong enough to conclude that there is no discarding on purpose, since even if there are several days are invalid, the validity could also be satisfied if the cut-off for number of daily missing values is satisfied. As the cut-off is 27 days, I plot it by months with 31 days and months with 30 days separately. From the month level, as shown in Figure B.3 and Figure B.4, there is a decrease around the threshold. However, no conclusion could

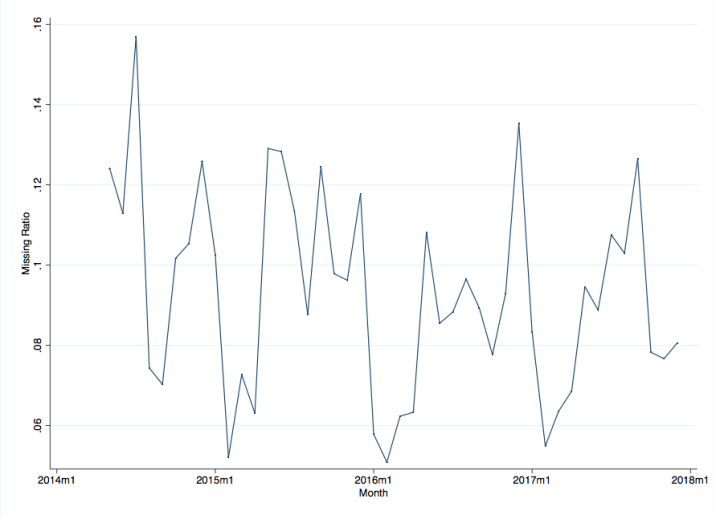
be made because it could be the distribution. Since the effect may be averaged by months, Figure B.5 shows the patterns by month. For the selected 9 cities, they show almost the same patterns. If plot for the cities with lowest missing ratio level and highest missing ratio level, the patterns keep for both of them(Figure B.14 and Figure B.15). For the year level cut-off which is 324 days, I also plot it by years with 365 days and year with 366 days(Figure B.6 and Figure B.7) and no obvious discontinuity is found.

In addition, considering that cities of different air pollution level may be pressured to different extend and the discontinuity may be weaken by cities with good air quality, I then plot the distribution by cities of different air quality level, which is represented by the average AOD level. For the general trend, no big difference between cities is found, and figures can be found in Appendix(Figure B.8,Figure B.9,Figure B.10, Figure B.11 and Figure B.12).

Also, the cities with different average missing ratios may differ in patterns. I plot these cut-offs by the level of average missing ratio as well(Figure B.13, Figure B.14, Figure B.15, Figure B.16 and Figure B.17). There is no strong evidence as well.

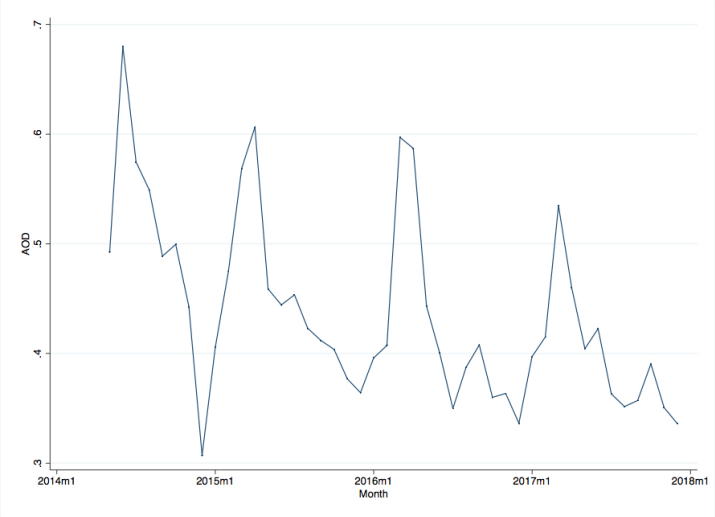
4.2 Missing Ratio and AOD

Figure 4.1 and Figure 4.2 are the monthly average of missing ratio and AOD. They both show the seasonality. For missing ratio, it tends to be lower during winter and higher during summer. For AOD, it is higher during winter and lower during summer. This makes sense because during winter, heating will lead to high level of air pollution. The decrease trend in Figure 4.2 is also reasonable as Chinese government took actions to reduce air pollution and these measures worked.



Note: This is monthly average of missing ratios of all the cities in this data set.

Figure 4.1: Monthly Average Missing Ratio

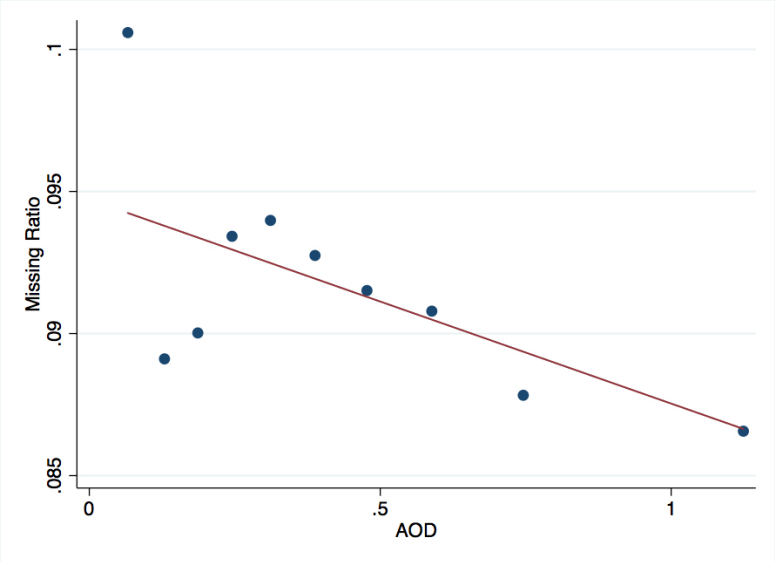


Note: This is monthly average of AOD of all the cities in this data set.

Figure 4.2: Monthly Average AOD

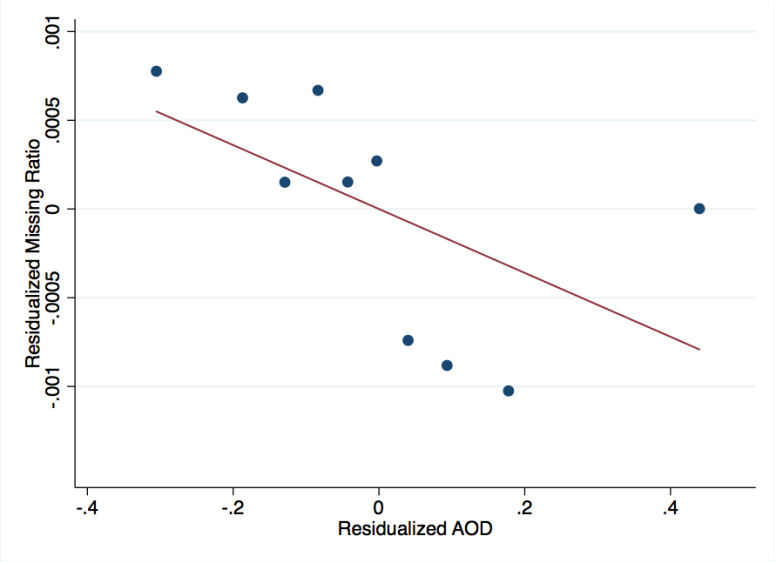
Based on these two figures, there seems to be a negative correlation between missing ratio and AOD. Figure 4.3 shows the negative relation between missing ratio and AOD, using raw data. The negative effect keeps after weather variables and fixed effects controlled(Figure 4.4), and the slope is about -0.002. If we look into the re-

lation by different air quality level which is represented by the average AOD of the city(Figure 4.5), it turns out cities of different air quality levels behave differently to the air pollution in terms of missing ratio. The slope for cities with better air quality is -0.004 while the slope for cities with worse air quality is -0.0006. The difference between their slopes indicates potential patterns of selective reporting. To test the significance of this trend, the empirical strategy should capture both the relation between missing ratio and AOD and the difference between cities of different air quality level.



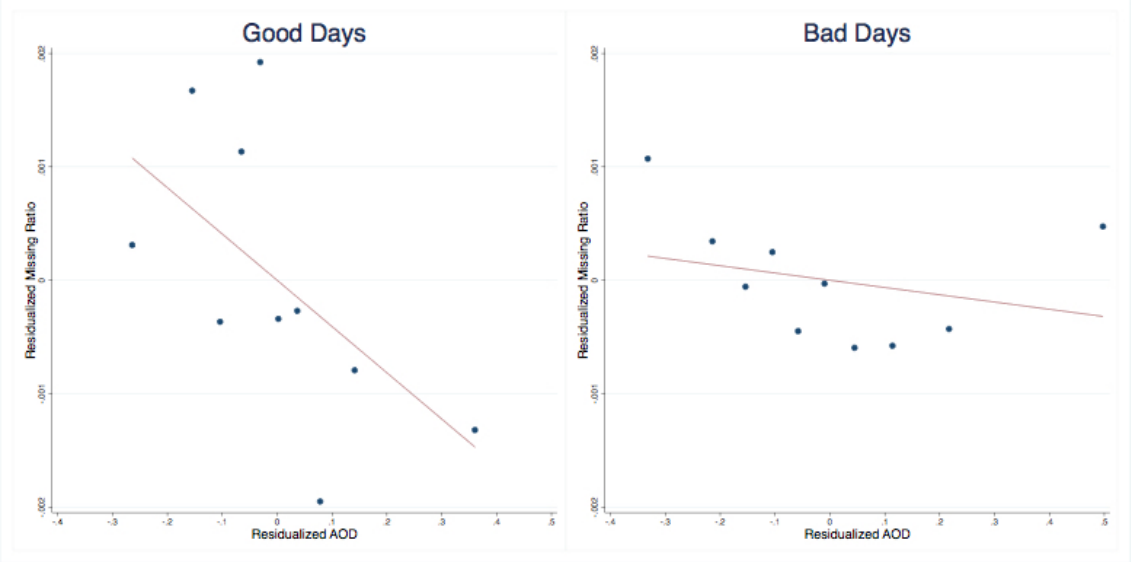
Note: Binscatter with n=10.

Figure 4.3: Missing Ratio and AOD



Note: Binscatter with n=10. Both are residualized with temperature, visibility, wind speed, precipitation, pressure, month_city FE, date FE and province_year FE.

Figure 4.4: Residualized Missing Ratio and Residualized AOD



Note: Binscatter with n=10. Both are residualized with temperature, visibility, wind speed, precipitation, pressure, month_city FE, date FE and province_year FE. Good days are the days with AOD no larger than the mean of AOD(0.43). Bad days are the days with AOD larger than the mean.

Figure 4.5: Residualized Missing Ratio and AOD during Good Days and Bad Days

CHAPTER 5

EMPIRICAL STRATEGY

5.1 Main Results

This section uses regressions to figure out whether the missing ratio is correlated to the air quality. As I mentioned above, since the missing values exist, the data is biased. As a result, I use AOD data to represent the air quality. Considering that the officials in cities with different air quality levels may differ in incentives to manipulate data, I also include an interaction of AOD and the average AOD which is by city level. Weather data, temperature, visibility, wind speed, precipitation and pressure are also added as explanatory variables to correct for the meteorological conditions of the application of AOD data.

The baseline specification is as following:

$$MissingRatio_{cd} = \alpha + \beta_1 AOD_{cd} + \beta_2 AOD_{cd} * AverageAOD_c + \beta_3 Temperature_{cd} + \beta_4 Visibility_{cd}$$

$$+ \beta_5 WindSpeed_{cd} + \beta_6 Precipitation_{cd} + \beta_7 Pressure_{cd} + \gamma_{cm} + \delta_d + \sigma_{py} + \epsilon_{cd},$$

where c is city, d is date, m is month, p is province and y is year. γ_{cm} is the city-by-month fixed effect to absorb month varying city characteristics. δ_d is the fixed effect to absorb the daily varying factors. σ_{py} is the fixed effect to absorb the year varying province characteristics.

Table 5.1 shows how I develop the specification. Column m1 is the regression of missing ratio on AOD. Column m2 is the regression with the interaction of AOD and average AOD level added. The average of variable AvgAOD is about 0.43, which

means for a city with the average air quality, the effect of AOD on missing ratio is around (β_1 + 0.43β_2)(Table 5.2). In Column m2, it is about -0.015 when the mean of missing ratio is 0.0915 and mean of AOD is 0.42. Columns m3-m6 all include weather data. And in Column m3, the effect of about -0.018, which doesn't change a lot from Column m2. Column m4-m6 is how I add fixed effects one by one into the model. β_1 and β_2 change a lot. The approximate effect of AOD on missing ratio for a city with the average air quality is -0.003(m4), -0.002(m5) and -0.003(m6), which almost keeps consistent. After adding fixed effects, effect of AOD on missing ratio is smaller. This is reasonable because with the fixed effects controlled, the variations caused by the factors that could be explained by these fixed effects are captured. For example, if the officials in some cities are more pressured to keep good air quality at the end of the year, then the variation in Column m3 may be partially caused by this. Without month-by-city fixed effect controlled, the coefficient β_1 and β_2 show upward biased effect of AOD on missing ratio.

Column m6 is the baseline specification. Considering the magnitude of weather data, they have very little impact on the missing ratio. In this specification, β_1 is -0.00928 and β_2 is 0.015. The effect of AOD on missing ratio is about -0.003, which mean when AOD increases 1 unit, missing ratio is about to decrease -0.003, which is about 3.3% of the average missing ratio. This is the approximate effect for the cities with average air quality. If we look into cities of different air quality level, the difference tells us more about manipulation. For cities like Beijing, Shijiangzhuang and Shanghai, with average AOD around 0.5, the effect of AOD on missing ratio is -0.001. For cities with higher pollution level, like Chengdu and Zhengzhou, with average AOD around 0.7, the effect is about 0.001. And for cities with lower pollution level, like Sanya and Kunming, with average AOD 0.2, the effect is around -0.006. Generally speaking, for the cities with worse air quality, the missing ratio is higher

when the pollution is higher while for the cities with better air quality, the missing ratio is lower when the pollution is higher. And the gap between -0.001 and 0.006 means if AOD changes 1 unit, the change of missing ratio will be 0.007 in difference for these two kinds of cities, about 7.7% of the average missing ratio.

Table 5.1: Regression Results with Fixed Effects(City-by-Day)

Missing Ratio(OLS)	m1	m2	m3	m4	m5	m6
AOD	-0.00718***	-0.0379*	-0.0542*	-0.0179***	-0.00575	-0.00928**
	(0.000693)	(0.0186)	(0.0218)	(0.00494)	(0.00444)	(0.00335)
AOD*AvgAOD		0.0532	0.0842	0.0336***	0.00797	0.0150*
		(0.0373)	(0.0437)	(0.00944)	(0.00822)	(0.00614)
Temperature			0.000377***	0.000282***	-0.0000405	-0.000190**
			(0.0000704)	(0.0000670)	(0.0000784)	(0.0000701)
Visibility			-0.000301	0.000146	0.000225	0.000199
			(0.000582)	(0.000187)	(0.000175)	(0.000160)
Wind Speed			-0.000649	-0.000201	0.000336*	0.000462***
			(0.000745)	(0.000182)	(0.000154)	(0.000134)
Precipitation			0.00446**	0.00639***	0.00509***	0.00480***
			(0.00159)	(0.000984)	(0.000915)	(0.000835)
Pressure			-0.0000806	-0.000503***	0.000143	-0.0000941
			(0.0000479)	(0.000102)	(0.000117)	(0.000111)
Month_City FE				Y	Y	Y
Date FE					Y	Y
Province_Year FE						Y
R-squared	0.000	0.001	0.006	0.288	0.616	0.631
Adjusted R-squared	0.000	0.001	0.006	0.280	0.610	0.626
Observations	374436	374436	333397	333397	333397	333397

Standard errors in parentheses ="* p0.05 ** p0.01 *** p0.001"

Standard errors are clustered by the city level.

AvgAOD is the city level average AOD, which is a variable to represent the air quality level.

Table 5.2: Joint Test for Effect of AOD on Missing Ratio

	m2	m3	m4	m5	m6
$\beta_1 + 0.43\beta_2$	-0.015**	-0.018***	-0.003*	-0.002	-0.003*

Standard errors in parentheses ="* p0.05 ** p0.01 *** p0.001"

Mean of AvgAOD, 0.43 is used for the calculation.

Since there exists decrease around the cut-offs for the data to be valid, the following regression is also tested:

$$Valid_{sd} = \alpha + \beta_1 AOD_{cd} + \beta_2 Temperature_{cd} + \beta_3 Visibility_{cd} + \beta_4 WindSpeed_{cd}$$

$$+\beta_5 Precipitation_{cd} + \beta_6 Pressure_{cd} + \gamma_{sm} + \sigma_{py} + \epsilon_{cd},$$

where s is station, d is date, m is month, p is province and y is year. Valid is 1 when the daily data is valid for that day, 0 otherwise. γ_{sm} is the station-by-month fixed effect to absorb month varying station characteristics. σ_{py} is the fixed effect to absorb the year varying province characteristics.

The general trend is that when air quality is worse, the high possibility the daily data is valid. After fixed effects added in, the variation is smaller since part of the variation in Column m1 and Column m2 is due to the factors explained by the fixed effect, instead of air quality. However, the magnitude is very small since the mean of 'Valid' is about 0.9 and mean of AOD is about 0.43, which means although the coefficient in Column m6 is significant, AOD has very little effect on whether the daily data is valid.

Table 5.3: Regression Results for Daily Cut-off(Station-by-Day)

Valid(OLS)	m1	m3	m4	m5	m6
AOD	0.0132***	0.0171**	0.00402	0.00211	0.00328*
	(0.000771)	(0.00633)	(0.00294)	(0.00181)	(0.00146)
Temperature		-0.000505***	-0.000186	0.0000894	0.000241*
		(0.0000779)	(0.0000979)	(0.000108)	(0.0000970)
Visibility		0.000590	-0.0000119	-0.000189	-0.000131
		(0.000615)	(0.000256)	(0.000209)	(0.000180)
Wind Speed		0.000870	0.000441	-0.000547*	-0.000744***
		(0.000765)	(0.000288)	(0.000212)	(0.000184)
Precipitation		-0.0105***	-0.0119***	-0.00865***	-0.00833***
		(0.00247)	(0.00188)	(0.00167)	(0.00159)
Pressure		0.0000863	0.00125***	-0.000200	0.0000782
		(0.0000533)	(0.000162)	(0.000186)	(0.000174)
Month_Station FE			Y	Y	Y
Date FE				Y	Y
Province_Year FE					Y
R-squared	0.000	0.002	0.212	0.521	0.527
Adjusted R-squared	0.000	0.002	0.202	0.515	0.521
Observations	1620287	1427966	1427950	1427950	1427950

Standard errors in parentheses ="* p0.05 ** p0.01 *** p0.001"

Standard errors are clustered by the city level.

Valid is 1 when the daily data is valid, 0 otherwise.

5.2 Dynamic Effect

Considering that it may take time to take action, Table 5.4 is the specifications including the AOD lagged for 1 day, for 2 days and one week. β_1 and β_2 keep significant. And the effect of AOD on missing ratio is -0.003 for all the columns(Table 5.5 Row 1), which also keeps consistent and significant. For the coefficients of variables related to

lagged AOD(Table 5.5 Row 2-4), none of them is significant, as well as the joint test. In addition, they don't make a big difference to the missing ratio in the perspective of magnitude.

Table 5.4: Regression Results with Lagged AOD(City-by-Day)

Missing Ratio(OLS)	m0	LagDay1	LagDay2	Lag2Day	Lag1Week
AOD	-0.00928**	-0.0101***	-0.00984**	-0.00983***	-0.00915**
	(0.00335)	(0.00287)	(0.00311)	(0.00293)	(0.00337)
AOD*AvgAOD	0.0150*	0.0166**	0.0160**	0.0160**	0.0148*
	(0.00614)	(0.00506)	(0.00563)	(0.00522)	(0.00622)
AOD Lagged 1 Day		0.00179		-0.0000310	
		(0.00275)		(0.00201)	
AOD Lagged 1 Day*AvgAOD		-0.00329		0.000172	
		(0.00531)		(0.00382)	
AOD Lagged 2 Days			0.00354	0.00356	
			(0.00315)	(0.00288)	
AOD Lagged 2 Days*AvgAOD			-0.00661	-0.00669	
			(0.00605)	(0.00549)	
AOD Lagged 1 Week					-0.00223
					(0.00339)
AOD Lagged 1 Week*AvgAOD					0.00602
					(0.00652)
Weather	Y	Y	Y	Y	Y
Month_City FE	Y	Y	Y	Y	Y
Date FE	Y	Y	Y	Y	Y
Province_Year FE	Y	Y	Y	Y	Y
R-squared	0.631	0.632	0.632	0.632	0.632
Adjusted R-squared	0.626	0.626	0.626	0.626	0.626
Observations	333397	333184	332974	332974	331920

Standard errors in parentheses ="* p<0.05 ** p<0.01 *** p<0.001"

Standard errors are clustered by the city level.

The 5 variables for weather are included in all these five models as above.

Table 5.5: Joint Test for Effect of AOD or Lagged AOD on Missing Ratio

	m0	LagDay1	LagDay2	Lag2Day	Lag1Week
AOD	-0.003*	-0.003**	-0.003*	-0.003**	-0.003*
AOD Lagged 1 Day		0.0004		0.00005	
AOD Lagged 2 Day			0.0007	0.0007	
AOD Lagged 1 Week					0.0004

Standard errors in parentheses ="* p0.05 ** p0.01 *** p0.001"

5.3 Heterogeneous Effect

In Table 5.6, mayors' promotion pressure and education background are added to capture how the leader of a city affects the missing ratio. For most of the cities in China, the age of the position higher than their current position cannot be beyond 58 years old, which means whether the they are below 57 years old when they are eligible for promotion. This measures whether they are pressured for a better evaluation. For the provincial capital cities, the threshold is 62 years old. For Beijing, Shanghai, Tianjin and Chongqing, which is equal to province, the threshold is 66 years old. I also include whether their highest degree is bachelor, master or PhD to capture the education background of the city's leader. Triple interactions for AOD, average AOD level and these four dummies are also added into the models. The results suggest no significant impact of promotion threshold and education background on the missing ratio, in terms of either the coefficient alone or the joint test[1] of the two related coefficients. However, the magnitude of these four is relatively large. The reason for this may be that the highest degree for one person almost remains the same or the mayor doesn't change too much in such a short time(2014-2017) so that changes in several cities lead to the variation.

[1]Use mean of AOD and mean of average AOD for the test, which means this is based on a city of average air quality level and of average AOD on that day.

Table 5.6: Regression Results with Mayors(City-by-Day)

Missing Ratio(OLS)	m0	Young	Educ	All
AOD	-0.00928**	-0.00908*	-0.00930**	-0.00899*
	(0.00335)	(0.00351)	(0.00341)	(0.00348)
AOD*AvgAOD	0.0150*	0.0101	0.0146	0.00817
	(0.00614)	(0.0223)	(0.0157)	(0.0215)
Young		-0.00126		0.00444
		(0.00810)		(0.0173)
AOD*AvgAOD*Young		0.00486		0.0123
		(0.0207)		(0.0315)
Bachelor			0.0102	0.00609
			(0.0106)	(0.0181)
AOD*AvgAOD*Bachelor			0.00272	-0.00347
			(0.0185)	(0.0265)
Master			-0.0000555	-0.00415
			(0.00826)	(0.0171)
AOD*AvgAOD*Master			-0.00256	-0.00888
			(0.0163)	(0.0245)
PhD			-0.0210	-0.0250
			(0.0118)	(0.0186)
AOD*AvgAOD*PhD			0.00626	0.000262
			(0.0180)	(0.0260)
Weather	Y	Y	Y	Y
Month_City FE	Y	Y	Y	Y
Date FE	Y	Y	Y	Y
Province_Year FE	Y	Y	Y	Y
R-squared	0.631	0.631	0.633	0.633
Adjusted R-squared	0.626	0.626	0.627	0.627
Observations	333397	333397	333397	333397

Standard errors in parentheses ="* p<0.05 ** p<0.01 *** p<0.001"

Standard errors are clustered by the city level.

The 5 variables for weather are included in all these four models as above.

Young is 1 if the mayor is below 57 years old(otherwise, 0) for most of the cities.

For provincial capital cities, threshold for Young is 62 years old.

For Beijing, Shanghai, Tianjin and Chongqing, threshold for Young is 66 years old.

Bachelor is 1 if mayor's degree of is bachelor. The same for Master and PhD.

5.4 Robustness Check

In this data set, there are part of observations without $PM_{2.5}$ data. This is one kind of missing values. However, there is also another kind of missing. When a station is open, theoretically it should have data reported every hour. The truth is there are some hours that we have no observations. It is hard to figure out why there are these two cases and tell which one makes more sense than the other one. Summary statistics by different dimensions are shown in Appendix already, but this is not enough. So Table 5.7 is to compare these two types of missing ratio to see whether they make a big difference to our empirical analysis. As we can see, the results turn out to be very similar. The joint test show the same significance as well, for an effect around -0.003. There is no big difference how we define and calculate missing ratio.

Table 5.7: Regression Results by Different Missing Ratio(City-by-Day)

Missing Ratio(OLS)	m1	m2
AOD	-0.00928**	-0.00952**
	(0.00335)	(0.00341)
AOD*AvgAOD	0.0150*	0.0151*
	(0.00614)	(0.00625)
Temperature	-0.000190**	-0.000199**
	(0.0000701)	(0.0000729)
Visibility	0.000199	0.000200
	(0.000160)	(0.000163)
Wind Speed	0.000462***	0.000486***
	(0.000134)	(0.000138)
Precipitation	0.00480***	0.00490***
	(0.000835)	(0.000849)
Pressure	-0.0000941	-0.0000849
	(0.000111)	(0.000115)
Month_City FE	Y	Y
Date FE	Y	Y
Province_Year FE	Y	Y
R-squared	0.631	0.450
Adjusted R-squared	0.626	0.441
Observations	333397	333397

Standard errors in parentheses ="* p<0.05 ** p<0.01 *** p<0.001"

Standard errors are clustered by the city level.

m1 is the missing ratio including hours without observations.

m2 is missing ratio calculated only by observations without $PM_{2.5}$.

Table 5.8: Joint Test for Effect of AOD on Different Missing Ratios

	m1	m2
$\beta_1 + 0.43\beta_2$	-0.003*	-0.003*

Standard errors in parentheses ="* p0.05 ** p0.01 *** p0.001"

CHAPTER 6

CONCLUSIONS AND FUTURE WORK

This paper uses $PM_{2.5}$ data from 2014 to 2017 to test whether the local officials selectively report air quality data by discarding data. When I look into the cut-off for the number of missing values to make the data valid, there is little evidence of discontinuity around the threshold, as well as empirical evidence of the correlation between air pollution and the possibility of the daily data to be valid, in terms of magnitude. When applying AOD data as a proxy for air pollution, the empirical method shows there exists a little negative effect of air pollution on missing ratio. This relationship is weaker in dirty cities measured by the average AOD during the sample period and is reversed in very dirty cities. The dynamic check shows that the air quality one day before, two days before and one week before doesn't affect the missing ratio. In addition, I find no evidence for effect of officials' promotion pressure and education background on missing ratio as well.

About the negative general effect, the question is why when air pollution is higher, the missing ratio is lower for some cities? There may be two possible reasons. One is that the AOD data is the satellite data, which means it is not exactly the data represents the air pollution around the monitors which is much closer to the ground. Then the relations shown by this specification may be biased. And using AOD as a proxy for $PM_{2,5}$ could also lead to bias. However, the differences between cities of different air quality level tell us some story about manipulation, since the relative relations between these cities are not affected by the bias that much. Another possible reason is that when air pollution is high, it is easy for the public to feel the bad air quality. Then both the public and the government pay more attention to the data published, which leads to higher risk and allows less space for missing values. This may lead to

a negative effect. However, officials do have pressure when air quality is bad which means incentives to drop some high readings could also play an important role at the same time. This offsets the negative effect to some extent. This may be the reason why for the cities of extremely bad air quality, when air pollution is higher, there are more missing values.

There are also some directions this paper can be further developed in. First, Based on the fact that some measures are taken to prevent data misreporting, this paper works only on whether manipulation in terms of missing values exists. For future work, the discontinuity test for the air quality data after 2012 could also be done to check whether this kind of manipulation exist, since falsification by blocking monitors or just improving the environment around the monitor is still feasible. Besides, the data I use is day-by-city level. However, there is variation between different hours within a day(Figure C.1 and Figure C.2). The number of missing values is high around noon and low around evening. The pattern of this can be further test if the hour-by-city level data using as the proxy for $PM_{2.5}$ is available. It is possible that there are other patterns of manipulation by hour level, which is neglected in our day-by-city analysis. Last, about the officials' promotion pressure, the time period for the tenure should also be considered if the data is available, since before promotion to a higher level position, the official should has been on the current level(it could be different positions on the same level) for 5 years. This means the cut-offs of 57, 62 and 66 years old could only be a rough approximation to represent the promotion pressure. If the time period could be also be taken into consideration, the analysis of officials' promotion would make more sense. These three may be potential directions for future research on air quality data manipulation in China for the time period after 2012.

APPENDIX A

SUMMARY STATISTICS

A.1 Missing Ratio

Missing ratio is calculated by city-and-day level.

(1)Ratio of missing values

number of observations without values / number of total observations we have

(2)Ratio including the hours we don't have observations

number of observations without values and no observations/ number of total observations we should have

Table A.1: Summary Statistics of Missing Ratios of $PM_{2.5}$

Missing Ratio	Obs	Mean	Std	Min	Max
missing value	375,880	.0610	.1138	0	1
including no obs.	375,880	.0915	.1359	0	1

Table A.2: Missing Ratios by City Tiers

Tier	1	2	3	4	5	6	7
Obs	3,894	19,469	44,123	73,954	72,787	113,728	47,925
Mean							
(missing value)	.0541	.0749	.0660	.0744	.0465	.0511	.0760
(including no obs.)	.1027	.1054	.0967	.1047	.0774	.0813	.1053

Table A.3: Missing Ratios by Seasons

Season	Spring	Summer	Fall	Winter
Obs	84,071	100,587	100,651	90,571
Mean				
(missing value)	.0539	.0685	.0655	.0541
(including no obs.)	.0836	.1010	.0949	.0846

Table A.4: Missing Ratios by Location

Location	Northern	Southern
Obs	198,378	177,502
Mean		
(missing value)	.0621	.0600
(including no obs.)	.0923	.0907

Table A.5: Missing Ratios by Pollution Level

AOD Level	[0,1)	[1,2)	[2,3)	[3,4)	[4,5)	[5,6)	[6,7)
Obs	354,039	21,245	559	29	7	0	1
Mean							
(missing value)	.0611	.0585	.0676	.0816	.0234	-	0
(including no obs.)	.0918	.0862	.0949	.1659	.1375	-	.0833

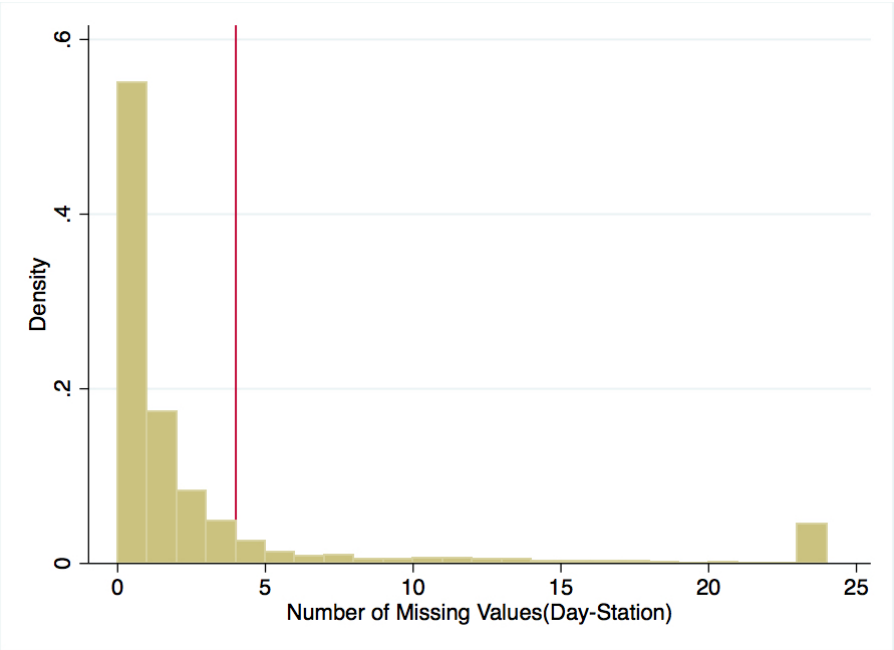
A.2 AOD

Table A.6: Summary Statistics of AOD

AOD	Obs	p25	Mean	p75	SD	# of Cities
All	441,228	0.18	0.42	0.58	0.32	332
Spring	97,940	0.25	0.50	0.69	0.33	332
Summer	122,176	0.19	0.42	0.55	0.33	332
Fall	120,848	0.17	0.39	0.54	0.30	332
Winter	100,264	0.13	0.38	0.55	0.31	332
2014	77,356	0.18	0.44	0.61	0.35	332
2015	121,180	0.19	0.44	0.61	0.33	332
2016	121,512	0.18	0.42	0.58	0.32	332
2017	121,180	0.18	0.40	0.55	0.29	332

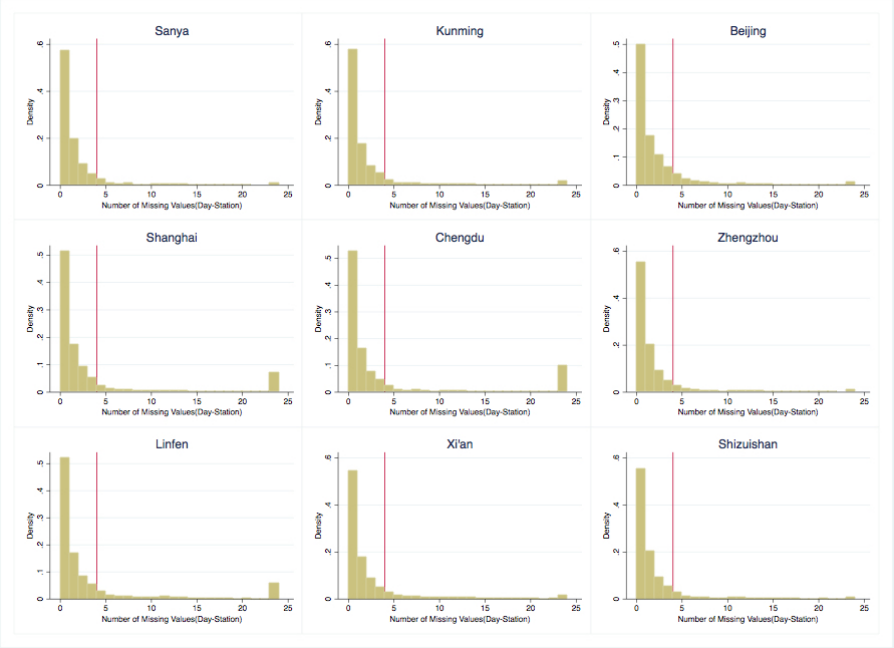
APPENDIX B

CUT-OFFS FOR VALID AIR QUALITY DATA



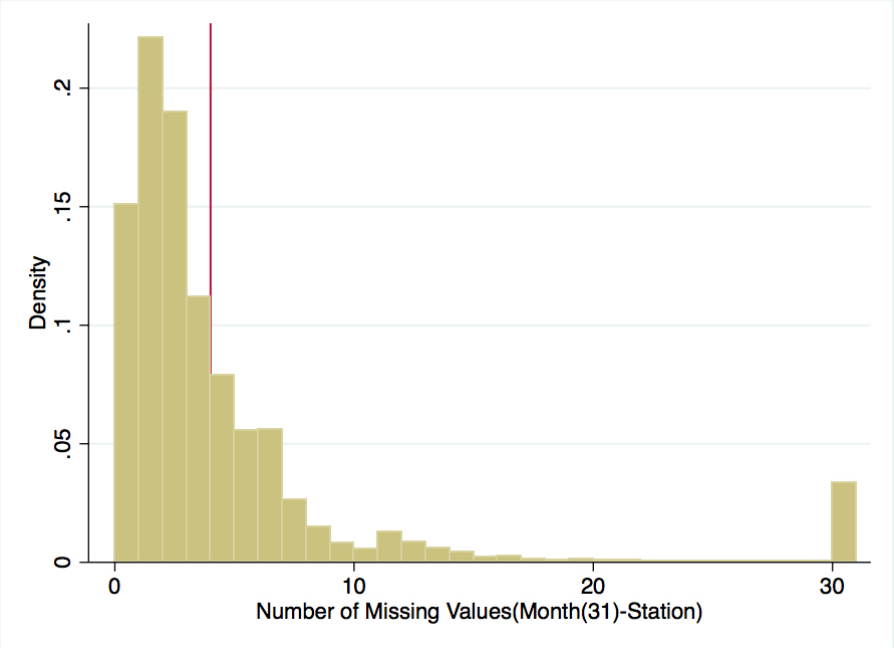
Note: To ensure the validity of daily $PM_{2.5}$ data, at least 20 hours should be reported. The threshold here is 4, which is denoted by the red line in the figure

Figure B.1: Number of Missing Hours(Station-by-Day)



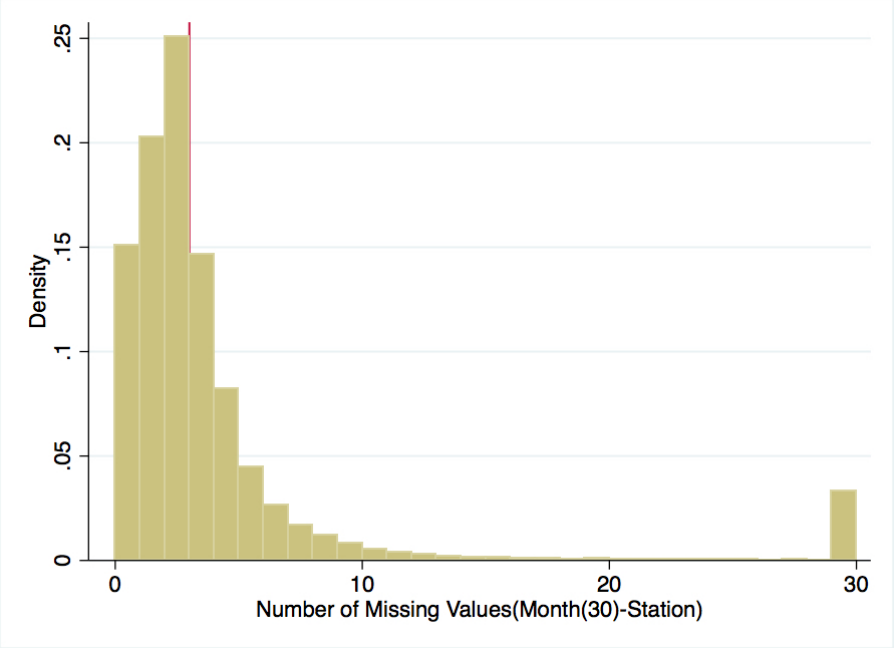
Note: To ensure the validity of daily $PM_{2.5}$ data, at least 20 hours should be reported. The threshold here is 4, which is denoted by the red line in the figure

Figure B.2: Number of Missing Hours(Station-by-Day) for Selected Cities



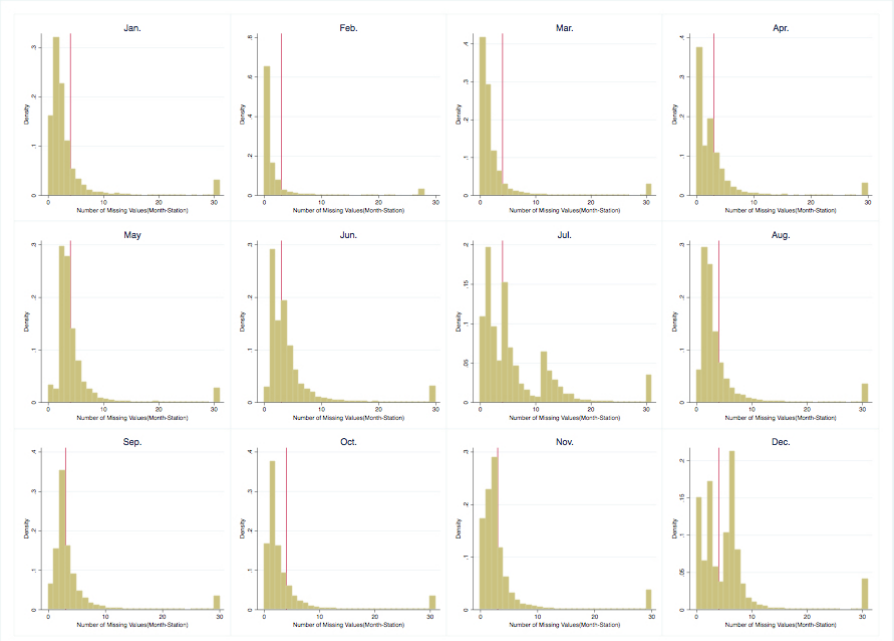
Note: To ensure the validity of $PM_{2.5}$ data for the month, at least 27 daily data should be reported. The threshold here is 4, which is denoted by the red line in the figure.

Figure B.3: Number of Missing Days(Station-by-Month with 31 days)



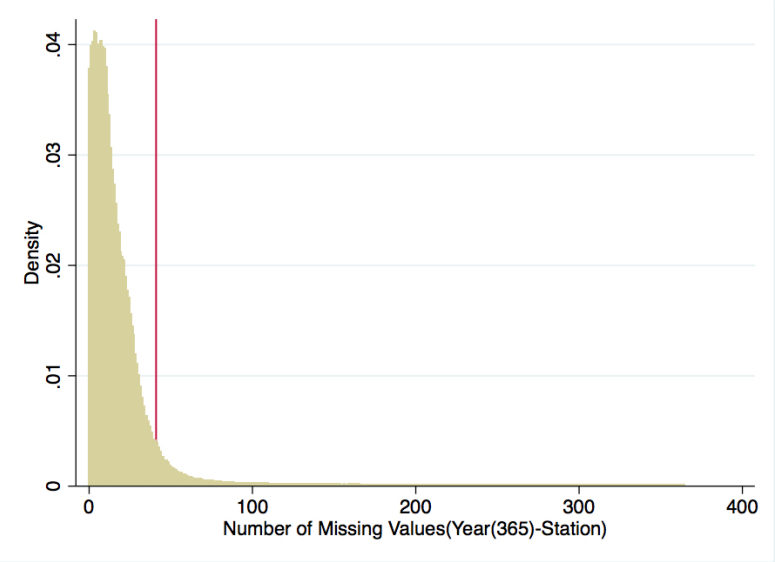
Note: To ensure the validity of $PM_{2.5}$ data for the month, at least 27 daily data should be reported. The threshold here is 3, which is denoted by the red line in the figure.

Figure B.4: Number of Missing Days(Station-by-Month with 30 days)



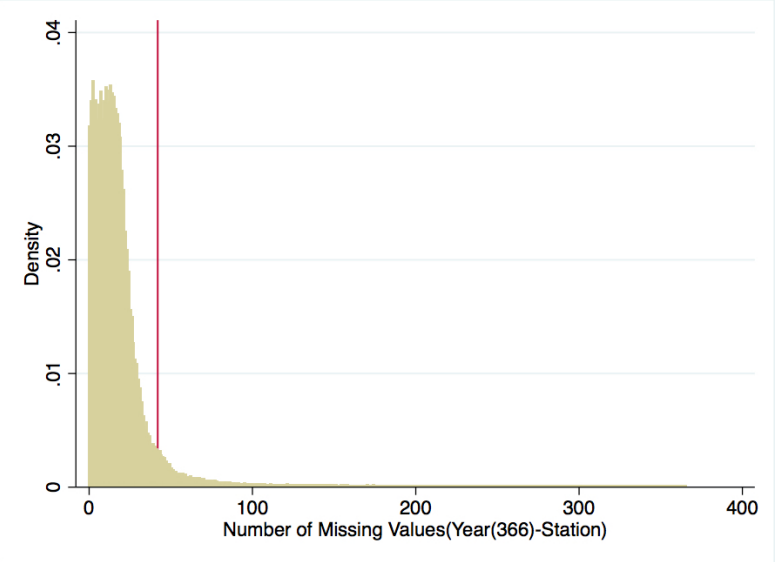
Note: To ensure the validity of $PM_{2.5}$ data for the month, at least 27 daily data should be reported. The threshold is denoted by the red line in the figure.

Figure B.5: Number of Missing Days(Station-by-Month)



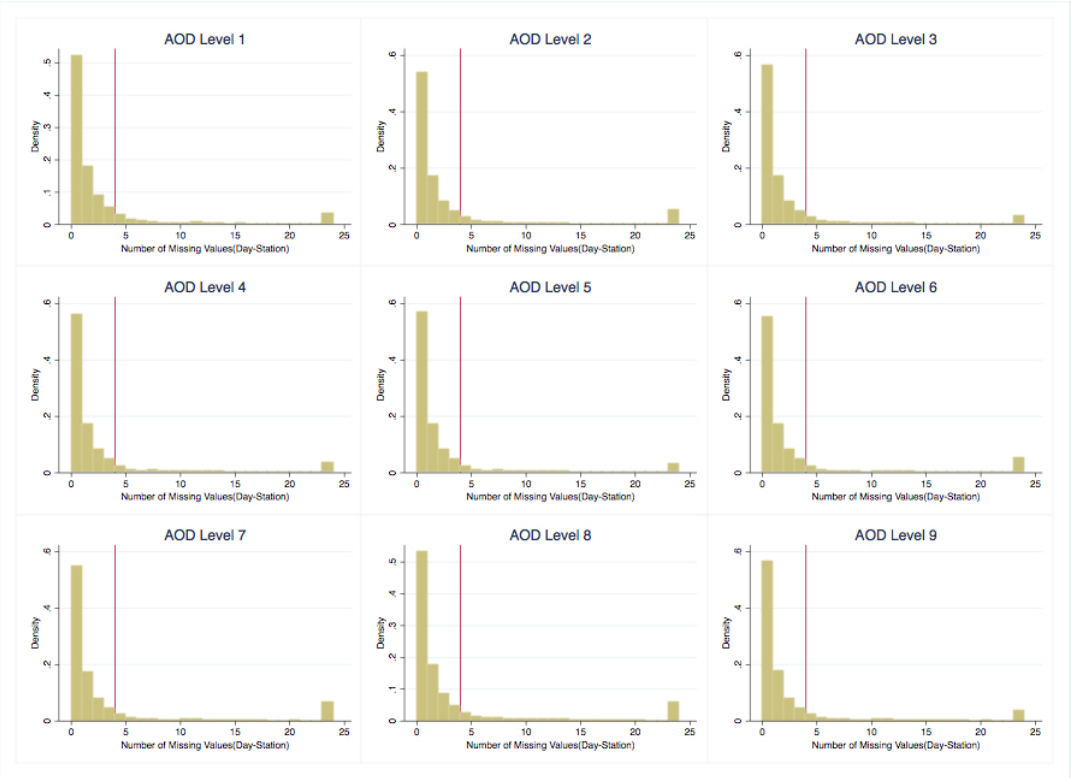
Note: To ensure the validity of $PM_{2.5}$ data for the year, at least 324 daily data should be reported. The threshold here is 41, which is denoted by the red line in the figure.

Figure B.6: Number of Missing Days(Station-by-Year with 365 days)



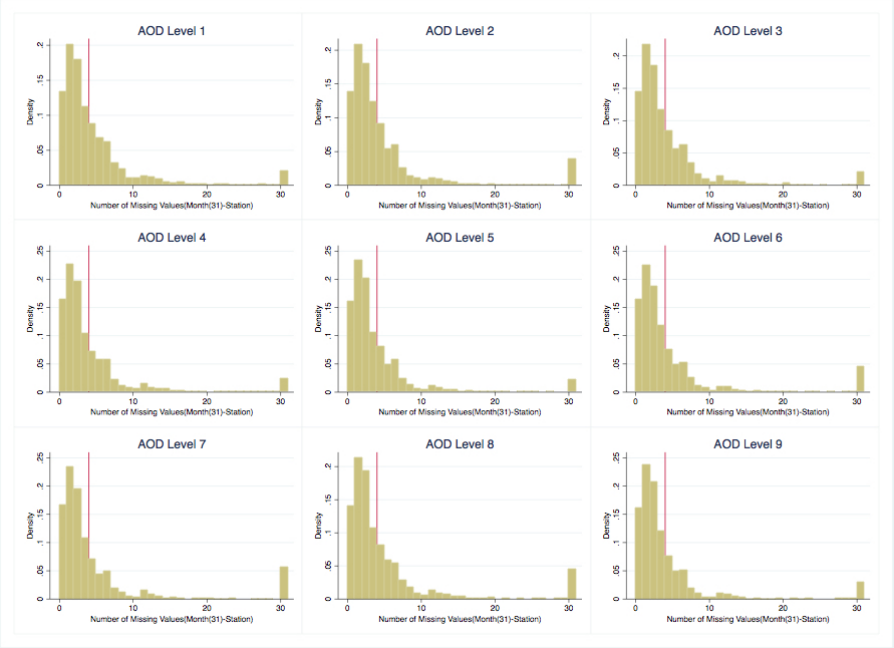
Note: To ensure the validity of $PM_{2.5}$ data for the year, at least 324 daily data should be reported. The threshold here is 42, which is denoted by the red line in the figure.

Figure B.7: Number of Missing Days(Station-by-Year with 366 days)



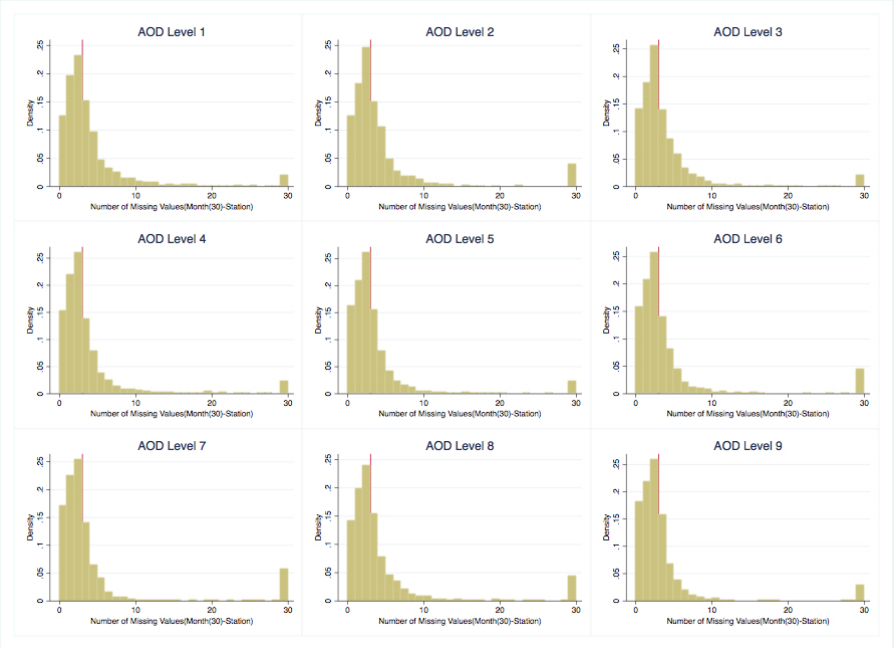
Note: Average AOD is calculated by city level, using the AOD data during the whole time period this data set covers. And they are divided equally into 9 levels, to show the variations between cities with better air quality and cities with worse air quality. Here, Level 1 indicates the lowest AOD level and Level 9 indicates the highest AOD Level. The threshold here is 4, which is denoted by the red line in the figure.

Figure B.8: Number of Missing Hours by Average AOD Level



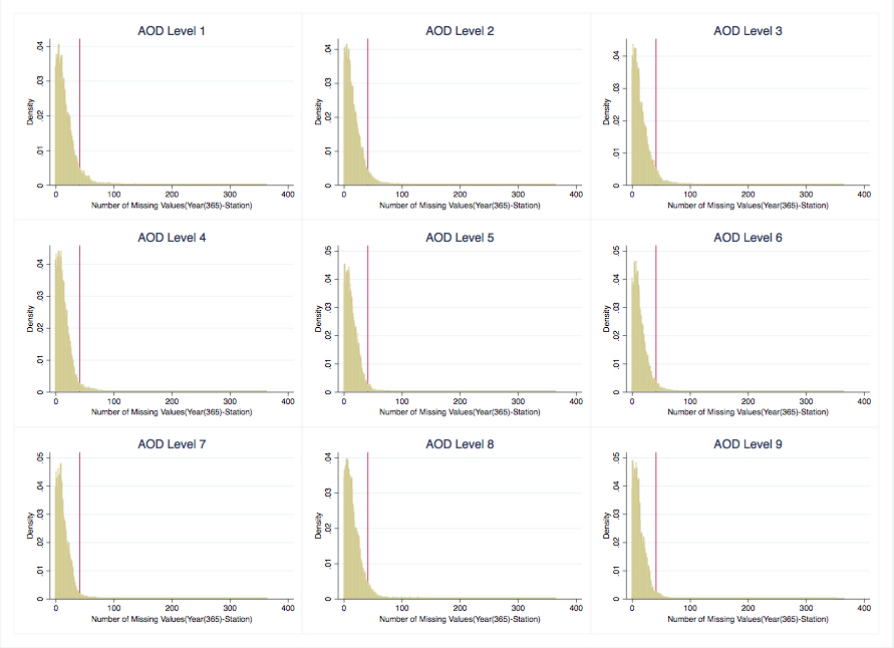
Note: AOD Level is defined in the same way in Figure B.8. The threshold here is 4, which is denoted by the red line in the figure.

Figure B.9: Number of Missing Days for Month(with 31 days) by Average AOD Level



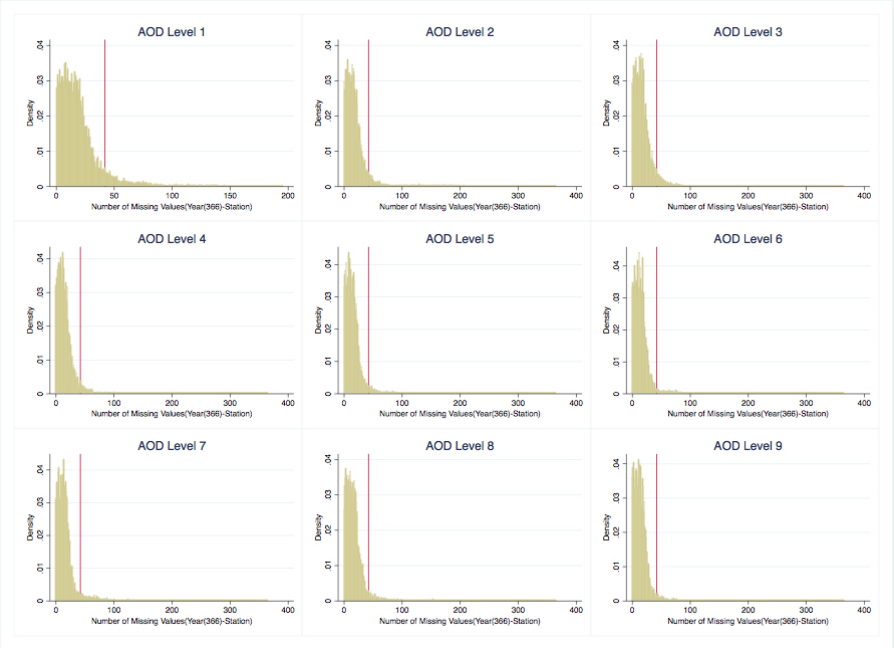
Note: AOD Level is defined in the same way in Figure B.8. The threshold here is 3, which is denoted by the red line in the figure.

Figure B.10: Number of Missing Days for Month(with 30 days) by Average AOD Level



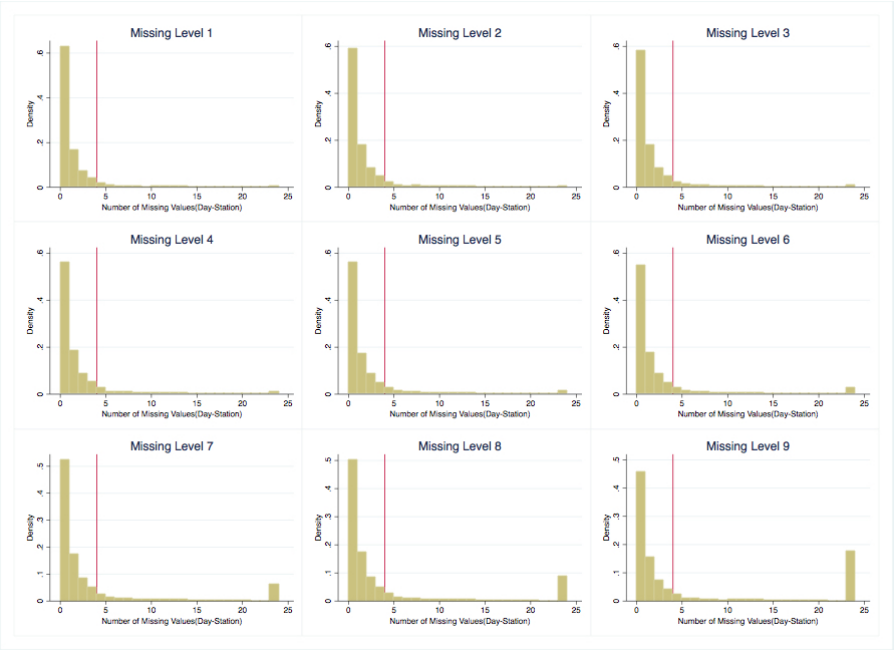
Note: AOD Level is defined in the same way in Figure B.8. The threshold here is 41, which is denoted by the red line in the figure.

Figure B.11: Number of Missing Days for Year(with 365 days) by Average AOD Level



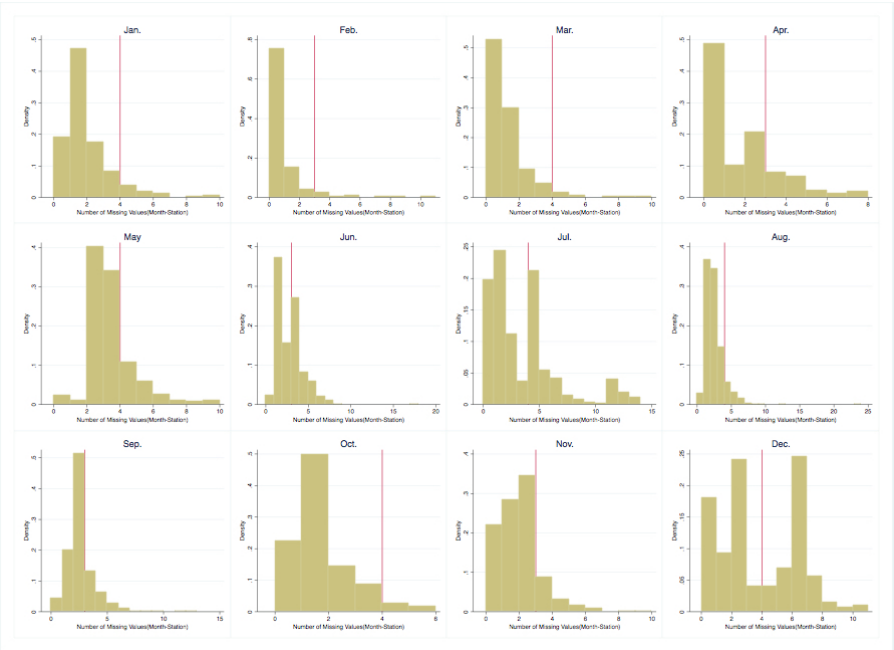
Note: AOD Level is defined in the same way in Figure B.8. The threshold here is 42, which is denoted by the red line in the figure.

Figure B.12: Number of Missing Days for Year(with 366 days) by Average AOD Level



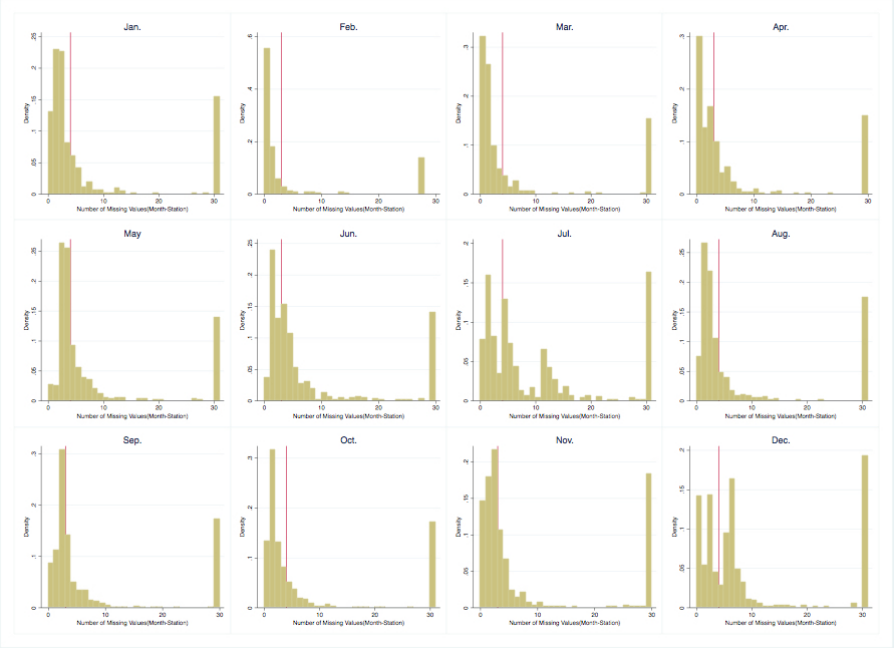
Note: Missing Level is calculated by city level, using the daily missing ratio during the whole time period this data set covers. And they are divided equally into 9 levels. The threshold here is 4, which is denoted by the red line in the figure.

Figure B.13: Number of Missing Hours(Station-by-Day) by Missing Ratio Level



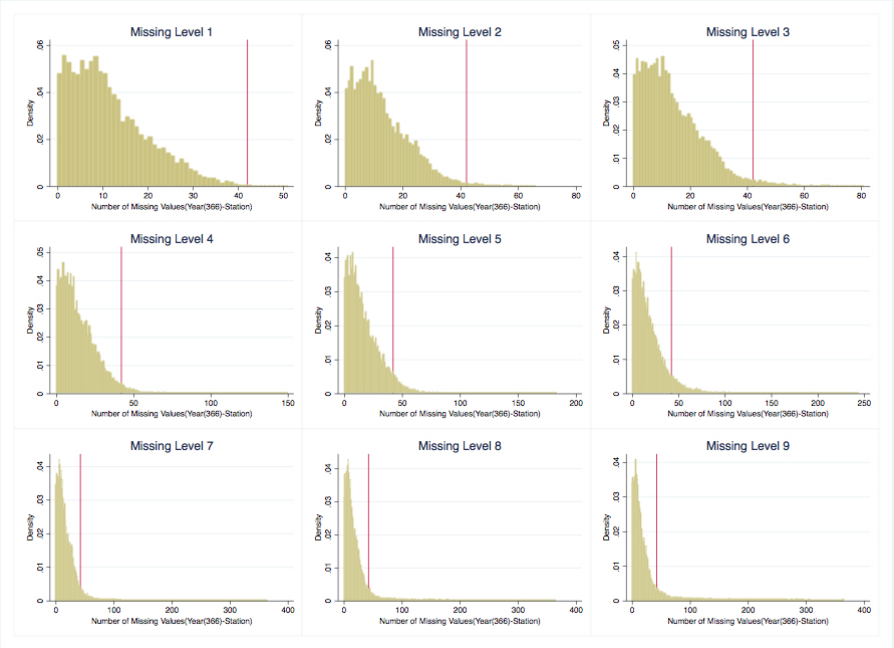
Note: Missing Level is defined in the same way in Figure B.13. The threshold here is denoted by the red line in the figure.

Figure B.14: Number of Missing Days(Station-by-Month) for Missing Level 1



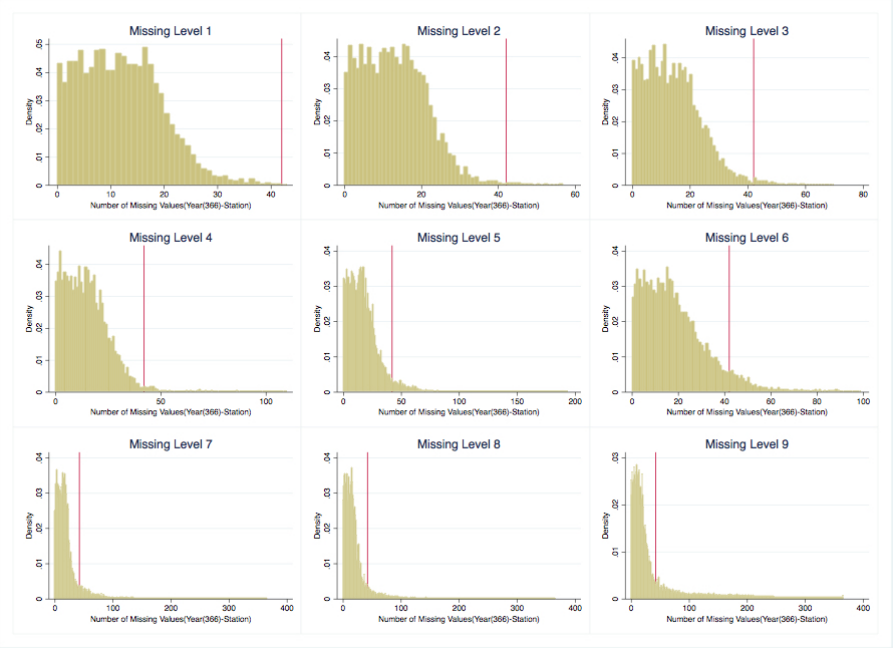
Note: Missing Level is defined in the same way in Figure B.13. The threshold here is denoted by the red line in the figure.

Figure B.15: Number of Missing Days(Station-by-Month) for Missing Level 9



Note: Missing Level is defined in the same way in Figure B.13. The threshold here is 41, which is denoted by the red line in the figure.

Figure B.16: Number of Missing Days for Year(with 365 days) by Missing Ratio Level



Note: Missing Level is defined in the same way in Figure B.13. The threshold here is 41, which is denoted by the red line in the figure.

Figure B.17: Number of Missing Days for Year(with 366 days) by Missing Ratio Level

APPENDIX C

VARIATION IN HOUR

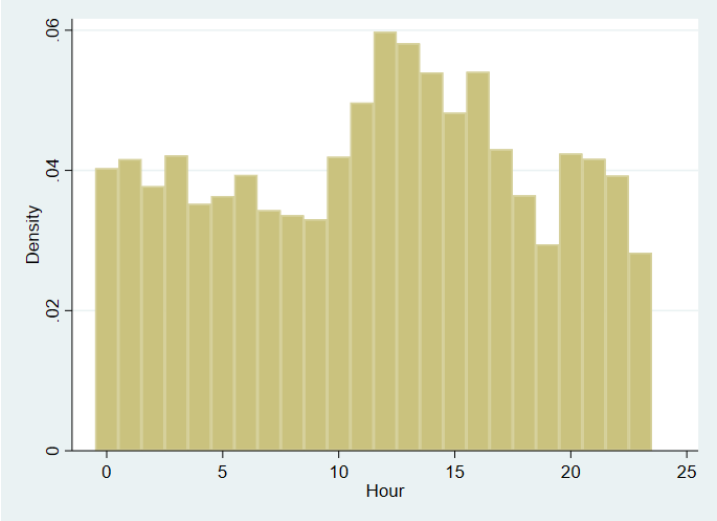


Figure C.1: Number of Missing Values by Hour(Missing Value)

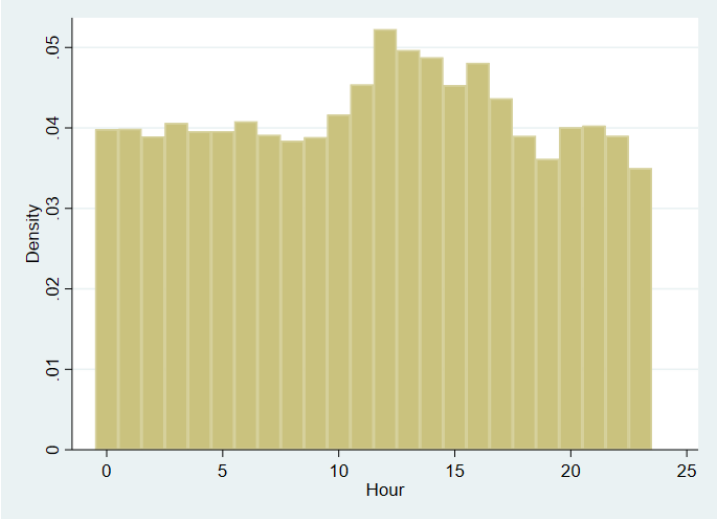


Figure C.2: Number of Missing Values by Hour(Including No Obs.)

BIBLIOGRAPHY

Brunekreef, Bert and Stephen T Holgate, "Air pollution and health," *The Lancet*, 2002, *360* (9341), 1233 – 1242.

Chan, Chak K. and Xiaohong Yao, "Air pollution in mega cities in China," *Atmospheric Environment*, 2008, *42* (1), 1 – 42.

Chen, Yuyu, Ginger Zhe Jin, Naresh Kumar, and Guang Shi, "Gaming in Air Pollution Data? Lessons from China," Working Paper 18729, National Bureau of Economic Research January 2013.

Chen, Zhu, Jin-Nan Wang, Guo-Xia Ma, and Yan-Shen Zhang, "China tackles the health effects of air pollution," *The Lancet*, 2013, *382* (9909), 1959 – 1960.

Chu, D. A., Y. J. Kaufman, C. Ichoku, L. A. Remer, D. Tanr, and B. N Holben, "Validation of MODIS aerosol optical depth retrieval over land," *Geophysical Research Letters*, 2002, *29* (12), MOD2–1–MOD2–4.

Donaldson, Dave and Adam Storeygard, "The View from Above: Applications of Satellite Data in Economics," *Journal of Economic Perspectives*, November 2016, *30* (4), 171–98.

Ghanem, Dalia and Junjie Zhang, "Effortless Perfection: Do Chinese cities manipulate air pollution data?," *Journal of Environmental Economics and Management*, 2014, *68* (2), 203 – 225.

Kampa, Marilena and Elias Castanas, "Human health effects of air pollution," *Environmental Pollution*, 2008, *151* (2), 362 – 367. Proceedings of the 4th International Workshop on Biomonitoring of Atmospheric Pollution (With Emphasis on Trace Elements).

Kan, Haidong and Bingheng Chen, "Particulate air pollution in urban areas of Shanghai, China: health-based economic assessment," *Science of The Total Environment*, 2004, *322* (1), 71 – 79.

Koelemeijer, R.B.A., C.D. Homan, and J. Matthijsen, "Comparison of spatial and temporal variations of aerosol optical thickness and particulate matter over Europe," *Atmospheric Environment*, 2006, *40* (27), 5304 – 5315.

Li, Hongbin and Li-An Zhou, "Political turnover and economic performance: the incentive role of personnel control in China," *Journal of Public Economics*, 2005, *89* (9), 1743 – 1762.

M.D., Xiping Xu, Douglas W. Dockery, David C. Christiani, Baoluo Li, and Huiying Huang, "Association of Air Pollution with Hospital Outpatient Visits in Beijing," *Archives of Environmental Health: An International Journal*, 1995, *50* (3), 214–220. PMID: 7618954.

Montinola, Gabriella, Yingyi Qian, and Barry R. Weingast, "Federalism, Chinese Style: The Political Basis for Economic Success in China," *World Politics*, 1995, *48* (1), 5081.

Rohde, Robert A. and Richard A. Muller, "Air Pollution in China: Mapping of Concentrations and Sources," *PLOS ONE*, 08 2015, *10* (8), 1–14.

Sullivan, Daniel M. and Alan Krupnick, "Using Satellite Data to Fill the Gaps in the US Air Pollution Monitoring Network," 2018. Working Paper. http://www.rff.org/valuables/research/publications/using-satellite-data-fill-gaps-us-air-pollution-monitoring-network.

van Donkelaar, Aaron, Randall V. Martin, Michael Brauer, Ralph Kahn, Robert Levy, Carolyn Verduzco, and Paul J. Villeneuve, "Global Estimates of Ambient Fine Particulate Matter Concentrations from Satellite-Based Aerosol Optical Depth: Development and Application," *Environmental Health Perspectives*, 2010, *118* (6), 847–855.

Wang, Jun and Sundar A. Christopher, "Intercomparison between satellite-derived aerosol optical thickness and PM2.5 mass: Implications for air quality studies," *Geophysical Research Letters*, 2003, *30* (21).

Xu, Chenggang, "The Fundamental Institutions of China's Reforms and Development," *Journal of Economic Literature*, 2011, *49* (4), 1076–1151.

Xu, Zhaoyi, Dogian Yu, Libin Jing, and Xiping Xu, "Air Pollution and Daily Mortality in Shenyang, China," *Archives of Environmental Health: An International Journal*, 2000, *55* (2), 115–120. PMID: 10821512.

Zou, Eric, "Unwatched Pollution: The Effect of Intermittent Monitoring on Air Quality," 2018. Working Paper. https://files.webservices.illinois.edu/7199/zoeric-jmp.pdf.