Information, Incentives and Air Quality: New Evidence from Machine Learning Predictions

Yue "Luna" Huang[*], Minghao Qiu[†]

Apr 16, 2018

Click here for link to the latest version

Abstract

In command-and-control regulations, information asymmetry between central regulators and local agents is often cited as a key issue leading to ineffective policies. We evaluate a policy in China, which built air quality monitoring stations and enforced automatic data reporting to the central government, effectively preventing data manipulations by local officials. Exploiting the staggered implementation of this policy across 367 cities, we examine the impacts of the policy on local air quality. However, before monitoring stations were set up and data were credibly reported, we cannot observe pre-treatment air quality data. To overcome this challenge, we leverage recent development in machine learning (specifically, extreme gradient boosting) and a rich set of satellite images from NASA and reconstruct a comprehensive air pollution dataset in China with almost 0.5 million observations spanning from 2005 to 2016. Our structural break estimates do not demonstrate significant program effects.

[*]yuehuang@berkeley.edu, Ph.D. Student in Department of Agricultural and Resource Economics, University of California, Berkeley.

[†]mhqiu@mit.edu, Ph.D. Student in Institute for Data, Systems and Society, MIT.

[‡]We thank Solomon Hsiang, Ted Miguel, Marco Gonzalez-Navarro, Michael Anderson, Joshua Blumenstock, Max Auffhammer, Thibault Fally, Ling Jin, Wei Lin, Yulei He, Danny Kannell, Robert Pickmans, Jay Sayre, Carly Trachtman, Molly Van Dop, James Sears, Hannah Druckenmiller, Katie Wright for helpful comments and suggestions, and Hilary Yu, Hannah Burak and David Contreras for assisting with data preprocessing and model training. We are especially grateful to Chiyu Jiang for generously providing technical support and computational resources.

1 Introduction

How important is information for guaranteeing regulatory enforcement and improving environmental governance in developing countries? A recent review article by Kosack and Fung (2014) illustrated that the provision of credible public information can lead to improvement of public services[1]. In the environmental economics literature, availability of information on environmental quality is also shown to be crucial for policy effectiveness. Assunção, Gandour, and Rocha (2013) attributed the 2000s deforestation slowdown in the Brazilian Amazon to satellite-based real-time deforestation monitoring. Cisneros, Zhou, and Borner (2015) showed that starting in 2008, the Brazilian Ministry of the Environment has regularly published blacklists of critical districts with high annual forest loss, which considerably reduced deforestation.

This research question is particularly relevant in China, where there are large information asymmetries between local and central governments. The intense inter-jurisdiction political competition in China also presents the potential for inducing competition amongst local regulators to improve environmental performance. We therefore may expect policies that seek to improve data reliability and transparency to have strong positive effects on local environmental quality. These expectations are echoed by China's recent surge in investments in monitoring equipments that amount to 6 billion CNY (approximately 0.95 billion USD) *in just one year* (2015)[2]. As a benchmark, EPA estimates that the Clean Air Act incurred approximately 65 billion USD *in total* from 1990 to 2020[3]. While the investments in monitoring technologies are large and growing, to our knowledge, there have not been a study that examined the effects of these investments on air quality.

Before 2013, the air pollution regulations in China have been a mix of command-and-

[1]See Supplemental Materials (Kosack and Fung 2014) for a list of 16 experimental studies examining whether transparency leads to better governance.

[2]See this article for more information.

[3]See the EPA report for more information.

control policies and tax and/or subsidies focused on emission reduction. This, however, is subject to severe mis-reporting issues on the part of local authorities. In 2013, the Chinese central government signed many "contracts"[4] with local governments, where local officials promised to reduce *ambient* $PM_{2.5}$ levels by 10%, 15% or 25% by 2017 (compared to 2013 levels). Specifically, very stringent targets were set for Beijing and the surrounding areas, as well as some other heavily polluted provinces. Perhaps surprisingly, while targets were set for almost all the major provinces, the baseline 2013 $PM_{2.5}$ levels were not measured in many parts of the country. National monitors were set up from 2013 to 2016, and we would expect that the increase in the capacity to monitor performance of local governments would lead to improvement in environmental quality.

We address two key challenges to answer this policy question.

First, investment in monitoring technologies and disclosure of environmental data often coincide with the availability of environmental data. In other words, researchers almost never have access to comparable and reliable data on the environmental quality *before* disclosure of air pollution data. Lack of pre-treatment data makes it virtually impossible for researchers to evaluate the validity of their empirical strategies.

With recent developments in machine learning methods and the availability of daily satellite images through NASA, we are now able to reconstruct a historical air pollution dataset. We train our data on hourly, automatically reported air pollution ground measurements in 2015 and 2016, and recovered balanced time series on about 1500 monitoring stations across China from 2005 to 2016. This dataset inherits its fidelity from the fidelity of satellite images, and ensures that our analysis is sufficiently powered.

Second, selection into or out of treatment may be non-random. Cities that were regulated first may be cleaner or dirtier, resulting in a classical form of omitted variable bias.

We exploit the variations in the roll-out of the mandatory reporting of $PM_{2.5}$ data across

[4]All the contracts can be found here in Chinese.

Chinese cities from 2013 to 2015. While we argue that data reporting in most of the cities is a result of regulations from the central government and is therefore largely exogeneous, this remains a key challenge throughout our analysis.

More concretely, our testable hypothesis is, **does building national monitoring stations reduce information asymmetry between central and local regulators, incentivize local regulators to reduce emission, and thus improve air quality?**

Surprisingly, we find no effects of this policy on any of the major air pollutants, despite its scale and its large costs. Our estimates are consistent across specifications. In line with Kosack and Fung's (2014) findings, we believe that the information is not provided in a way that sets up sufficient incentives for local officials to improve performance. This policy only provides information on outputs (air quality) rather than inputs (compliance with regulations, or reduction in emissions). This policy only presents absolute information on performance (air quality), rather than comparative information that allowed easy comparisons across city officials. This policy also does not recommend or imply clear actions for citizens to take in response to the information.

The remainder of the paper is structured as follows. Section 2 relates this paper to several strands of literature. Section 3 describes the policy that generates variation in data transparency across Chinese cities. Section 4 describes how we constructed our dataset from satellite images with a machine learning algorithm. Section 5 describes our event study identification strategy and tests for lack of differential pre-trends. Section 6 presents our main results. Section 7 discusses threats to validity of our results. Section 8 concludes.

2 Literature

This paper is very much in line with the literature looking at inter-jurisdiction competition in China and its consequences for regulations and environmental quality. Kahn, Li, and Zhao

(2015) show that more precise measurement of water pollution on province borders changed the incentives for local officials and caused a reduction in water pollution for upstream provinces. Y. J. Chen, Li, and Lu (2018) show that a shift in the evaluation criteria of politicians induces meaningful reduction in SO_2 emissions in targeted cities. In contrast to these policies, the policy that this paper looks at is more opaque in terms of how it enters the performance evaluation of local officials, and that may be the reason why we do not find meaningful changes in air quality.

This paper is also related to a burgeoning literature using remote sensing data in policy evaluations to avoid issues with cross-country data consistency or poor quality of official statistics. Donaldson and Storeygard (2016) provides an excellent review of various applications of remote sensing data. Instead of directly using remote sensing observations, which can contain many missing values and be very noisy, we match these data to ground-level measurements and construct a predictive model to recover past air pollution levels. Our paper is similar in spirit to Di et al. (2017), which builds a neural network to predict high-resolution particulate matter levels in the U.S., and uses the derived dataset for estimating the health effects of air pollutants among sensitive populations.

Finally, in environmental sciences and atmospheric sciences, many studies have attempted to predict ground-level particulate matter concentrations with satellite-derived aerosol products and meteorological information, using either tradition methods like chemical transport models or state-of-the-art machine learning algorithms. Chu et al. (2016) offers an overview of the advantages and disadvantages of different types of models and their relative performances. We borrow heavily from this literature, especially in terms of choices of features and satellite data products.

3 Policy

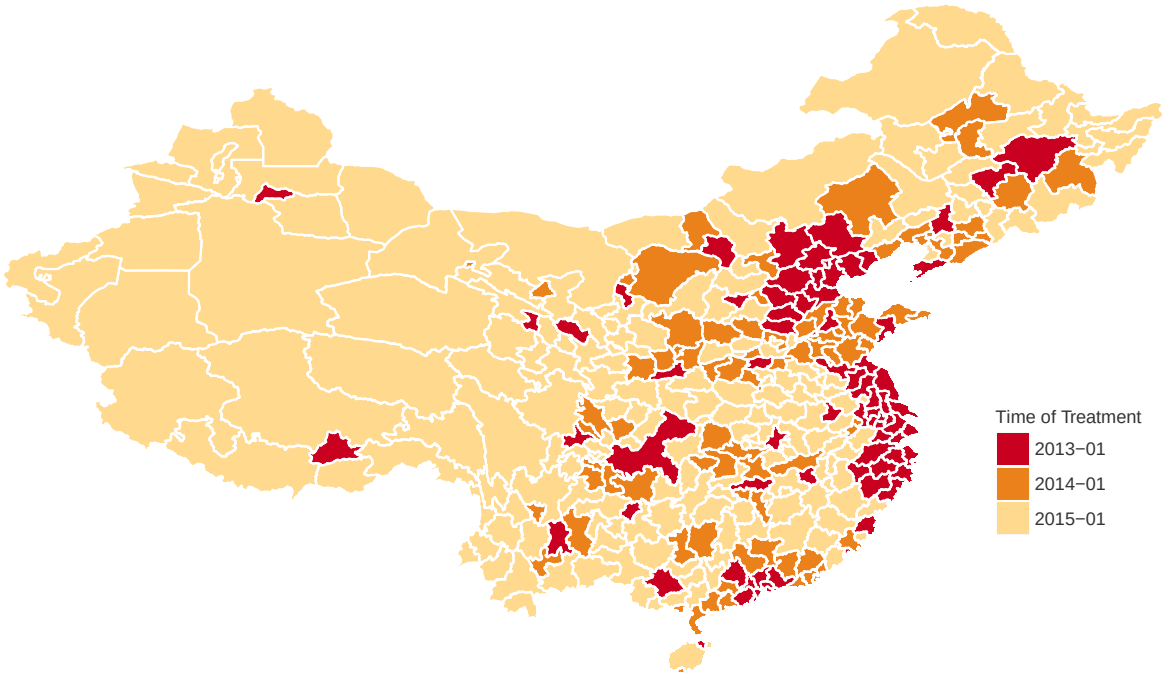
Despite the fact that air pollution has long been a serious public health hazard in China, people knew and cared very little about it before 2010. In 2008, the U.S. Embassy in Beijing built a monitoring station on the roof and started publishing $PM_{2.5}$ readings on Twitter. This quickly attracted media attention when these readings "went off the chart"[5] for several days in early 2012. Back then, $PM_{2.5}$ was not even reported in China's official air pollution data reporting system, and the data that existed were an aggregate daily reading combining PM_{10}, NO_2 and SO_2, which is widely considered to be unreliable and susceptible to human manipulation.

In response to accumulating pressure on social media, in 2012, the central government launched a policy for building a national monitoring network. The policy consists of modifying existing monitoring stations for measure three new pollutants ($PM_{2.5}$, O_3 and CO), setting up new monitoring stations, and stipulating that these monitoring stations must automatically report real-time hourly data to a national air quality database, making the data available online to the public. This significantly reduced information asymmetry between central regulators and local agents.

We define the treatment as **reporting of Fine Particulate Matter ($PM_{2.5}$) monitoring data to the central government.**. The national policy treats all the cities in three waves, in Jan 2013, Jan 2014 and Jan 2015, respectively. Figure 1 shows the roll-out of this policy across country. We leverage the rich variation in policy changes to identify the impacts of this policy on local air quality.

[5]The Air Quality Index exceeded 500, beyond which the index is not defined.

Figure 1: Time of Treatment: Dates when Cities Start Reporting $PM_{2.5}$ Values



Notes: (i) The cities shown in the figure do not exactly match the cities in our dataset. The former is taken from the prefecture-level city boundaries from the GADM dataset and consists of 344 cities (with one duplicate city). The latter is as defined in the Air Quality Index dataset as mentioned above and consists of 367 cities, including both prefecture-level cities and county-level cities (which are typically smaller than prefecture-level cities). Most of the mismatched cities are smaller cities that are likely to be treated on Jan 1, 2015 with the rest of the country. We use this assumption when plotting this graph. All the remaining analyses do not require linking cities to geographical boundaries, so this would not impact our other results. (ii) We impute these dates based on an NGO report and official policy documents (first, second, third). (iii) While the west part of China seems to have less geographical variation in terms of policy, those cities are typically not very densely populated.

4 Data

In this paper, we use a novel source of data: machine learning predictions. We overcome two major challenges.

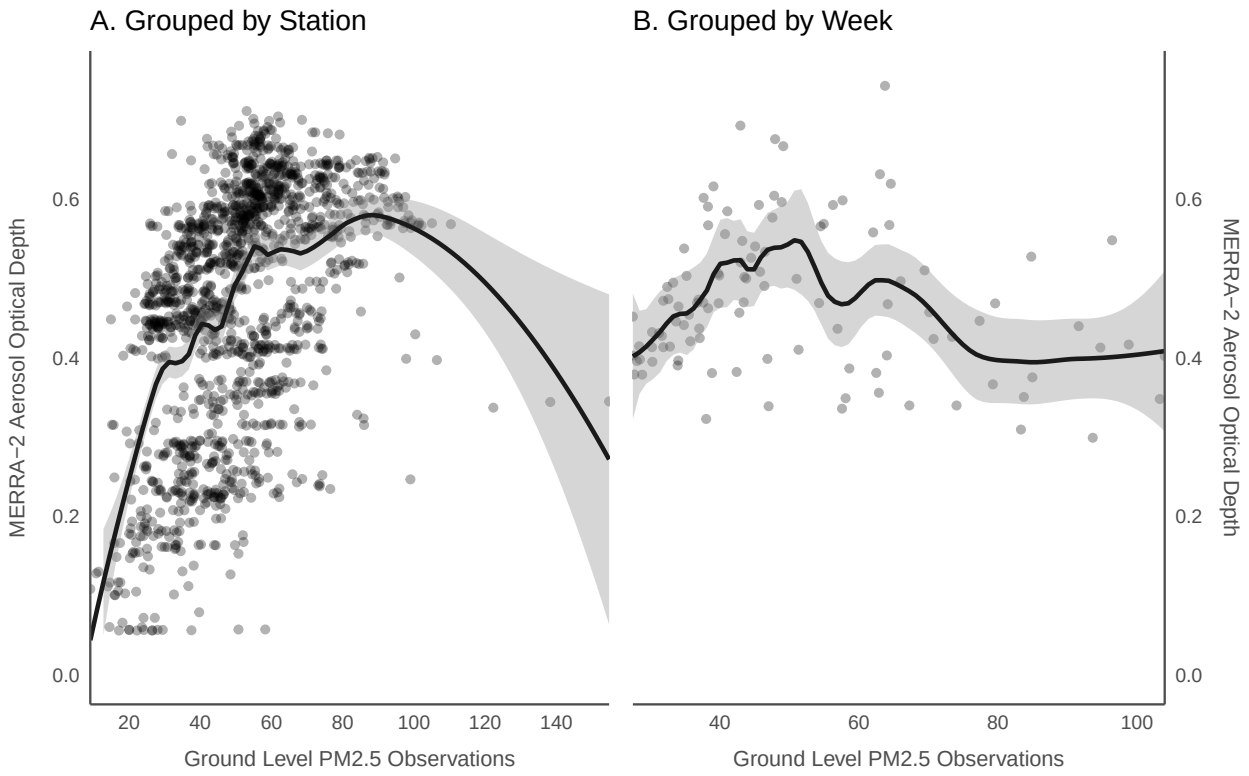
First, we reconstruct a dataset from 2005 to 2016, when official data were either non-

existent (for $PM_{2.5}$, O_3 and CO) or shown to be subject to human manipulation (for PM_{10}, SO_2 and NO_2, as we will also show in Figure 5, 11 and 12). Our predictions only infer from satellite images from NASA, which are objective measurements. We train our model on data in 2015 to 2016, which are considered to be more reliable because they are directly and automatically reported from monitors and publicized in real time on data-center websites.

Second, we improve upon directly using satellite data products (see, for example, Gendron-Carrier et al. (2018) for using aerosol optical depth as a proxy for particulate matter). Our approach takes into consideration that satellite data often contain many non-random missing values. We use Extreme Gradient Boosting, which conducts surrogate split to impute missing values flexibly. The surrogate split strategy leverages the high dimensionality of our dataset, and groups observations that are "similar" to conduct imputations. Importantly, we recover **ground-level concentrations**, which have more direct welfare and health consequences, whereas raw satellite products report **column concentrations**. Figure 2 shows that despite the fact the the column AOD concentrations are fairly strongly correlated with surface $PM_{2.5}$ concentrations when observations are grouped by monitoring station, the correlations are really weak when observations are grouped by week. Column concentrations capture much of the geographical variations but not the temporal variations, which is ultimately the variations used for doing statistical inference.

This section is structured as follows. Section 4.1 introduces the satellite data products that we use as our input data. Section 4.2 describes our machine learning model which establishes an input-output mapping between satellite images and ground-level observations. Section 4.3 reports the performance of our model. Section 4.4 validates our predictions against officially reported statistics. Section 4.5 describes our limitations.

Figure 2: Correlation between Aerosol Optical Depth and $PM_{2.5}$ in China, 2015–2016



Notes: (i) In the left panel, each dot represents a station (out of over 1400 stations across the country). (ii) In the right panel, each dot represents a week (out of over 100 weeks in 2015–2016).

4.1 Satellite Images

We leverage data from a variety of sources and match the observations in satellite data products onto all of the monitoring stations in our sample. We use observations from 2015 to 2016 as training data and predict air pollution levels for 2005 to 2016. Additionally, we use data in 2014 (from May to December) to examine how well our model perform out-of-sample. Table 1 shows all the feature and target variables that we extracted from the satellite images. We include a set of meteorological variables that are shown to be important for the physical and chemical processes through which air pollutants form and interact in

the atmosphere[6]. See Section SM.1 for a more detailed description of each of our datasets.

Table 1: Targets, Features and Data Sources

Targets (2015–2016 for Training, 2014 for Test)	Dataset	Source
Monitoring Station Measurements ($PM_{2.5}$, PM_{10}, NO_2, SO_2, O_3, CO)	AQI	Harvard Dataverse
Reconstructed Air Pollution Index		
Features (2005–2016)	Dataset	Source
Day of Year		
Aerosol Optical Depth (Aqua and Terra)	MODIS	NASA EarthData
SO_2, NO_2, O_3 Column Concentrations	OMI	NASA EarthData
CO, O_3 and AOD Reanalysis Product	MERRA2	NASA EarthData
Temperature, Relative Humidity, Pressure, Eastward and Northward Wind Speed, Planetary Boundary Layer Height	MERRA2	NASA EarthData

Notes: (i) Air Pollution Index is reconstructed with the official aggregation rules using observations from AQI: first piecewise linearly transform PM_{10}, SO_2 and NO_2, then take the max of the subindices. (ii) AQI: Air Quality Index. (iii) MODIS: Moderate Resolution Imaging Spectroradiometer. (iv) OMI: Ozone Monitoring Instrument. (v) MERRA2: Modern-Era Retrospective analysis for Research and Applications Version 2.

Table 2 presents our summary statistics. Because of cloud coverage and satellite malfunction, missing values in high-frequency satellite data are very pervasive. Table 2 shows the data coverage of several of our key features on one particular day is rather poor. To address this challenge, we first leverage spatial and temporal neighbors to provide more information to the model. We add a spatial Epanechnikov kernel with a radius of 1.5 degrees[7] and 3-day and 7-day moving averages for the variables with many missing values (mainly from OMI

[6]Relative humidity is shown to be important for the formation of particulate matters. Wind speed determines how air pollutants are transmitted in the atmosphere. Planetary boundary layer height are positively correlated with column concentrations.

[7]This is roughly how far air pollutants travel in a day, calculated from the average wind speed in China.

Table 2: Descriptive Statistics of Feature and Target Variables

Description	Kernel	Moving Average	Adjustment	Mean	(Std. Dev.)	Missing
Full Sample, 1497 Stations, 4383 Days						
Temperature at 2m				286.79	(11.45)	0.07%
Relative humidity at 985–1000 hPa				0.60	(0.22)	0.0%
Pressure				95.74	(7.98)	0.0%
Planetary boundary layer height				594.52	(625.27)	0.05%
Eastward wind speed at 2m				-0.37	(1.42)	0.02%
Northward wind speed at 2m				-0.10	(1.82)	0.0%
Day of year				183.13	(105.44)	0.0%
AOD (MODIS Terra)				0.54	(0.51)	94.68%
AOD (MODIS Terra)	Yes			0.54	(0.61)	80.93%
AOD (MODIS Terra)	Yes	3-day		0.55	(0.6)	57.86%
AOD (MODIS Terra)	Yes	7-day		0.55	(0.55)	36.29%
AOD (MODIS Terra)	Yes		Yes	0.14	(0.26)	80.93%
AOD (MODIS Aqua)				0.52	(0.49)	95.03%
AOD (MODIS Aqua)	Yes			0.53	(0.59)	81.75%
AOD (MODIS Aqua)	Yes	3-day		0.53	(0.57)	59.09%
AOD (MODIS Aqua)	Yes	7-day		0.54	(0.53)	37.13%
AOD (MODIS Aqua)	Yes		Yes	0.14	(0.26)	81.76%
AOD reanalysis (MERRA2)				0.53	(0.42)	0.02%
NO_2 (OMI)				0.29	(0.44)	66.1%
NO_2 (OMI)	Yes			0.24	(0.41)	54.75%
NO_2 (OMI)	Yes	3-day		0.25	(0.39)	16.9%
NO_2 (OMI)	Yes	7-day		0.25	(0.35)	1.62%
O_3 (OMI)				295.47	(40.33)	24.46%
O_3 (OMI)	Yes			296.30	(40.83)	23.85%
O_3 (OMI)	Yes	3-day		296.07	(39.63)	0.6%
O_3 (OMI)	Yes	7-day		296.05	(38.1)	0.14%
O_3 reanalysis (MERRA2)				296.00	(40.1)	0.0%
SO_2 (OMI)				0.28	(0.68)	63.83%
SO_2 (OMI)	Yes			0.24	(0.68)	61.93%
SO_2 (OMI)	Yes	3-day		0.23	(0.58)	31.47%
SO_2 (OMI)	Yes	7-day		0.22	(0.48)	12.44%
CO reanalysis (MERRA2)				257.41	(262.75)	0.0%
Training Sample, 1497 Stations, 731 Days						
Reconstructed daily Air Pollution Index				66.68	(38.81)	3.9%
24-h mean CO ground level				1.08	(0.75)	3.35%
24-h mean NO_2 ground level				31.88	(20.32)	3.32%
24-h mean O_3 ground level				56.99	(31.56)	3.34%
24-h mean PM_{10} ground level				86.53	(73.65)	3.58%
24-h mean $PM_{2.5}$ ground level				50.17	(42.66)	3.34%
24-h mean SO_2 ground level				24.26	(29.23)	3.28%

Notes: (i) The spatial kernel is an Epanechnikov kernel with a radius of 1.5 degrees. (ii) Both the spatial kernel and 3-day and 7-day moving averages omit missing values and average over observed data points with given weights. (iii) The adjusted AOD is $AOD \times (1 - RelativeHumidity)/PlanetaryBoundaryHeight$ (Zheng et al. 2017).

and MODIS datasets). We also calculated adjusted aerosol optical depth based on Zheng et al. (2017) taking into account relative humidity and planetary boundary layer height.

As shown in Table 2, even after imposing spatial and temporal kernels, we still have significant proportions of missing data. We then take advantage of the reanalysis products in the MERRA2 dataset that are produced by climate models. While they have been shown to be inaccurate in particularly high levels of air pollution, the machine learning model should be flexible enough to correct for this bias.

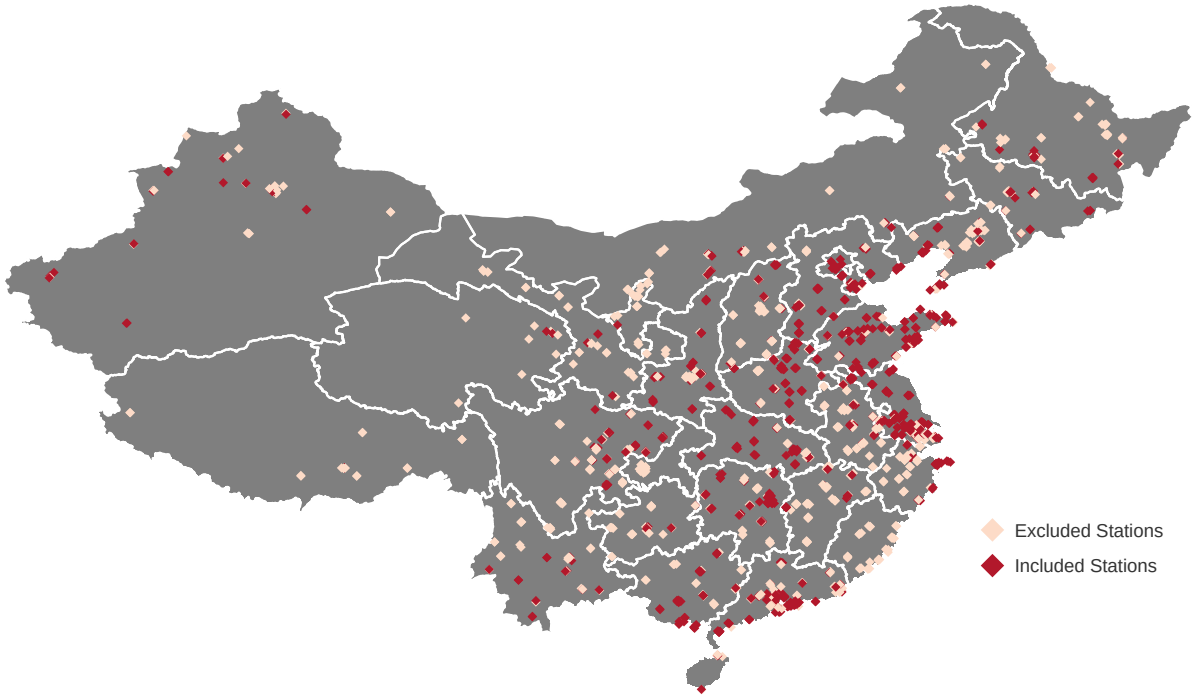
4.2 Model

Machine learning models are useful in predicting air quality because these models are flexible enough to represent and learn complex non-separable and non-linear relationships, which are typically challenging to model but extremely important in the physical and chemical processes of air pollution formation and transmission. Compared with chemical transport models, machine learning models are easier to implement, less likely to suffer from misspecified physical or chemical relationships, and can predict air quality in a data-driven manner.

We use a regression-tree based algorithm, Extreme Gradient Boosting[8], for training our pollution prediction model. Instead of training a model for all the stations in our sample, we train a separate model for every single monitoring station. This takes into account the possibility that the data generating process in different stations may be different in nature (e.g. in high altitudes, chemical reactions follow a different pattern than places in low altitudes). This also makes sure that our model will be trained to capture the temporal variations over time, making it more suitable for reconstructing past datasets. We describe more details on the training process in Section SM.2.

[8] A short introduction can be found in the xgboost documentation.

Figure 3: The Geographical Distribution of Included and Excluded Monitoring Stations



Notes: (i) We drop different sets of stations for different target variables, based on the cross validated daily R^2 for that particular target variable. This graph shows the station split for $PM_{2.5}$.

We discard stations where our predictions are not satisfactory. This can either be due to heavily missing features in certainly areas where it is usually cloudy, or where the target variables are also sparse (a few monitoring stations enter our sample much later than the beginning of 2015, leaving us with few informative observations). Unobserved factors can also cause our predictions to be inaccurate. More concretely, for each of the seven target variables, we evaluate our performance and discard half of the stations that have R^2 below the median. Discarding half of the stations does not reduce the rich variations in the policy that we need for identification. As shown in Figure 3, while we do exclude certain areas (presumably because of consistently poor satellite data quality), overall the sample is still representative, especially in densely populated areas (in the eastern part of China).

Finally, Extreme Gradient Boosting conducts surrogate splits, which internally addresses the issue of non-random missing-ness and places these observations into appropriate leaf nodes based on information from other dimensions. T. Chen and Guestrin (2016) offers a more formal treatment.

4.3 Performance

To measure performance, we report 5-fold cross-validated daily R^2 (Figure 3) and weekly R^2 (Figure 4) on training data[9]. While it is extremely difficult to predict daily pollution levels accurately, our model performance on weekly average pollution data is much better. All of our subsequent analysis are therefore based on weekly means. Performance on test data are reported in the Supplementary Materials in Section SM.2.

Importantly, unlike canonical methods, we do not shuffle our data when we cross validate on our training data. This is key to making sure that our performance reported here is comparable to out-of-sample prediction performances.

Table 3: Predictive Performance: Cross Validated Daily R^2 for Included Stations

Target Variable	Overall R^2	Station-Specific R^2 Percentiles				
		5%	10%	50%	90%	95%
API	0.55	0.28	0.29	0.36	0.47	0.51
PM_{10}	0.51	0.28	0.28	0.35	0.46	0.50
$PM_{2.5}$	0.50	0.32	0.33	0.39	0.50	0.52
O_3	0.62	0.43	0.44	0.55	0.69	0.73
SO_2	0.57	0.16	0.19	0.36	0.59	0.65
NO_2	0.60	0.29	0.31	0.41	0.53	0.56
CO	0.53	0.14	0.16	0.30	0.51	0.56

Notes: (i) Overall R^2 are calculated across all observations from the included stations. (ii) Station-specific R^2 are calculated within each station and thus have a distribution. (iii) we use 5-fold cross validation on training data to obtain predicted and true value pairs.

[9]Performance is only reported for included stations.

Table 4: Predictive Performance: Cross Validated Weekly R^2 for Included Stations

Target Variable	Overall R^2	Station-Specific R^2 Percentiles				
		5%	10%	50%	90%	95%
API	0.82	0.38	0.42	0.54	0.68	0.72
PM_{10}	0.80	0.37	0.40	0.52	0.66	0.70
$PM_{2.5}$	0.87	0.42	0.46	0.57	0.70	0.73
O_3	0.92	0.54	0.56	0.69	0.84	0.86
SO_2	0.86	0.19	0.24	0.48	0.76	0.81
NO_2	0.85	0.34	0.39	0.56	0.71	0.74
CO	0.92	0.16	0.21	0.43	0.69	0.73

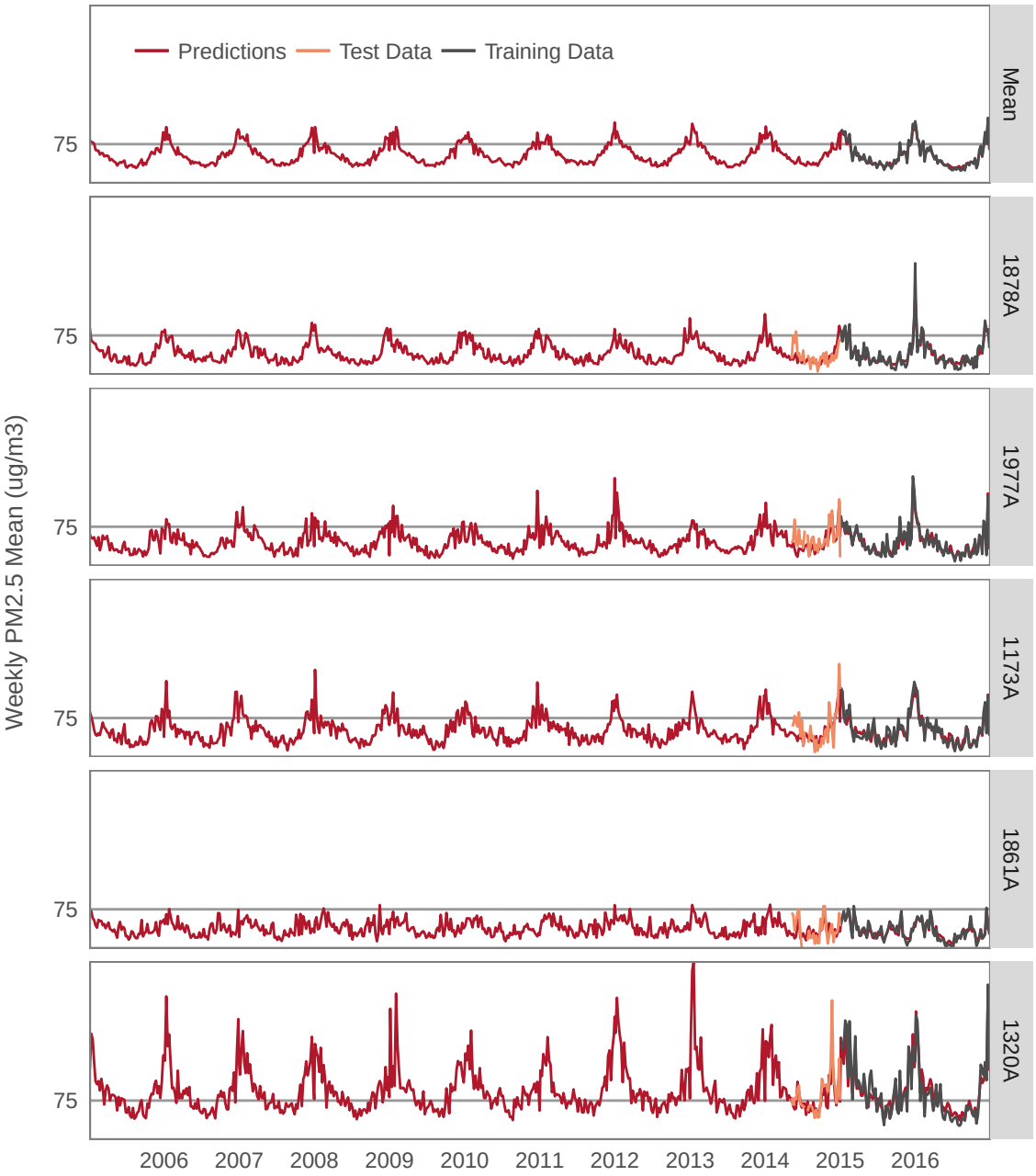
Notes: (i) Overall R^2 are calculated across all observations from the included stations. (ii) Station-specific R^2 are calculated within each station and thus have a distribution. (iii) we use 5-fold cross validation on training data to obtain predicted and true value pairs.

4.4 Validating Machine Learning Predictions

One major concern in using machine learning predictions for policy evaluations is whether these predictions are consistent over time so as to facilitate inference that relies heavily on variances within a unit over time[10]. In our model, each input variable is from the same satellite data product, so both input and output data should be consistent over time. As shown in Figure 4, although our predictions are quite a bit noisier, there does not seem to be any systematic biases. Notably, we cannot seem to predict "spikes" very well. This is a common problem for either traditional or machine learning predictive models—while there can be a relatively localized air pollution "spike" on the ground, the column concentration of air pollutants (i.e. the total amount of air pollution in the atmosphere "column") may not change by much, and satellite images cannot capture it well.

[10]Satellite decay may also be a problem but these systematic biases should impact all stations equally and thus be controlled for when we control for general time trends in our regressions.

Figure 4: $PM_{2.5}$ Trend in Predicted, Training and Test Data



Notes: (i) The five stations shown here are randomly drawn from our sample. (ii) Weekly means are shown here for display purposes. (iii) The machine learning model is trained on training data. Predictions in 2015–2016 here do not involve cross validation. (iv) For some stations (built after 2014), test data do not exist.

While the premise for building a machine learning model is that there are no reliable ground-level measurements, we still hope to present some suggestive evidence to validate predictions from our model.

First, Figure 4 and Table 6 shows that our predictions match test data reasonably well, which our model has never seen.

Second, we compare our results to the official statistics "Air Pollution Index" before 2013. Studies have shown that some city officials manipulate reported data by changing values above 100 to be below 100, in order to score a "blue sky day"[11] (Ghanem and Zhang 2014). Figure 5 shows the trends of predicted and reported API in China, and Figure 11 and Figure 12 show the trends for Beijing and Shanghai, respectively. The trends match up fairly well, but it is important to keep in mind that the reported data are not necessarily accurate. In Figure 11 (for Beijing), for example, there is a large discontinuity at 100 for Air Pollution Index in the density function, indicating data manipulations. In Figure 5, this is smoothed out because many cities do not manipulate their reported data.

For future work, we hope to use data collected in monitoring stations from Hong Kong and Taiwan, which contain a long time series covering 2005 up to 2016, to validate our methods and justify the extrapolation from existing training dataset.

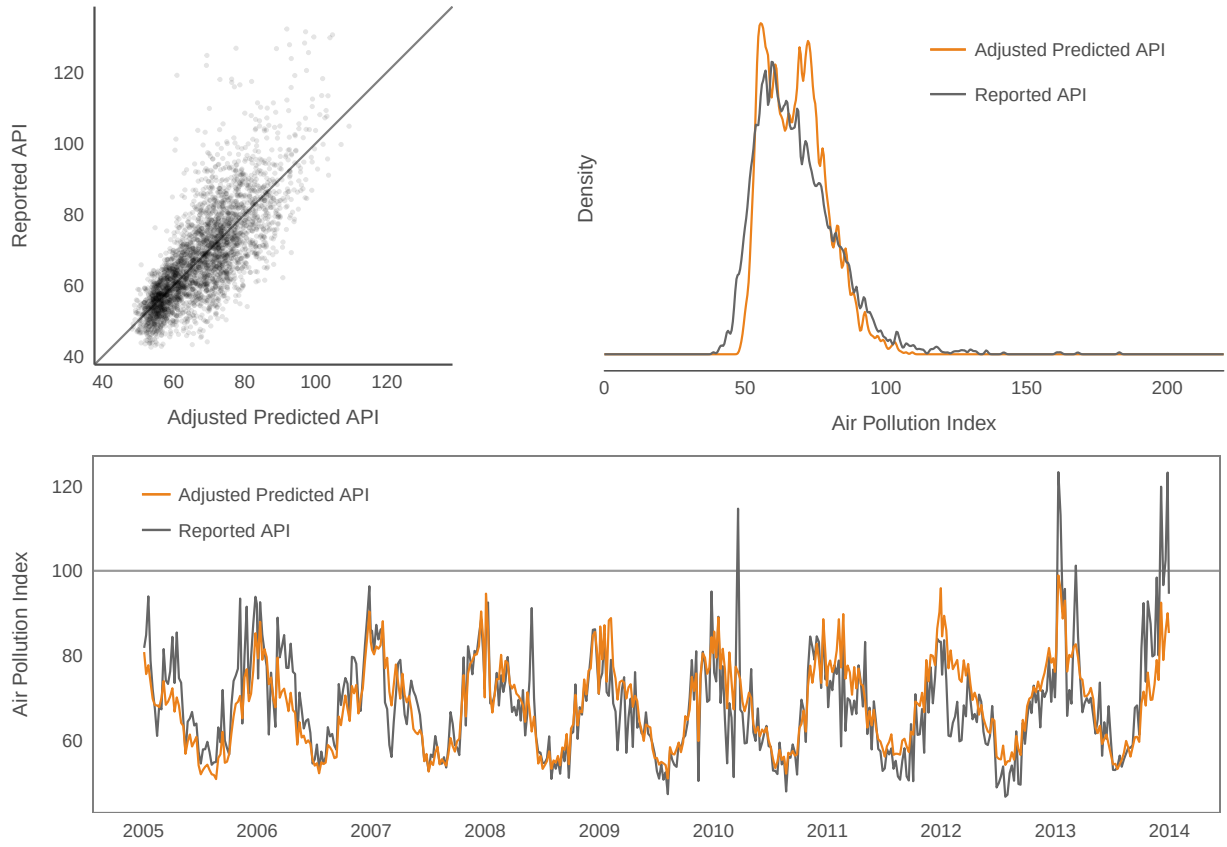
4.5 Limitations

One limitation of this project is that the underlying data generating process may be different for different years. There may be changes in compositions of air pollutants, which may then change the functional relationships that we are trying to model.

Another limitation is that bad predictions tend to be "clustered" in our time series. A close examination of cross-validated R^2 reveals that there are large variances within R^2 for

[11] This is defined to be days with an Air Pollution Index smaller than 100.

Figure 5: Comparing Predicted and Reported Air Pollution Index in China



Notes: (i) We build our predictions by setting the target variable to be API, which is calculated from ground-level measurements in 2015–2016. (ii) The upper panels plot daily API whereas the lower panel plot weekly means for display purposes. (iii) Adjusted predicted API is calculated by regressing reported API on predicted API and taking the fitted values. This is to account for differences in composition of monitoring stations in predicted and reported (city-level mean) API. (iv) Because API is the maximum of a piecewise linear transformation of the raw observations, discontinuities in the density graph are expected at the 50 and 100 cutoff. However, the point mass should be accumulated at slightly above 50 or 100. So this is still clear evidence of human manipulation.

different folds. These could have unclear consequences on our results.

Also, our inability to predict "spikes" will have no effect on our results if we define the outcome variable to be air pollution levels, but could bias our results if the outcome variable is number of polluted days or weeks.

5 Identification Strategy

We test for structural breaks in the time series using the following estimating equation.

$$Y_{iwy} = \alpha_{iw} + \beta_y + \tau_j \mathbf{1}\{K_{iwy} \geq j\} + \epsilon_{iwy} \tag{1}$$

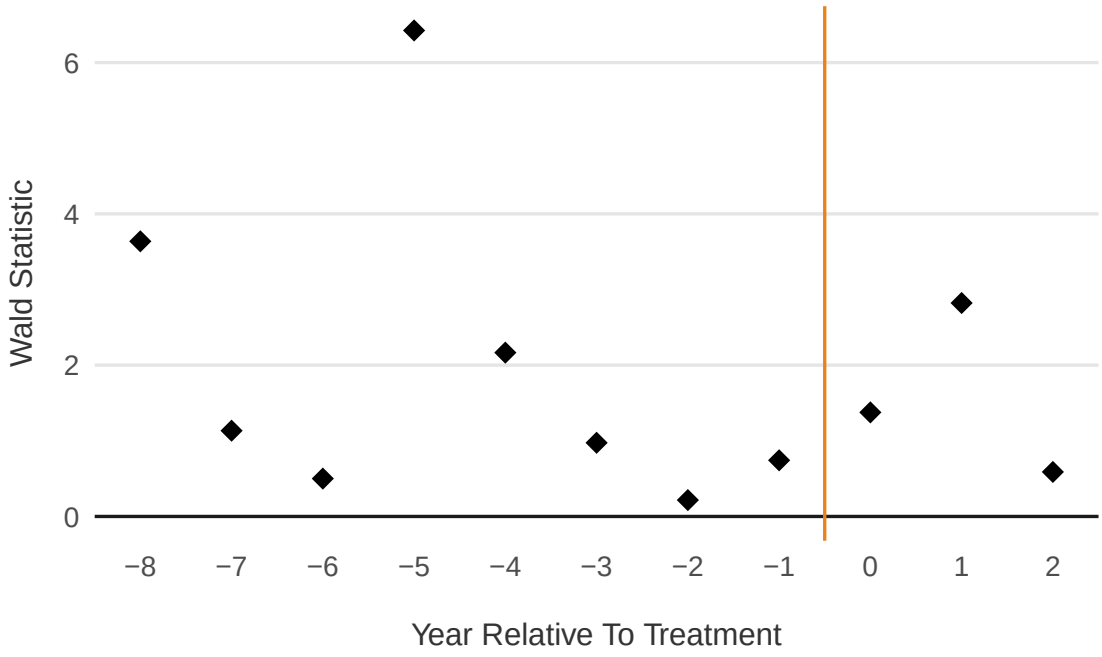
where

- Each i indicates one monitoring station;
- Each w indicates one week, each y indicates one year;
- K_{iwy} is the year relative to treatment;
- $j \in [-8, 2]$ is the placebo or actual treatment time;
- Y_{iwy} is average weekly air pollution levels;
- ϵ_{iwy} is clustered at the city level.

There is obvious selection in the roll-out of the policy—heavily polluted cities tend to get treated first, and well-developed cities also tend to get treated early. We think it is plausible that treatment and control cities may have differential pre-trends in air pollution levels. Therefore, we test for a structural break in the air pollution trends with the actual policy treatment (when $j = 0$), as well as placebo policy treatments (when $j \neq 0$).

As Figure 6 shows, we plot Wald Statistics for testing whether τ_j is significant, against j, the hypothesized treatment time. For $j = 0$, this can be understood as implementing a event-study design and testing for whether the coefficient for the treatment indicator is significant. For $j \neq 0$, this is essentially doing a number of placebo tests, where a placebo treatment is assumed to be j years after (or before, if $j < 0$) the actual treatment time. If the treatment has significant effects on outcomes of interests, then all the Wald Statistics

Figure 6: Structural Break Estimates: Machine Learning Predictions for $PM_{2.5}$



before the actual treatment should be relatively small, and the one for $j = 0$ should be large, whereas for $j > 0$ the Wald Statistics should be large yet declining. In Figure 6, the Wald Statistics estimates are fairly noisy and do not show consistent patterns.

6 Results

Our main results are shown in Figure 7 and Figure 8. Consistent with Figure 6, treatment does not seem to have large effects on air quality. In Figure 7, we look at the air quality data predicted by our machine learning model, which reflect surface concentrations that people are exposed to. In Figure 8, we run the same regressions with satellite data, which gives us more confidence that the null results are robust to the outcome variables that we use. In the lower right panel, it does seem that there is a small jump of SO_2 column concentrations from the raw satellite observations after the treatment, indicating that building national monitoring stations have reduced SO_2 column concentrations to some extent. However, it is

purely suggestive.

We also run an alternative specification with the following estimating equation.

$$Y_{iwy} = \alpha_i + \gamma T_{wy} + \delta \mathbf{1}\{K_{iwy} \geq j\} + \tau_j \mathbf{1}\{K_{iwy} \geq j\} \times T_{wy} + \epsilon_{iwy} \tag{2}$$

where

- Each i indicates one monitoring station;
- Each w indicates one week, each y indicates one year;
- T_{wy} is a continuous variable indicating calendar time;
- K_{iwy} is the year relative to treatment;
- $j \in [-8, 2]$ is the placebo or actual treatment time;
- Y_{iwy} is average weekly air pollution levels;
- ϵ_{iwy} is clustered at the city level.

Figure 9 and Figure 10 show the results for the alternative specification. Compared to our main specification, this specification tests for whether policy treatment changes the trends for these air pollutants, instead of levels. These specifications are more parametric and less flexible (assuming linear pre- and post- trends). However, they provide consistent evidence that policy treatment does not have systematic effects on air quality.

7 Threats to Validity

7.1 Biases and Noises in Machine Learning Predictions

Despite some suggestive evidence that we present in Section 4, there remains the possibility that there are systematic biases in our machine learning predictions. The fact that day of

Figure 7: Structural Break Estimates: Machine Learning Predictions

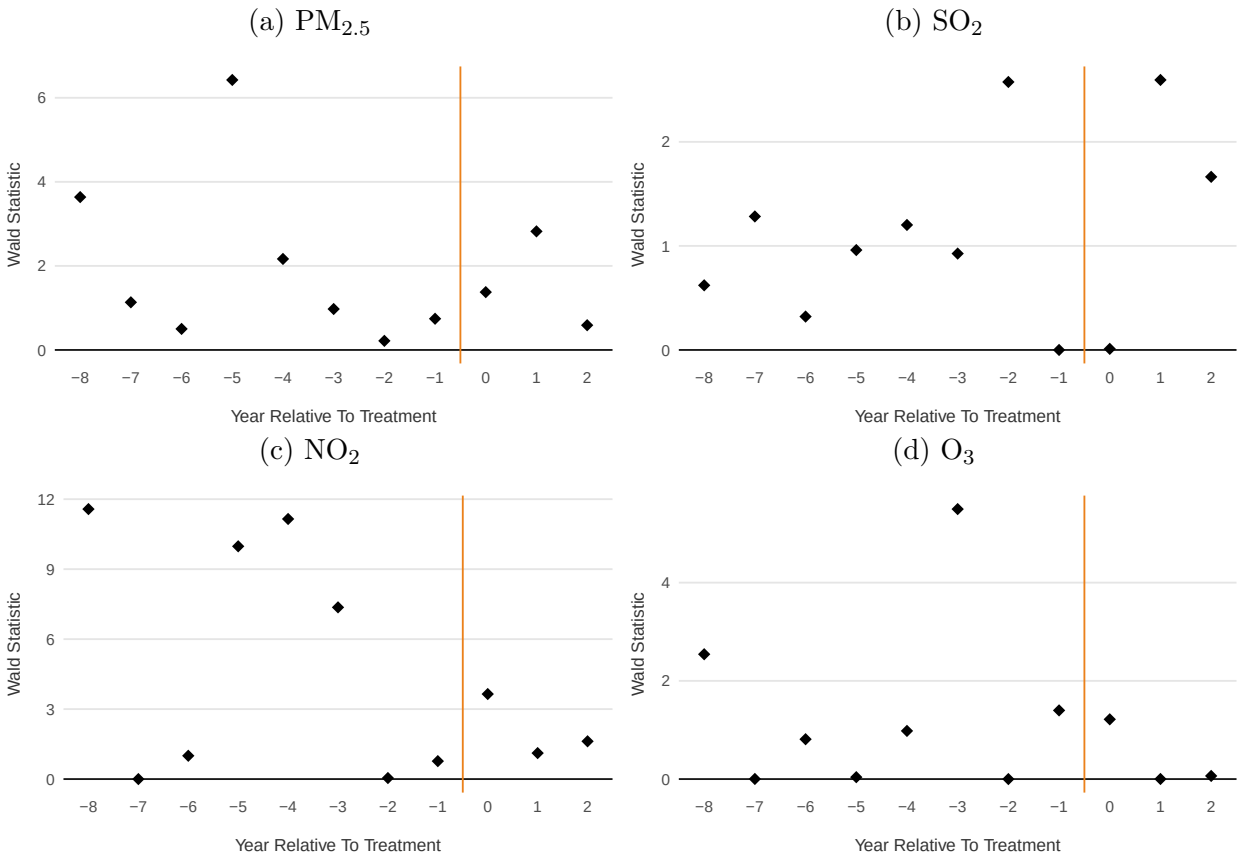
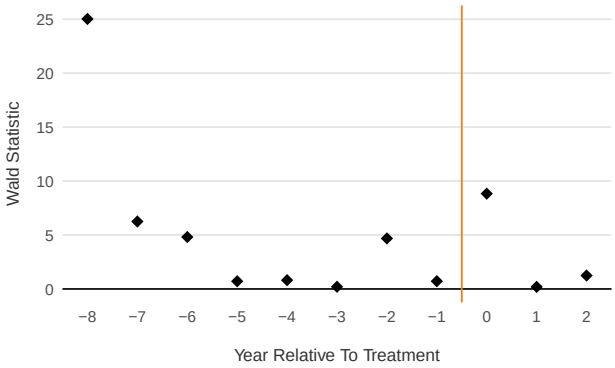
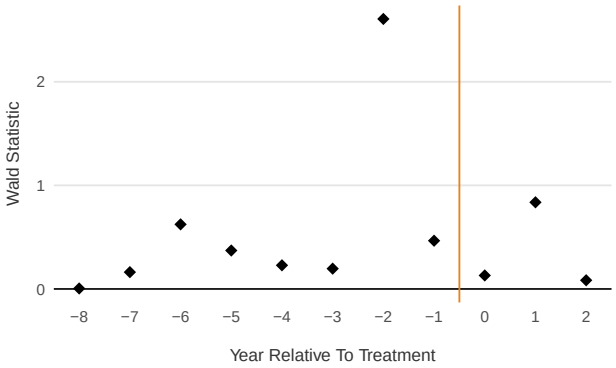


Figure 8: Structural Break Estimates: Satellite Observations

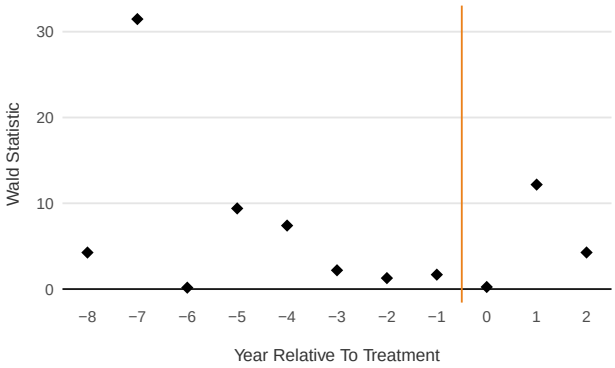
(a) AOD (MERRA2)



(b) AOD (Terra)



(c) O_3 (Aura)



(d) SO_2 (Aura)

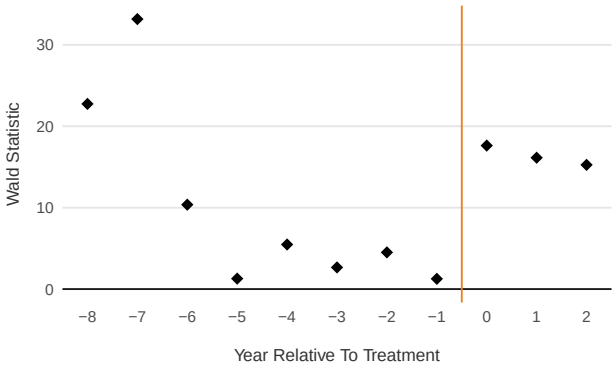


Figure 9: Structural Break Alternative Specification: Machine Learning Predictions

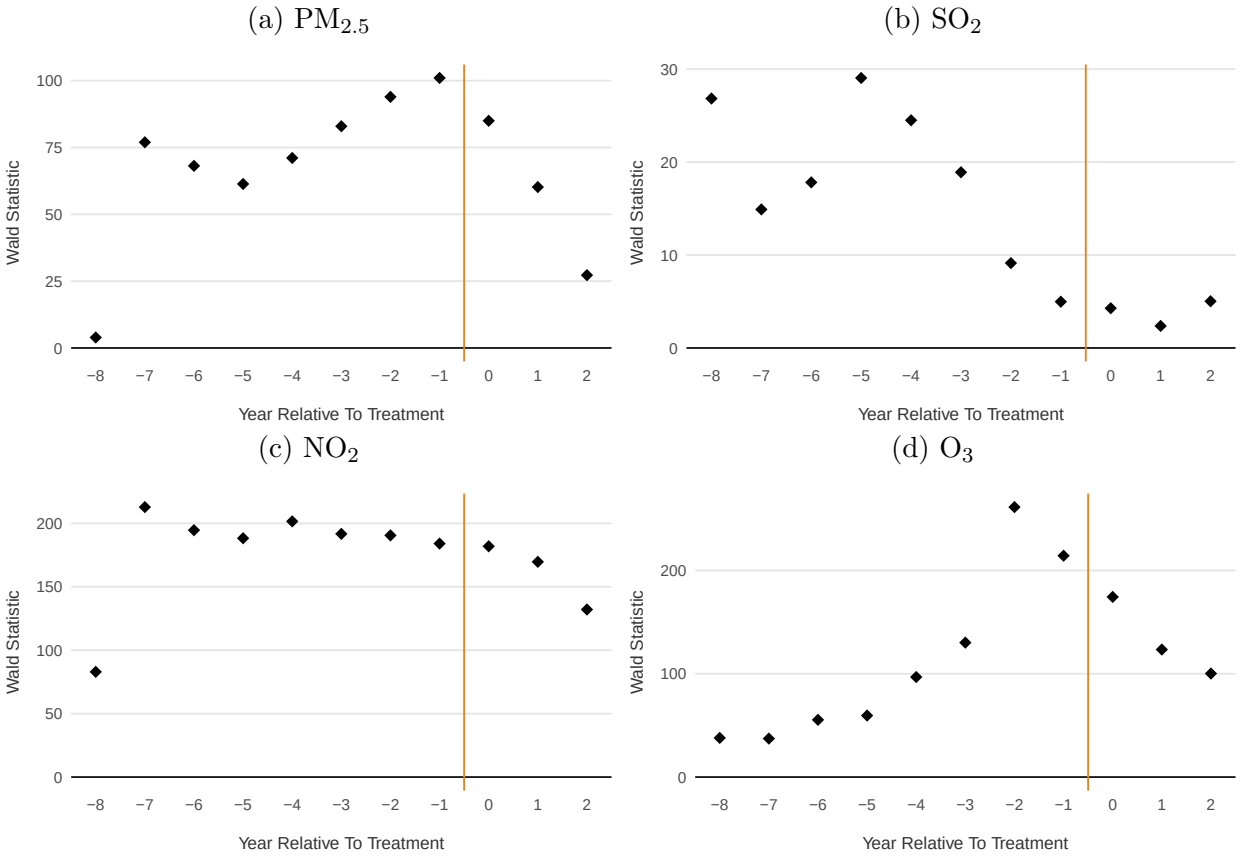
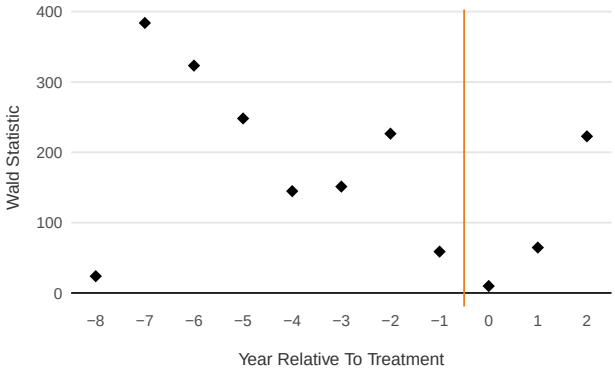
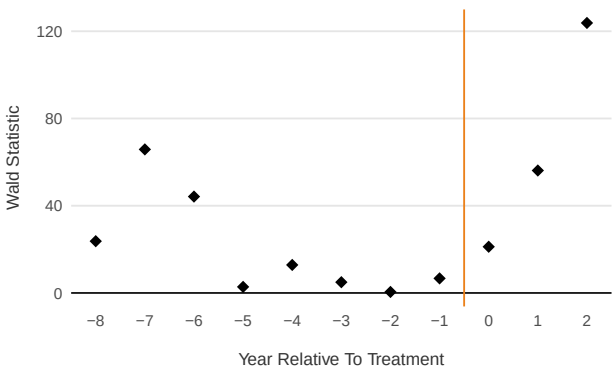


Figure 10: Structural Break Alternative Specification: Satellite Observations

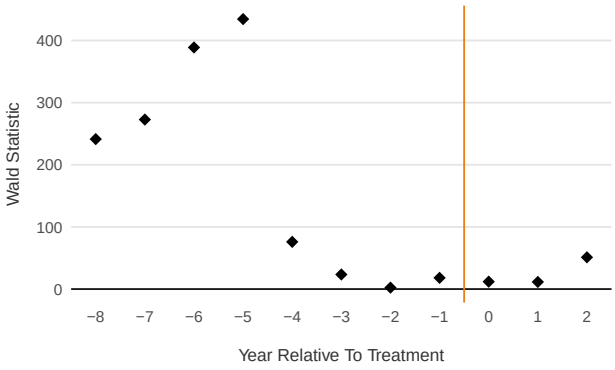
(a) AOD (MERRA2)



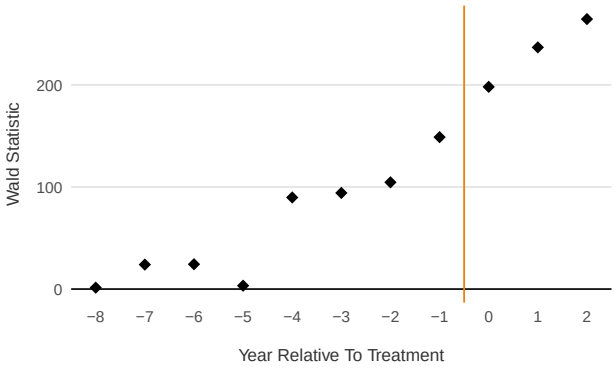
(b) AOD (Terra)



(c) O_3 (Aura)



(d) SO_2 (Aura)



year is a strong predictor in our model could indicate that much of the variances that we are capturing is seasonal, and our ability to predict long-term trends are not as good as indicated by our cross-validation results.

7.2 Contamination of Controls

The issue with every information treatment is that our controls are likely to be contaminated. Local officials may feel a greater pressure to improve environmental performance when other cities start reporting $PM_{2.5}$ data. Importantly, it is common knowledge that all the regulated provinces and cities should achieve their target by 2017, and the exact time when monitoring activity starts may then be less crucial for incentivizing emission reduction efforts.

Since the policy was announced before it was implemented, there could also be strong anticipatory effects.

These contaminations may reduce the differences in "intensity" of treatment for our different treatment groups and thus bias our estimates towards zero.

8 Conclusion

While the investments in monitoring technologies are large and growing in China, improving data quality and transparency alone seems insufficient to generate meaningful improvements in air quality. We use a novel dataset to evaluate a policy aimed at reducing information asymmetry between central regulators and local agents, and find that despite seemingly stringent air quality targets and large incentives for local officials, the policy has no significant effects on air quality.

References

Assunção, Juliano, Clarissa Gandour, and Romero Rocha (2013). "DETERring Deforestation in the Brazilian Amazon: Environmental Monitoring and Law Enforcement". In:

Cai, Wenju et al. (2017). "Weather conditions conducive to Beijing severe haze more frequent under climate change". In: *Nature Climate Change* 7.4, pp. 257–262.

Chen, Tianqi and Carlos Guestrin (2016). "Xgboost: A scalable tree boosting system". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, pp. 785–794.

Chen, Yvonne Jie, Pei Li, and Yi Lu (2018). "Career concerns and multitasking local bureaucrats: Evidence of a target-based performance evaluation system in China". In: *Journal of Development Economics*. URL: https://www.sciencedirect.com/science/article/pii/S0304387818300245.

Chu, Yuanyuan et al. (2016). "A Review on Predicting Ground PM2.5 Concentration Using Satellite Aerosol Optical Depth". In: *Atmosphere* 7.10. ISSN: 2073-4433. DOI: 10.3390/atmos7100129. URL: http://www.mdpi.com/2073-4433/7/10/129.

Cisneros, Elias, Sophie Lian Zhou, and Jan Borner (2015). "Naming and shaming for conservation: evidence from the Brazilian Amazon". In: *PloS one* 10.9, e0136402.

Di, Qian et al. (2017). "Air pollution and mortality in the Medicare population". In: *New England Journal of Medicine* 376.26, pp. 2513–2522.

Donaldson, Dave and Adam Storeygard (2016). "The view from above: Applications of satellite data in economics". In: *Journal of Economic Perspectives* 30.4, pp. 171–98.

Gendron-Carrier, Nicolas et al. (2018). *Subways and Urban Air Pollution*. Working Paper 24183. National Bureau of Economic Research. DOI: 10.3386/w24183. URL: http://www.nber.org/papers/w24183.

Ghanem, Dalia and Junjie Zhang (2014). "Effortless Perfection: Do Chinese cities manipulate air pollution data?" In: *Journal of Environmental Economics and Management* 68.2, pp. 203–225. ISSN: 0095-0696. DOI: http://dx.doi.org/10.1016/j.jeem.2014.05.003. URL: http://www.sciencedirect.com/science/article/pii/S0095069614000400.

Kahn, Matthew E., Pei Li, and Daxuan Zhao (2015). "Water Pollution Progress at Borders: The Role of Changes in China's Political Promotion Incentives". In: *American Economic Journal: Economic Policy* 7.4, pp. 223–42. DOI: 10.1257/pol.20130367. URL: http://www.aeaweb.org/articles?id=10.1257/pol.20130367.

Kosack, Stephen and Archon Fung (2014). "Does Transparency Improve Governance?" In: *Annual Review of Political Science* 17.1, pp. 65–87. DOI: 10.1146/annurev-polisci-032210-144356. URL: https://doi.org/10.1146/annurev-polisci-032210-144356.

Wang, Julian XL and James K Angell (1999). "Air stagnation climatology for the United States". In: *NOAA/Air Resource Laboratory ATLAS* 1.

Zheng, C. et al. (2017). "Analysis of influential factors for the relationship between $PM_{2.5}$ and AOD in Beijing". In: *Atmospheric Chemistry and Physics* 17.21, pp. 13473–13489. DOI: 10.5194/acp-17-13473-2017. URL: https://www.atmos-chem-phys.net/17/13473/2017/.

Supplementary Materials

SM.1 Data

SM.1.1 Air Quality Index

This dataset mirrors the officially-reported real-time air quality data on the Ministry of Environmental Protection website. It is downloaded from National AQI Beijing Air Archive Dataverse. It spans over two and a half years (May 2014 to December 2016) and contains hourly observations from 1497 national monitoring stations across China. Six air pollutants are included in this dataset: SO_2, NO_2, $PM_{2.5}$, PM_{10}, O_3, CO. We take the 24-hour mean of these hourly observations to match with our features. We split the data into two parts. We use data from 2015–2016 for training and data from part of 2014 for testing.

SM.1.2 Air Pollution Index

This dataset is used primarily for comparison with our predicted air quality indices before more monitoring stations were set up and higher air quality standards were enforced from 2012 to 2015. The Air Pollution Index (API) dataset is scraped from the Ministry of Environmental Protection website. It contains daily reported API levels for about 120 cities across the country.

Air Pollution Index is the aggregated measure of severity of air pollution in China before 2013. All three pollutants included in API (PM_{10}, NO_2, SO_2) are transformed with piecewise linear functions to map onto $[0, 500]$, and API is then taken to be the maximum of all subindices. **We cannot observe the subindices in publicly available data.** PM_{10} usually dominate other subindices in Air Pollution Index because SO_2 and NO_2 standards are rather lax before 2013.

SM.1.3 OMI

The Ozone Monitoring Instrument (OMI) Aura dataset provides column concentrations for ozone (O_3), SO_2 and NO_2. All the data are obtained from Goddard Earth Sciences Data and Information Services Center. The total column density of SO2 in the Planetary Boundary Layer (PBL) is calculated based on an improved Band Residual Difference Algorithm (BRD). The column density of NO2 in troposphere is calculated based on a troposphere-stratosphere separation algorithm. All observations are in Dobson Units.

SM.1.4 MODIS

The MODIS level-2 atmospheric aerosol product provides full global coverage of aerosol properties from the Dark Target (DT) and Deep Blue (DB) algorithms. We obtained two sources of AOD dataset from MODIS products: MODIS/Aqua Aerosol data (MYD04) from EarthData and MODIS/Terra Aerosol data (MOD04) from EarthData. The original dataset has a spatial resolution of 10km.

SM.1.5 MERRA-2

The second Modern-Era Retrospective analysis for Research and Applications (MERRA-2) dataset is used for providing relevant meteorological variables and reanalysis air pollution data products. It is obtained from Goddard Earth Sciences Data and Information Services Center. The original dataset has the resolution of 0.5 degree latitude by 0.625 degree longitude.

We extract six relevant meteorological variables: relative humidity, planetary boundary layer height, air temperature at 2m above ground, wind speed at 2m above ground (eastward and northward) and surface pressure at 985–1000 hPa. We also extract carbon monoxide, ozone and aerosol optical depth reanalysis data products.

In future work, we also hope to analyze the meteorological variables associated with stagnation conditions. In particular, we analyze variables used to construct the air stagnation index (ASI) from Wang and Angell (1999) and haze weather index (HWI) from Cai et al. (2017). For ASI, we extract wind speed at mid-troposphere (500 hpa), differences between surface temperature and temperature at 850 hpa. For HWI, we furthur extract differences between temperature at 850 hpa and 250 hpa and the latitudinal differences of mid-troposphere wind speed (500 hpa).

Importantly, the features are all sampled to be between 10AM to 2PM (at noon) to better match with other satellite observations (such as MODIS and OMI)[12]. This is based on the fact that many meteorological variables behave very differently during day and night, and the measurement errors at night are much larger. Planetary boundary layer height, for example, is not accurately measured at night. However, the target variables are all 24-h averages. This is because empirically, we tend to predict 24-h averages better, presumably because it is less noisy.

We are particularly careful when using the reanalysis data products within the MERRA-2 dataset. MERRA-2 is produced by combining three components: (i) GEOS-5 atmospheric model (with little chemistry), (ii) data assimilation system, (iii) three dimensional variational data analysis (interpolation). An intuitive (and simplistic) way to understand MERRA-2 is to consider it as an atmospheric model with numerous (but often sparse) observations as the prior. The MERRA-2 system generates a "weighted mean" of the model forecast and observations based on the uncertainty of observations and model forecast.

In addition to the general setting in MERRA-2 modeling system, the aerosol product of MERRA-2 uses a chemical transport model (Goddard Chemistry, Aerosol, Radiation, and Transport model, GOCART) and emission inventories to simulate aerosols. The anthropogenic emission inventory which concerns us the most, is taken from EDGARv4.2 and

[12]Although the exact time may vary depending on the availability of the original dataset.

AeroCom Phase II. The observations used in generating AOD data are multiple satellite measurements of reflectance (AOD) and ground-based measurements (AERONET) including two sites in China. We are not concerned about misreporting and human manipulations for these input data, although the measurement errors could be large.

The biggest concern in using MERRA-2 reanalysis products is the potential time inconsistency of the model outputs given that they assimilate inputs from different sources over time. However, MERRA-2 is designed to use a no-changed assimilation system to compute all years. Certainly, as more observations are available for more recent years, the estimates for different years might differ in unclear ways. Our study period is from 2005 to 2016, where satellite data have already become available, and changes in data availability may be less of a concern.

SM.1.6 Data Pre-processing

All the input datasets are satellite images with values recorded on regular grids. We first match them with coordinates of monitoring stations and extract values corresponding to certain stations with bilinear interpolation. We record a value as being missing if amongst the four closest centroids in the raw raster, no observations are recorded.

To include relevant spatial and temporal information, we also apply an Epanechnikov kernel with a spatial bandwidth of 1.5 degree (the approximate distance over which air pollutants can travel in one day), discard missing values, and extract a variable of the reweighted mean of all the values within the bandwidth of the kernel. 3-day and 7-day moving averages are also added to the model to reduce numbers of missing values.

The bounding box for our raster is (16, 137, 56, 72) (S, E, N, W). The time range is from 2005-01-01 00:00:00 to 2016-12-31 23:59:59. The resolution is 0.1 degree longitude by 0.1 degree latitude. If the original datasets are not recorded on the desired 0.1 by 0.1 degree grid, we resample the data. If the observations are coarser, then bilinear interpolation is

used.

SM.2 Model

SM.2.1 Implementation

We use Extreme Gradient Boosting to learn the input-output mapping between remote sensing observations and ground-level measurements. Our target variables are the six air pollutants in Air Quality Index, plus the reconstructed Air Pollution Index. We use the `Python xgboost` package for training. The loss function that we specified is mean squared error.

Importantly, when we evaluate our performance with cross validation on our training data, **we do not shuffle our training data to create random train-test splits within the training data.** This is because we want cross-validated R^2 to accurately reflect our performance on a period of consecutive days, rather than on a set of randomly chosen days. The former is more informative for evaluating whether our extrapolation out-of-sample is valid.

We tune the models for each target variables separately, but all the stations in our sample share the same hyper-parameters to prevent hyper-parameter overfitting.

We compared the performance of Extreme Gradient Boosting with other models, such as linear regression, conventional gradient boosting and multilayer perceptron. We find that Extreme Gradient Boosting performs best, with fewer assumptions made about how to impute missing values.

We also examined feature importances to gain insight into the structure of the model. Day of year is highly predictive (dropping it results in a reduction in R^2 but the model performance remains reasonable), so are meteorological satellite reanalysis data products (from MERRA2). Some direct observations (such as AOD), on the contrary, is not very predictive, presumably because of pervasive missing values. After adding spatial and temporal kernels, their feature importances improve.

SM.2.2 Performance

We use held-out data from May 2014 to December 2014, which the model has not been trained on, as test data. Table 5 and 6 report the predictive performance of our model on test data, for daily predictions and weekly predictions, respectively.

Table 5: Predictive Performance: Test Daily R^2 for a Subset of Stations

Target Variable	Overall R^2	Station-Specific R^2 Percentiles				
		5%	10%	50%	90%	95%
API	0.45	-0.13	-0.00	0.28	0.50	0.55
PM_{10}	0.44	-0.12	0.01	0.28	0.47	0.54
$PM_{2.5}$	0.34	-0.07	0.01	0.30	0.51	0.58
O_3	0.55	-0.47	-0.01	0.53	0.72	0.75
SO_2	0.47	-0.42	-0.19	0.28	0.57	0.63
NO_2	0.48	-0.35	-0.08	0.34	0.56	0.60
CO	0.28	-0.66	-0.39	0.19	0.47	0.53

Notes: (i) Overall R^2 are calculated across all observations from the included stations. (ii) Station-specific R^2 are calculated within each station and thus have a distribution. (iii) we fit the model on training data in 2015 and 2016 and test it on test data from 2014.

We see no systematic differences in performance between overall cross-validated R^2 and test R^2, although the station-specific R^2 becomes noisier, especially for weekly predictions. These will likely overstate the variance in R^2 for the whole time series because this is a very short time period (about half a year) and there are simply few weekly observations for each station. Because of the autocorrelation structure in our data, R^2 over a short period of time is almost definitely noisier than R^2 over a longer period of time. In other words, good predictions and bad predictions tend to be "clustered" in terms of time. Also, only about 900 out of 1500 monitoring stations have data that traces back to 2014, which restricts our sample. We believe that cross-validated R^2 reported in the main text is more representative

Table 6: Predictive Performance: Test Weekly R^2 for a Subset of Stations

Target Variable	Overall R^2	Station-Specific R^2 Percentiles				
		5%	10%	50%	90%	95%
API	0.57	-0.34	-0.14	0.35	0.67	0.72
PM_{10}	0.53	-0.40	-0.09	0.34	0.65	0.70
$PM_{2.5}$	0.65	-0.36	-0.12	0.38	0.68	0.74
O_3	0.80	-0.89	-0.25	0.59	0.80	0.83
SO_2	0.55	-0.97	-0.51	0.35	0.74	0.80
NO_2	0.83	-0.89	-0.43	0.39	0.73	0.78
CO	0.86	-1.36	-0.73	0.22	0.66	0.74

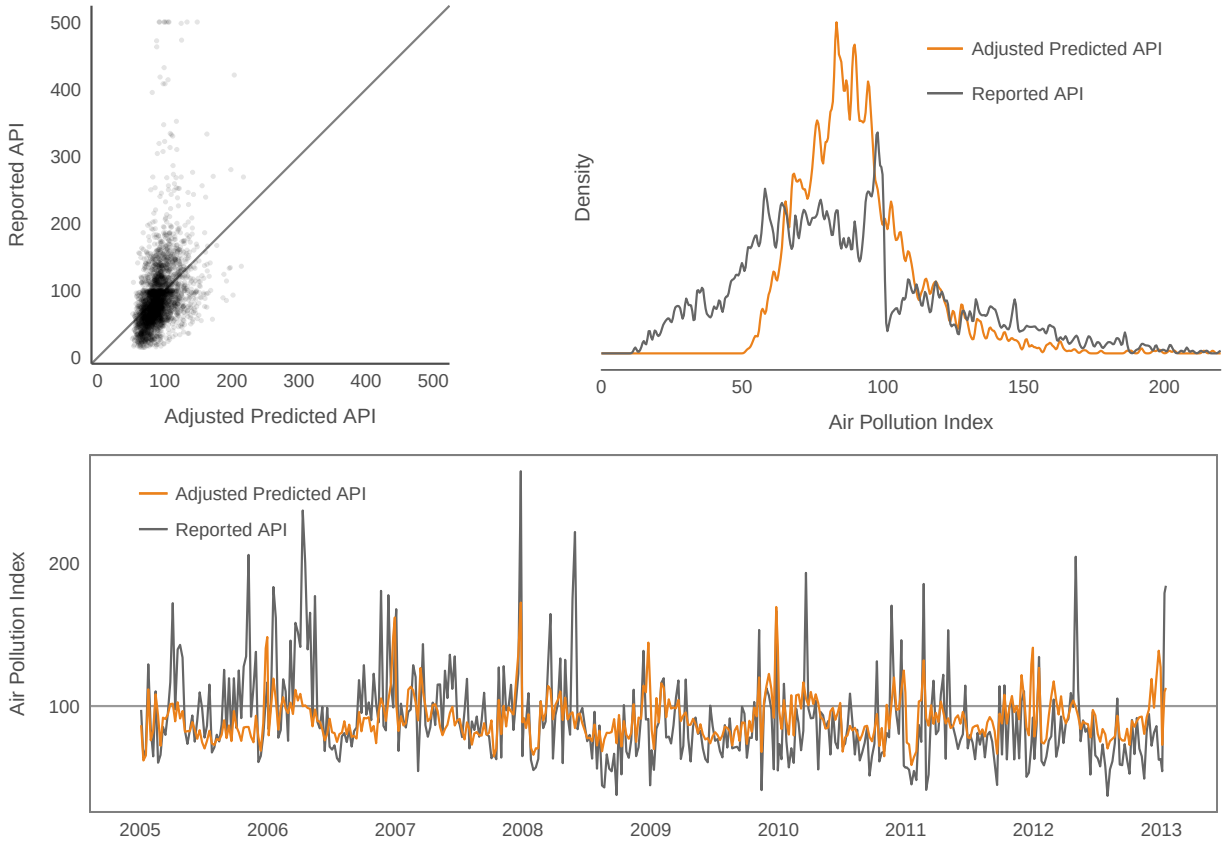
Notes: (i) Overall R^2 are calculated across all observations from the included stations. (ii) Station-specific R^2 are calculated within each station and thus have a distribution. (iii) we fit the model on training data in 2015 and 2016 and test it on test data from 2014.

of our model performance, although these results do indicate that we may have too small a sample size for each model that we are training[13].

[13]One viable solution is to purchase scraped data from 2017 on from the black market.

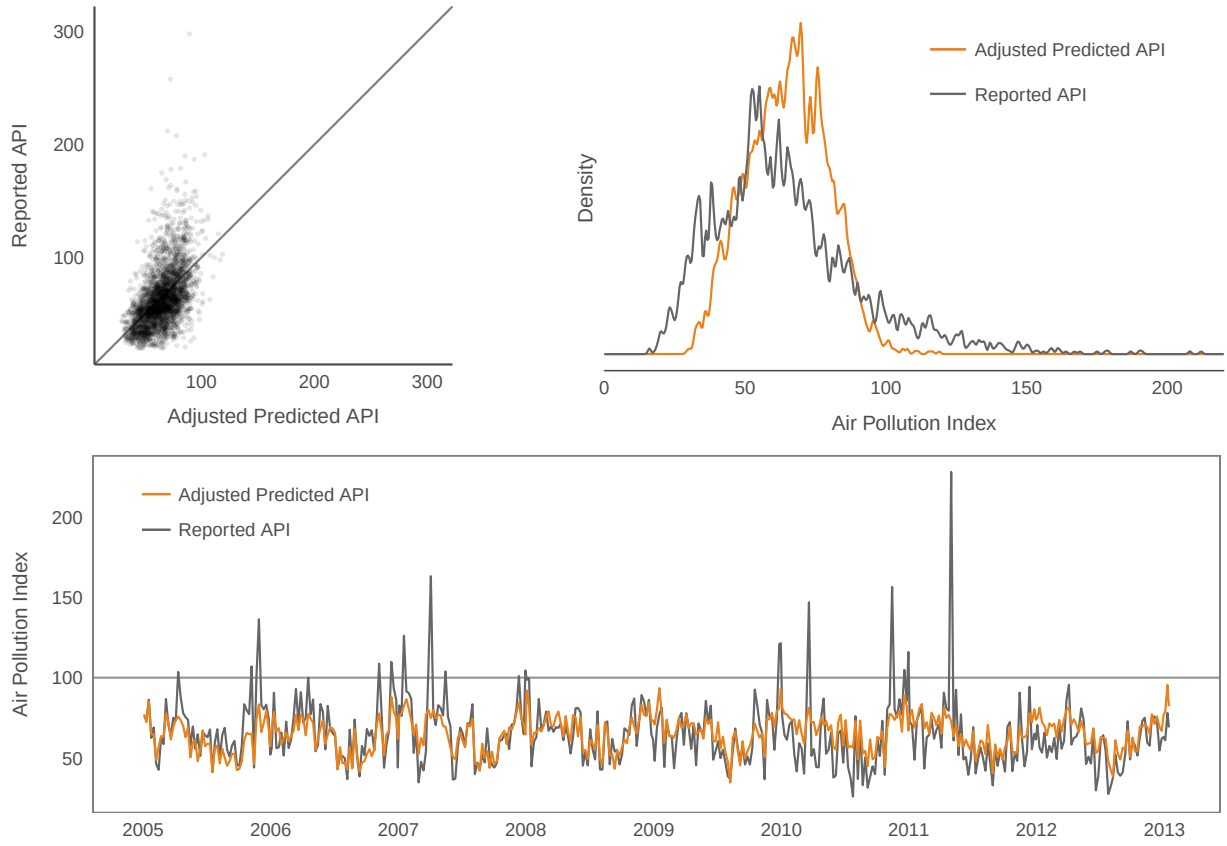
SM.3 Tables and Figures

Figure 11: Comparing Predicted and Reported Air Pollution Index in Beijing



Notes: (i) We build our predictions by setting the target variable to be API, which is calculated from ground-level measurements in 2015–2016. (ii) The upper panels plot daily API whereas the lower panel plot weekly means for display purposes. (iii) Adjusted predicted API is calculated by regressing reported API on predicted API and taking the fitted values. This is to account for differences in composition of monitoring stations in predicted and reported (city-level mean) API. (iv) Because API is the maximum of a piecewise linear transformation of the raw observations, discontinuities in the density graph are expected at the 50 and 100 cutoff. However, the point mass should be accumulated at slightly above 50 or 100. So this is still clear evidence of human manipulation.

Figure 12: Comparing Predicted and Reported Air Pollution Index in Shanghai



Notes: (i) We build our predictions by setting the target variable to be API, which is calculated from ground-level measurements in 2015–2016. (ii) The upper panels plot daily API whereas the lower panel plot weekly means for display purposes. (iii) Adjusted predicted API is calculated by regressing reported API on predicted API and taking the fitted values. This is to account for differences in composition of monitoring stations in predicted and reported (city-level mean) API. (iv) Because API is the maximum of a piecewise linear transformation of the raw observations, discontinuities in the density graph are expected at the 50 and 100 cutoff. However, the point mass should be accumulated at slightly above 50 or 100. So this is still clear evidence of human manipulation.