Information, Incentives and Air Quality:

New Evidence from Machine Learning Predictions

GARESC Conference

Luna Yue Huang[1] Minghao Qiu[2]

April 14, 2018

[1]University of California, Berkeley

[2]Massachusetts Institute of Technology

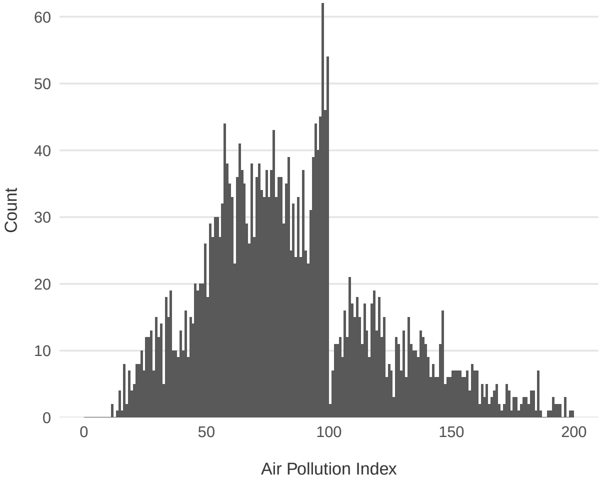
Stylized Fact 1: Rampant Data Manipulation

- Anecdotal evidence: *The Ministry of Environmental Protection inspected 8,500 businesses in Beijing and surrounding areas and found that over 3,100 factories had tampered with their emission monitoring equipment and altered reported data*[1]
- Statistical evidence: Local government officials also manipulate air quality data to satisfy targets assigned by the central government

[1] Source: Caixin Global News Article

Stylized Fact 1: Rampant Data Manipulation

Figure 1: Histogram of Reported Air Pollution Index in Beijing, 2005–2013



Stylized Fact 2: Recent Improvement in Data Quality

- Recent surge in investments in monitoring equipments in China that amount to approximately 0.95 billion USD in just 2015 (Clean Air Act incurred approximately 65 billion USD in 30 years).
- Much more stringent regulations on maintaining the fidelity of air quality data:
  - Require real-time hourly data to be automatically publicized on data center websites and mobile apps
  - Employees at local environmental protection bureaus cannot have keys to monitoring stations[2]

[2]Source: Jinchu News Article

Testable Hypothesis

Testable Hypothesis

Does building national monitoring stations reduce information asymmetry between central and local regulators, incentivize local regulators to reduce emission, and thus improve air quality?

Institutional Context:

- Lack of any $PM_{2.5}$ information before 2012
- Intensive inter-jurisdiction competition for political promotion
- In 2013, the central government signed separate "contracts" with provincial leaders promising reduction in ambient $PM_{2.5}$ levels of up to 25% in five years

Policy

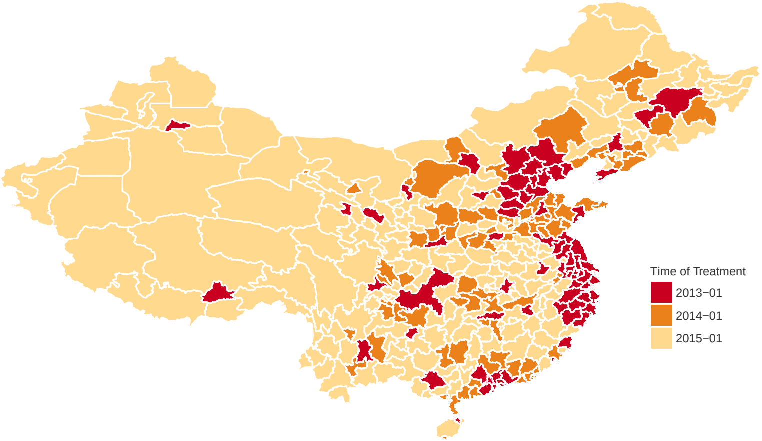
Testable Hypothesis

Does building national monitoring stations reduce information asymmetry between central and local regulators, incentivize local regulators to reduce emission, and thus improve air quality?

- Treatment: Reporting of Fine Particulate Matter ($PM_{2.5}$) monitoring data to the central government.
- "Contracts" were signed between central and local government to reduce $PM_{2.5}$ by a specific target value (ranging from 5% to 25%) by 2017
- The central government imposed the regulation on 74 cities in Jan 2013, over 100 cities in Jan 2014, and the rest in Jan 2015.

Policy

Figure 2: Time of Treatment: Dates when Cities Start Reporting $PM_{2.5}$ Values

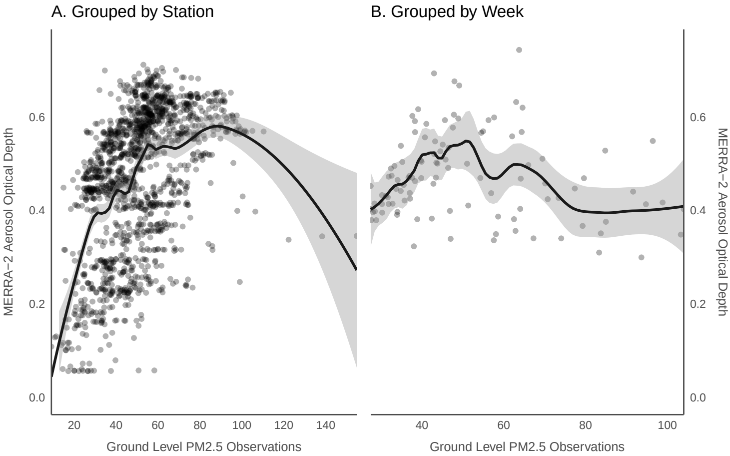


Key Contribution: Data

- Challenge: Data did not exist before monitoring stations were built—pre-treatment data are unavailable
- Solution: Recent development in machine learning, combined with satellite images collected by NASA, allows us to reconstruct historical air pollution datasets
- Compared to directly using satellite observations, we recover ground-level concentrations, with real welfare and health consequences, whereas raw satellite products report column concentrations

Column Concentrations Capture Little Temporal Variation

Figure 3: Aerosol Optical Depth and $PM_{2.5}$ in China, 2015–2016



Data: Overview

- We feed our machine learning model with satellite data throughout 2005–2016 as features, train our model on 2015–2016 ground-level observations, and use it to predict 2005–2016 ground-level concentrations, 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).
- We train a different model for every single station amongst about 1500 stations, and drop half of the stations which do not yield satisfactory performance.
- We use Extreme Gradient Boosting, which is a variant of Random Forest and a regression-tree-based algorithm. It conducts surrogate splits to do “smart” imputations for observations with missing features.

Data: Targets and Features

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) Reconstructed Air Pollution Index	AQI	Harvard Dataverse
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

Data: Performance

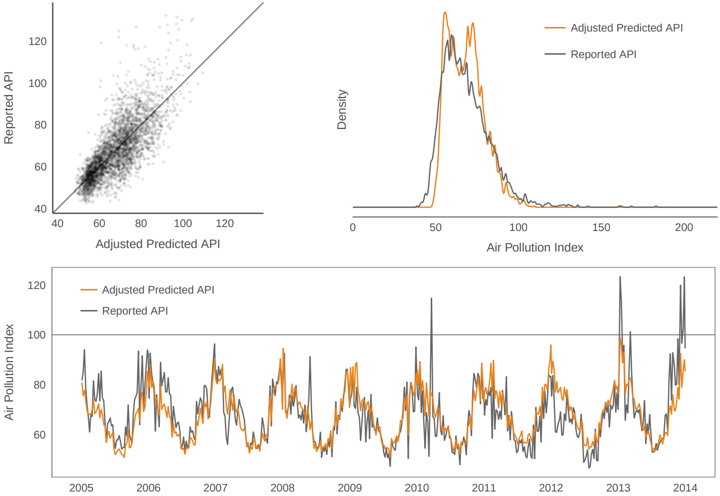
Table 2: Predictive Performance: Cross Validated Weekly R^2

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) We use 5-fold cross validation on training data to obtain predicted and true value pairs. (ii) We include only half of all the stations.

Data: Performance

Figure 4: Comparing Predicted and Reported Air Pollution Index in China



Empirical Strategy: Event Study

$$Y_{iwy} = \alpha_i + \beta_{wy} + \sum_{k \in [-10,4] \setminus \{-8,-1\}} \tau_k 1\{K_{iwy} = k\} + \epsilon_{iwy} \qquad (1)$$

- Each i indicates one monitoring station;
- Each t indicates one week;
- K_{iwy} is the year relative to treatment;
- $Y_{i,t}$ is average weekly air pollution levels;
- $\epsilon_{i,t}$ is clustered at the city level.

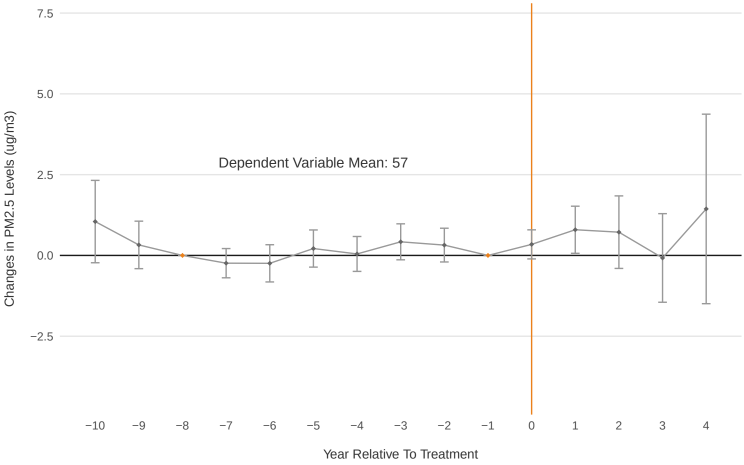
Empirical Strategy: Structural Break

$$Y_{iwy} = \alpha_{iw} + \beta_y + \tau_j 1\{K_{iwy} \geq j\} + \epsilon_{iwy} \tag{2}$$

- 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.

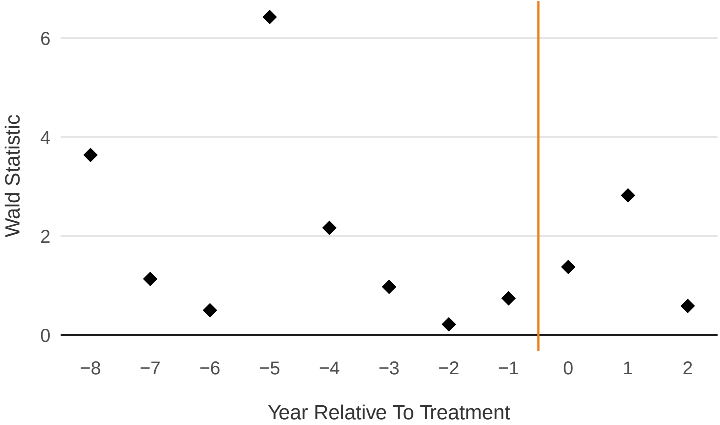
Treatment Effects are Tightly Bounded Around Zero

Figure 5: Event Study Estimates: $PM_{2.5}$ Levels



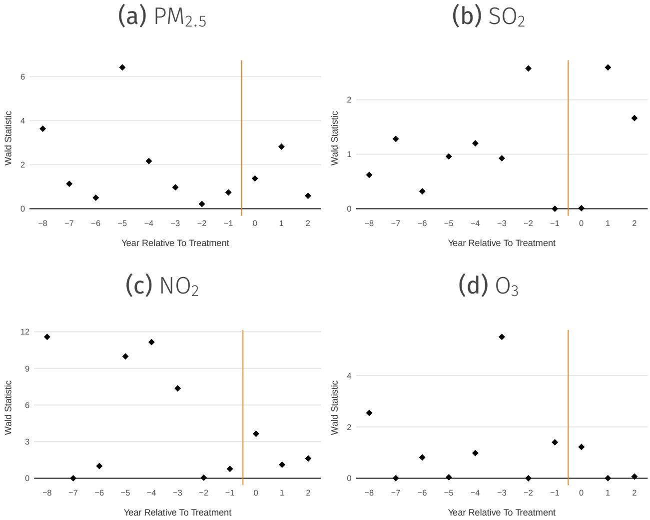
Treatment Has No Effects on Air Quality

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



Treatment Has No Effects on Air Quality

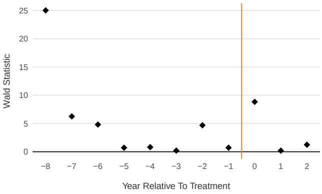
Figure 7: Structural Break Estimates: Machine Learning Predictions



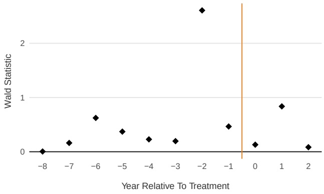
Treatment Has No Effects on Air Quality

Figure 9: Structural Break Estimates: Satellite Observations

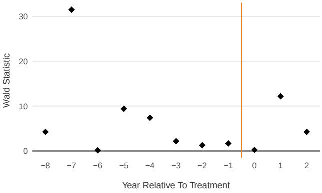
(a) AOD (MERRA2)



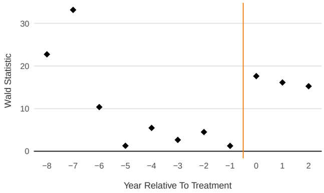
(b) AOD (Terra)



(c) O_3 (Aura)



(d) SO_2 (Aura)



A Conceptual Model to Reconcile Results

Local regulators are evaluated with either emissions (which may be mis-reported)

$$\underbrace{b(e+l)}_{\text{benefit of reported emission reduction}} - \underbrace{c(e)}_{\text{cost of effort}} - \underbrace{p(l)}_{\text{punishment for being caught}} \tag{3}$$

or ambient concentrations

$$\underbrace{q(e,\epsilon)}_{\text{(air) quality depends on effort but is uncertain}} - \underbrace{c(e)}_{\text{cost of effort}} \tag{4}$$

The relative effectiveness of the two regulations depend on the extent of information asymmetry $p(\cdot)$ and the uncertainty in ambient concentrations ϵ, conditional on emissions.