Do Data Matter for Air Quality?
New Evidence from Machine Learning Predictions

Yue 'Luna' Huang[1], Minghao Qiu[2]
[1]University of California, Berkeley, [2]Massachusetts Institute of Technology

Research Question

- Will improving data transparency and quality effectively improve air quality?

Motivation

- Does transparency improve environmental governance in developing countries?—Active research agenda in public health and education; largely unexplored in the environment literature.
- Recent surge in investments in monitoring equipment in China that amount to approximately 0.95 billion USD in just 2015 (Clean Air Act incurred approximately 65 billion USD in 30 years).
- Expectation for monitoring to work in China: (i) lack of any $PM_{2.5}$ information before 2012; (ii) intensive inter-jurisdiction competition for political promotion.

Policy

- Treatment: disclosure of Fine Particulate Matter ($PM_{2.5}$) data by city officials. Cities either (semi-) voluntarily started publicizing data or started reporting when the central government policy went into effect.
- Real-time hourly data were automatically reported (to data center websites), leaving little room for manipulations.

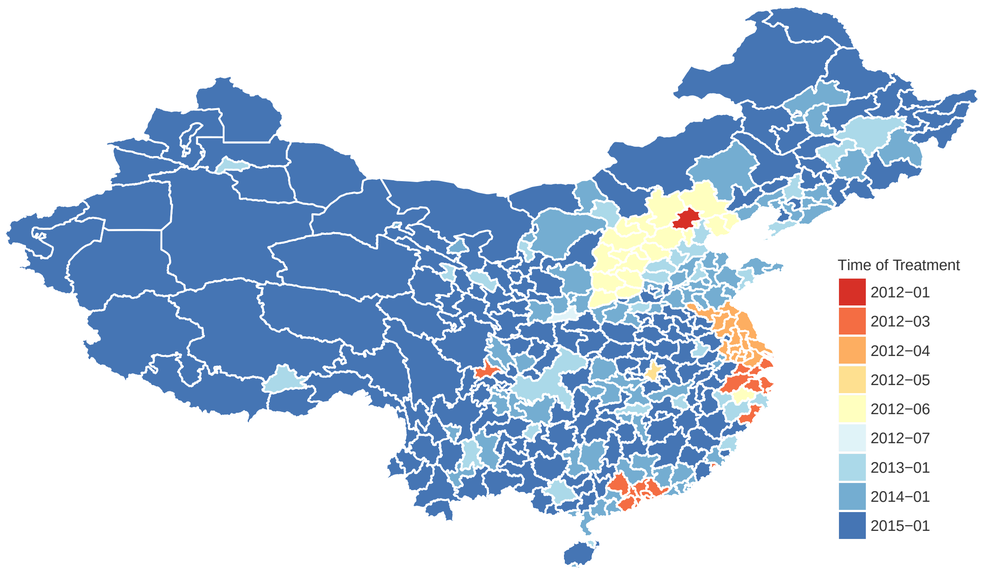

Figure: Time of Treatment: Dates when Cities Start Reporting $PM_{2.5}$ Values

Identification

- Main estimation equation:

$$Y_{i,t} = \alpha_i + \beta_t + \sum_{k=0}^{4} \tau_k \mathbf{1}\{K_{i,t} = k\} + \epsilon_{i,t} \quad (1)$$

- Identification assumption: Absent treatment, the treatment and control cities have parallel trends.

Contribution

- Challenge: Data did not exist before monitoring stations were built—pre-treatment data are unavailable.
- Solution: Recent development in machine learning models, combined with satellite images collected by NASA (surface reflectance at multiple bandwidths), allows us to reconstruct historical air pollution datasets.
- Compared to directly using satellite observations: Addressing the issue of non-random missing values; obtaining ground-level concentrations that can be more directly linked to health consequences.

Data

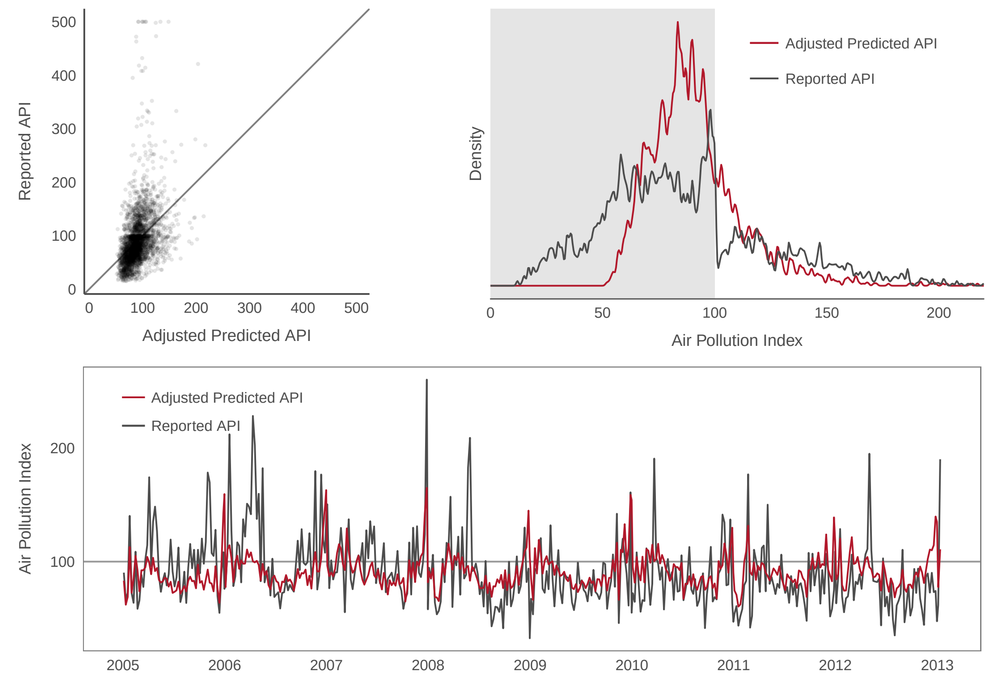

Figure: Comparing Predicted and Reported Air Pollution Index in Beijing

Table: 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

- 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.

Results

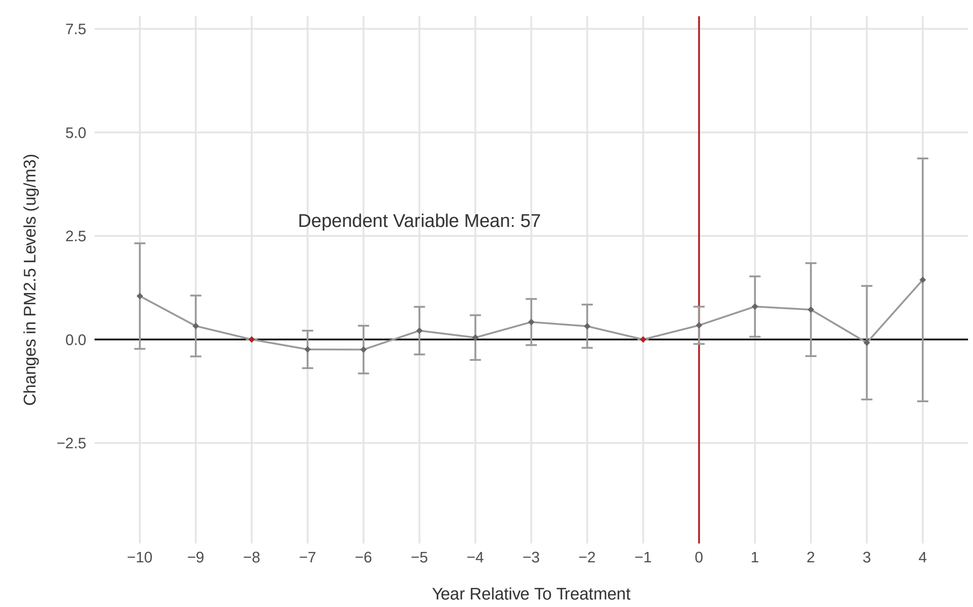

Figure: The Impacts of $PM_{2.5}$ Monitoring on Air Pollution are Small and Statistically Insignificant