Courtesy: Weiwei Lu.

Effortless Perfection

Do Chinese Local Governments Manipulate Air Pollution Data?

Dalia Ghanem Junjie Zhang

UC Davis UC San Diego

March 30, 2015

Motivation

- impact of China's economic growth on air quality
- central government's efforts to curb air pollution
  - air quality standards by China's central government since 1982
  - daily air pollution index (API) and publicizing whether a day is a blue-sky (API$\leq$ 100) since 2001
  - performance evaluations of local government officials based on meeting target number of blue-sky days
- incentive for manipulation

central government

data$\uparrow$ $\downarrow$evaluate

local governments

Motivation

Citizens are also watching...

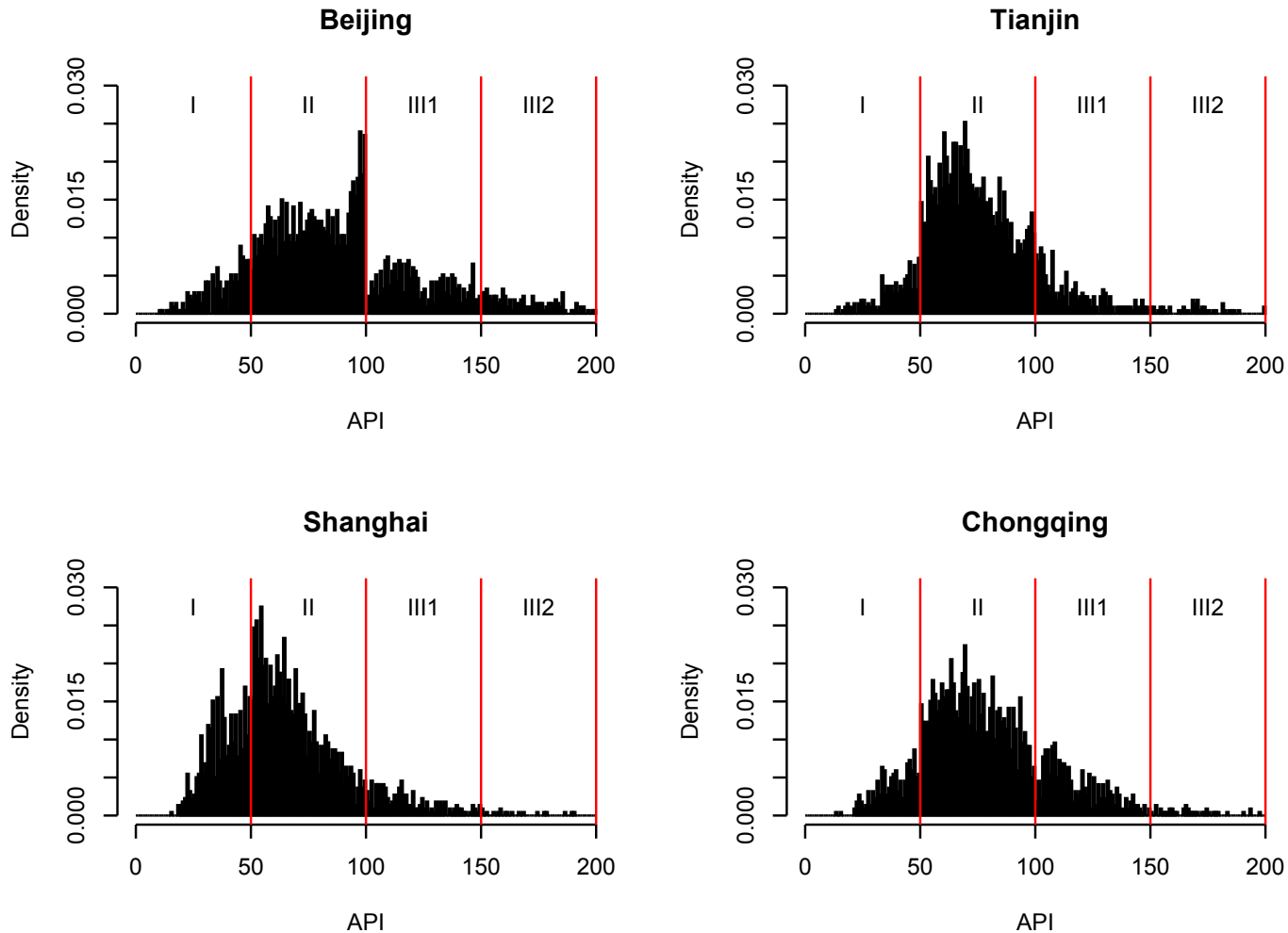
Weiwei Lu's Pictures of Blue-sky Days in Beijing

...threatening social stability.

Motivation

"Suspicious" Behavior around the cut-off for Blue-Sky Days

Figure: Histogram of Air Pollution Index (API)



This Paper

Empirical Questions tackled using *nonparametric* methods

(1) Do cities manipulate their air pollution data around the cutoff?

(2) What are the patterns of manipulation conditional on weather variables?

This Paper

Empirical Questions tackled using *nonparametric* methods

(1) Do cities manipulate their air pollution data around the cutoff?

(2) What are the patterns of manipulation conditional on weather variables?

Preview of Results

(1) *Effortless Perfection*
sharp discontinuities around the cut-off consistent with manipulation

(2) *Patterns of Manipulation*
high visibility
low wind speed

Disclaimer
We will use manipulation short for 'evidence consistent with manipulation'.

This Paper

Beyond Air Pollution in China

public policy/political economy

- manipulation of self-reported data that are used for performance evaluation

econometrics

- new identification strategy of manipulation
- measurement error resulting from such manipulation
- measurement error's relationship with exogenous variables (weather)

Roadmap

Data

Empirical Question 1: Effortless Perfection
- Identification Strategy
- Results
- Caveats

Empirical Question 2: Patterns of Manipulation
- Identification Strategy
- Results
- Caveats

Conclusion and Directions for Future Research

Data

Air quality

113 cities, daily, 2001-2010

API (public), criteria pollutant concentrations (confidential)

Weather

every three hours, 2001-2010

visibility, temperature, atmospheric pressure, precipitation, wind speed

Economy

yearly, 2001-2010

economic growth, industrial structure, emissions, etc.

Data: Air Pollution Index

Construction of Daily API

1. average pollutant concentrations for SO2, NO2, and PM10
2. piece-wise linear transformation for each pollutant p
   Individual pollutant index I, Pollutant concentration C, concentration breakpoints C_u and C_l, and index classes I_u and I_l

$$API_p = \frac{I_u - I_l}{C_u - C_l}(C - C_l) + C_l$$

3. Air pollution Index

$$\text{API} = \max\{API_{\text{SO}_2}, API_{\text{NO}_2}, API_{\text{PM}_{10}}\}$$

Data: Air Pollution Index

Air Pollution Index and Blue-Sky Day

API	Grade	Air Quality Conditions	Health Concerns
0-50	I	Excellent (Blue Sky)	Good
51-100	II	Good (Blue Sky)	Moderate
101-200	III	Lightly polluted	Unhealthy for sensitive groups
201-300	IV	Moderately polluted	Unhealthy
>300	V	Heavily polluted	Hazardous

City's annual environmental assessment

20 points out of 100

20/20 if the share of blue sky days $\geq 85\%$
0/20 if the share $\leq 30\%$

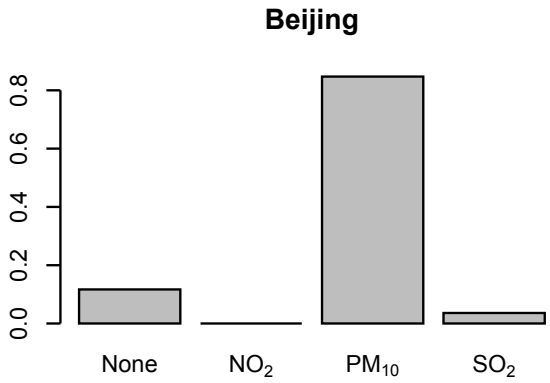
Data: Summary Statistics

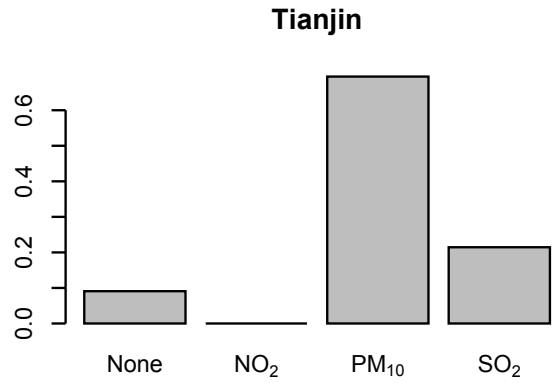
Table: Summary statistics of the variables used in the analysis

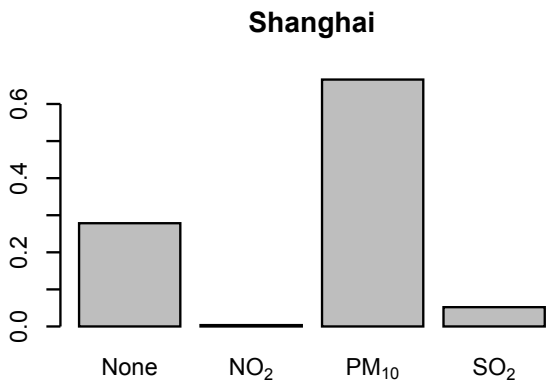
	Mean	Std	Min	Max
API (max)	76.315	37.680	0	625
Blue-Sky Days	0.846			
Primary Pollutant				
SO_2	0.092			
NO_2	0.002			
PM_{10}	0.737			
Pollution Concentration (mg/m^3)				
SO_2	0.054	0.055	0.001	2.147
NO_2	0.036	0.020	0.001	0.353
PM_{10}	0.100	0.066	0.001	2.721
API (pollutant level)				
SO_2	43.262	27.544	0	409
NO_2	22.850	14.116	0	226
PM_{10}	74.731	38.277	0	501

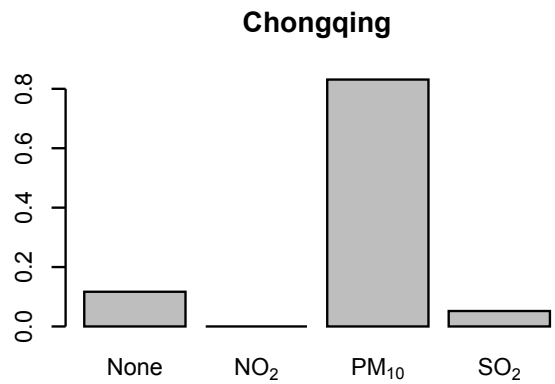
Data

PM_{10} Responsible for 74% of Pollution Days









Back to our Roadmap

Data

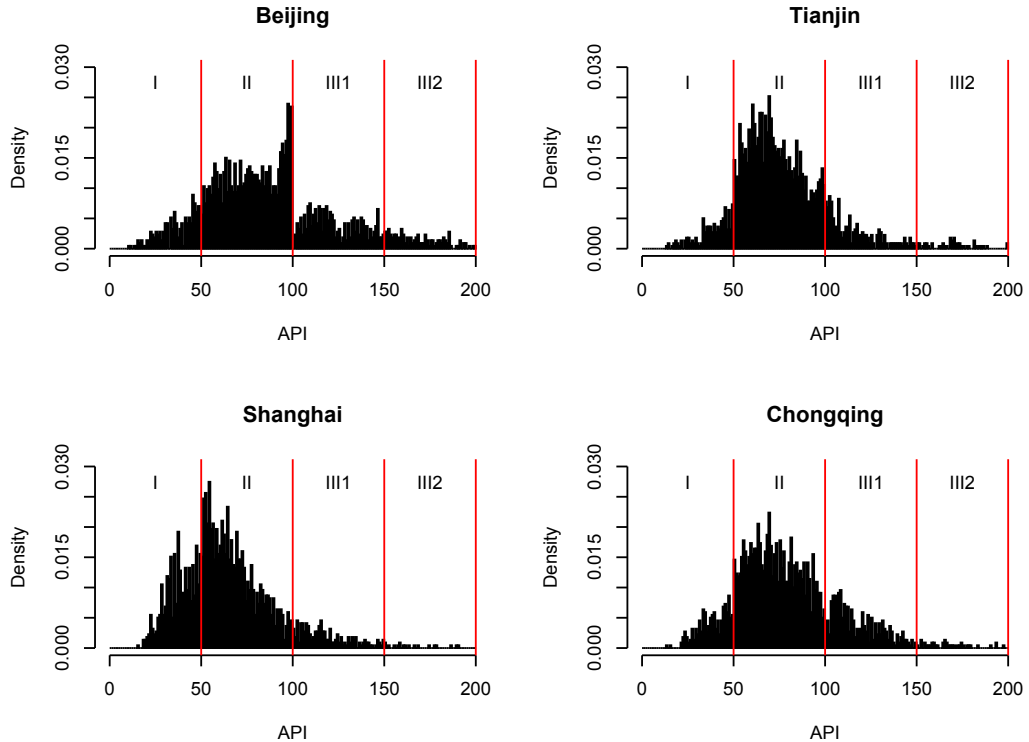
Empirical Question 1: Effortless Perfection
- Identification Strategy
- Results
- Caveats

Empirical Question 2: Patterns of Manipulation
- Identification Strategy
- Results
- Caveats

Conclusion and Directions for Future Research

Empirical Question 1: Effortless Perfection - Intuition

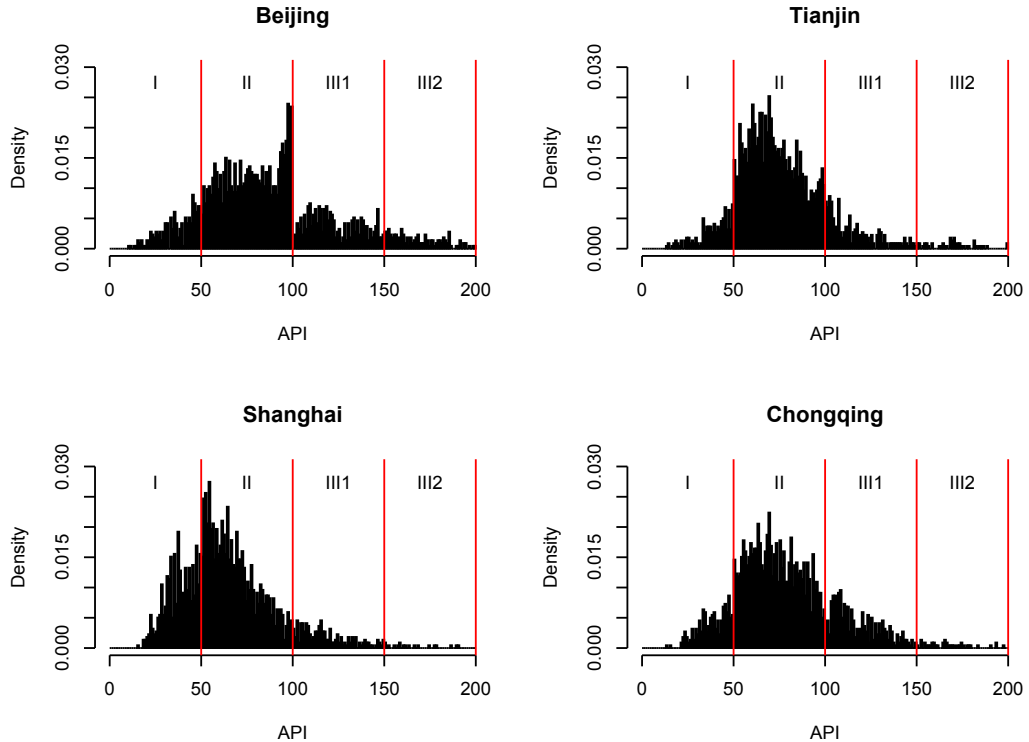
Figure: Histogram of Air Pollution Index: Effortless Perfection



Discontinuity at 100 (cutoff for blue-sky day)

Empirical Question 1: Effortless Perfection - Intuition

Figure: Histogram of Air Pollution Index: Effortless Perfection



Discontinuity at 100 (cutoff for blue-sky day)

***Assumption*: no discontinuity absent manipulation!**

Empirical Question 1: Effortless Perfection - Identification Strategy

Test: McCrary (2008)

$$\theta = \ln \lim_{p \downarrow c} f(p) - \ln \lim_{p \uparrow c} f(p)$$

where p pollutant concentration, assumed to be continuous, c cutoff, and $f()$ pdf

$$H_0 \quad : \quad \theta \geq 0$$

Empirical Question 1: Effortless Perfection - Identification Strategy

Test: McCrary (2008)

$$\theta = \ln \lim_{p \downarrow c} f(p) - \ln \lim_{p \uparrow c} f(p)$$

where p pollutant concentration, assumed to be continuous, c cutoff, and $f()$ pdf

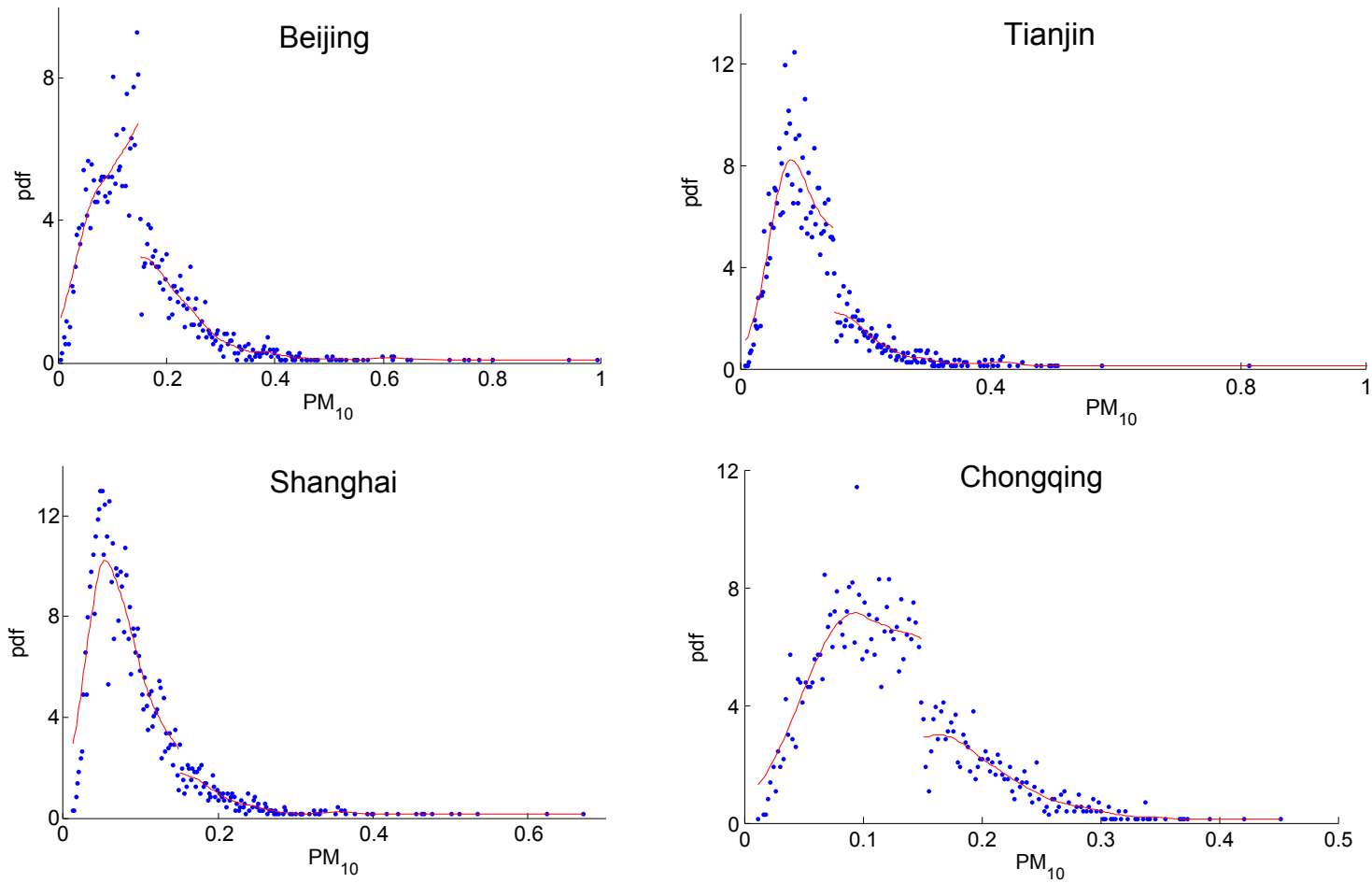
$$H_0 \quad : \quad \theta \geq 0$$

Chen et al (2012): API
discontinuous density due to the piece-wise linear transformation
McCrary (2008) continuity assumption violated

This Paper: Pollutant Concentrations (PM_{10}, SO_2, NO_2)
McCrary (2008) continuity assumption fulfilled

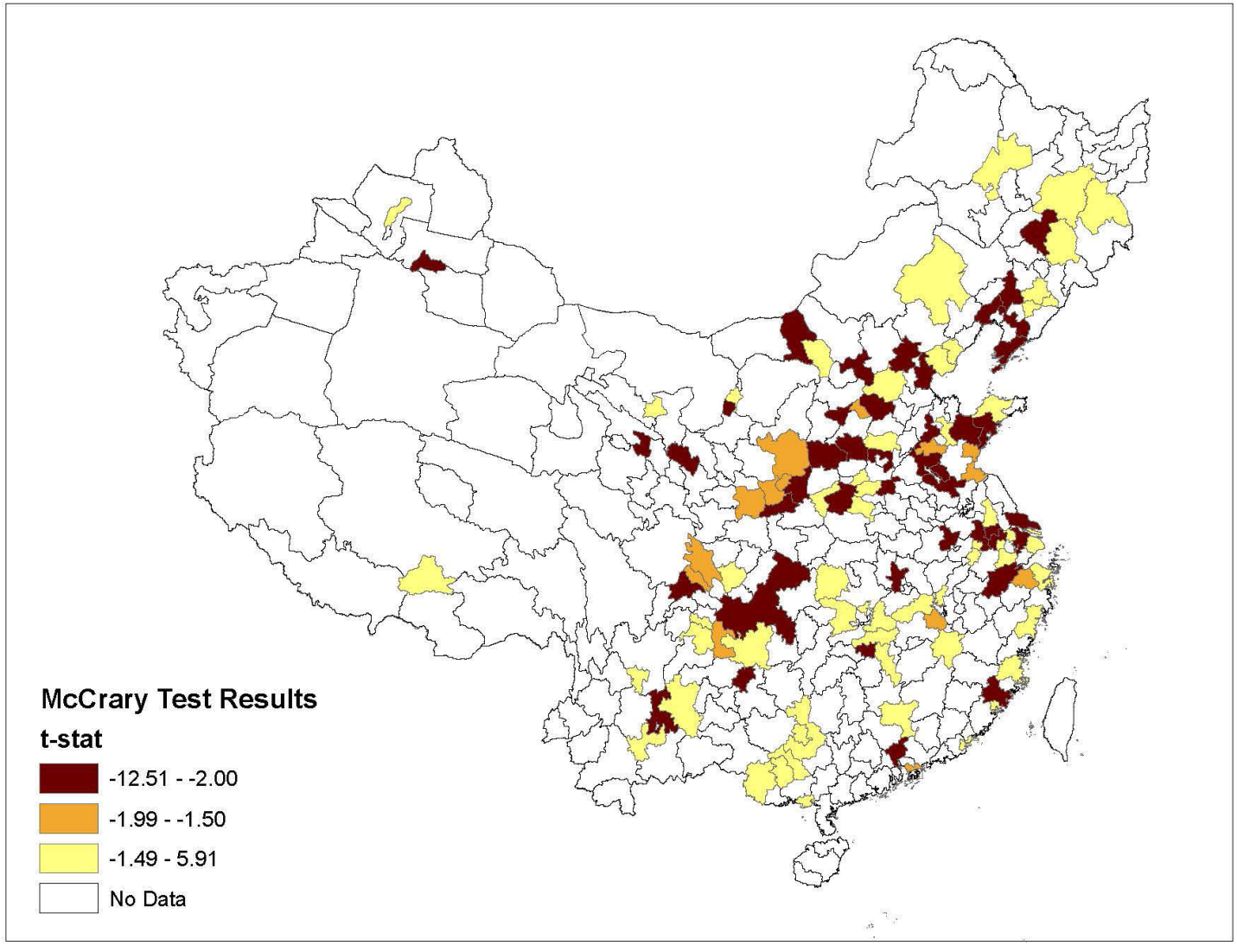
Empirical Question 1: Effortless Perfection - Results

Figure: McCrary (2008) Test Illustrated



Empirical Question 1: Effortless Perfection - Results

Figure: McCrary Test Results for PM_{10} Concentrations



Empirical Question 1: Effortless Perfection - Results

Summary

PM_{10}: evidence of manipulation for about 46% of cities

SO_2 : evidence of manipulation for about 17% of cities

NO_2 : evidence of manipulation for about 4% of cities

Empirical Question 1: Effortless Perfection - Results

Robustness Checks

different bandwidths

"artificial" cut-offs

comparison with results using API at the pollutant level

Empirical Question 1: Effortless Perfection - Caveats

detection of manipulation leading to discontinuities only

- other types of manipulation less likely

bunching resulting from command-and-control policies?

- pollutant concentrations not perfectly controllable
- does not explain the sharp discontinuities we find

Back to Roadmap

Data

Empirical Question 1: Effortless Perfection

- Identification Strategy
- Results
- Caveats

Empirical Question 2: Patterns of Manipulation

- Identification Strategy
- Results
- Caveats

Conclusion and Directions for Future Research

Empirical Question 2: Patterns of Manipulation - Intuition

Gist: find two cities that should be the same in every way except for manipulation behavior! (*matching*)

Empirical Question 2: Patterns of Manipulation - Intuition

Gist: find two cities that should be the same in every way except for manipulation behavior! (*matching*)

Ideal Experiment: Twin Cities (A and B)
API distribution for A and B should be the same

$$\text{API}_A \overset{d}{=} \text{API}_B$$

$\overset{d}{=}$ denotes identity of distribution.

Empirical Question 2: Patterns of Manipulation - Intuition

Construct "Twin" Cities

API distribution for A and B conditional on some variables being the same

$$\text{API}_A | W_A = w \quad \stackrel{d}{=} \quad \text{API}_B | W_B = w$$

Empirical Question 2: Patterns of Manipulation - Intuition

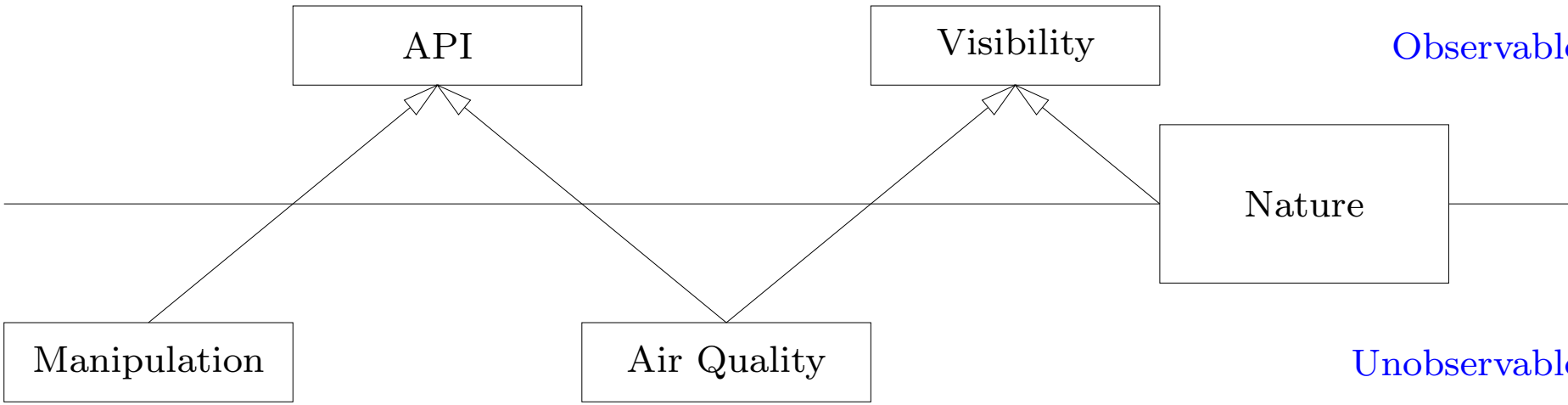
Construct "Twin" Cities

API distribution for A and B conditional on some variables being the same

$$\text{API}_A | W_A = w \quad \stackrel{d}{=} \quad \text{API}_B | W_B = w$$

Figure: API: Empirical CDFs of City A and B when $W_A = W_B = w$

Empirical Question 2: Patterns of Manipulation - Intuition



Controlling for visibility and other weather variables ensures we are using comparable days in terms of air quality

$$\Downarrow$$

In the absence of manipulation, API distributions for "twin" cities should be the same.

Empirical Question 2: Patterns of Manipulation - Identification Strategy

$$\text{API}_{tji} = \xi(W_{tji}, \mathcal{A}_j, \mathcal{U}_{tji})$$

t, j, i	time, city pair, city
API_{tji}	air pollution index
W_{tji}	control variables such as visibility and weather conditions
$\mathcal{A}_j$	unobservable city-pair heterogeneity
$\mathcal{U}_{tji}$	idiosyncratic shocks
$\xi()$	unknown function

Testing manipulation

$$\begin{aligned} \mathbf{H}_0 : \mathcal{U}_{tj1} | W_{tj1}, \mathcal{A}_j &\overset{d}{=} \mathcal{U}_{tj2} | W_{tj2}, \mathcal{A}_j \\ \Rightarrow F(\text{API}_{tj1} | W_{tj1} = w) &= F(\text{API}_{\tau j2} | W_{\tau j2} = w) \end{aligned}$$

Empirical Question 2: Patterns of Manipulation - Identification Strategy

Why Not a Linear Model?

No Manipulation	$\text{API}_{tji} = W'_{tji}\beta + \mathcal{A}_j + \mathcal{U}_{tji}$
Under Manipulation	$\text{API}_{tji} = W'_{tji}\beta + \mathcal{A}_j + \mathcal{C}_i + \mathcal{U}_{tji}$
Testing manipulation	H_0: $\mathcal{C}_1 = \mathcal{C}_2$

1. effects of W_{tji} and unobservables are separable
2. heterogeneity is just an intercept
3. manipulation is a fixed intercept, regardless of W_{tji} and *unobservable incentive* to manipulate is not affected by which primary pollutant, degree of visibility, etc.

Empirical Question 2: Patterns of Manipulation - Identification Strategy

Statistical Tests

Two-sample Kolmogorov-Smirnov (KS) test

$$KS_j(w) = \sup_{y \in \mathcal{Y}} \left| \frac{\sum_{t=1}^{T} 1\{API_{tj2} \leq y\} 1\{W_{tj1} = w\}}{\sum_{t=1}^{T} 1\{W_{tj1} = w\}} - \frac{\sum_{t} 1\{API_{tj2} \leq y\} 1\{W_{tj2} = w\}}{\sum_{t=1}^{T} 1\{W_{tj2} = w\}} \right|$$

t-test

$$t_j(w) = \frac{\sum_{t} API_{tj2} 1\{W_{tj1} = w\}}{\sum_{t} 1\{W_{tj1} = w\}} - \frac{\sum_{t} API_{tj2} 1\{W_{tj2} = w\}}{\sum_{t} 1\{W_{tj2} = w\}}$$

Correction for multiple testing using the step-up procedure (Romano and Shaikh, 2006)

Empirical Question 2: Patterns of Manipulation - Identification Strategy

Implementation

Step 1: form city pairs

same province: administrative characteristics

nearest neighbors: geographical characteristics

Step 2: discretize conditional variables

same level of visibility

same weather conditions

▸ Discretization

Step 3: within-pair comparisons by KS test and t-test for PM_{10} days

Empirical Question 2: Patterns of Manipulation - Results

Step 1 yields following 13 city pairs

Table: City Pairs

City 1	City 2	Region	Coastal
Kaifeng	Zhengzhou	Central	0
Zhuzhou	Xiangtan	Central	0
Yinchuan	Shizuishan	Northwest	0
Xian	Xianyang	Northwest	0
Huehaote	Baotou	North	0
Jilin	Changchun	Northeast	0
Shenyang	Fushun	Northeast	0
Wuhu	Maanshan	East	0
Zhenjiang	Yangzhou	East	0
Changzhou	Wuxi	East	0
Jinan	Taian	East	0
Quanzhou	Xiamen	East	1
Hangzhou	Shaoxing	East	1

For each pair, we apply Steps 2-3.

▸ Example

Empirical Question 2: Patterns of Manipulation - Results

Table: Summary and Comparison with McCrary Test Results

City-pair	Matching Rejections	McCrary (2008) City 1	McCrary (2008) City 2
Wuhu-Maanshan	NO	NO	NO
Zhenjiang-Yangzhou	NO	Borderline	NO
Changzhou-Wuxi	NO	YES	NO
Jilin-Changchun	NO	NO	YES
Kaifeng-Zhengzhou	YES	YES	NO
Zhuzhou-Xiangtan	YES	NO	YES
Quanzhou-Xiamen	YES	YES	NO
Hangzhou-Shaoxing	YES	YES	NO
Shenyang-Fushun	YES	YES	NO
Yinchuan-Shizuishan	YES	NO	YES
Xian-Xianyang	YES	YES	Borderline
Jinan-Taian	YES	YES	NO
Huhehaote-Baotou	YES	NO	YES

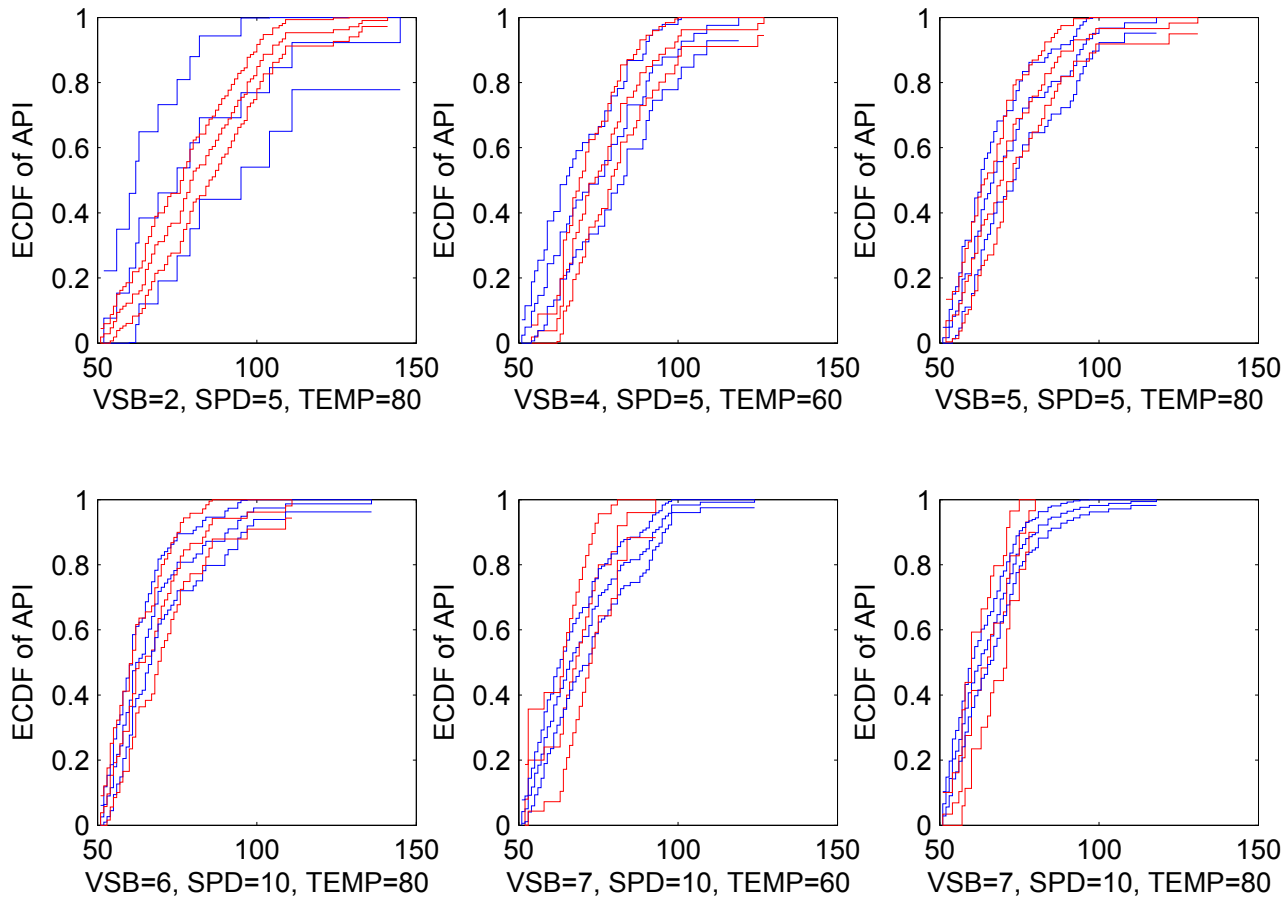
Rejections more likely with...

- high visibility (*citizens are watching*)
- low wind speed (*pollution not gone with the wind*)

Note: KS test is somewhat conservative due to dependence in API data!

Empirical Question 2: Patterns of Manipulation - Results

Figure: Conditional Empirical CDF Comparison (Wuhu-Maanshan)

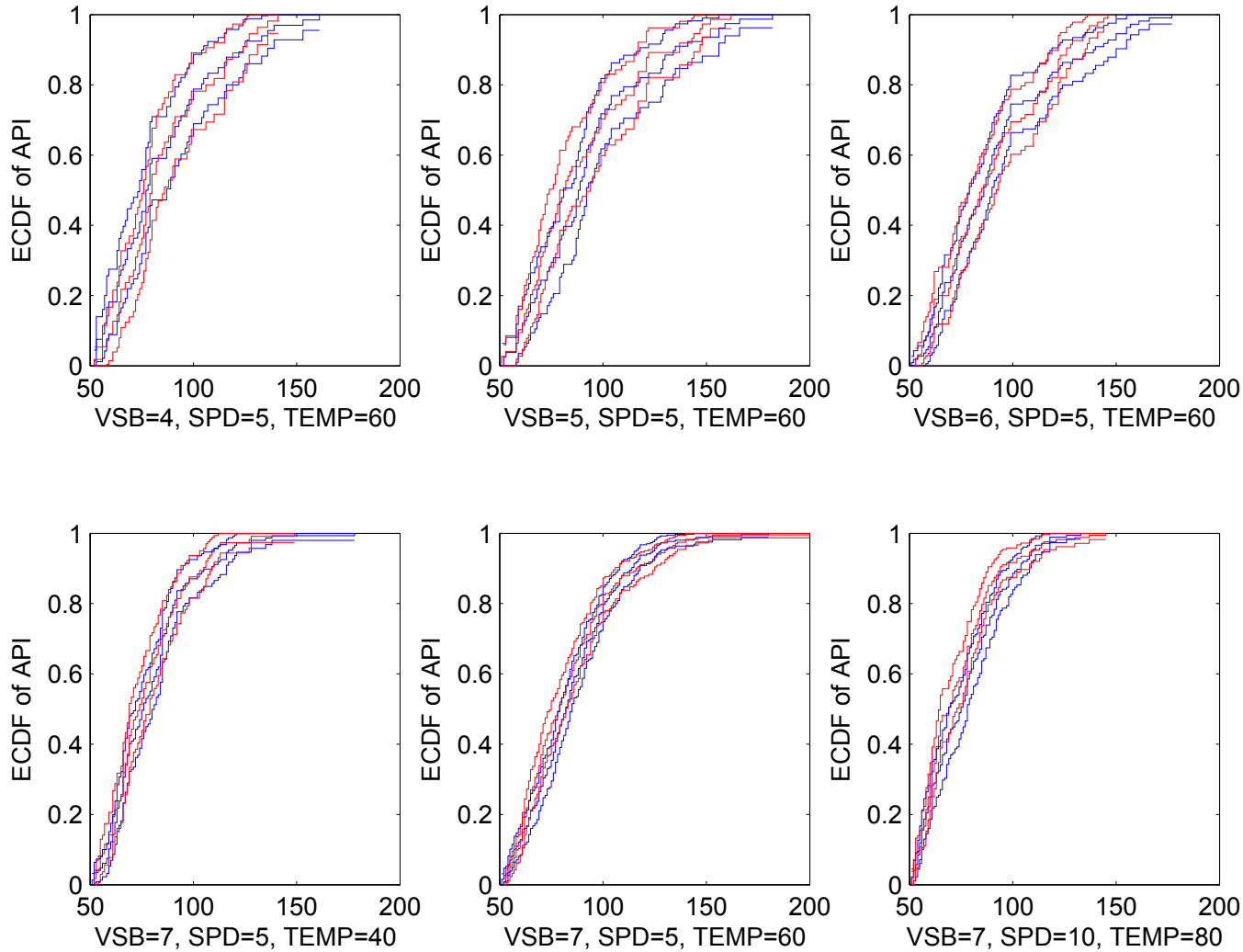


NO REJECTIONS!

McCrary Test: Neither exhibit manipulation.

Empirical Question 2: Patterns of Manipulation - Results

Figure: Conditional Empirical CDF Comparison (Changzhou-Wuxi)



NO REJECTIONS!

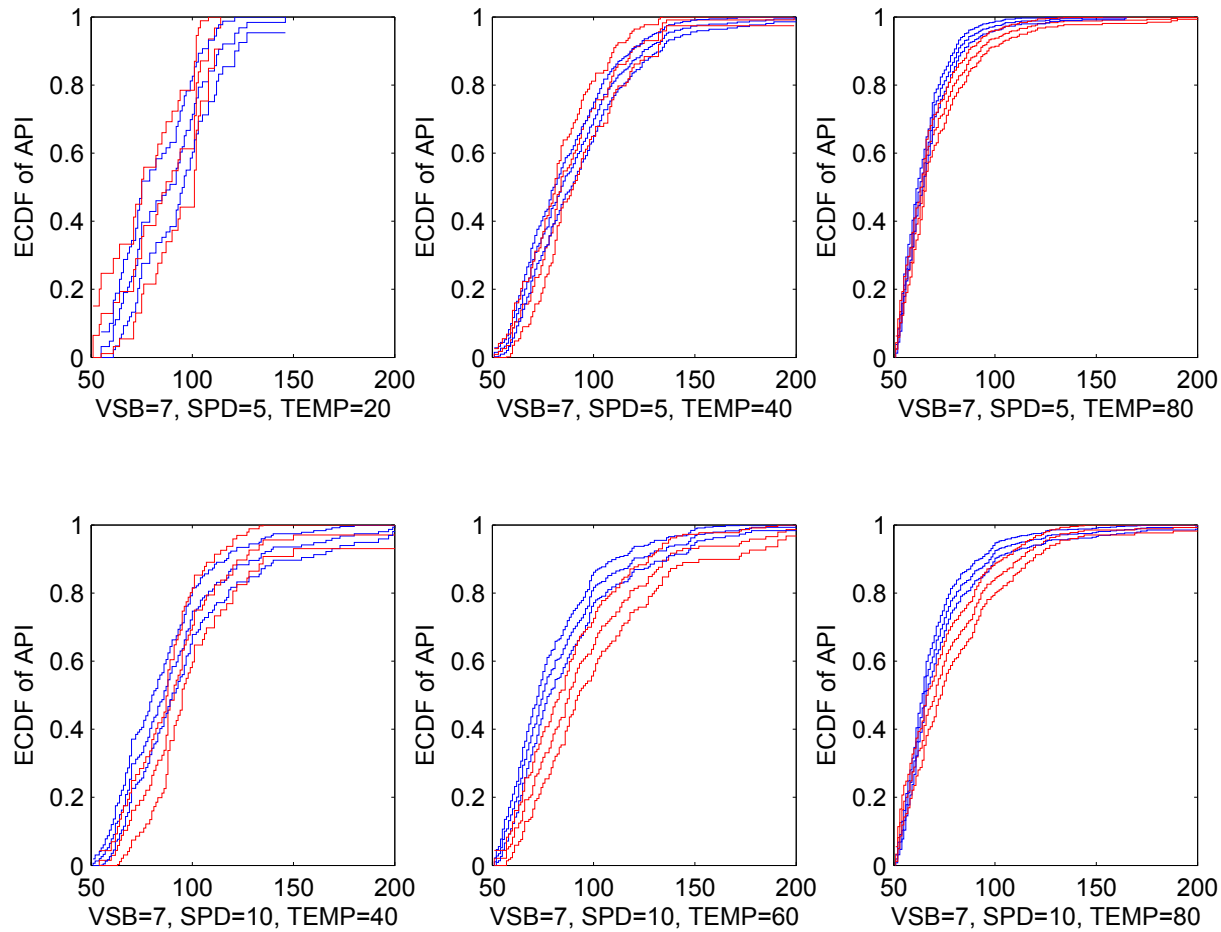
Empirical Question 2: Patterns of Manipulation - Results

Figure: Conditional Empirical CDF Comparison (Zhuzhou-Xiangtan)

McCrary evidence of manipulation: Xiangtan.

Empirical Question 2: Patterns of Manipulation - Results

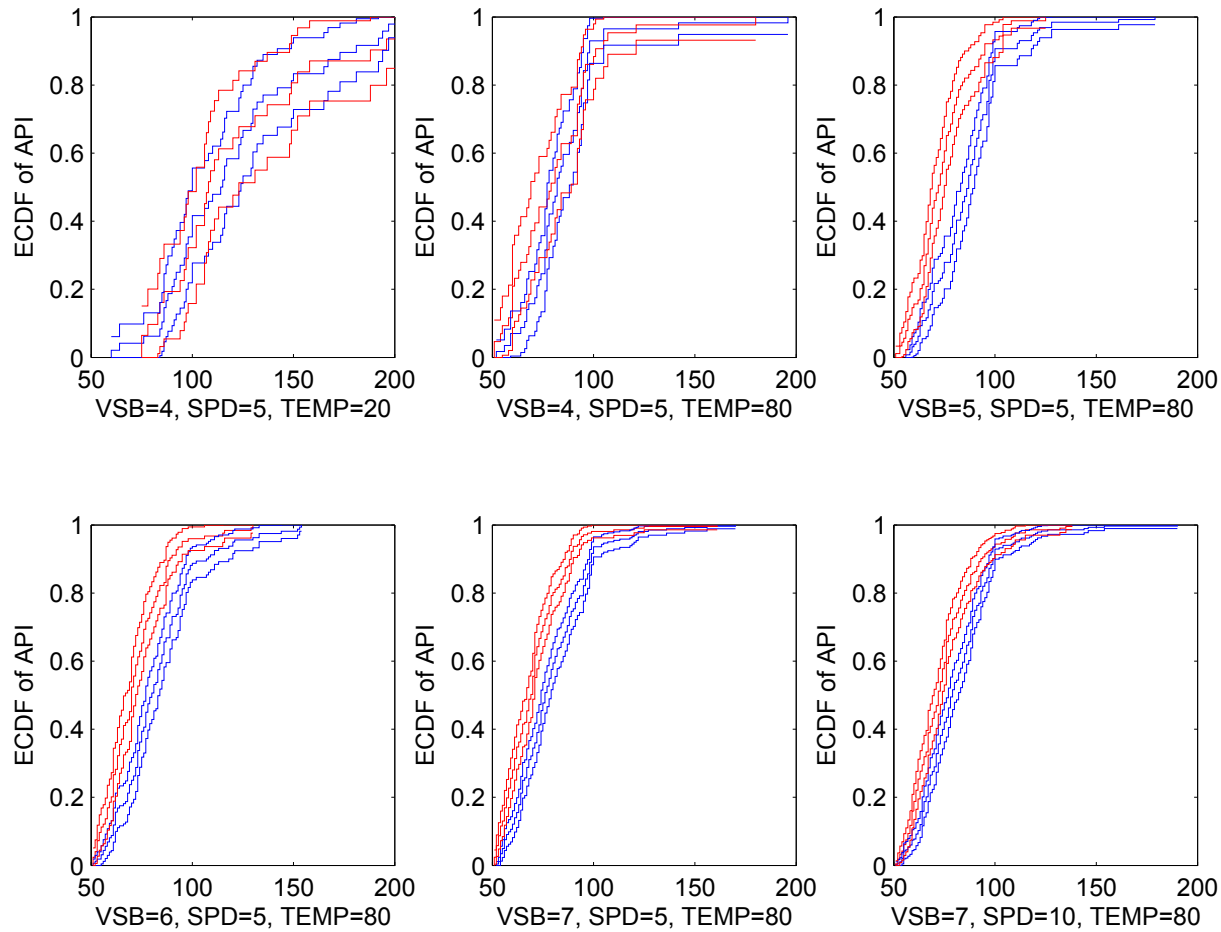
Figure: Conditional Empirical CDF Comparison (Yinchuan-Shizuishan)



McCrary evidence of manipulation: Shizuishan

Empirical Question 2: Patterns of Manipulation - Results

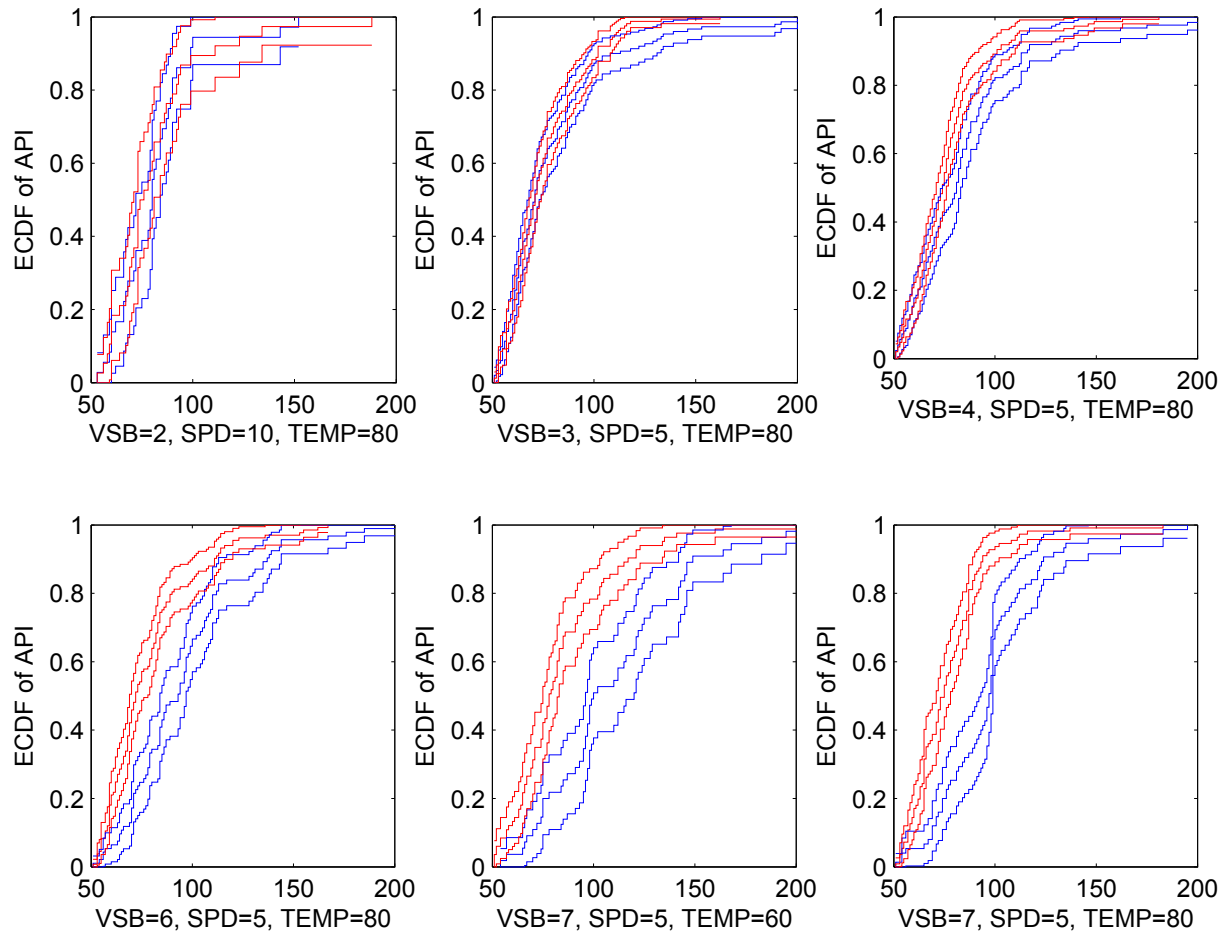
Figure: Conditional Empirical CDF Comparison (Shenyang-Fushun)



McCrary evidence of manipulation: Shenyang

Empirical Question 2: Patterns of Manipulation - Results

Figure: Conditional Empirical CDF Comparison (Kaifeng-Zhengzhou)



McCrary evidence of manipulation: Kaifeng

Empirical Question 2: Patterns of Manipulation - Results

Summary

intuitive finding: manipulation tends to occur with high visibility and low wind speed

matching approach and McCrary results lead to consistent results

Empirical Question 2: Patterns of Manipulation - Caveats

Weather variables have to be good controls for true air quality *at all their various levels and combinations*!
Recall $API_A|W_A = w \overset{d}{=} API_B|W_B = w$

- coastal pairs
  different air quality when wind speed is high, e.g. Quanzhou-Xiamen

rejections resulting from different command-and-control policies

- again, pollutant concentrations not perfectly controllable
- same province $\Rightarrow$ broad administrative characteristics
- air quality in cities in the same pair should be highly correlated given their proximity; hence policy responses should be similar

Conclusion and Directions for Future Research

Summary of Results

- evidence of manipulation of the primary pollutant concentration around the cutoff by many cities
- evidence of manipulation of the primary pollutant concentration around the cutoff by many cities
- manipulation is more likely on days with higher visibility and lower wind speed
- implications for measurement error and its dependence on weather variables (invalidated as instruments)

Directions for Future Research

- discontinuity test that are more robust to noise (hourly data)
- embassy data
- follow-up study with data after 2010 change in data reporting
- impact of measurement error in self-reported data on IV results

THANK YOU!

Discretization of Weather Variables

Visibility: rounded to the nearest integer

Precipitation: divided into three equi-space bins
$[0, 0.33]$, $(0.33, 0.67]$ and $(0.67, 1]$

Wind speed: divided into 5 bins, $[0, 5]$, $(5, 10]$, $(10, 15]$, $(15, 20]$, $(20, 25]$.

Temperature: divided into 5 bins ≤ 20, $(20, 40]$, $(40, 60]$, $(60, 80]$, and $(80, 100]$.

Back

Empirical Question 2: Patterns of Manipulation - Results

Example of City-Pair Results

Table: Yinchuan-Shizuishan

VSB	SPD	Temp	PCP	Subs 1	Subs 2	KS	Z	T	KS	Z	T
5	5	40	0.0	25	11	0.691	0.416	0.422	DNR	DNR	DNR
6	5	40	0	39	14	0.968	0.862	0.863	DNR	DNR	DNR
7	5	20	0	63	31	0.950	0.573	0.575	DNR	DNR	DNR
7	5	40	0	334	115	0.448	0.630	0.630	DNR	DNR	DNR
7	5	60	0	348	205	0.043	0.124	0.125	DNR	DNR	DNR
7	5	80	0	593	416	0.043	0.001	0.001	DNR	R	R
7	5	100	0	69	41	0.851	0.530	0.532	DNR	DNR	DNR
7	10	20	0	22	14	0.420	0.642	0.645	DNR	DNR	DNR
7	10	40	0	154	68	0.010	0.398	0.398	DNR	DNR	DNR
7	10	60	0	289	145	0.000	0.002	0.002	R	R	DNR
7	10	80	0	521	270	0.002	0.009	0.010	DNR	DNR	DNR
7	10	100	0	40	16	0.211	0.100	0.106	DNR	DNR	DNR
7	15	60	0	35	14	0.461	0.844	0.845	DNR	DNR	DNR

Back