

What's Missing in Environmental (Self-)Monitoring: Evidence from Strategic Shutdowns of Air Quality Monitors

September 2021

Yingfei Mu (JHU)

Edward Rubin (Oregon)

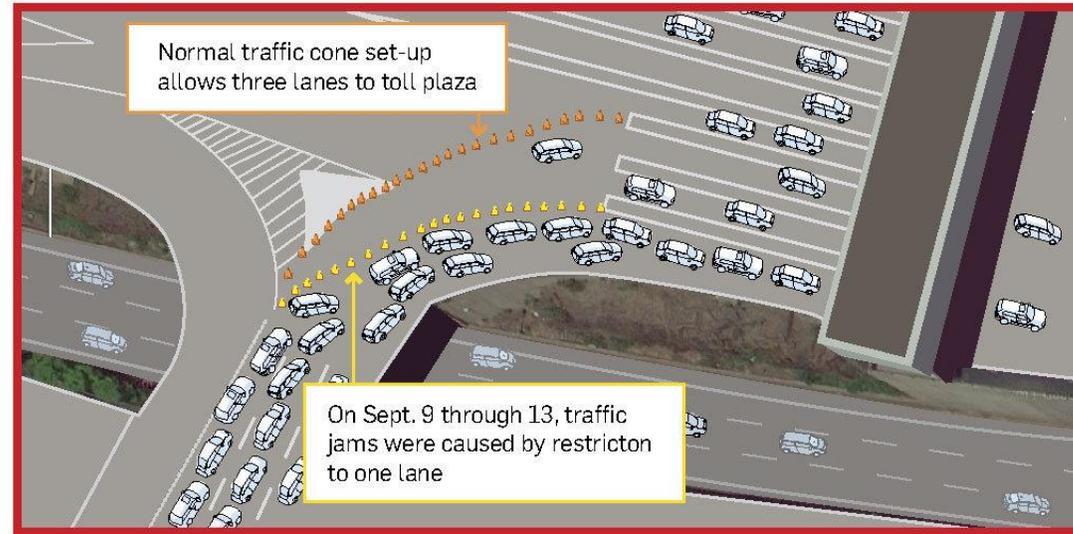
Eric Zou (Oregon)

Motivation

- Environmental regulations rely on the **regulated** to record compliance monitoring data
 - Cap-and-trade participants are charged with monitoring emissions
 - States operate pollution monitoring stations to show compliances to federal standards
 - Country self-monitor GHGs to demonstrate adherence to climate commitments
- Self-monitoring is a common practice when federal regulators face high monitoring requirements
 - Police officers are responsible for turning on/off body cameras
 - Doctors catalog what happens in the operating room
 - Tax liability assessment sometimes relies on self-reported income and expenses
- This paper
 - Studies U.S. Clean Air Act's outdoor air quality monitoring rule
 - Shows federal EPA's tolerance for gaps in monitoring data may have incentivized strategic timing of state gov's compliance monitoring

FORT LEE SEPTEMBER TRAFFIC SNARL

The change in access lanes to the George Washington Bridge toll plaza, caused traffic tie-ups on every street in Fort Lee and spreading south to Edgewater.



FRANK CECALA/THE STAR-LEDGER

Source: Vox



Source: The New York Times



Source: New Jersey Department of Environmental Protection; Google



UNITED STATES ENVIRONMENTAL PROTECTION AGENCY
REGION 2
290 BROADWAY
NEW YORK, NY 10007-1898

FEB 28 2014

Mr. Jeff Ruch
Executive Director
Public Employees for Environmental Responsibility
2000 P Street NW, Suite 240
Washington, DC 20036

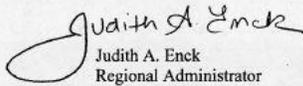
Dear Mr. Ruch:

This is in response to your letter dated January 31, 2014 to the U.S. Environmental Protection Agency's Office of the Inspector General requesting that the EPA investigate the operational status of air monitoring equipment during the lane closures of the George Washington Bridge from September 9, through September 13, 2013. Your letter was referred to EPA Region 2 for a response.

The EPA has conducted a review of the air quality monitoring information and data in the EPA's Air Quality System database for monitors located in the vicinity of the George Washington Bridge from September 7 to September 14, 2014. Based on this review, the EPA has concluded that the ambient air quality monitoring network equipment in question was operated by the New Jersey Department of Environmental Protection in accordance with the EPA's rules. Details of the EPA's review are enclosed. In addition, the measured air quality concentrations were in compliance with the EPA's National Ambient Air Quality Standards during this time period.

If you have any further questions, please contact me at 212-637-5000 or have your staff contact Richard Ruvo, Chief of our Air Programs Branch at 212-637-4014.

Sincerely,


Judith A. Enck
Regional Administrator

Enclosure

cc: Bill Wolff, PEER
Douglas Zmorzenski, OIG
Clay Brown, OIG

Internet Address (URL) • <http://www.epa.gov>
Recycled/Recyclable • Printed with Vegetable Oil Based Inks on Recycled Paper (Minimum 50% Postconsumer content)

Jersey City, NJ Site at 355 Newark Avenue is 10 miles from the Bridge

- PM_{2.5} sampler collecting on a daily schedule
 - No data collected from September 7, 2013 to September 13, 2013 due to reported equipment malfunction
- PM_{2.5} sampler (for quality assurance purposes) collecting on a one in six day schedule
 - Data collection
 - On September 7, 2013, 24 hour average was 7.5 ug/m³
 - On September 13, 2013, 24 hour average was 5.2 ug/m³
- PM_{2.5} sampler collecting continuously
 - Data collection
 - On September 7, 2013, 24 hour average was 5.9 ug/m³
 - On September 8, 2013, partial data was collected, average was 6.8 ug/m³
 - No data September 9, 2013 due to reported wireless router malfunction
 - No data September 10, 2013 due to reported wireless router malfunction
 - On September 11, 2013, partial data was collected, average was 26.3 ug/m³
 - On September 12, 2013, 24 hour average was 17.0 ug/m³
 - On September 13, 2013, 24 hour average was 2.3 ug/m³

Source: U.S. EPA

Why Do People Worry about This Incident?

- Reflects an underappreciated challenge for environmental self-monitoring
 - **Incentive:** states self-monitor air quality compliance, and suffers regulatory penalties when their own data suggest violation of EPA air quality standards (“NAAQS”)
 - **Discretion:** up to 25% missing data permissible per quarter
 - **Ability:** states’ weather department often run air quality forecasting
 - **No adequate detection mechanism:** regulator ignores missing data when assessing compliance

Why Do People Worry about This Incident?

- Reflects an underappreciated challenge for environmental self-monitoring
 - **Incentive:** states self-monitor air quality compliance, and suffers regulatory penalties when their own data suggest violation of EPA air quality standards (“NAAQS”)
 - **Discretion:** up to 25% missing data permissible per quarter
 - **Ability:** states’ weather department often run air quality forecasting
 - **No adequate detection mechanism:** regulator ignores missing data when assessing compliance
- Do local governments skip monitoring in expectation of a looming air quality deterioration?

This Paper

- Goal: Provide a framework to detect strategic shutdowns of pollution monitors
- Idea: Look for abnormal missing patterns around **pollution alerts**
 - E.g. “High Pollution Advisory”, AZ
 - Alerts are based on state gov’s own pollution forecasting (**expectation**)
 - Test if monitors’ sampling rates fall around alert days
- Large-scale inference: test JCF monitor first, then apply the method to over 1,300 monitors in counties with similar alert programs
 - Address false discovery with multiple-testing tools
 - Come up with a list of **“interesting” monitors** that responded to alerts
- Policy: Discuss imputation methods that may deter strategic shutdowns

Primary Data Sources

1. EPA's ground monitoring data 2004-2015
 - Daily Summary File: pollution value for each monitor-day
 - Focus on “criteria” pollutant monitors (PM_{2.5}, PM₁₀, O₃, NO₂, SO₂, CO)

2. AirNow.gov compilation of air pollution alerts 2004-2015
 - Ex: “Spare the Air” (CA Bay Area), “High Pollution Advisory” (AZ)
 - 33,357 alerts issued by 342 jurisdictions (city, county, or metro areas)
 - Aggregate to county-day events

- Final study pool includes 1,359 monitors
 - These are continuous monitors scheduled to sample everyday
 - Span 167 counties that have pollution alert programs

Outline

- **Institution**
- **Main Results**
- **Discussion**
 - Mechanisms?
 - Economic importance?
 - Policy alternatives?

The National Ambient Air Quality Standards

- .. or NAAQS: safety standards for **outdoor air quality** established under the U.S. Clean Air Act
 - E.g., most recent standards for PM_{2.5}: 3-year avg mean ≤ 12 ug/m³; 3-year daily 98th percentile ≤ 35 ug/m³
 - Standards exist for “criteria pollutants”: Particulate matter (PM_{2.5}, PM₁₀, Pb) and trace gas (O₃, NO₂, SO₂, CO)
- EPA uses states’ submitted monitoring data to categorize jurisdictions (mostly counties) into three groups
 - “**Nonattainment**”: violating the standards
 - “**Attainment**”: adhering to the standards
 - “**Unclassifiable**”: not sufficient data; *de facto* “**attainment**” (more later)

The National Ambient Air Quality Standards

- “Nonattainment” areas face substantially elevated regulatory scrutiny
 - State needs to develop a State Implementation Plan (SIP) that details regulatory actions: adoption of expensive pollution abatement tech (“LAER”) and emission limits
- Large **fiscal burden** to the state and local governments in addition to direct compliance costs
 - Lost manufacturing sector productivity (Greenstone, List, Syverson, 2012), labor market transition costs (Walker, 2013), etc.

Rules for Incomplete Monitoring

- To demonstrate compliance with NAAQS, states' monitoring data must satisfy completeness goals
 - Varies across pollutants, but the typical requirement is for each monitor to take **at least 75% of required samples per quarter** of the year

Rules for Incomplete Monitoring

- What does the regulator do if states' data fall below the completeness requirement?
 - Calculate compliance statistics (annual mean, 98th percentile, etc.) using the incomplete data *anyway*
- If calculated statistic < regulatory threshold: county is “unclassifiable” (*de facto* “attainment”)
- If calculated statistics > regulatory threshold: assign the county with “nonattainment” status
 - EPA has authority to do this with *very* limited data: as few as 11 samples per quarter are sufficient to designate nonattainment
 - If fewer than 11 samples are available, can use alternative data such as “nearby concentrations”

Rules for Incomplete Monitoring

- Implication: the (75%) completeness goal *per se* is not subject to gaming
 - A **violating area** cannot bring itself out of nonattainment simply by reducing sampling rate below 75% (because EPA can use very few data points to determine nonattainment)
 - For a **non-violating area**, makes little difference if its sampling rate is above or below 75%
- Point of this paper: because the regulator uses the **incomplete data** directly to calculate compliance statistics, strategic response *can* arise when local monitoring agencies skip high-pollution days to water down the average (or whatever relevant statistics) of measured pollution

Outline

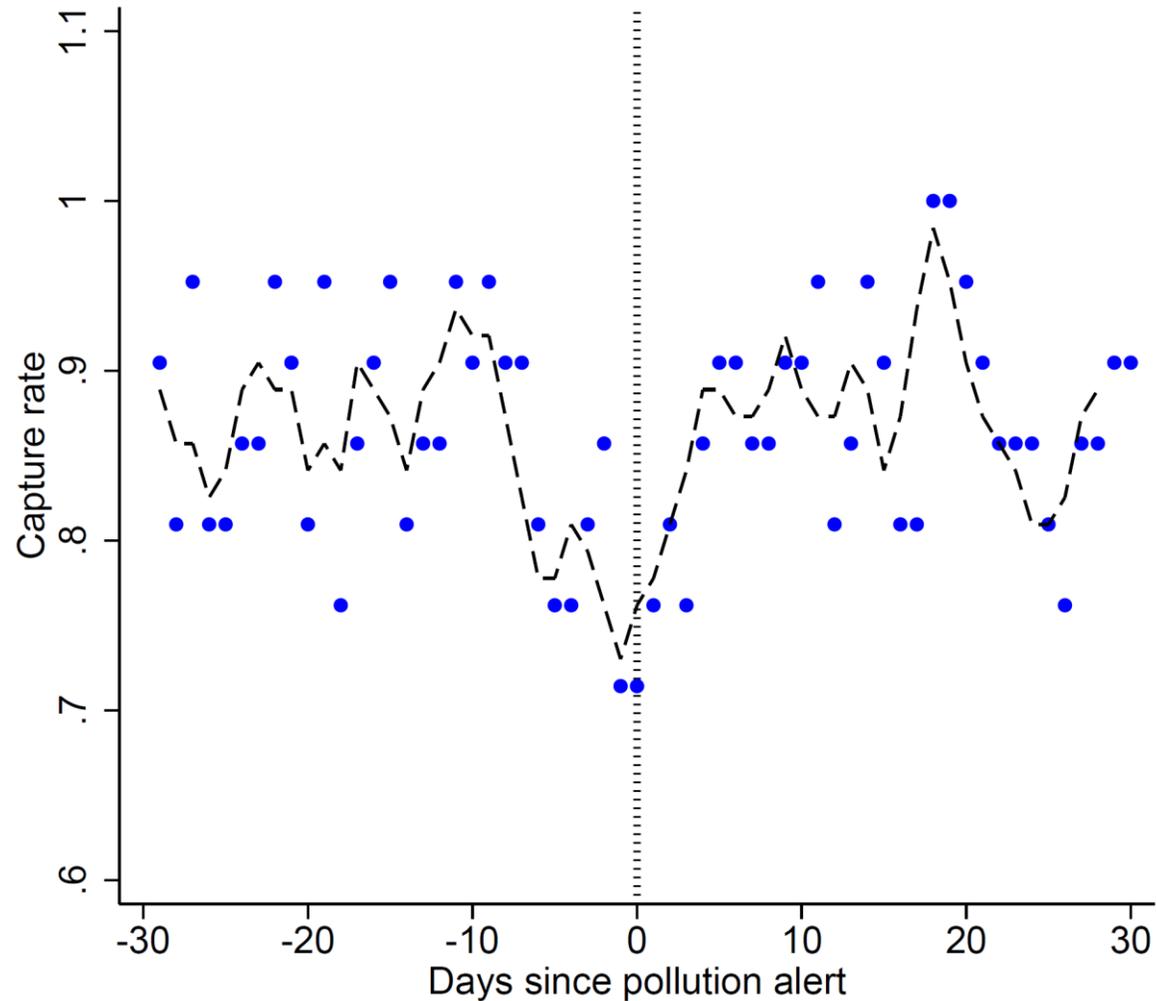
- Institution
- Main Results
- Discussion
 - Mechanisms?
 - Economic importance?
 - Policy alternatives?

Event Study (one monitor)

- Describe how we test for strategic shutdowns for a single monitor, using the Jersey City Firehouse (JCF) monitor as an example
- Event study estimation equation:
$$\text{CaptureRate}_t = 1 - 1(\text{missing PM2.5 data})_t = \sum_{\tau \in [-30, 30]} \beta_\tau \cdot 1(t = \tau) + \epsilon_t$$
 - Jersey City issued 21 alerts during our study period
 - We look at JCF monitor's capture rate in the 30 days before and 30 days after an alert, forming an event study dataset of $21 \cdot 61 = 1,281$ observations
- Coefficients of interest:
$$\hat{\beta}_\tau = \text{the capture rate } \tau\text{-day relative to the alert day}$$

Event study: Do monitors shutdown around pollution alerts?

Data capture rate of the JCF monitor around pollution alerts:



Notes: Estimation equation: $1 - 1(\text{missing PM2.5 data})_t = \sum_{\tau \in [-30, 30]} \beta_{\tau} \cdot 1(t = \tau) + \epsilon_t$.

Dashed line shows 3-day moving average. Total 21 alert events.

Inference (one monitor)

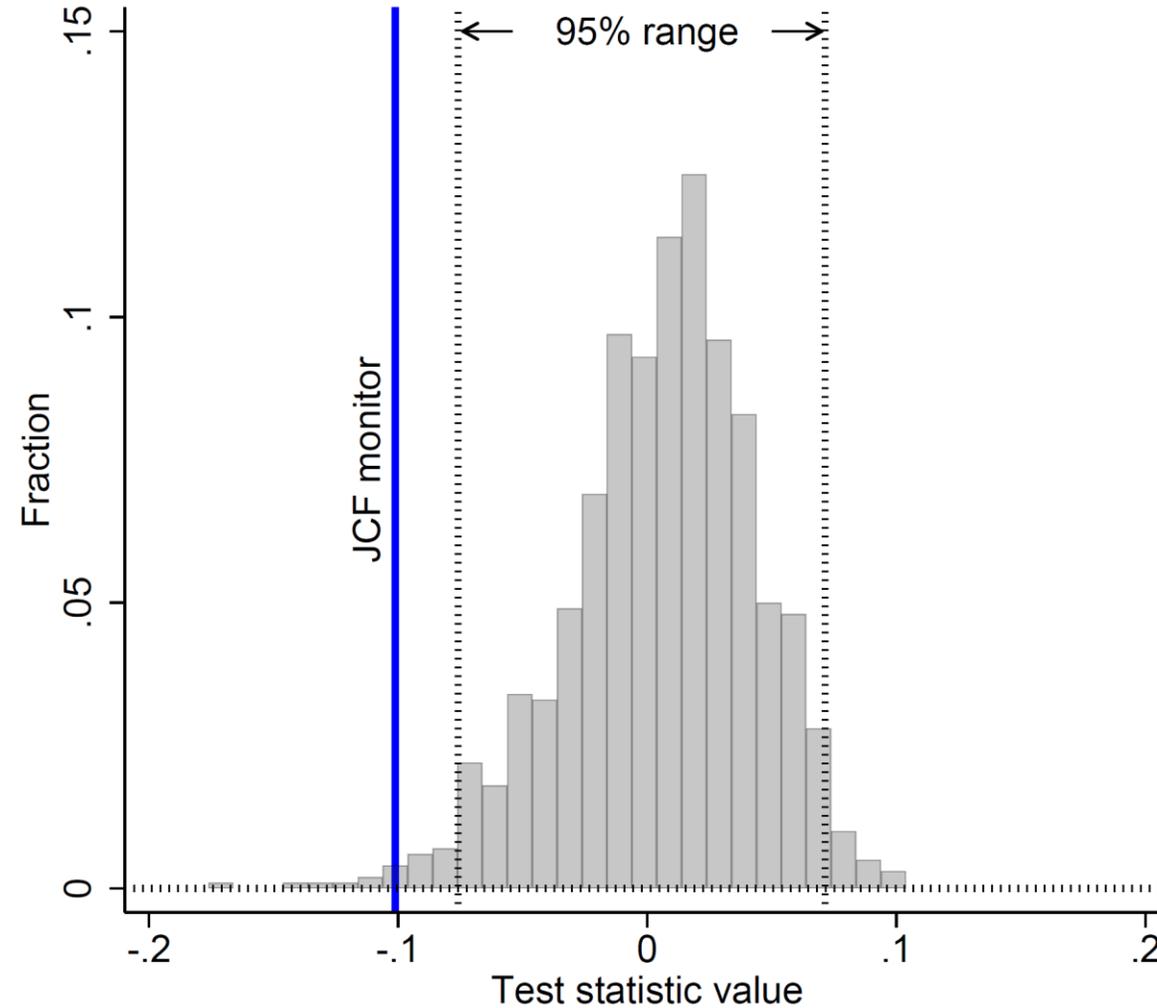
- Inference goal: test whether the $\hat{\beta}_\tau$'s have lower values around $\tau = 0$, i.e., more missing data near pollution alerts
- Define test statistic: donut difference-in-means estimator

$$T = \frac{1}{7} \sum_{\tau \in [-3,3]} \hat{\beta}_\tau - \frac{1}{40} \sum_{\tau \in [-30,-11] \cup [11,30]} \hat{\beta}_\tau$$

- Mean of probably treated period – mean of probably untreated period, with some buffer
- Null: $T = 0$; Alternative: $T \neq 0$
- Randomized inference:
 - Generate 5,000 hypothetical scenarios, each with 21 randomly-dated pollution alerts
 - Obtain 5,000 “placebo” test statistics $\{\tilde{T}\} \Rightarrow$ “empirical null distribution”
 - p -value of actual $\hat{T} =$ proportion of the empirical null that is more extreme than \hat{T}

Randomized inference: How statistically significant is the dip?

Distribution of effect sizes across 5,000 placebo alert scenarios for the JCF monitor:



Notes: Test statistic = $\frac{1}{7} \sum_{\tau \in [-3,3]} \hat{\beta}_{\tau} - \frac{1}{40} \sum_{\tau \in [-30,-11] \cup [11,30]} \hat{\beta}_{\tau}$.

Vertical line shows true test statistic for the JCF monitor.

Simultaneous test (all monitors)

- Repeat JCF exercise to the entire pool of 1,359 monitors, testing a collection of hypotheses at once

$\{H_i : \text{Monitor } i\text{'s capture rate is not affected by pollution alerts}\}_{i=1}^{1,359}$

- Output:

$\{\hat{\beta}_\tau\}_i$: event study coefficients for each monitor

$\{\hat{T}\}_i$: test statistic based on $\hat{\beta}_\tau$'s

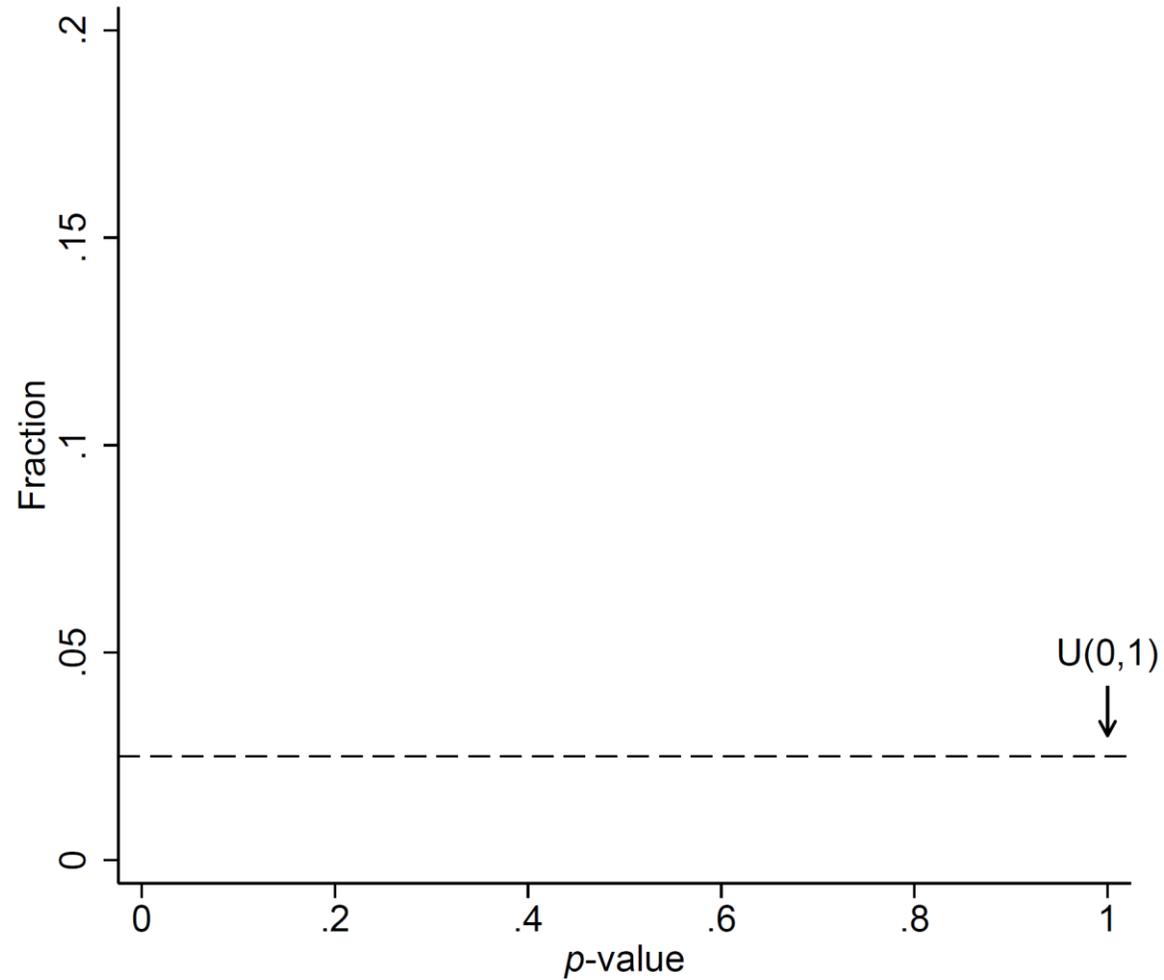
$\{\text{p_value}\}_i$: permutation-based two-tail p-value based on \hat{T} 's

- Main challenge is **over-rejection**:

- At any chosen rejection threshold α , about $100 \cdot \alpha\%$ false positives even if alerts have no effect whatsoever

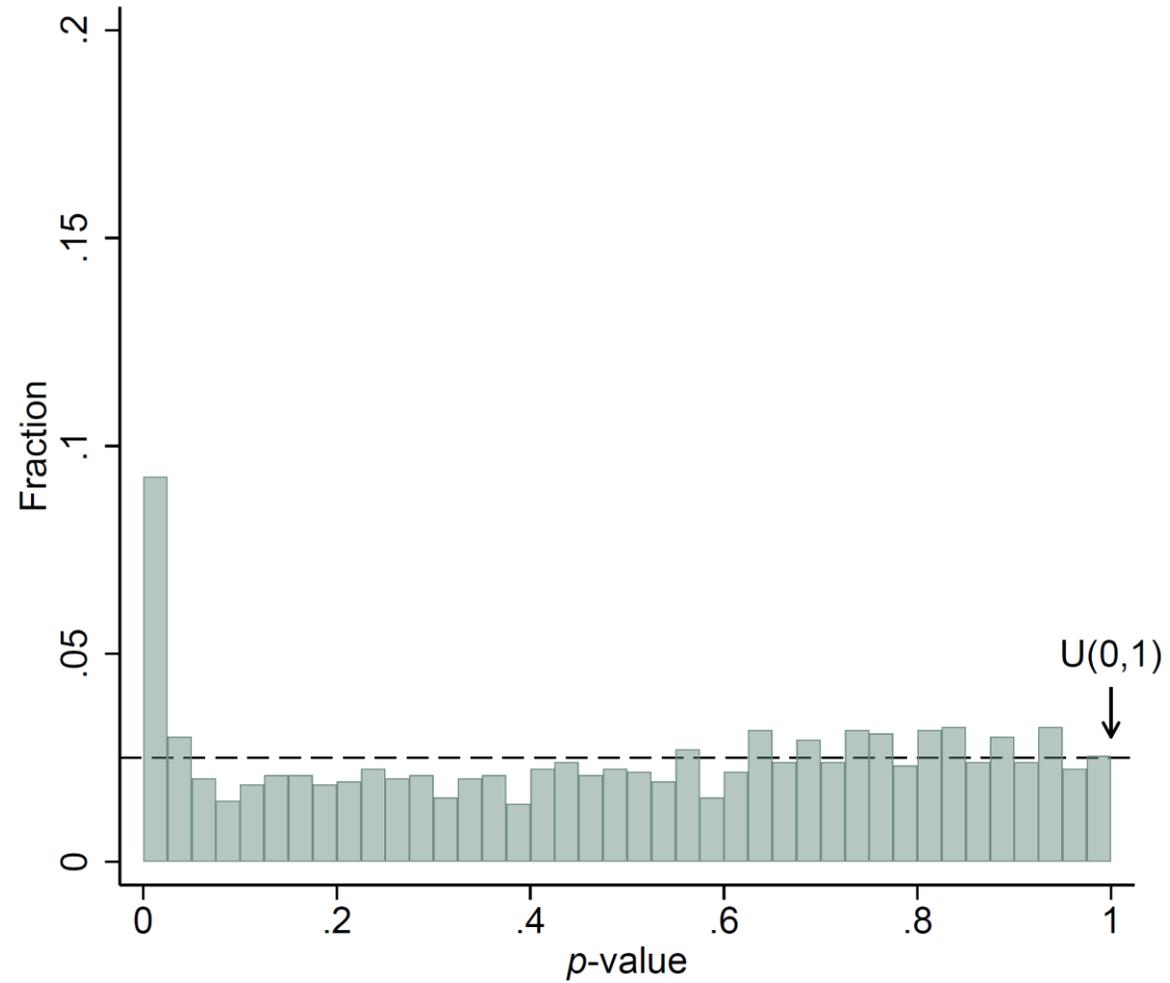
p-Curve: Distribution of permutation-based p-values

Under the null that alerts do not affect missingness, p-values should follow $U(0,1)$:



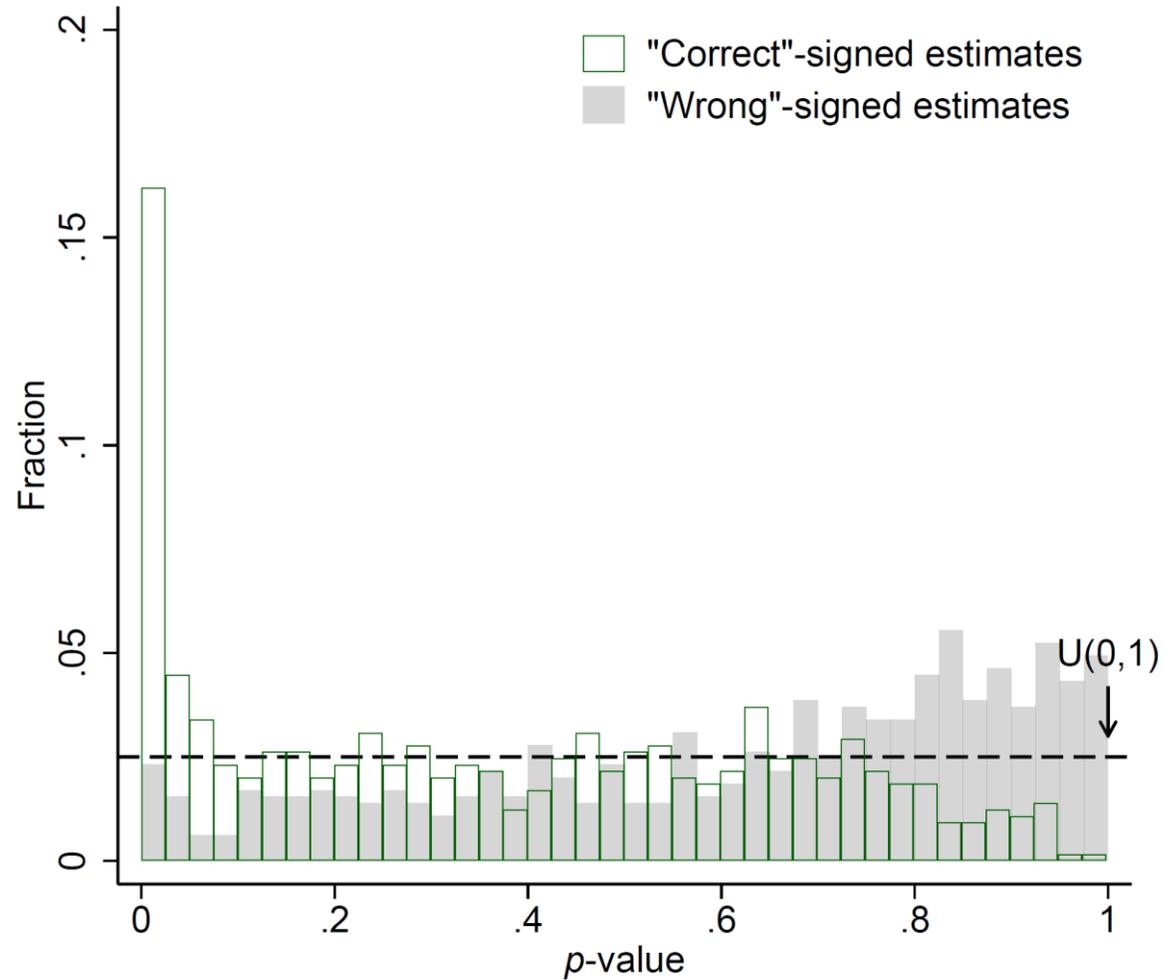
p-Curve: Distribution of permutation-based p-values

Instead, we find over-abundance of small p-values



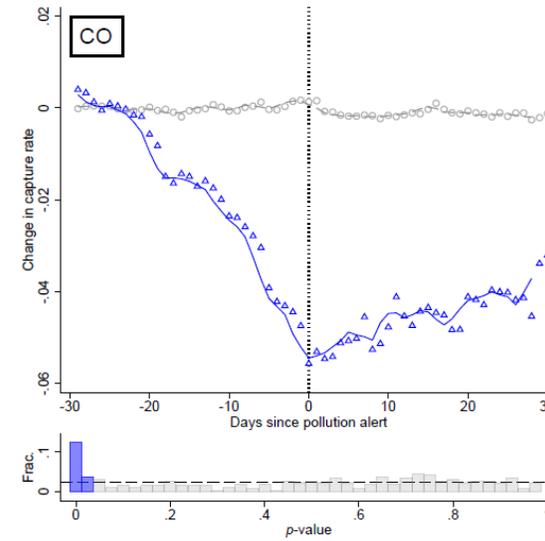
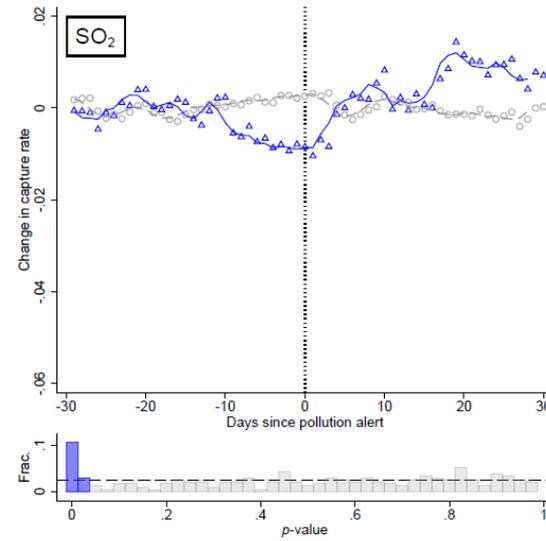
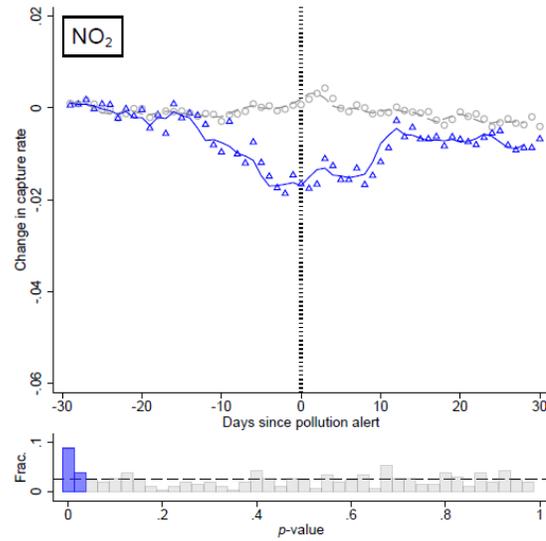
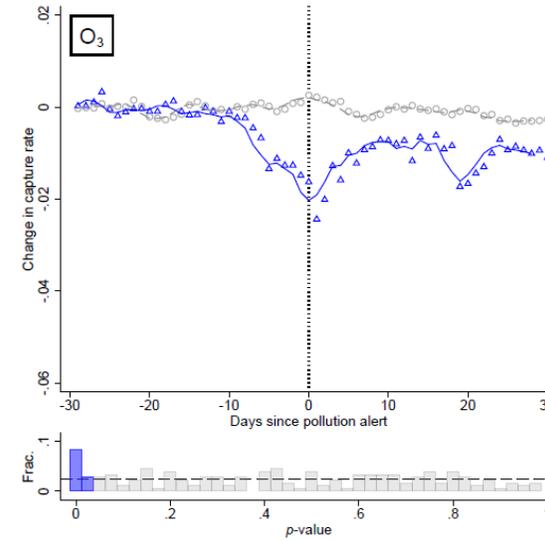
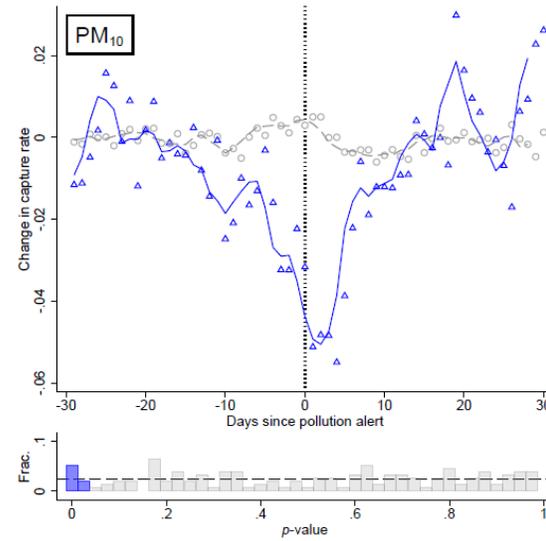
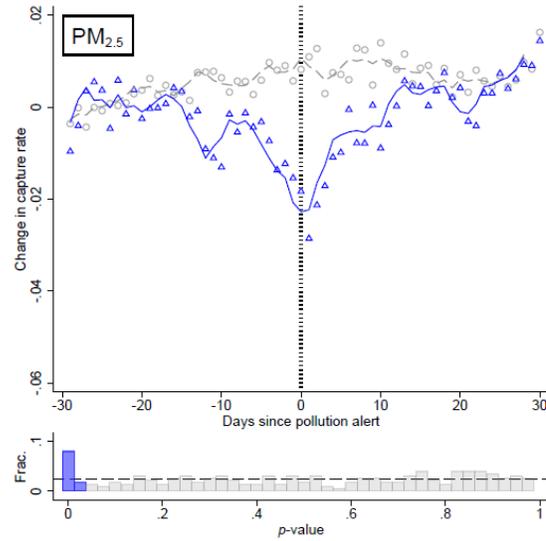
p-Curve: Distribution of permutation-based p-values

... and the spike of small p-values are driven by “correct”-signed estimates (“dips”)



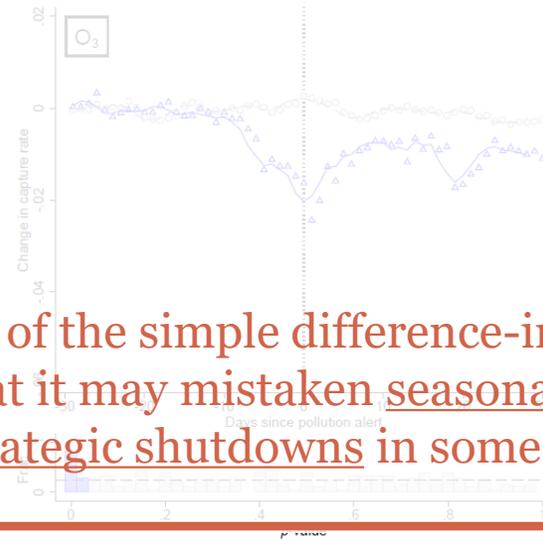
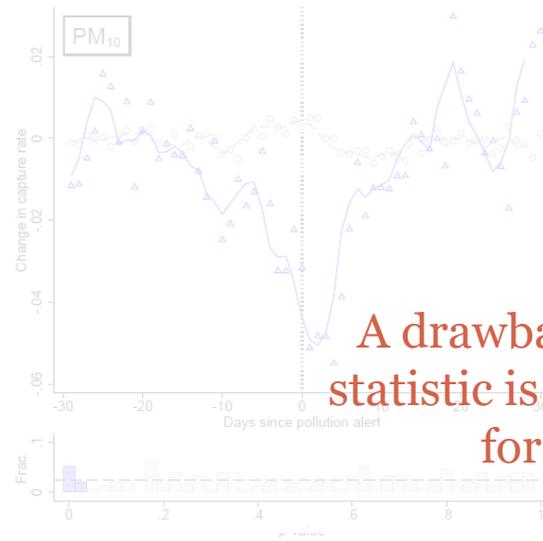
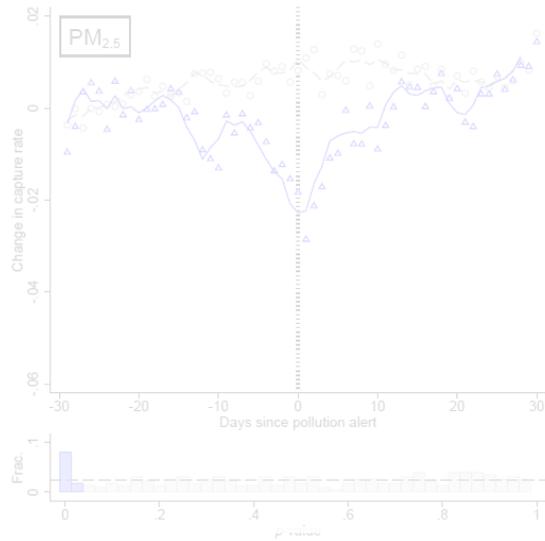
Notes: “Correct”-signed means $\hat{T} < 0$ (i.e., data capture rate drops around pollution alerts)

“Interesting” Monitors (Δ) and Other Monitors (\circ)

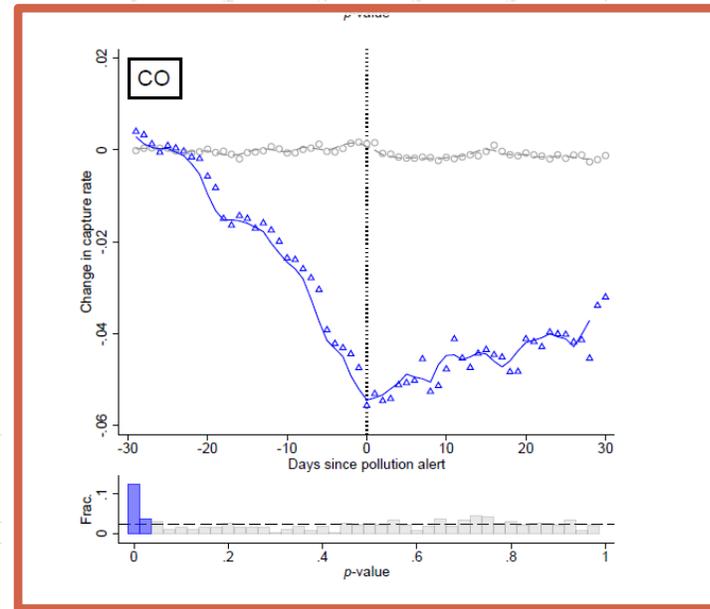
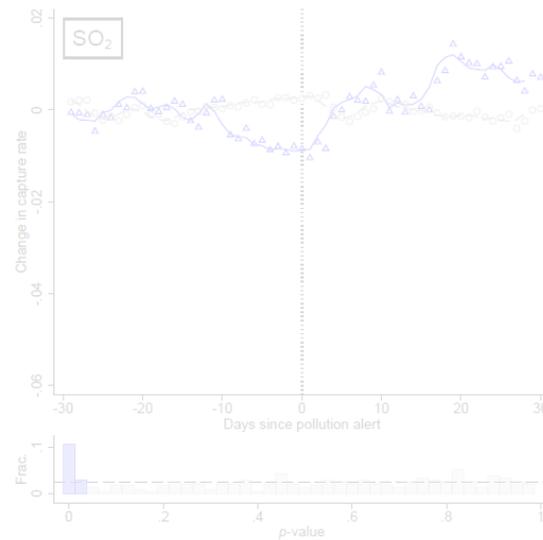
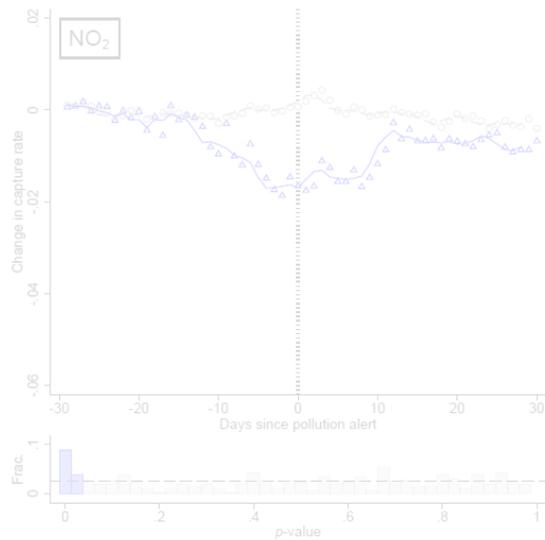


Notes: “Interesting” monitors are those with p-values ≤ 0.05

“Interesting” Monitors (Δ) and Other Monitors (\circ)

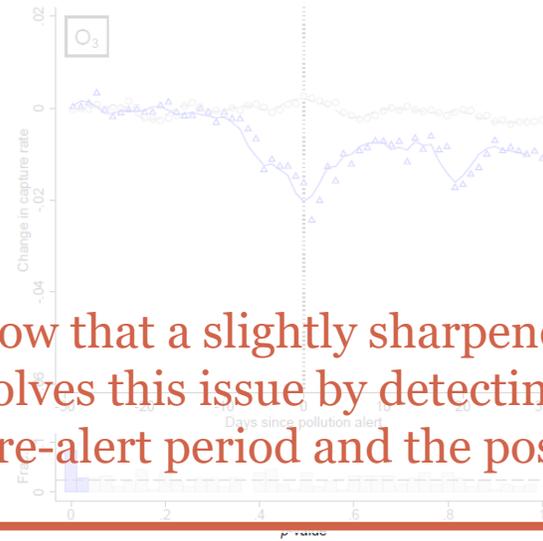
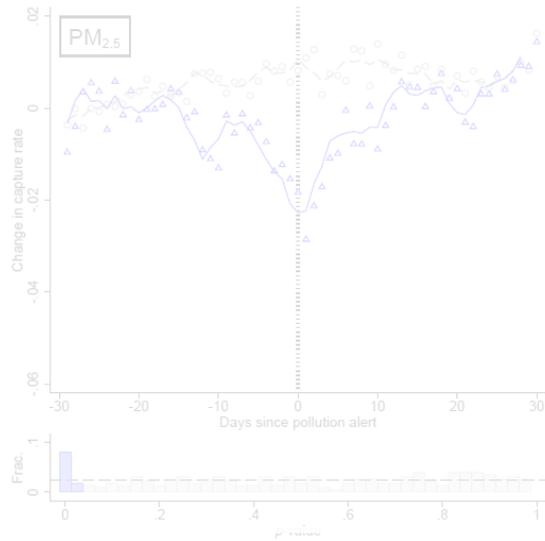


A drawback of the simple difference-in-mean test statistic is that it may mistaken seasonal monitoring for strategic shutdowns in some cases

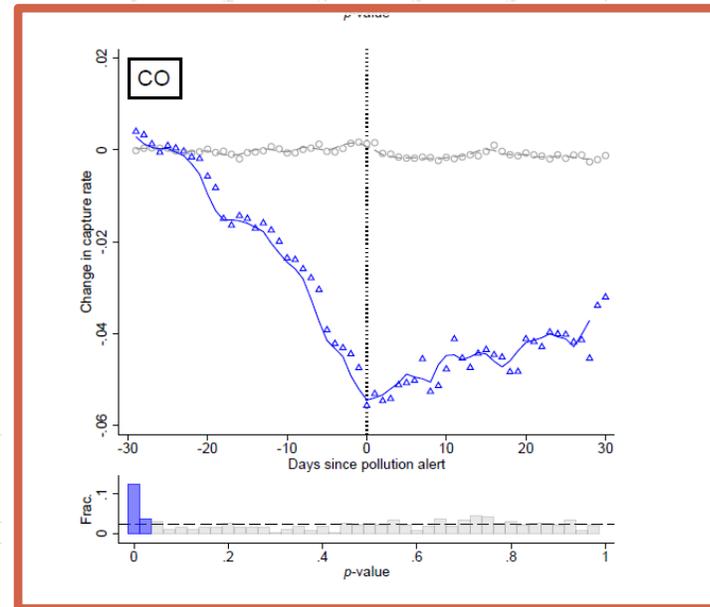
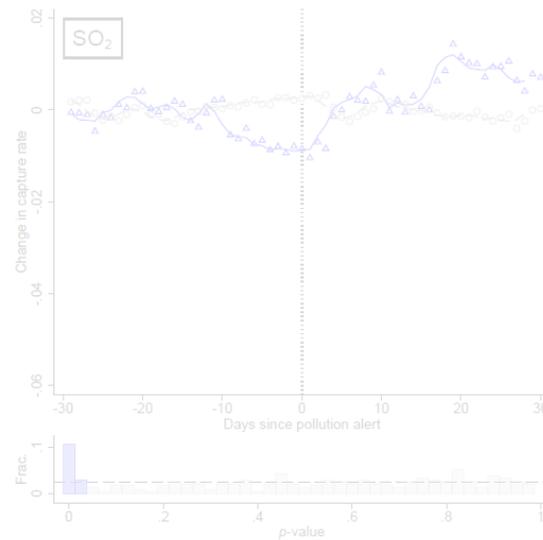
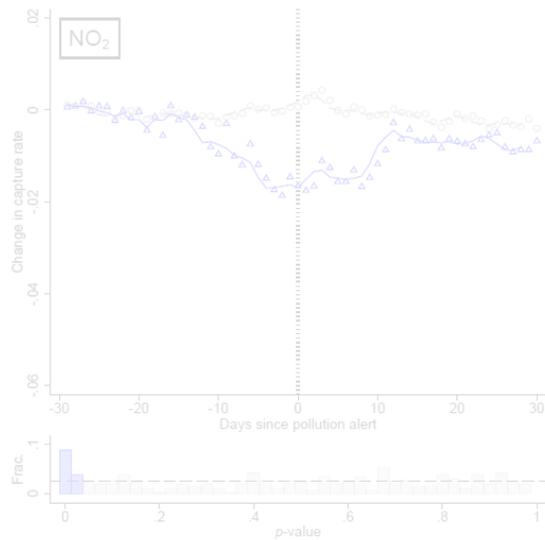


Notes: “Interesting” monitors are those with p-values ≤ 0.05

“Interesting” Monitors (Δ) and Other Monitors (\circ)

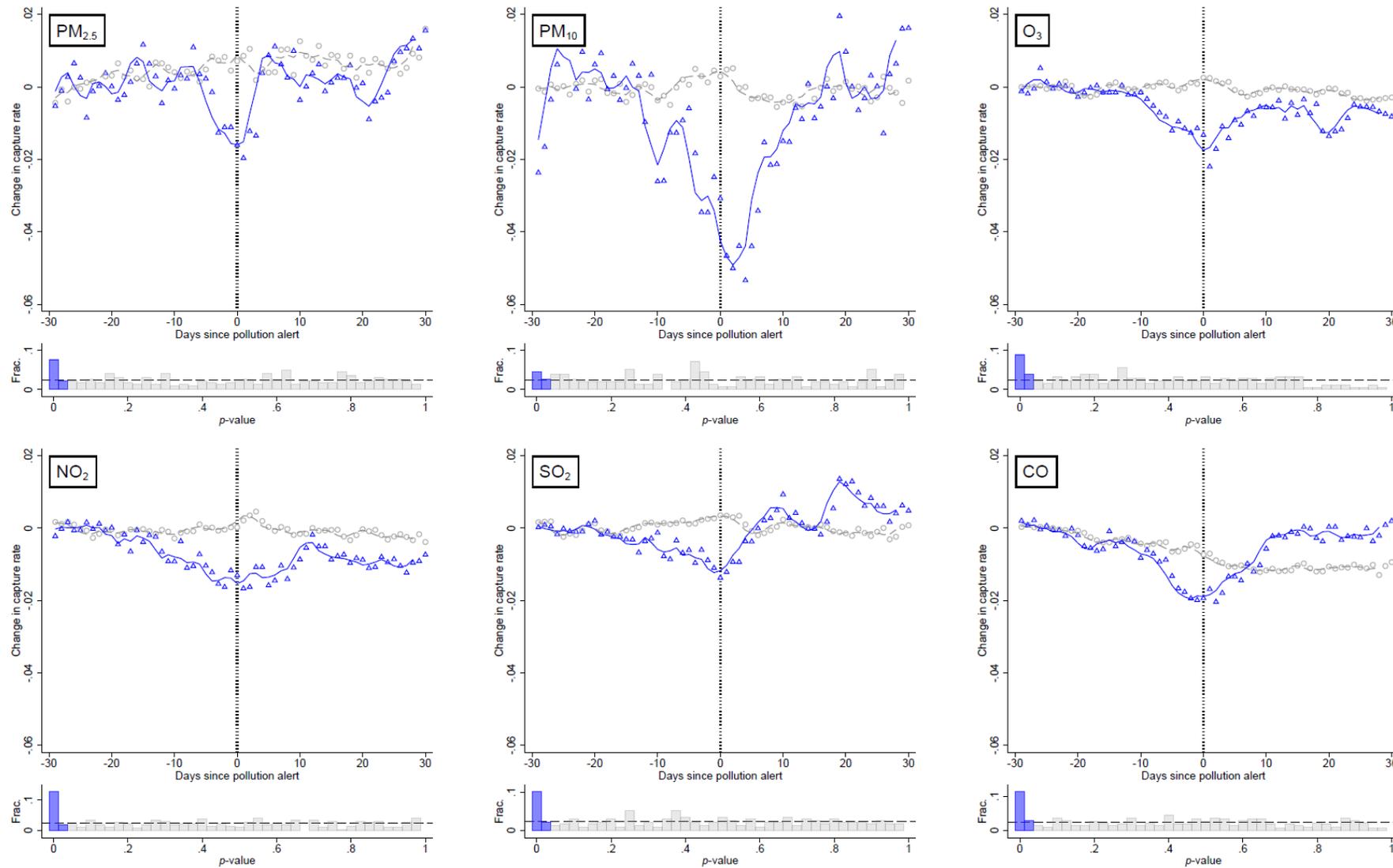


In paper, we show that a slightly sharpened version of the test statistic resolves this issue by detecting the dip relative to *both* the pre-alert period and the post-alert period



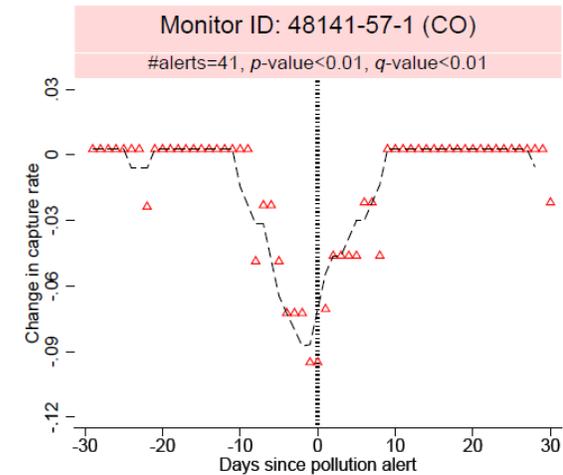
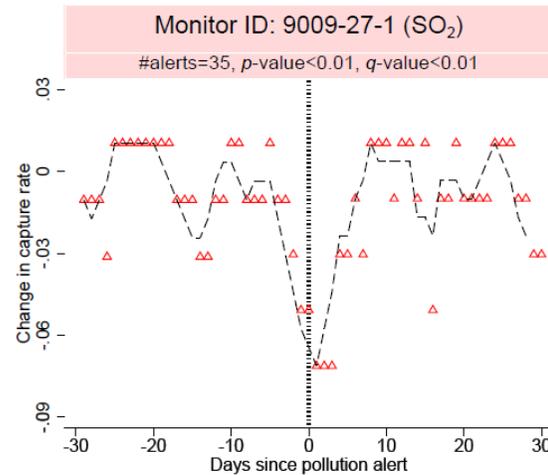
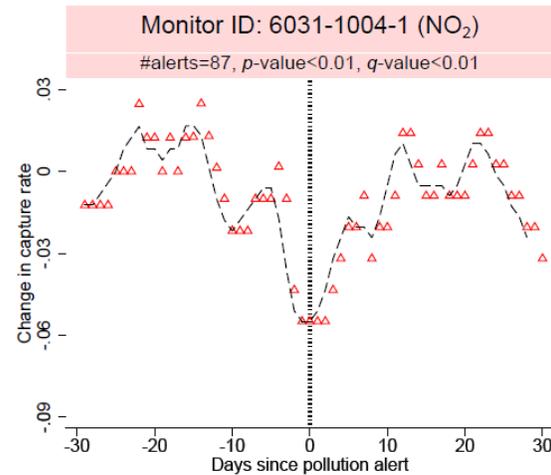
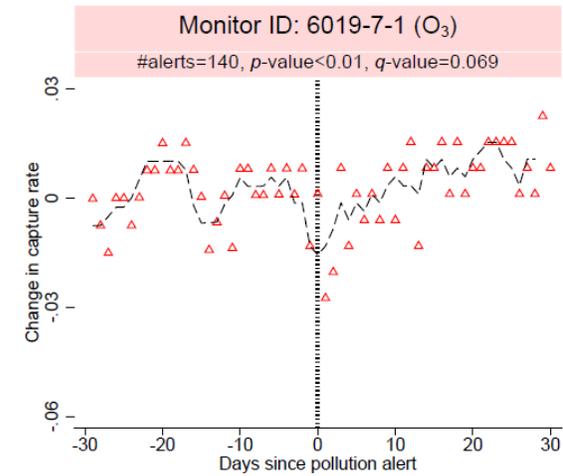
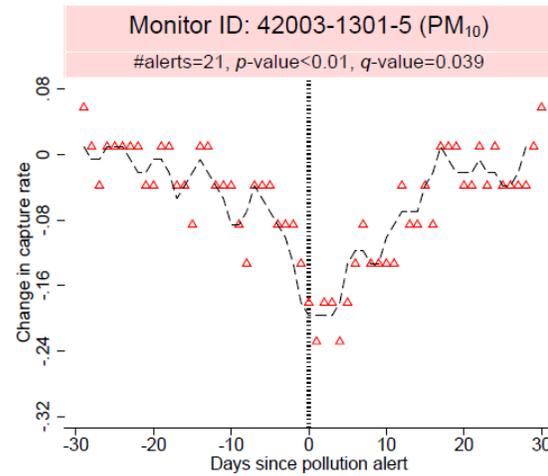
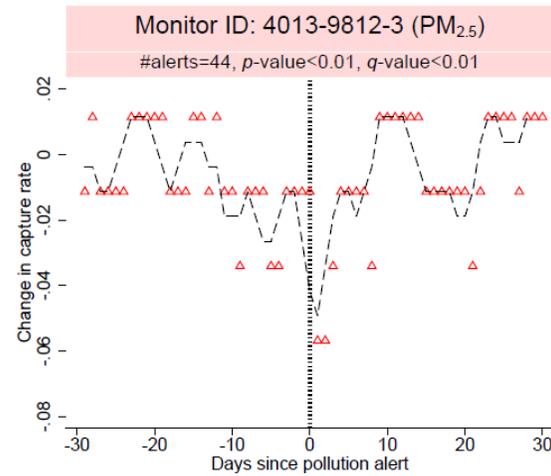
Notes: “Interesting” monitors are those with p-values ≤ 0.05

“Interesting” Monitors (Δ) and Other Monitors (\circ): “Sharpened” Test Statistic



Notes: The sharpened, two-sided test rejects the null if the capture rate around time zero is lower than *both* the pre-period and the post-period. See paper for more details.

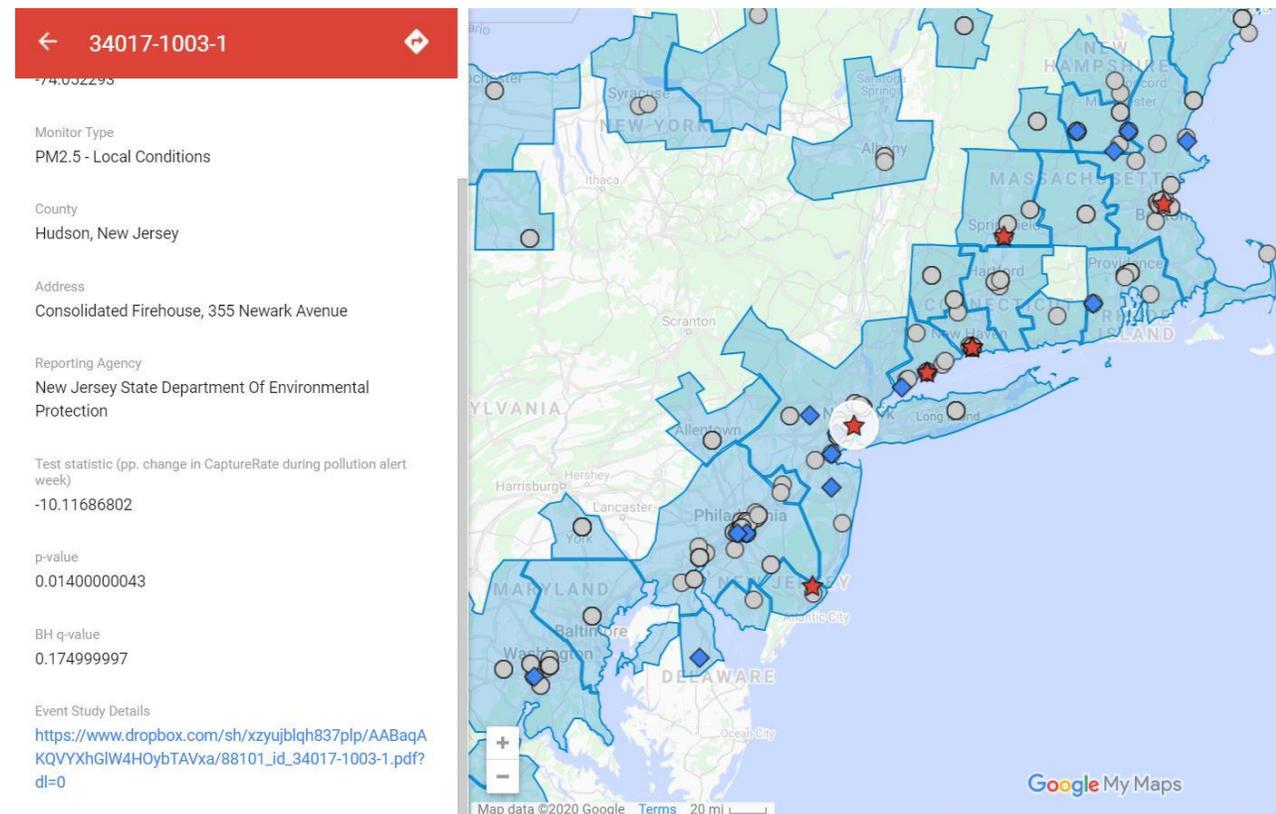
Examples of “Very Interesting” Monitors (☆)



Notes: “Very Interesting” monitors are manually selected for illustration purpose, not used in any formal analysis

Study Website (BETA)

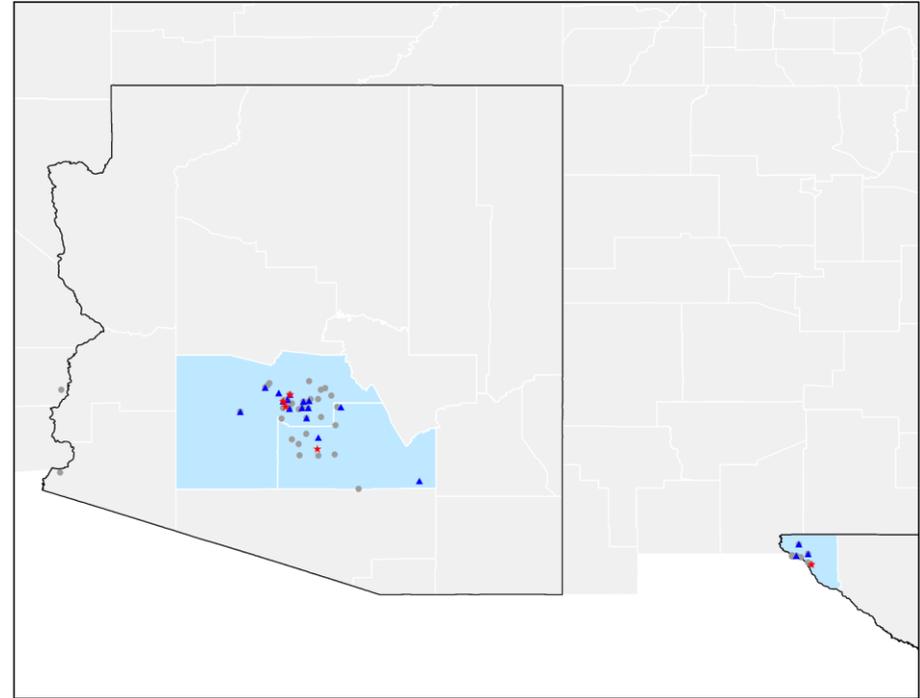
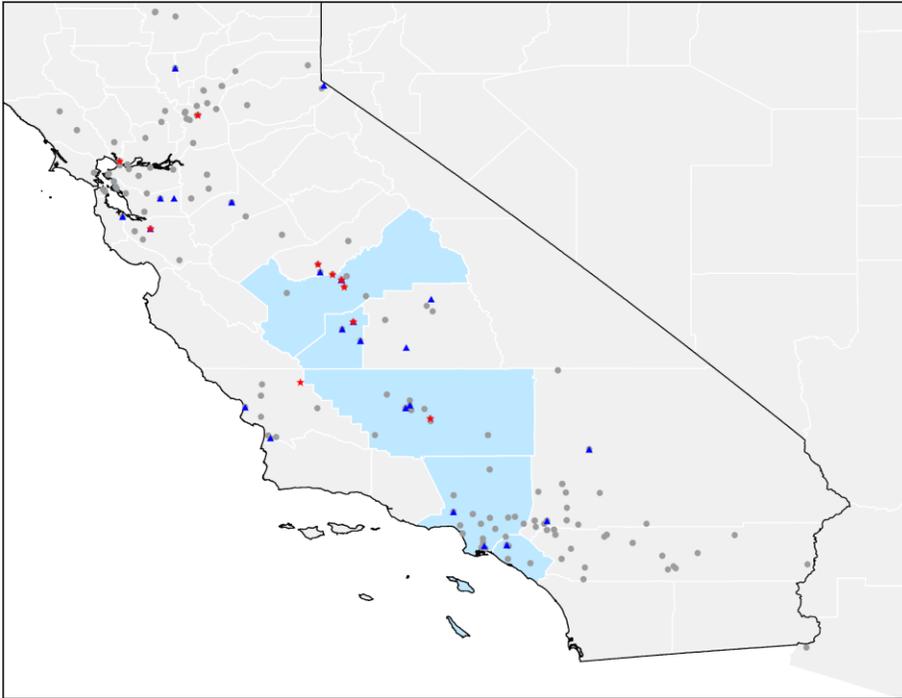
- Estimation results available at the individual monitor level:



Link: https://www.google.com/maps/d/u/o/edit?mid=1e6vuA_OXa-QfCMrYanwkWV7XiGl5od1q&usp=sharing

Hotspot Regions

- 14 CBSAs across the U.S. house 60% of all “interesting monitors”
- Examples: CA & AZ



Outline

- Institution
- Main Results
- Discussion
 - Mechanisms?
 - Economic importance?
 - Policy alternatives?

Mechanism

1. How?
2. Why?

Mechanism

- Paper discusses monitoring protocols and why *might* monitors miss data
 - Key reference: *Quality Assurance Handbook for Air Pollution Measurement Systems* (U.S. EPA, 2013)
- Consider possibility of missing data in three stages of monitoring
 1. Measurement Acquisition
 2. Quality Control
 3. Data Submission

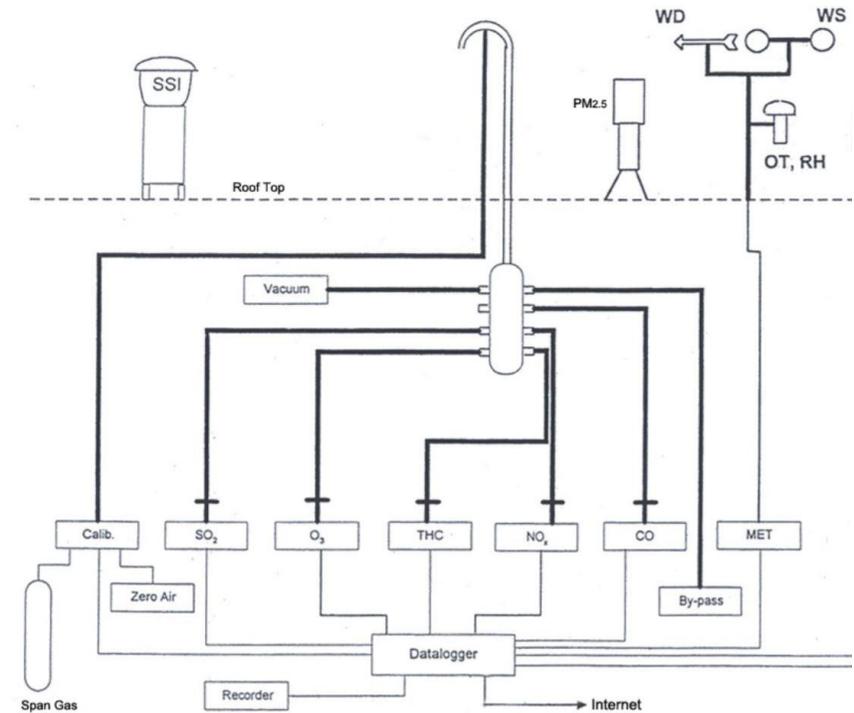
1. Measurement Acquisition

Monitoring site



Source: U.S. EPA

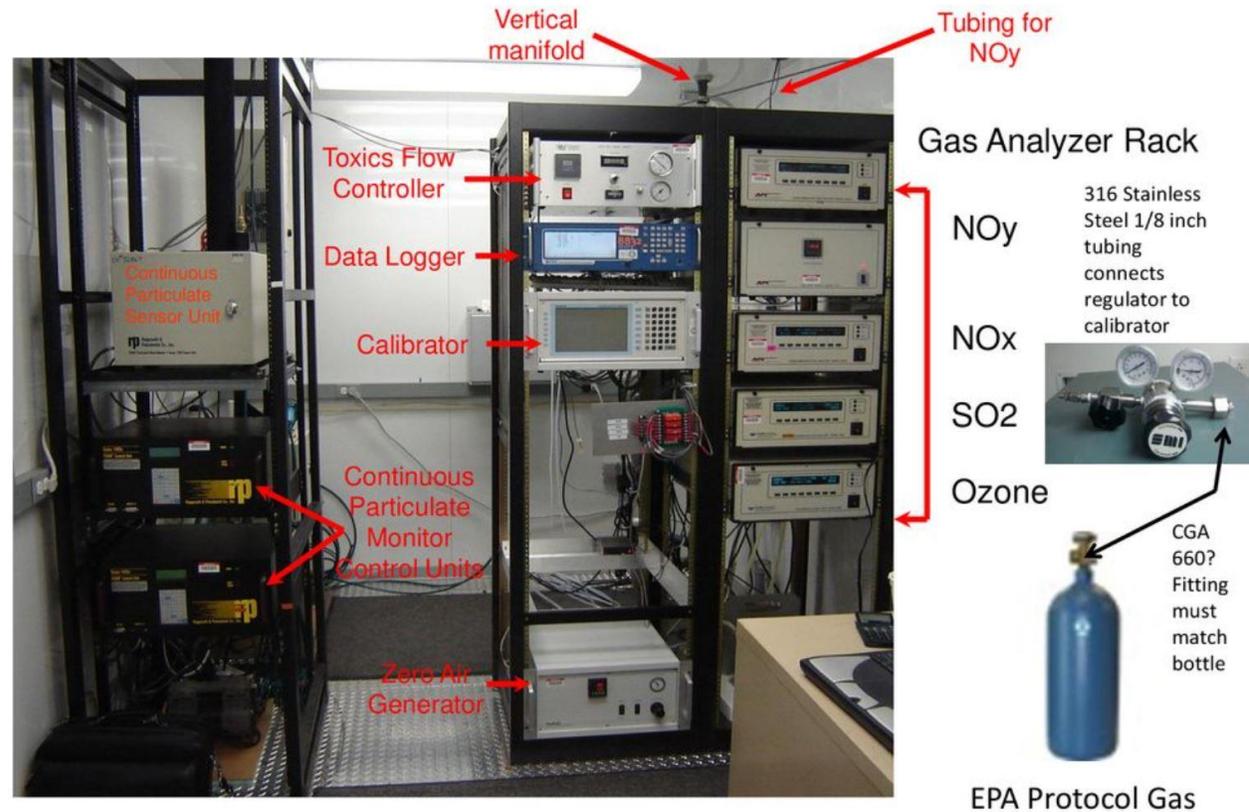
Shelter design



Source: California Resources Board

1. Measurement Acquisition

Look inside a shelter



Source: Glenn Gehring

1. Measurement Acquisition

- Missing data problem may arise at measurement acquisition stage
 - Instrument malfunction, sample contamination, preventive maintenance, staff shortage, power outage ..
 - .. and **strategic non-sampling**, as we argue in this paper

1. Measurement Acquisition

- Missing data problem may arise at measurement acquisition stage
 - Instrument malfunction, sample contamination, preventive maintenance, staff shortage, power outage ..
 - .. and **strategic non-sampling**, as we argue in this paper
- Can these **alternative reasons** explain the finding?

1. Measurement Acquisition

- Missing data problem may arise at measurement acquisition stage
 - Instrument malfunction, sample contamination, preventive maintenance, staff shortage, power outage ..
 - .. and **strategic non-sampling**, as we argue in this paper
- Can these **alternative reasons** explain the finding?
 - Probably not.
 - Most pollution analyzers are placed inside the HVAC-controlled shelter
 - Federal FRM/FEM-certified monitoring technologies should stand up to the range of pollution conditions seen in the U.S. (over 99% daily observations < 100 ug/m³)
 - We train machine learning models; find outdoor weather elements are not predictive of missingness at all

2. Quality Control

- Missing data problem may also arise if a monitor fails **periodic QC tests** conducted by the state agency
 - Example: one-point QC check. An ozone monitor is exposed to a gas of known concentration; if measured ozone exceeds true concentration by 7%, the monitoring agency should voids all previous readings extending back to the date when the monitor passed the previous one-point QC check
 - Done once every two weeks
- What about **extreme values**?
 - EPA guideline encourages manual inspections of all data to spot unusual values to “indicate a gross error in the data collecting system”
 - But, an outlier is considered valid until there is an explanation for why the data should be invalidated, e.g. if the monitor fails a subsequent one-point QC test.
- Bottom line: QC failures typically result in the invalidation of large chunks of data, which we believe is unlikely to explain short-term missingness as we identify in this paper

3. Data Submission

- Processed, QC-ed data are submitted by the state to the federal EPA's Air Quality System
- EPA has the ultimate authority to decide whether it will use the submitted data in determining NAAQS compliance
- Very occasionally, EPA has invalidated states' data after failures in federal audits
 - Example: A contract lab's audit failure led data from four states to be suspended from NAAQS comparison (<https://www.epa.gov/air-trends/pm25-data-omitted-airtrends-assessment>)
- These cases also tend to invalidate large swaths of data; unlikely to be relevant for this paper

Mechanism

1. How?
2. Why?

Mechanism

- The incentive to avoid falling (back) into non-attainment appears to be the primary driver of our findings
 - Perhaps not entirely surprising given large fiscal costs of NAAQS violation (e.g., Greenstone, List, Syverson 2012; Walker 2013)
 - ... and evidence on states' efforts to achieve localized air quality improvements near monitors (e.g., Bento, Freedman, Lang, 2015; Auffhammer, Bento, Lowe, 2019)

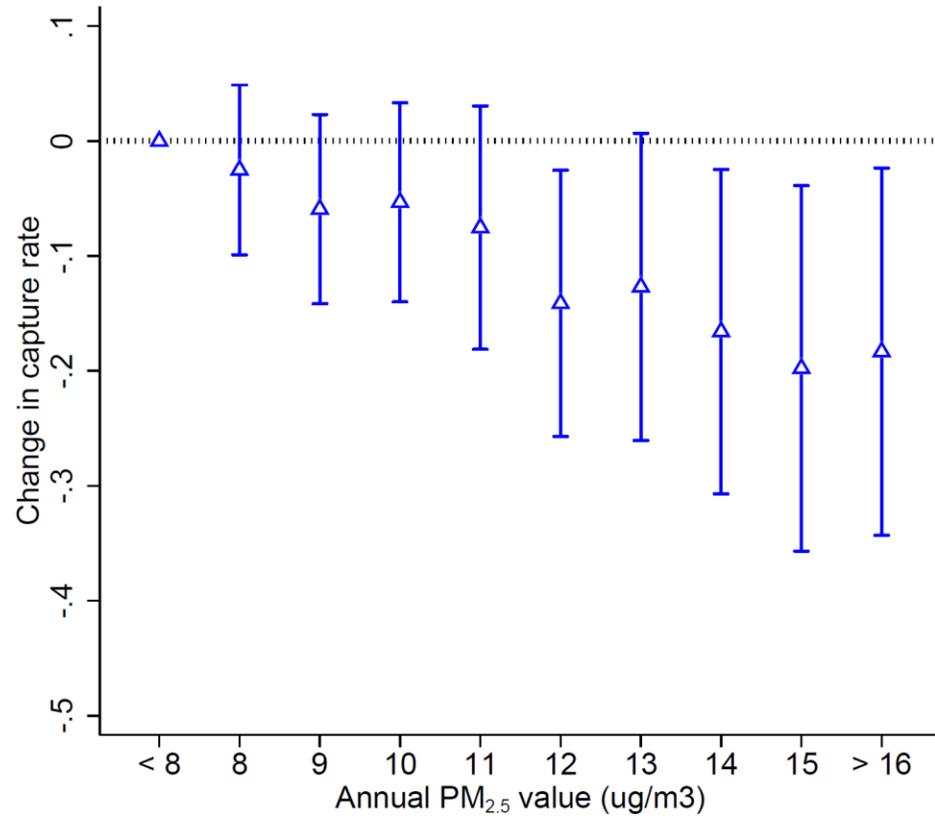
County's NAAQS violation status is a strong predictor for having “interesting” monitors

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Dep. var.:	1(p -value \leq 0.05)				1(q -value \leq 0.05)			
Non-attainment	0.066** (0.030)				0.039* (0.021)			
Non-attainment \times 1(“wrong” sign)		-0.014 (0.033)	0.011 (0.034)	-0.001 (0.041)		-0.002 (0.024)	0.012 (0.024)	0.022 (0.030)
Non-attainment \times 1(“correct” sign)		0.203*** (0.055)	0.220*** (0.055)	0.223*** (0.061)		0.111*** (0.039)	0.124*** (0.039)	0.129*** (0.044)
Above median Democrats			-0.022 (0.027)				-0.014 (0.019)	
Above median LCV score			-0.023 (0.027)				-0.021 (0.019)	
Above median government size			0.007 (0.017)				-0.001 (0.012)	
Above median corruption			0.035* (0.018)				0.008 (0.013)	
State fixed effects				✓				✓
Mean dep. var.	0.117	0.117	0.117	0.117	0.052	0.052	0.052	0.052
Observations	1,359	1,359	1,359	1,359	1,359	1,359	1,359	1,359

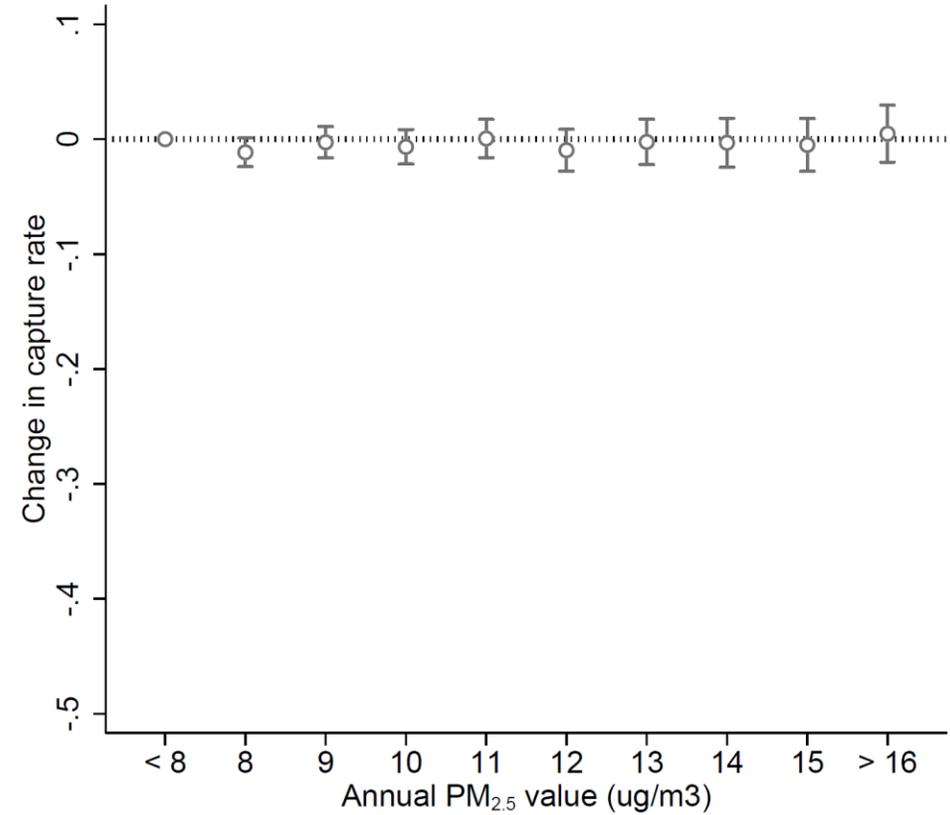
Notes: “ q -value” is false discovery adjusted significance level a la [Benjamini and Hochberg \(1995\)](#), [Storey \(2013\)](#), [Anderson \(2008\)](#).

For “interesting” monitors, data capture rates are lower during bad years in general, not just around pollution alerts

A. “Interesting” monitors



B. Other monitors

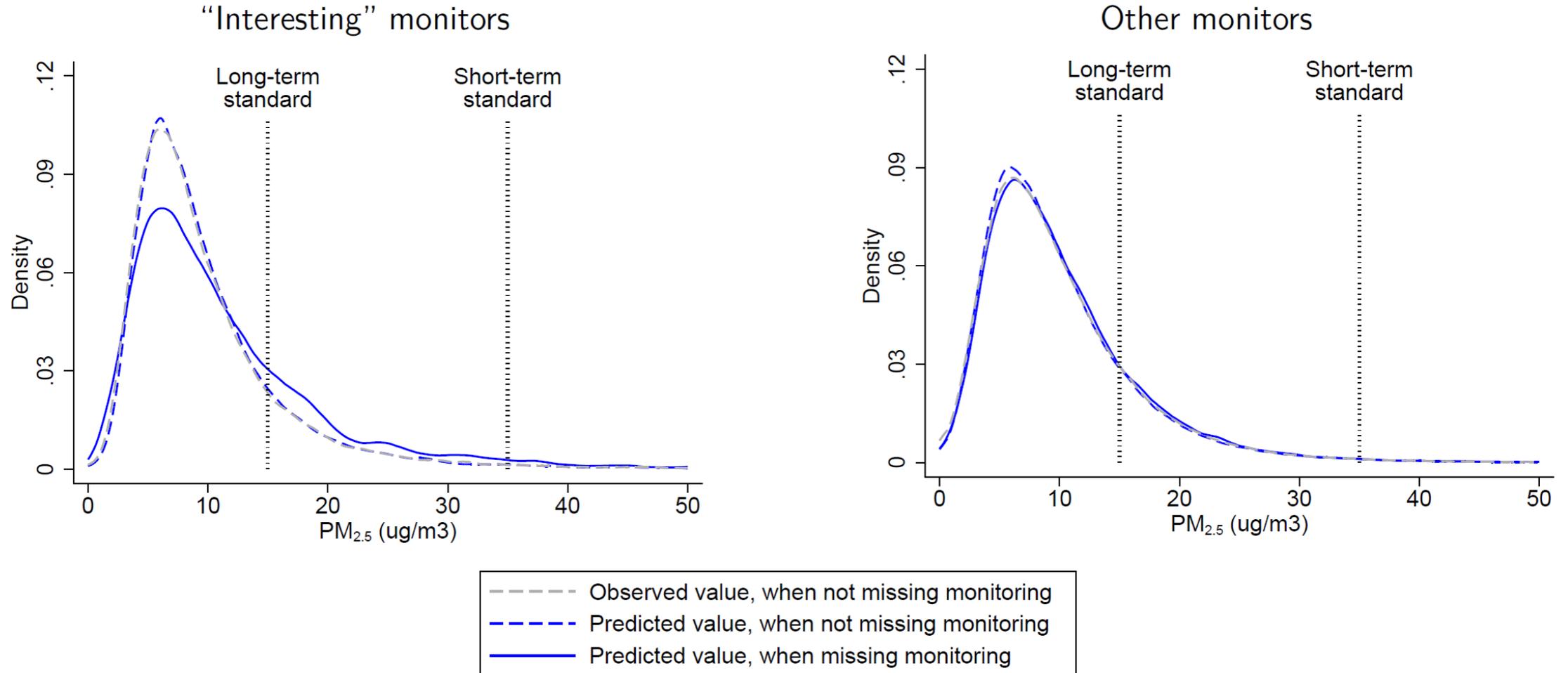


Economic Importance

- Risk of **understating pollution** in a county if high-pollution days are under-sampled
 - This can lead to **foregone health value** ...
 - ... due to regulation-induced air quality improvements that the county would otherwise have enjoyed without strategic monitoring
 - See this idea in Sullivan and Krupnick (2018) and Fowlie, Rubin, Walker (2019)
- To illustrate this effect, use **inverse distance weighting (IDW)** to characterize distribution of $PM_{2.5}$ when monitoring data are missing
 - Impute monitor i 's reading as inverse-distance-weighted average of readings from all monitors within 20 miles

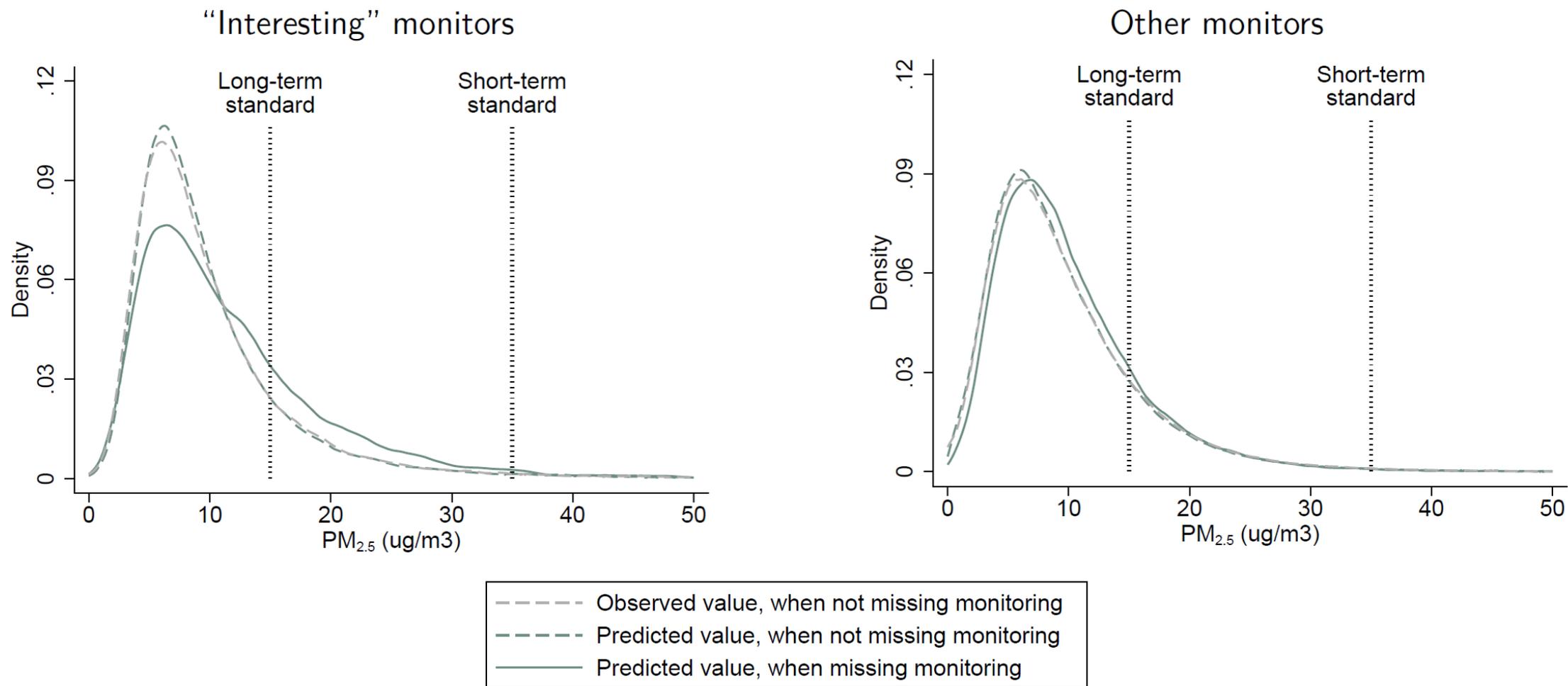
Imputation method: Inverse distance weighting (IDW)

Impute missing PM_{2.5} from “donor” monitors within 20-mile radius



Imputation method: Atmospheric modeling ([Di et al., 2019](#))

Pattern replicates almost exactly using ML-based predictions instead



Notes: Di, Qian, et al. "An ensemble-based model of PM_{2.5} concentration across the contiguous United States with high spatiotemporal resolution." *Environment International* 130 (2019): 104909.

Economic Importance

- Our imputation suggests:
 - 23.1% of the missing days would have $> 15 \text{ ug/m}^3$ (6.6 extra days of annual-standard violation)
 - 2.7% of the missing days would have $> 35 \text{ ug/m}^3$ (0.8 extra day of daily-standard violation)
- This works to an annual foregone health (mortality) VSL of \$67 million per interesting monitor:
 - 1) Each 24-hour exceedance = 9.6 pp. increase in the chance of nonattainment status within next three years
 - 2) Nonattainment = 1.6 ug/m^3 reduction in $\text{PM}_{2.5}$ per year (Sanders, Barreca, Neidell, 2020)
 - 3) 10 ug/m^3 $\text{PM}_{2.5}$ = 6% change in all-cause adult mortality (Krewski et al., 2009)
 - 4) VSL of \$8.9 million 2020 USD

Alternative Institutions

- How to prevent strategic non-monitoring?
- Don't just *ignore* missing values
 - Substitute missingness with something that better approximate the truth
- Can learn from the **U.S. EPA Acid Rain Program (ARP)**
 - Cap-and trade program that monitors power plants SO₂ and NO_x emissions through CEMS
 - If a unit's data capture rate falls below 90%, impute with **maximum** value in the past 30 days
 - ARP Data capture rate: > 90%
- Probably too conservative in context of ambient air monitoring; but more stringent data substitution rule can probably help

Conclusion

- An example of large-scale inference problems where the research goal is to credibly identify a small amount of interesting units among a sea of null
- Many applications in other fields; relatively few in economics
 - High-throughput screening for drug discovery
 - Genomics/proteomics data analysis

Thank you!

Yingfei Mu ([JHU econ Ph.D. candidate](#))

Edward Rubin ([edrub.in](#))

Eric Zou ([eric-zou.com](#))

Appendix

p-Curve: Distribution of permutation-based p-values

Robustness to alternative test-statistic specifications

