Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01477

What's *Missing* in Environmental Self-Monitoring: Evidence from Strategic Shutdowns of Pollution Monitors*

Yingfei Mu
Edward Rubin
Eric Yongchen Zou

May 2024

Abstract

Regulators often rely on regulated entities to self-monitor compliance, creating strategic incentives for endogenous monitoring. This paper builds a framework to detect whether local governments skip air pollution monitoring when they expect air quality to deteriorate. The core of our method tests whether the timing of monitor shutdowns coincides with the counties' air quality alerts – public advisories based on local governments' own pollution forecasts. Applying the method to a monitor in Jersey City, NJ, suspected of a deliberate shutdown during the 2013 "Bridgegate" traffic jam, we find a 33% reduction of this monitor's sampling rate on pollution-alert days. Building on large-scale inference tools, we then apply the method to test more than 1,300 monitors across the U.S., finding 14 metro areas with clusters of monitors showing similar strategic behavior. We assess geometric imputation and remote-sensing technologies as potential solutions to deter future strategic monitoring.

JEL Codes: C12, H77, Q53, Q58

* Mu: Department of Economics, Johns Hopkins University (email: ymu7@jhu.edu); Rubin: Department of Economics, University of Oregon (email: edwardr@uoregon.edu); Zou: Department of Economics, University of Michigan and NBER (email: ericzou@umich.edu). We thank Michael Anderson, Trudy Ann Cameron, Jonathan M.V. Davis, Eric Edwards, Dave Evans, Mary Evans (thanks to the Adopt-a-Paper program), Meredith Fowlie, Cynthia Giles, Corbett Grainger, Andreas Hagemann, Alex Hollingsworth, Nicolai Kuminoff, Shanjun Li, Julian Reif, Michelle M. Rubin, Ivan Rudik, Jay Shimshack, William Wheeler, Jianwei Xing, officials at the U.S. Environmental Protection Agency, and seminar participants at various conferences and seminars for helpful comments. All errors are our own.

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01477

1. Introduction

When regulators face substantial monitoring requirements, they commonly ask regulated entities to monitor their own compliance. Police officers are charged with turning on/off body cameras that verify their maintenance of ethical behavior; hospital staff transcribe operation events to catalog surgeons' regulatory compliance; countries self-monitor greenhouse gas emissions to demonstrate adherence to climate commitments. This system of self-monitoring is particularly common within environmental regulation, where local entities – such as state governments and individual firms – assume the roles of both the subject of regulation *and* the recorder of pollution data that demonstrate compliance. Can federal regulators rely on the regulated to provide complete, representative self-monitoring data? We study this question in the context of U.S. air quality regulation, where state and local governments monitor air pollution to demonstrate compliance with federally set air quality standards. We show that state agencies' leeway to decide when (not) to monitor, combined with the ability to anticipate pollution events in the near future, results in strategic timing in monitoring activities at some locations. We begin with a motivating anecdote, followed by an econometric analysis of the general prevalence of strategic monitoring.

On September 9th, 2013, two of three lanes to the George Washington Bridge closed for five days at the toll plaza connecting Fort Lee, New Jersey and Manhattan, New York for what was initially said to be a traffic study. The event was later found to be a deliberate act of political retribution.[1] Coincidentally, at the time of the Bridgegate-induced traffic jams, a nearby fine particulate matter ($PM_{2.5}$) air pollution monitor on the rooftop of the Jersey City Firehouse stopped collecting data. The monitor, placed by the state government to continuously monitor compliance

[1] See *Wikipedia, The Free Encyclopedia*, s.v. "Fort Lee lane closure scandal," (accessed November 13, 2020), https://en.wikipedia.org/wiki/Fort_Lee_lane_closure_scandal

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01477

with federal Clean Air Act mandates, was later found to have been inoperative for 13 days (September 6th–18th), the longest inactive period recorded in the decade since its installation. While an investigation by the U.S. Environmental Protection Agency (EPA) blamed "equipment malfunction,"[2] the timing of the incident has raised concerns that the monitor was intentionally disabled so that it would not record the spike in air pollution caused by the Bridgegate traffic jam.

This incident raises a general question of whether local officials are able to deliberately halt pollution monitoring when they anticipate the monitor will record elevated pollution levels. First, state governments have the **incentive** to "game." While the federal EPA sets the national air quality standards (NAAQS), states and local governments self-monitor compliance with these standards. When state governments' own monitoring indicates a lack of compliance, they bear the regulatory penalties including elevated requirements of expensive emission-reduction investments. Second, state governments have the **discretion** to game. While the federal EPA encourages states to stick to their monitoring schedules as strictly as possible, states have significant leeway, with every monitor typically allowed to miss up to 25% of its scheduled data during each quarter. Third, state governments have the **ability** to game. In many states, the same agencies that carry out monitoring also run advanced air quality forecasting – providing these agencies with the best data and forecasts of air quality in the near future. Despite these concurrent factors, the current system is not set up to detect strategic monitoring. Missing days are ignored by the federal regulators, implicitly assuming that pollution levels on monitored days are equal to pollution levels on

[2] Enck, Judith. Regional Administrator of the U.S. Environmental Protection Agency Region 2. Letter to Jeff Ruch, Executive Director of the Public Employees for Environmental Responsibility. February 28, 2014.

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01477

unmonitored days. This tolerance for gaps in compliance monitoring data may induce strategic timing in state and local agencies' self-monitoring activity.[3]

We propose an econometric framework that assesses whether air pollution monitors strategically shut down to avoid sampling on high-pollution days – specifically focused on identifying individual monitors whose pattern of shutdowns suggests gaming. Our framework has three components. The first component infers the government's *expectation* of high-pollution events through locally issued air quality alerts. These public advisories calling for citizens to reduce outdoor activities and vehicle use are often issued when forecasts predict that air pollution will exceed the Clean Air Act standards. We use an event study to assess whether a monitor's sampling rate falls when pollution alerts are in place. As a motivating example, our analysis begins with the sampling patterns from the $PM_{2.5}$ monitor at the Jersey City Firehouse (JCF). Our analysis focuses on the JCF monitor's *data capture rate*: the share of scheduled monitoring days in which the monitor produces readings. Analyzing 21 alerts sent by Jersey City from 2007 to 2014, we show that the data capture rate of the JCF monitor drops significantly during pollution alert weeks (declining by 10 percentage points from a mean of 88%) and especially during the alert day itself (declining by 28 percentage points). Though we do not directly address the reasons for the failure of the JCF air pollution monitor during the Bridgegate incident, our analysis indicates that the JCF monitor's sampling pattern over the seven-year period is consistent with strategic shutdowns during times of high pollution.

The second component of our framework incorporates simultaneous inference. We repeat the Jersey City Firehouse monitoring exercise to analyze 1,359 monitors that are set up to

[3] In Appendix A, we present a stylized model of self-monitoring that illustrates the challenges a regulator faces in eliciting complete and unbiased monitoring results from a regulated entity.

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01477

continuously sample air quality compliance for six different pollutants ($PM_{2.5}$, PM_{10}, O_3, NO_2, SO_2, and CO) throughout the contiguous United States. These monitors are located in 167 counties with similar pollution alert programs. Importantly, our task is not to estimate the response of the *average* monitor, but instead to pinpoint *which* monitors are gaming the regulatory design by excluding days likely to have high pollution levels. This inference problem poses two challenges, which we address with large-scale inference tools (Efron, 2012). First, for each individual monitor, the event-study test for strategic shutdowns likely uses a small sample due to the limited number of alerts and/or short time series for the monitor. Consequently, traditional inference comparing the test statistic with its theoretical (asymptotic) null distribution is likely invalid. We remedy this issue with a randomization inference scheme, which allows us to generate an *empirical* null distribution based upon "placebo" event studies that each use randomly dated pollution alerts (Rosenbaum, 2002). We then calculate *p*-values, for each monitor, as the proportion of the empirical null distribution that is more extreme than the observed effect. Second, by testing a large number of monitors, there is risk of overstating the confidence of rejection for any individual monitor. We address this risk in several ways, including an assessment of the *p*-value histogram (we find an overabundance of tests with small *p*-values; see e.g., Hung, O'Neill, Bauer and Kohne, 1997; Simonsohn, Nelson, and Simmons, 2014), a standard false discovery control strategy (Benjamini and Hochberg, 1995), and an "eye-ball" screening of monitors whose patterns of strategic missingness are the most visually apparent. Following these steps, we generate a list of "interesting" monitors whose distinctive monitoring patterns warrant further regulatory attention.[4] We post detailed estimation results for all monitors on a publicly available website. Together, these first

[4] We follow Efron (2012)'s language on the large-scale inference goal of detecting "interesting" units.

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01477

two steps of detection and inference offer researchers and policymakers the ability to narrow in on a potentially small subset of *gamers* within a much larger population of monitors/agencies.

The third component of our framework is economic characterization. We document key characteristics of these interesting monitors, and we shed light on underlying mechanisms for the patterns we see. We discover two primary features. First, we map the locations of the interesting monitors, and find 14 metro areas with clusters of interesting cases. Because the statistical procedure to determine interesting monitors does not use geographic proximity as an input, the fact that interesting cases cluster in specific regions suggests state- and/or local-government influences. Second, we use regression analysis to characterize counties with interesting monitoring patterns, and we find that a county's Clean Air Act compliance status plays a major role. For example, our state fixed effects regression suggests that being located in a noncompliant (nonattainment) county raises the probability a monitor is "interesting" by 64 percent, compared to other counties *within the same state*. Regressions with additional county-level characteristics, such as environmental friendliness, government size, and corruption, show limited explanatory power conditional on nonattainment status. Together, these test results support our hypothesis that strategic shutdowns arise from state and local governments' incentives to avoid or alleviate nonattainment penalties.

One possible way federal regulators could deter strategic shutdowns would be filling in missing monitoring data with values that better approximate the true air quality conditions, rather than omitting the missing days from records. We first build a $PM_{2.5}$ pollution dataset with imputed values based on inverse distance weighting (IDW), a spatial-averaging prediction method commonly used in the epidemiology and the economics literatures to infer air quality at an unmonitored location using available data from nearby monitors (e.g., Shepard, 1968; Schwartz,

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01477

2001; Currie and Neidell, 2005). We adapt this idea to our study context in which data are *temporally* incomplete; we impute a monitor's missing value on a given day by using the inverse distance-weighted average of data from a set of nearby "donor" monitors on that day. Because donor monitors that are closer to the monitor of interest are more heavily weighted, we use a liberal, 20-mile search windows for donor monitors. This allows the IDW to provide substantial coverage while still preserving local variations in pollution concentration. We find that the IDW is able to explain 81.4% of observed $PM_{2.5}$ variation and provide predictions for 38.6% of the missing values.[5] In a complementary exercise, we consider an alternative imputation method that uses newly available atmospheric modeling-based $PM_{2.5}$ products (Di et al., 2019) thanks to the increasing availability of satellite observations of air pollution. This second imputation is methodologically more complex, but is able to provide imputation values for *all* days. We use these imputed datasets to illustrate that among the aforementioned interesting monitors, the distribution of pollution on "unobserved" days exhibits a longer right tail – a pattern that replicates in both the IDW data and the modeling data. No such pattern is observed for non-interesting monitors, where the distribution of pollution across observed and unobserved days are indistinguishable from each other. In other words, although our quasi-experimental framework detects strategic monitors using a specific indicator – low levels of data capture rate around pollution alerts – these monitors turn out to the ones, and likely the *only* ones, that are generally strategic in sampling air quality. Had the measurements been taken for the interesting monitors, $PM_{2.5}$ levels would have exceeded the 15 ug/m^3 *annual* standard on 23% of these missing days and would have exceeded the 35 ug/m^3 *daily* standard on 2.7% of the unmonitored days. These

[5] The remaining missing observations are too far from non-missing monitors to use IDW with any confidence.

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01477

findings suggest strategic shutdowns could have misled federal compliance status designations. We hope our method may provide the regulator with a tractable route to assessing strategic shutdowns beyond the scope of this study – such as monitors located in areas without pollution alert programs.

We believe that the strategic self-monitoring problem highlighted in this paper is an underappreciated challenge for environmental compliance. We reported our findings to members of the federal EPA's ambient air quality monitoring group.[6] Officials with whom we spoke reacted that the shutdowns may be explained by local agencies' benevolent actions to prepare for incoming pollution episodes by taking the monitors offline and conducting maintenance.[7] In fact, we take away from the conversation that federal regulators tend not to worry about strategic responses in ambient air quality monitoring programs in which the entity of regulation is the state/local *government* – at least much less so than they would worry about point-source monitoring where the entity of regulation is often a *company*. The officials do agree with the importance of identifying interesting monitors. In their language, while these patterns do not necessarily suggest the local agencies are doing something "wrong", it is worth informing the corresponding agencies that their data look "different" from the data generated by others. We hope our analysis can raise awareness about monitoring and enforcement challenges associated with the tension between the imperative of national environmental protection and individual states' compliance incentives (Giles, 2020).

[6] We held a 1-hour meeting with a senior staff scientist and a statistician, both with expertise in ambient air quality monitoring and enforcement. Our discussion primarily focused on Figures 1A, 3, B.8, and B.9 of this paper.
[7] We believe these explanations do not fit the data. Our findings suggest that, if anything, such maintenance actions have caused the monitors to miss out the incoming pollution peaks.

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01477

The existing literature on environmental federalism emphases the role of decentralization on policy decisions such as inter-regional competition of environmental standards (e.g., Oates, 2001; Levinson, 2003; Millimet, 2014). Our work contributes to an emerging literature on monitoring and enforcement (e.g., Gray and Shimshack, 2011; Shimshack, 2014; Evans and Stafford, 2019) which emphasizes the fact that federalism in environmental legislations – and in many other regulatory contexts too as we mentioned at the beginning of this paper – often comes with the decentralization of the responsibility of monitoring and enforcement as well, creating potential principle-agent type of incentive misalignment in local agencies' self-monitoring activities. We believe this paper is among the first to examine selective monitoring as local agencies' strategy to help achieve environmental compliance. We corroborate an emerging literature that reveals strategic actions that contribute to an underrepresentation of high-pollution observations in states' self-monitored air quality data in the U.S. (Fowlie, Rubin, and Walker, 2019; Sullivan and Krupnick, 2019). Examples of strategic actions include states' decisions as of where to locate pollution monitoring sites (Grainger, Schreiber, and Chang, 2017) and where to locate polluters (Morehouse and Rubin, 2021); in a related paper, Zou (2021) presents evidence of strategic polluting suppression in places where pollution monitoring follows pre-scheduled on-and-off cycles (the effect of monitoring on strategic polluting behavior). Our paper presents the converse setting in which monitors that are scheduled to operate *continuously* choose to strategically shut down in response to expected high pollution events (the effect of pollution on strategic monitoring behavior).

Similar phenomenon has been observed in developing country settings as well, where local officials' desire to demonstrate air quality achievements has impaired truthfulness in pollution monitoring (Andrews, 2008; Chen, Jin, Kumar, and Shi, 2012; Duflo, Greenstone, Pande, and

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01477

Ryan, 2013; Duflo, Greenstone, Pande, and Ryan, 2018; Ghanem and Zhang, 2014; Karplus, Zhang, and Almond, 2018; Ghanem, Shen, and Zhang, 2020; Greenstone, He, Jia, and Liu, 2020; He, Zhang, Wang, 2020; Yang, 2020).[8] Our analysis is also inspired by Bennear, Jessoe, and Olmstead (2009) who showed that, in the context of drinking water regulation, local agencies undertook more testing to avoid the appearance of violating environmental standards.

Methodologically, we demonstrate that basic econometric tools can assist in making *individualized* conclusions about where strategic behavior occurs. The empirical literature typically identifies the average extent of an activity and subgroup heterogeneity, but rarely seek to provide insight into exactly where "interesting" behaviors may merit a closer look. Policymakers, on the other hand, often care more about actionable evidence than of broader characterizations of the extent of a problem. Our paper aims to provide concrete evidence for regulatory responses. On this front, we are related to recent development in the application of large-scale inference tools, where the research goal is to credibly detect a relatively small group of interesting units among a sea of null (Efron, 2012).[9]

[8] A related literature analyzes emission test cheating and collusion behavior in the vehicle sector (e.g., Oliva, 2015; Reynaert, 2020; Ale-Chilet et al., 2021; Reynaert and Sallee, 2021). Our paper is also related to the broader forensic economics literature that aims at detecting hidden, socially undesirable actions, often by modeling honest behavior and testing for deviations. We are grateful for Jay Shimshack who pointed us to this strand of literature. See Zitzewitz (2012) for a review.

[9] See applications in bioinformatics, such as high-throughput screening for drug discovery (Malo et al., 2006), and genomics/proteomics data analysis (Dudoit, Shaffer, and Boldrick, 2003). Within economic applications, we are most closely related to the literature on permutation inference (e.g., Barrios, Diamond, Imbens, and Kolesár, 2012; Buchmueller, Miller, and Vujicic, 2016; Young, 2016; Hagemann, 2019), multiple hypothesis testing (e.g., Anderson, 2008; Heckman et al., 2010; Finkelstein et al., 2012; Christensen and Miguel, 2018; Jones, Molitor, and Reif, 2019; List, Shaikh, and Xu, 2019; Kline and Walters, 2021; Kline, Rose, and Walters, 2022) and heterogeneous treatment effects estimation (e.g., Athey and Imbens, 2016; Chernozhukov, Demirer, Duflo, and Fernandez-Val, 2018; Davis and Heller, 2020).

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01477

2. Background and Data

2.1. Clean Air Act and Ambient Air Quality Monitoring

The National Ambient Air Quality Standards (NAAQS). The U.S. Clean Air Act (CAA) delegates the U.S. Environmental Protection Agency (EPA) to set up safety standards in the form of maximum concentration levels for outdoor air pollution. These are the National Ambient Air Quality Standards (NAAQS). Since the 1970s, the EPA has set up NAAQS for "criteria" air pollutants including particulate matter ($PM_{2.5}$ and PM_{10}), ozone (O_3), nitrogen dioxide (NO_2), sulfur dioxide (SO_2), lead (Pb), and carbon monoxide (CO). The CAA charges state governments with monitoring air quality within their own jurisdictions. The federal EPA uses states' submitted data to categorize counties into "attainment" (adhering to the standards) and "nonattainment" (violating the standards) groups. Most criteria pollutants have two standards: a 24-hour standard and an annual standard; ozone's standard is based upon an 8-hour period. For example, a county falls into $PM_{2.5}$ nonattainment if its three-year average for $PM_{2.5}$ exceeds 15 ug/m^3, and/or if the three-year average of annual 98th percentile concentration values exceeds 35 ug/m^3. The most updated NAAQS for all criteria pollutants are listed in the federal EPA's NAAQS Table (https://www.epa.gov/criteria-air-pollutants/naaqs-table).

Nonattainment counties face substantially elevated regulatory costs for both existing and prospective entities. The state is required to develop a State Implementation Plan (SIP) that details plant-specific regulations to bring the county back into compliance. These regulations typically involve the adoption of expensive pollution abatement technologies and emission limits on existing factories. Factories planning new production capacity in nonattainment jurisdictions must adopt technologies with the "lowest achievable emission rate," irrespective of the cost of doing

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01477

so.[10] Local governments and individual polluters occasionally receive direct penalties from the EPA in cases of sustained nonattainment. The NAAQS provision functions as the CAA's ultimate safeguard for outdoor air quality. Its regulatory incentives for the state economy – with respect to the compliance costs, firms' productivity changes, and labor market implications – have been widely documented in the literature (e.g., Greenstone, List, and Syverson, 2012; Walker, 2013; Blundell, Gowrisankaran, and Langer, 2018; Shapiro and Walker, 2020). A separate strand of literature finds evidence that by directing regulatory resources toward sources in high-pollution areas, local governments have been able to achieve localized air quality improvements near the violating monitors (e.g., Bento, Freedman, and Lang, 2015; Auffhammer, Bento, and Lowe, 2019).

EPA Rules for Incomplete Monitoring. To demonstrate compliance with NAAQS, states' monitoring data must satisfy completeness goals. Appendix Figure B.1 tabulates the EPA's completeness goals for each of the criteria pollutants (U.S. EPA, 2013). The typical requirement is for each monitor to take at least 75% of required samples per quarter of the year. What happens if monitoring data fall below the completeness goals? In principle, incomplete data cannot be used to demonstrate compliance, and the areas is thus designated as "unclassifiable." In practice, an unclassifiable county is treated just as an attainment county. However, if statistics computed from incomplete data suggest a potential violation of NAAQS, then the EPA can invoke rights to assign "nonattainment" status using limited data available. For example, in the case of $PM_{2.5}$ monitors, only 11 days of observations per quarter are needed for the EPA to designate violation – if, for example, the average of the available observations exceeds the annual standard of 15 ug/m^3. If the

[10] Lowest Achievable Emission Rate, or LAER, refers to technologies that achieve the lowest possible emission rate in practice without cost consideration. In contrast, new sources in attainment jurisdictions comply with the Best Available Control Technology, which is often much less strict and allows for considerations of energy, environmental, and economic impacts and other costs.

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01477

monitor collects even fewer than 11 samples per quarter, the CAA gives the EPA the right to use alternative data. The federal regulation states that the EPA administrator "*may consider factors such as monitoring site closures/moves, monitoring diligence, the consistency and levels of the daily values that are available, and nearby concentrations*" in determining attainment / nonattainment status.[11]

These rules imply that the completeness goal *per se* is not subject to gaming. A violating area cannot bring itself out of nonattainment simply by reducing its data capture rate below 75% per quarter because nonattainment can be designated using very limited data (11 observations); for a non-violating area, it makes little difference if its quarterly capture rate is above the 75% level (attainment) or below (unclassifiable). However, strategic responses *can* arise when local monitoring agencies skip high-pollution days to water down the average (or whatever relevant statistics) of captured pollution, which is the focus of this study. Appendix B provides additional information for interested readers on various forms of missing pollution data and how they may be linked to strategic factors.

2.2. Pollution Alerts

Pollution alerts are based on air quality forecasts made by state and local agencies using chemistry transport models, such as the Community Multiscale Air Quality Modeling System (CMAQ).[12] An alert is issued when unfavorable local weather events (thermal inversions, light winds, high pressure zones, etc.) and emission events (traffic congestion, wildfires, etc.) are

[11] 40 C.F.R. Appendix N to Part 50 - Interpretation of the National Ambient Air Quality Standards for PM2.5.
[12] https://www.epa.gov/cmaq/cmaq-models-0

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01477

expected to push air pollution to unhealthy levels as defined by the NAAQS nonattainment standards.[13]

Appendix Figure B.3 shows the distribution of the forecasted Air Quality Index associated with the alerts. The distribution exhibits substantial pileup at the AQI cutoff of 100 (at which point the AQI code moves from "Moderate" to "Unhealthy for Sensitive Groups") and the cutoff of 150 (at which point the AQI code becomes "Unhealthy"). We are unaware of any institutional reasons for a *mechanical* link between alerts and missing monitoring data. To the extent that a forecasting algorithm uses monitoring data as predictors for future pollution, lower data capture on higher pollution days, if anything, would *decrease* the odds of pollution alerts, generating a *positive* correlation between capture rate and alerts. Alerts are associated with changes in general atmospheric conditions, such as temperature and precipitation, which could influence monitors' data capture due to equipment and/or staff performance. However, such mechanical association would affect all monitors, and it should not be specific to "interesting" monitor groups. Appendix B offers a more detailed discussion on this point.

2.3. Data

Pollution monitoring data come from the EPA's Air Quality System (AQS) database. We use AQS Daily Summary Data which contain information from every monitor for each day from 2004 to 2015. A daily summary record is the aggregate of all sub-daily measurements, typically 24 hourly samples taken by the monitor. Our primary variable of interest is a monitor-by-day level indicator for missing data, i.e., none of the sub-daily measurements being available.

[13] Pollution alerts are often salient. Previous research has shown that alerts suppress outdoor activities and influence transportation choices (Neidell, 2009; Cutter and Neidell, 2009; Graff Zivin and Neidell, 2009).

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01477

We obtain air pollution alerts data through the EPA AirNow (airnow.gov) Action Day Program. Action Day provides a tracker of all air quality alert programs implemented by state and local agencies.[14] The database we use contains a total of 33,357 pollution alerts issued by 342 jurisdictions between 2004 and 2015. An advisory is often issued one day ahead of the actual alert day. We use the alert day itself to define the timing of pollution alert events.

3. Framework and Evidence

This section describes the three components of our framework and presents our findings. We begin in Section 4.1 with the Jersey City Firehouse (JCF) $PM_{2.5}$ monitor and explain how we test for strategic shutdowns for a single monitor. Section 4.2 describes the simultaneous testing problem where we scale up the exercise in Section 4.1 to all 1,359 monitors. Section 4.3 presents an econometric analysis of the characteristics of monitors that are deemed "interesting" by the testing process.

3.1. Test of Individual Monitor: The Jersey City Firehouse Monitor as an Example

Using an event study framework, we model the JCF monitor's "capture rate" of $PM_{2.5}$ data – an indicator variable equaling "1" when scheduled monitoring occurs – around the timing of pollution alerts (the "events"). There are a total of 37 pollution alert days in Jersey City during our study period. Alerts are sometimes issued for several consecutive days, in which case we keep the first day of the episode to avoid overlapping windows and focus on the alert issuance effect. This leaves us with 21 pollution alert *events*. For each alert event, we pull the JCF monitor's operational

[14] For example, this includes the "Spare the Air" program in the Bay Area of California (https://www.sparetheair.org/), and the "High Pollution Advisory" program managed by the state of Arizona (https://ein.az.gov/keywords/high-pollution-advisory). A full list of programs contained in the database is here: https://www.airnow.gov/aqi/action-days/.

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01477

status 30 days before and 30 days after the alert day, forming an event study dataset with 1,281 observations (21 alert events multiplied by the 61-day event study window for each alert). The estimation equation is:

$$\text{Capture Rate}_t = 1 - \mathbb{I}(\text{Missing PM}_{2.5}\text{ Data})_t = \sum_{\tau\in[-30,30]} \beta_\tau \cdot \mathbb{I}(t=\tau) + \varepsilon_t \quad (1)$$

where $\mathbb{I}(\cdot)$ represents the indicator function. Note that the $\hat{\beta}_\tau$ estimates are simply the capture rate τ-day relative to the pollution alert day, averaged across all 21 events.

Our goal is to assess whether the $\hat{\beta}_\tau$'s have lower values around $\tau = 0$, i.e., a lower capture rate (more missing values) near the time when a pollution alert is issued. We specify a donut difference-in-means estimator as our test statistic:

$$T = \frac{1}{7}\sum_{\tau\in[-3,3]} \hat{\beta}_\tau - \frac{1}{40}\sum_{\substack{\tau\in[-30,-11]\\ \cup[11,30]}} \hat{\beta}_\tau \quad (2)$$

which is the average capture rate within a seven-day event window around the pollution alert day, subtracted by the average capture rate outside that window, with a seven-day buffer on each side of the event window. In Appendix Figure B.5, we show that our findings are insensitive to alternative choices of the buffer window.[15] Our identification premise is the standard zero-trend

[15] Another potential issue with the simple difference-in-means test statistic is that it may falsely categorize monitors as strategic when a long-term shutdown occurs near the pollution alert day. For example, imagine an always-active monitor that shuts down at $\tau = 0$ and remains inactive for a month. From equation (1), this event is associated with a test statistic of -7.1 percentage points. In practice, we will find that such possibility is only relevant with carbon monoxide (CO) monitors that often experience seasonal shutdowns (Section 4.2). We will show that a slightly sharpened version of the test statistic

$$T = \max\left(\frac{1}{7}\sum_{\tau\in[-3,3]} \hat{\beta}_\tau - \frac{1}{20}\sum_{\tau\in[-30,11]} \hat{\beta}_\tau \;,\; \frac{1}{7}\sum_{\tau\in[-3,3]} \hat{\beta}_\tau - \frac{1}{20}\sum_{\tau\in[11,30]} \hat{\beta}_\tau\right)$$

can successfully detect strategic monitors in the presence of seasonal shutdowns, as the null hypothesis is rejected when the capture rate around the alert day is lower than that of *both* the pre-alert period and the post-alert period.

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01477

assumption: under the null hypothesis that the pollution alert has no impact on the capture rate, we should have $T = 0$, and, alternatively, if alerts do affect the capture rate, we expect $T \neq 0$. One departure from the standard event study framework is that our treatment (pollution alert) reflects the monitoring agency's belief about future pollution, and thus the shutdown of monitors may well occur before the issuance of pollution alerts. Such anticipation underlies our choice to allow the test statistic to capture potential change in shutdown rates several days before the actual alert day.

Figure 1, panel A plots the $\hat{\beta}_\tau$ coefficients for the JCF monitor. The graph features a clear drop of the monitor's capture rate around the pollution alert day. The corresponding T estimate is -0.101, meaning the capture rate within the seven-day window around a pollution alert is 10.1 percentage points lower than the outside-window average of 88 percent (an 11.5% reduction). Note that the largest change in the monitor's capture rate occurs on the alert day and the day before, with a 28.6 percentage points reduction (a 32.5% reduction).

An important feature of Figure 1, panel A is that the decline in the data capture rate began days before the actual pollution alert day. This is a pattern that we repeatedly observe among "interesting" monitors. Note that pollution not only increases on the actual alert day, but that it tends to rise leading up to the alert-day peak. Thus, the capture rate pattern is likely a consequence of forward-looking agencies changing monitoring effort in anticipation of a future high-pollution *episode* (Malani and Reif, 2015). This pattern also suggests data absences are not mechanically linked with alert issuances (Section 2.2), in which case one would expect to see a change in capture on the alert day only.

We are now ready to conduct inference on whether T is statistically different from zero. In a large-sample setting, we could implement a *t*-test of $T = 0$ via an OLS regression of Capture Rate_t on an indicator for the seven-day window around the alert day. This approach has

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01477

several limitations in our setting. First, it relies on the distributional assumption that the *t*-statistic of T under the null hypothesis will be normally distributed N(0,1), which may not be true in our finite-sample setting. Second, with a small sample, the magnitude and precision of T could be sensitive to specification choices. Therefore, rather than relying on distributional and specification assumptions, we build on the idea that, under the (sharp) null hypothesis that pollution alerts have no effects whatsoever on the monitor's capture rate, variable T does not depend on whether a pollution alert occurs; therefore, we can generate the null distribution of T from the data by randomly shuffling the timing of pollution alerts. In practice, we assign 21 dates as the "placebo" alert days. We restrict the randomization so that the placebo day does not occur within one month of the true alert day. In Appendix Figure B.5, we report that our results are robust to using alternative randomization buffers such as 15 days or 7 days. We repeat the process 1,000 times, each iteration generating a placebo test statistic. We then compute a two-tail *p*-value of the observed T as the proportion of the null distribution that is more extreme (in absolute value) than T. Note that we employ two-tail testing, allowing T to be significant for the "wrong" sign. In Section 4.2, we will show this "wrong" tail provides us with an opportunity for sanity checks.

Figure 1, panel B reports the inference exercise. The histogram plots the empirical null distribution of T across 1,000 randomization of pollution alerts. The vertical solid line marks the true estimate which lies outside of the 95 percent range of the null distribution, with a two-tail *p*-value of 0.014. Evidence thus points to a statistically significant reduction of the JCF monitor's capture rate around pollution alerts.

Causal Interpretation and "Strategic" Shutdowns. Before proceeding, we discuss the causal interpretation of the T estimate. Taken at face value, patterns of Figure 1 suggest evidence of *selective* shutdowns. That is, more missing data are occurring around pollution alerts with

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01477

significant deterioration of air quality. Note that selective shutdown *per se* is an undesirable feature of monitoring data that is worth documenting: if missing rate is differentially higher around high-pollution alerts, the resulting monitoring data will understate the true pollution concentration. Improving continuity of monitoring near these pollution events will thus increase accuracy of the monitoring data regardless of *why* such selective shutdowns were occurring.

But to interpret such selective shutdowns as *strategic* behavior – a term that implies intentionality – we need the assumption to hold that a local government's expectation of bad air quality *causes* the reduction in monitor's capture rate. In other words, the strategic interpretation relies on the identification assumption that there will be no changes in the monitor's capture rate in the absence of pollution expectation. Here we discuss two concerns for potential violation of this identification assumption.

A first concern is selection. Equation (1) may be mis-specified if monitors' capture rates and pollution alerts are both correlated with some unobserved factors. We note that the permutation inference should purge the influence of unobserved factors except for those that are systematically correlated with the timing of alert issuance. Moreover, if systemic omitted variable bias exists, it likely applies broadly to many monitors. In contrast, we report in Section 4.3 that strategic shutdowns tend to occur in regions with higher risk of violating the clean air standards. Section 4.3 also shows that interesting monitors are also more likely to miss monitoring during bad pollution years in general, not just around pollution alerts.

We also note that we have little prior reasons to expect any "spontaneous" relationship between monitors' sampling rate and socioeconomic/atmospheric conditions. As we mentioned in Section 2.1, monitoring techniques certified by the EPA have stringent technological standards and can robustly operate under various meteorological and pollution conditions. The periodic

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01477

quality control procedures we described in Section 2.1 are also precisely designed to make sure monitors are functioning properly. Nevertheless, in order to assess this point more directly, we consider an exercise that tries to predict daily monitoring missingness using weather conditions. Weather is a candidate for confounding that could affect both when a monitor misses observations and when local agencies issue air alerts (e.g., if bad weather events influence the functioning of monitoring devices and, at the same time, affect polluting activities such as road traffic). For this confounding to occur, some function of weather must predict missing monitoring data. We train several flexible machine-learning (ML) models that use contemporaneous and lagged weather data to predict whether monitors missed planned observations.[16] None of the weather-based ML models successfully improve upon a "null model" that predicts a region's majority class ("not missing"). In fact, the models functionally ignore the weather inputs and replicate the null model – always predicting "not missing." This fact remains true even when we oversample missing days to control for missingness days' relative infrequency.

A deeper concern is reverse causality. Because pollution forecasts such as the CMAQ use contemporaneous monitoring data as input, one might be concerned that the natural (non-strategic) absence of monitors' data may interact with the issuance of pollution alerts in ways that could generate a reversely causal relationship between monitor shutdowns and alerts. This is unlikely

[16] Specifically, we use three different ML algorithms: (i) lasso-penalized OLS regression, (ii) lasso-penalized logistic regression, and (iii) random forest. The outcome for each model is a binary indicator for whether the monitor-day's observation is missing. The predictors include contemporaneous and lagged weather features—temperature (mean, minimum, and maximum), precipitation, dew point, pressure, visibility, wind speed, wind gust, and indicators for extreme weather events. We tune the models' hyperparameters using 5-fold cross validation and ultimately assess performance on final, held out test set. The daily weather data come from NOAA's Global Summary of the Day (GSOD) 2005-2014. We use inverse-distance weighting to estimate each monitor-day's weather based upon the monitor's distance to each of the 4,579 NOAA weather stations in the GSOD data.

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01477

the case. Note that if missing data occur randomly, then the distribution of missing pollution data should mirror that of the observed data. Thus, natural missingness should not affect forecasts or alert issuances. On the other hand, if the data capture rate does fall on high pollution days (but for reasons unrelated to pollution alerts), then one would expect a *decrease* in the odds of pollution alerts because fewer high-pollution days are being captured. This thus creates a *positive* relationship between capture rates and alerts (i.e., the lower the capture rates → the smaller likelihood for alert issuance) instead of a negative one that we find in the data.

3.2. Simultaneous Test of All Monitors

We now repeat the exercise in Figure 1 with all other monitors. We make the following sample restrictions: First, we restrict to monitors located in counties that have issued at least two pollution alerts during our study period. Second, we restrict to monitors that are designated to sample air quality every day. For $PM_{2.5}$ and PM_{10} monitors, this means restricting to monitors sampling on a "1-in-1-day" basis.[17] For O_3, NO_2, SO_2, and CO monitors, seasonal monitoring is often practiced (e.g., ozone is often deemed a problem only during the summer months), and we restrict to monitor-months for which at least one day of monitoring data was available.[18] Our final pool of tests includes 1,359 pollution monitors (including the Jersey City Firehouse monitor) for

[17] Particulate pollution monitoring is often done intermittently (once every three or six days) for sites that still adopt manual sampling technologies. Intermittent monitoring is typically allowed by the Environmental Protection Agency in jurisdictions that are not in immediate danger of violating the NAAQS. We identify 1-in-1-day monitors using the Air Quality System database's "required day count" field. We do not test lead (Pb) particulate monitors because virtually all lead monitors follow an intermittent monitoring schedule.

[18] Gaseous pollutant monitoring uses chemiluminescent technologies, and are by default conducted continuously. Our sample selection primarily reflects the fact that monitoring seasons may differ across monitors.

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01477

$PM_{2.5}$, PM_{10}, O_3, NO_2, SO_2, and CO located in 167 counties operating between 2004 and 2015. We begin with a collection of null hypotheses that we test at once:

$$\{H_i: \text{Monitor i's operation schedule is not affected by pollution alerts}\}_{i=1}^{N}$$

and the corresponding mean-difference test statistics $\{T_i\}_{i=1}^{N}$ analogously defined as in equation (1). For each monitor, we use randomization inference to obtain its two-tail *p*-value p_i measuring the degree to which the observed T_i contradicts H_i.

We next turn to the simultaneous testing problem. At any given chosen rejection threshold α, the test will falsely reject the null approximately $100\alpha\%$ of the time. With a large number of hypotheses, a substantial number of monitors will be falsely considered to be "gaming." We introduce several measures to approach this issue.

First, we present the *p*-value histogram. By construction, the *p*-value histogram should feature a uniform distribution U(0,1) if the null hypothesis holds true (i.e., alerts have no effect on data availability) for every monitor i. In practice, the histogram of $\{p_i\}_{i=1}^{N}$ is potentially a mixture of cases where the null hypothesis is true and cases where the null is false. If enough monitors are gaming, one would expect a deviation from U(0,1); more specifically, the *p*-value histogram would exhibit an overabundance of small *p*-values (<0.05). Figure 2, panel A presents the *p*-value histogram. We see a clear spike in small *p*-values in the p<0.05 range. When test statistics are further partitioned into $T_i \leq 0$ ("correct"-signed test statistic) and $T_i > 0$ ("wrong"-signed test statistic) groups, we find that the spike in small *p*-values are driven by tests with the "correct" signs, i.e., those with drops in the capture rate, rather than increases around pollution alerts (Figure 2, panel B). Figure 2 also shows that significant cases tend to emerge at the smallest *p*-values. This pattern may be consistent with (a) strategic behavior being concentrated with extreme cases rather

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01477

than a large number of monitors being "slightly" strategic, and (b) the distribution of *p*-values under the alternative hypothesis is steeply right-skewed with high statistical power (Hung, O'Neill, Bauer and Kohne, 1997).

Second, we employ the Benjamini-Hochberg procedure to control for false discovery rates (Benjamini and Hochberg, 1995). This method is closely related to the *p*-value histogram. Large *p*-values on the *p*-value histogram mostly represent observations from the null hypothesis, and thus can be used to estimate the proportion of small *p*-values that also come from the null hypothesis. More formally, we order $\{p_i\}_{i=1}^{N}$ in an increasing order $p_{(1)} \leq \cdots \leq p_{(N)}$, and for a choice of target false discovery rate $\alpha = 0.05$, we find the largest value of k such that $p_{(k)} \leq \alpha k/N$, and reject the null for $i = 1, \ldots, k$. For each T_i, we also compute a *q*-value equals to the minimum false discovery rate that can be attained when T_i is considered significant (Storey, 2003; Anderson 2008). We follow the literature and give *q*-value a Bayesian posterior significance level interpretation (i.e., false discovery adjusted significance level).

In Appendix Figure B.6, we present a placebo exercise in which we *randomly* assign a placebo (fake) alert profile to each monitor and then replicate the main analysis steps. As anticipated, the distribution closely resembles U(0,1) – the theoretical distribution of *p*-values under the null hypothesis that alerts have no effect on monitor data capture rates. In this case, the Benjamini and Hochberg procedure correctly indicates that there are no monitors of interest after adjustment for multiple comparisons (minimum false discovery rate adjusted *q*-value = 0.544).

Finally, recent applied econometrics has demoted sole reliance on *p*- or *q*-values and promoted weights on the degree to which the data patterns are visually compelling. In our case, because monitoring agencies have no incentive to pull capture rates way down (Section 2.1), one would expect that interesting monitors would exhibit a T-shaped response where otherwise stable

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01477

monitoring operation shows a sharp drop just around the days of high pollution episodes. To operationalize this test, we aggregate the $\hat{\beta}_\tau$'s estimates and compute average event study patterns separately for two groups: interesting monitors and other monitors. Figure 3 displays the findings. First, except for the case of PM_{10}, the overabundance of small *p*-values is apparent for each type of monitor. Second, non-interesting monitors show a flat and stable operation pattern around pollution alerts; this suggests the average monitor is not strategically shutting down around pollution alerts. Third, except for the case of CO, visual evidence is strong for interesting monitors, with a sharp but transient drop in capture rates around pollution alert days. As we noted, the L-shaped response for the CO monitors is likely driven by seasonal monitor shutdowns that are picked up by the simple difference-in-means test statistic of equation (2). In Appendix Figure B.7, we present results using a "sharpened" version of the test statistic with which the null hypothesis is rejected when the capture rate during the -3 to 3 day event window is lower than that of *both* the pre- and post-alert periods (Section 4.1). We find that this approach successfully identifies the T-shaped pattern for CO monitors, while the event-time patterns for other pollutants remain almost the same. For the rest of the paper, we choose to stick with the simpler test statistic as specified in equation (2).[19]

Going one step further, we manually screen the event study patterns among all interesting monitors, and pinpoint those with "very interesting", T-shaped pattern. Appendix Figure B.8 presents one example for each type of pollution monitor. Of course, this visual screening process

[19] Note that the patterns observed in Figure 3 are not necessarily mechanical. Due to two-sided testing, one could, in principle, observe an increase in capture rate around alerts for the "interesting" monitors, or an overall flat curve that represents a composite of dips and jumps. However, Figure 3 predominantly shows dips for the "interesting" monitors. It should be noted that Figure 3 is, to some extent, a repetition of Figure 2, Panel B, but it allows us to examine what the event-time patterns look like.

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01477

is subjective and hence we do not use the "very interesting" status in any of the subsequent statistical analyses. We do note, however, that visual screening is likely the most directly accessible approach to regulators and practitioners in our context. We have made our estimation results for all monitors publicly available on a website. Figure 4 provides an illustration. The interactive map presents all tested areas, interesting monitors, very interesting monitors, and other tested monitors. For each monitor, we report the test statistic, the *p*- and *q*-values, and a link to the event study graph.

3.3. Features of "Interesting" Monitors

The statistical procedure in the previous two sections generates a list of monitors whose patterns of missing data are consistent with strategic shutdowns. In this subsection, we present two exercises that document characteristics of these interesting monitors that speaks to underlying mechanisms.

Location. Appendix Table B.1 tabulates total number of pollution alerts, tested monitors, interesting monitors, and very interesting monitors by all 54 Core Based Statistical Areas (CBSAs) in our data. We find that interesting monitors tend to cluster in certain regions of the country. For example, among the 86 pollution monitors that we test in the Phoenix-Mesa-Scottsdale metro area in Arizona, 23 show up as interesting. Appendix Figure B.9 maps the locations of monitors in the 14 CBSAs that in total house 60% of all these interesting monitors. The clustering pattern is not an artifact of some CBSAs simply having disproportionately more monitors. Several large metro areas we examined – such as Chicago-Naperville-Elgin (IL-IN-WI), Sacramento-Roseville-Arden-Arcade (CA), and Philadelphia-Camden-Wilmington (PA-NJ-DE-MD) – have many monitors but very few interesting cases. Because the statistical procedure we use to determine

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01477

interesting monitors does not use geographic proximity as an input, the fact that interesting cases cluster in certain places is informative, and suggests regional government influences.

The clustering pattern also implies that the decision to strategically monitor is spatially correlated, and some of the local variation in monitors' interesting/non-interesting status is due to the use of a sharp statistical decision criterion (i.e., monitor is interesting if its *p*-value is less than 0.05). For example, among all non-interesting $PM_{2.5}$ monitors within 20 miles of interesting $PM_{2.5}$ monitors, over 18% have permutation *p*-values between 0.05 and 0.15. This fraction is 7% and 3% for non-interesting monitors within 20-50 mile and 50-100 mile distance, respectively. Put differently, some non-interesting monitors in fact do exhibit strategic monitoring patterns like their interesting neighbors, and they would have been considered as interesting if we were to use a less conservative decision rule in hypothesis testing.

County Characteristics. A key premise of our analysis is that state and local governments avoid sampling high-pollution days in an effort to either avoid nonattainment status of the federal air quality standards (NAAQS) or, in the case of counties already in violation, to move out of nonattainment. We now use cross-sectional regressions to test whether being located in counties currently in or with a history of NAAQS nonattainment in fact increased a monitor's likelihood to operate strategically. Table 1, column 1 reports a simple linear regression of an indicator variable for being labeled interesting (*p*-value ≤ 0.05) on an indicator for the NAAQS nonattainment status of the county in which the monitor is located. This is a cross-sectional regression with 1,359 underlying monitors, 11.7% of which are interesting cases. Our estimate suggests a county's nonattainment is associated with a 6.6 percentage point increase (or a 6.6/11.7=56 percent increase) in the odds of the monitor being interesting. In column 2, we repeat the "correct-vs-wrong sign" breakdown, finding that the nonattainment correlation is driven by cases with the "correct" sign

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01477

(i.e., the capture rate *decreases* around pollution alerts). In column 3, we further control for several state-level regulatory/political characteristics including party affiliation,[20] an index for environmental friendliness,[21] government size,[22] and a proxy for corruption.[23] We find that nonattainment is still a predominant predictor for monitor's "interesting" status. In column 4, we include state fixed effects, comparing monitors within the same state but locating in attainment versus nonattainment counties, thus purging of the influence of any observable or unobservable characteristics that might differ across states. The results again indicate a robust role of nonattainment status. In columns 5-8, we repeat the same set of regressions now using the FDR-adjusted significance, i.e., an indicator variable for q-value ≤ 0.05, as the dependent variable. We obtain similar results from these alternative specifications.

While we have focused on the importance of nonattainment *history*, a monitor's strategic incentive may rise as pollution levels approximate, but not yet exceed, the regulatory standards. Figure 5 provides evidence on the role of such nonattainment *risk* using $PM_{2.5}$ monitors as an example. The chart documents the relationship between a monitor's quarterly capture rate (number of days with creditable sample as a fraction of required days of sampling) and 1-unit bins of annual $PM_{2.5}$ concentration (the "design value") where an exceedance of 15 ug/m^3 corresponds to a higher risk of violation. The underlying regression controls for monitor fixed effects and year fixed effects, so that the underlying variation comes from year-of-year changes in recorded pollution levels

[20] Share of Democratic Party affiliation according to 2006 Gallup Pool.
[21] League of Conservation Voters score, which is based on state representatives' voting records on environmental issues. A higher score indicates to a stronger environmental preference (Dietz et al., 2015).
[22] Government-sector (two-digit NAICS: 92) employment as a share of total employment. Data are sourced from the Bureau of Economic Analysis.
[23] Per capita number of federal convictions among state and local public officials. Data are sourced from the Report to Congress on the Activities and Operations of the Public Integrity Section (Glaeser and Saks, 2006; Leeson and Sobel, 2008; Grooms, 2015).

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01477

within the same monitor. Panel A reports that for interesting monitors, years with high levels of pollution correspond to low data capture rates. Panel B shows that for the other, non-interesting monitors, the capture rate is largely independent of the design value.

Two important messages emerge from this analysis. First, strategic monitoring may occur beyond situations that involve pollution alerts. Figure 5 panel A shows that, for interesting monitors, capture rates are *generally* lower when the monitor is closer to noncompliance. Second, while the identification of interesting monitors is based on the specific context of pollution alerts, the fact that Figure 5 panel B shows precise zero response from non-interesting monitors suggest the detection framework is able to capture most, if not all, $PM_{2.5}$ monitors that exhibit strategic monitoring. We will see a similar set of results in Section 5.1, where we use $PM_{2.5}$ imputation data to show that a deviation of pollution distribution on unobserved days from observed days exists *only* among the interesting monitors.

4. Missing Data Imputation

We begin by estimating what the distribution of pollution readings would have been had monitoring been done on the missing days. The **first** imputation method we use is a simple and transparent prediction procedure known as the inverse distance weighting (IDW). The IDW builds on the idea that atmospheric conditions such as air pollution are often spatially correlated. The approach predicts the pollution in a given location as the average of readings from nearby "donor" monitors; each donor reading is weighted by the inverse of the donor monitor's distance to the location of interest.[24] Because donor values that are closer to the monitor of interest are more

[24] Formally, at any given point in time, the IDW pollution imputation for a monitor x given a set of nearby donor monitors $\{x_i\}_{i=1}^{N}$ is $x = \frac{\sum_{i=1}^{N}[d(x,x_i)]^{-1}x_i}{\sum_{i=1}^{N}[d(x,x_i)]^{-1}}$ where $d(x, x_i)$ is the distance between the monitor of interest and the donor monitor i. The IDW is commonly used in epidemiology and

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01477

heavily weighted, we use a liberal, 20-mile search window for donor monitors, which allows the IDW to provide substantial coverage while still preserving local variations in pollution concentration. Note that IDW "imputation" can be done even if x is *not* missing, given us an opportunity to conduct in-sample validation.

A disadvantage of IDW is that it only works when at least one donor monitor exists within 20-mile radius to the monitor of interest. In our data, IDW provides imputation values for 38.6% of missing data. Further, because strategic monitoring behavior exhibits spatial clustering (Appendix Figure B.9), one might worry the availability of donor monitors' imputation data *per se* is endogenous. We therefore also consider a **second** imputation method that relies upon machine-learned predictions of $PM_{2.5}$ from Di et al. (2019) who provide predictions of daily $PM_{2.5}$ concentrations for the contiguous United States on a grid of approximately 1km by 1km resolution. These predictions result from machine-learning algorithms trained on more than 100 variables that should be predictive of ground-level $PM_{2.5}$, including satellite-measurements of aerosol optical depth, simulation outputs from two chemical-transport models, meteorological data, physical variables like elevation, and land-use data (e.g., road density and industry). A pro of this second method is that it provides imputation values for *all* missing observations as the atmospheric modeling covers all place and time; a con is that the data result from complex modeling, thus being relatively less transparent than simple IDW. Below we present findings using these two approaches side by side.

environmental economics studies to improve spatial coverage of data as ground monitoring of weather and pollution is often sparse (e.g., Schwartz, 2001; Currie and Neidell, 2005). Here we adapt the same idea to the context where data are temporally incomplete.

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01477

Figure 6 presents results from the imputation exercise. Figure 6A presents data from the IDW method, while Figure 6B repeats the exact same exercises using the atmospheric modeling method. There are four panels. Each panel displays three distributions: observed $PM_{2.5}$ (of course, for when monitoring is not missing), predicted $PM_{2.5}$ when monitoring is *not* missing, and predicted $PM_{2.5}$ when monitoring is missing. Hence, the two dashed lines tell us how closely the predicted $PM_{2.5}$ tracks observed $PM_{2.5}$ levels, and the solid line indicates what the distribution of $PM_{2.5}$ would have looked like had monitoring been done on the missing days.

Take Figure 6A, left panel that summarizes data for interesting monitors using the IDW method. First, we find that IDW does a reasonable job predicting actual $PM_{2.5}$ when pollution monitoring is not missing. A simple linear regression of observed $PM_{2.5}$ on predicted $PM_{2.5}$ yields an R-squared of 0.814. Second, our prediction exercise suggests that, compared to observed $PM_{2.5}$, the distribution of "missed" $PM_{2.5}$ (solid line) features a longer right-tail. About 23.1% of the missing days would have shown $PM_{2.5}$ exceeding 15 ug/m^3 had the measurements been taken, and about 2.7% of the missing days would have exceeded 35 ug/m^3. These fractions convert to about 6.6 days per year of annual standard exceedance and 0.8 days per year of 24-hour standard exceedance. Figure 6A, right panel shows no such discrepancy between observed and missed $PM_{2.5}$ exists for non-interesting monitors. More broadly, we hope the IDW provides the regulator with a tractable tool to assess strategic shutdowns beyond the scope of this study. Evidence of Figure 6A reveals the difference in interesting monitors' $PM_{2.5}$ distributions on observed days and missed days. The fact that we find the *same* group of monitors that respond to pollution alerts also exhibit a distribution-wide, selective pattern in the timing of absent data, suggesting that strategic monitoring goes beyond just the context of pollution alerts. Figure 6B shows that the same

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01477

patterns replicate almost exactly using the modeling data method. In Appendix B, we further discuss potential health costs associated with strategic monitoring shutdowns.

Another important feature of Figure 6 is that the distributional deviation of "missed" $PM_{2.5}$ manifests only for the interesting monitors and not for other monitors (which include non-interesting monitors and untested monitors in counties that do not have alert programs). In other words, although our quasi-experimental framework detects strategic monitors using a specific indicator – low levels of data capture rate around pollution alerts – these monitors turn out to the ones, and likely the only ones, that are *generally* strategic in sampling air quality.

5. Conclusion

We investigate the operating patterns of air quality monitors in the U.S. using a framework that we create to test for evidence of shutdowns that are strategically timed to avoid periods when forecasts predict high levels of pollution. We identify clusters of monitors in at least 14 metropolitan areas whose patterns of operation show strong evidence of the use of such strategic timing and, thus, warrant further regulatory attention. (We make the list of these monitors available at a public website.) Our findings show that the monitors that display such operating patterns are predominantly located in federal nonattainment areas that face the likelihood of costly penalties for violations of US Clean Air Act standards.

Our work suggests that current regulatory practices that ignore gaps in compliance-monitoring data collection may incentivize strategic changes in local agencies' monitoring diligence. We propose two key ways to deter such behavior: detection and incentives. The statistical framework we have devised could detect monitors that show a pattern of skipping high-pollution days. We also suggest ways to disincentivize the use of strategic shutdowns. Regulators

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01477

could consider revising the current practice of ignoring missing data when they determine compliance. For example, inverse distance weighting, a method used successfully by the research community, is one possible solution to provide imputed values. Imputation methods are imperfect, but their output may act as a trigger for regulatory investigation, and may thus serve as reasonable deterrence to strategic shutdowns.

We believe that concrete evidence can help level the playing field for environmental regulations, improve accuracy of air quality data, and motivate better design of monitoring and enforcement schemes in the future to better achieve the wider aims of improved public health from having less-polluted air. More broadly, we hope that the new possibilities made possible by large-scale inference tools can extend to other research contexts where the detection of a small group of units that evidence distinct patterns (among a sea of nulls) is important.

References

Alé-Chilet, Jorge, Cuicui Chen, Jing Li, and Mathias Reynaert. "Colluding against environmental regulation." Working paper (2021).

Anderson, Michael L. "Multiple inference and gender differences in the effects of early intervention: A reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects." *Journal of the American Statistical Association* 103, no. 484 (2008): 1481-1495.

Andrews, Steven Q. "Inconsistencies in air quality metrics: 'Blue Sky' days and PM10 concentrations in Beijing." *Environmental Research Letters* 3, no. 3 (2008): 034009.

Athey, Susan, and Guido Imbens. "Recursive partitioning for heterogeneous causal effects." *Proceedings of the National Academy of Sciences* 113, no. 27 (2016): 7353-7360.

Auffhammer, Maximilian, Antonio M. Bento, and Scott E. Lowe. "Measuring the effects of the Clean Air Act Amendments on ambient PM10 concentrations: The critical importance of a

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01477

spatially disaggregated analysis." *Journal of Environmental Economics and Management* 58, no. 1 (2009): 15-26.

Bennear, Lori S., Katrina K. Jessoe, and Sheila M. Olmstead. "Sampling out: regulatory avoidance and the total coliform rule." *Environmental Science & Technology* (2009): 5176-5182.

Barrios, Thomas, Rebecca Diamond, Guido W. Imbens, and Michal Kolesár. "Clustering, spatial correlations, and randomization inference." *Journal of the American Statistical Association* 107, no. 498 (2012): 578-591.

Benjamini, Yoav, and Yosef Hochberg. "Controlling the false discovery rate: a practical and powerful approach to multiple testing." *Journal of the Royal Statistical Society: Series B (Methodological)* 57, no. 1 (1995): 289-300.

Bento, Antonio, Matthew Freedman, and Corey Lang. "Who benefits from environmental regulation? Evidence from the Clean Air Act Amendments." *Review of Economics and Statistics* 97, no. 3 (2015): 610-622.

Blundell, Wesley, Gautam Gowrisankaran, and Ashley Langer. "Escalation of scrutiny: The gains from dynamic enforcement of environmental regulations." *American Economic Review* 110, no. 8 (2020): 2558-85.

Buchmueller, Thomas, Sarah Miller, and Marko Vujicic. "How do providers respond to changes in public health insurance coverage? Evidence from adult Medicaid dental benefits." *American Economic Journal: Economic Policy* 8, no. 4 (2016): 70-102.

Chernozhukov, Victor, Mert Demirer, Esther Duflo, and Ivan Fernandez-Val. Generic machine learning inference on heterogenous treatment effects in randomized experiments. No. w24678. National Bureau of Economic Research, 2018.

Currie, Janet, and Matthew Neidell. "Air pollution and infant health: what can we learn from California's recent experience?." *The Quarterly Journal of Economics* 120, no. 3 (2005): 1003-1030.

Chen, Yuyu, Ginger Zhe Jin, Naresh Kumar, and Guang Shi. "Gaming in air pollution data? Lessons from China." *The BE Journal of Economic Analysis & Policy* 13, no. 3 (2012).

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01477

Christensen, Garret, and Edward Miguel. "Transparency, reproducibility, and the credibility of economics research." *Journal of Economic Literature* 56, no. 3 (2018): 920-80.

Cutter, W. Bowman, and Matthew Neidell. "Voluntary information programs and environmental regulation: Evidence from 'Spare the Air'." *Journal of Environmental Economics and Management* 58, no. 3 (2009): 253-265.

Davis, Jonathan M.V. and Sara B. Heller. "Rethinking the Benefits of Youth Employment Programs: The Heterogeneous Effects of Summer Jobs." *The Review of Economics and Statistics*, 102, no. 4 (2020): 664–667.

Di, Qian, Heresh Amini, Liuhua Shi, Itai Kloog, Rachel Silvern, James Kelly, M. Benjamin Sabath et al. "An ensemble-based model of PM2.5 concentration across the contiguous United States with high spatiotemporal resolution." *Environment International* 130 (2019): 104909.

Dietz, Thomas, Kenneth A. Frank, Cameron T. Whitley, Jennifer Kelly, and Rachel Kelly. "Political influences on greenhouse gas emissions from US states." *Proceedings of the National Academy of Sciences* 112, no. 27 (2015): 8254-8259.

Dudoit, Sandrine, Juliet Popper Shaffer, and Jennifer C. Boldrick. "Multiple hypothesis testing in microarray experiments." *Statistical Science* (2003): 71-103.

Duflo, Esther, Michael Greenstone, Rohini Pande, and Nicholas Ryan. "Truth-telling by third-party auditors and the response of polluting firms: Experimental evidence from India." *The Quarterly Journal of Economics* 128, no. 4 (2013): 1499-1545.

Duflo, Esther, Michael Greenstone, Rohini Pande, and Nicholas Ryan. "The value of regulatory discretion: Estimates from environmental inspections in India." *Econometrica* 86, no. 6 (2018): 2123-2160.

Efron, Bradley. *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*. Vol. 1. Cambridge University Press, 2012.

Evans, Mary F., and Sarah L. Stafford. "The Clean Air Act Watch List and federal oversight of state enforcement efforts." *Journal of Environmental Economics and Management* 93 (2019): 170-184.

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01477

Finkelstein, Amy, Sarah Taubman, Bill Wright, Mira Bernstein, Jonathan Gruber, Joseph P. Newhouse, Heidi Allen, Katherine Baicker, and Oregon Health Study Group. "The Oregon health insurance experiment: evidence from the first year." *The Quarterly Journal of Economics* 127, no. 3 (2012): 1057-1106.

Fowlie, Meredith, Edward Rubin, and Reed Walker. "Bringing satellite-based air quality estimates down to earth." In *AEA Papers and Proceedings*, vol. 109, pp. 283-88. 2019.

Ghanem, Dalia, and Junjie Zhang. "'Effortless Perfection:' Do Chinese cities manipulate air pollution data?" *Journal of Environmental Economics and Management* 68, no. 2 (2014): 203-225.

Ghanem, Dalia, Shu Shen, and Junjie Zhang. "A censored maximum likelihood approach to quantifying manipulation in China's air pollution data." *Journal of the Association of Environmental and Resource Economists* 7, no. 5 (2020): 965-1003.

Giles, Cynthia. "Next Generation Compliance: Environmental Regulation for the Modern Era." Harvard Law School Environmental and Energy Law Program, Cambridge, Massachusetts (2020).

Glaeser, Edward L., and Raven E. Saks. "Corruption in America." *Journal of Public Economics* 90, no. 6-7 (2006): 1053-1072.

Graff Zivin, Joshua, and Matthew Neidell. "Days of haze: Environmental information disclosure and intertemporal avoidance behavior." *Journal of Environmental Economics and Management* 58, no. 2 (2009): 119-128.

Grainger, Corbett, Andrew Schreiber, and Wonjun Chang. " Do Regulators Strategically Avoid Pollution Hotspots when Siting Monitors? Evidence from Remote Sensing of Air Pollution." Working paper (2017).

Gray, Wayne B., and Jay P. Shimshack. "The effectiveness of environmental monitoring and enforcement: A review of the empirical evidence." *Review of Environmental Economics and Policy* 5, no. 1 (2011): 3-24.

Greenstone, Michael, Guojun He, Ruixue Jia, and Tong Liu. Can Technology Solve the Principal-Agent Problem? Evidence from China's War on Air Pollution. No. w27502. National Bureau of Economic Research, 2020.

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01477

Greenstone, Michael, John A. List, and Chad Syverson. The effects of environmental regulation on the competitiveness of US manufacturing. No. w18392. National Bureau of Economic Research, 2012.

Grooms, Katherine K. "Enforcing the Clean Water Act: The effect of state-level corruption on compliance." *Journal of Environmental Economics and Management* 73 (2015): 50-78.

Hagemann, Andreas. "Placebo inference on treatment effects when the number of clusters is small." *Journal of Econometrics* 213, no. 1 (2019): 190-209.

He, Guojun, Shaoda Wang, and Bing Zhang. "Watering down environmental regulation in China." *The Quarterly Journal of Economics* 135.4 (2020): 2135-2185.

Heckman, James, Seong Hyeok Moon, Rodrigo Pinto, Peter Savelyev, and Adam Yavitz. "Analyzing social experiments as implemented: A reexamination of the evidence from the HighScope Perry Preschool Program." *Quantitative Economics* 1, no. 1 (2010): 1-46.

Hung, HM James, Robert T. O'Neill, Peter Bauer, and Karl Kohne. "The behavior of the p-value when the alternative hypothesis is true." *Biometrics* (1997): 11-22.

Jones, Damon, David Molitor, and Julian Reif. "What do workplace wellness programs do? Evidence from the Illinois workplace wellness study." *The Quarterly Journal of Economics* 134, no. 4 (2019): 1747-1791.

Karplus, Valerie J., Shuang Zhang, and Douglas Almond. "Quantifying coal power plant responses to tighter SO2 emissions standards in China." *Proceedings of the National Academy of Sciences* 115.27 (2018): 7004-7009.

Kline, Patrick, and Christopher Walters. "Reasonable Doubt: Experimental Detection of Job‐Level Employment Discrimination." *Econometrica* 89, no. 2 (2021): 765-792.

Kline, Patrick, Evan K. Rose, and Christopher R. Walters. "Systemic discrimination among large US employers." *The Quarterly Journal of Economics* 137, no. 4 (2022): 1963-2036.

Leeson, Peter T., and Russell S. Sobel. "Weathering corruption." *Journal of Law and Economics* 51, no. 4 (2008): 667-681.

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01477

Levinson, Arik. "Environmental regulatory competition: A status report and some new evidence." *National Tax Journal* (2003): 91-106.

List, John A., Azeem M. Shaikh, and Yang Xu. "Multiple hypothesis testing in experimental economics." *Experimental Economics* 22, no. 4 (2019): 773-793.

Malani, Anup, and Julian Reif. "Interpreting pre-trends as anticipation: Impact on estimated treatment effects from tort reform." *Journal of Public Economics* 124 (2015): 1-17.

Malo, Nathalie, James A. Hanley, Sonia Cerquozzi, Jerry Pelletier, and Robert Nadon. "Statistical practice in high-throughput screening data analysis." *Nature Biotechnology* 24, no. 2 (2006): 167-175.

Millimet, Daniel L. "Environmental federalism: a survey of the empirical literature." *Case Western Reserve Law Review* 64 (2013): 1669.

Morehouse, John, and Edward Rubin. "Do polluters strategically locate near borders? The geography of US power plants and their emissions." Working paper. 2021.

Neidell, Matthew. "Information, avoidance behavior, and health the effect of ozone on asthma hospitalizations." *Journal of Human Resources* 44, no. 2 (2009): 450-478.

Oates, Wallace E. A reconsideration of environmental federalism. Resources for the Future Discussion Paper 01-54. 2001.

Oliva, Paulina. "Environmental regulations and corruption: Automobile emissions in Mexico City." *Journal of Political Economy* 123, no. 3 (2015): 686-724.

Reynaert, Mathias. "Abatement strategies and the cost of environmental regulation: Emission standards on the European car market." *The Review of Economic Studies* 88, no. 1 (2021): 454-488.

Reynaert, Mathias, and James M. Sallee. "Who Benefits When Firms Game Corrective Policies?." *American Economic Journal: Economic Policy* 13, no. 1 (2021): 372-412.

Rosenbaum, Paul R. "Overt bias in observational studies." In *Observational Studies*, pp. 71-104. Springer, New York, NY, 2002.

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01477

Schwartz, Joel. "Air pollution and blood markers of cardiovascular risk." *Environmental Health Perspectives* 109, no. suppl 3 (2001): 405-409.

Shapiro, Joseph S., and Reed Walker. Is Air Pollution Regulation Too Stringent? No. w28199. National Bureau of Economic Research, 2020.

Shepard, Donald. "A two-dimensional interpolation function for irregularly-spaced data." In *Proceedings of the 1968 23rd ACM national conference*, pp. 517-524. 1968.

Shimshack, Jay P. "The economics of environmental monitoring and enforcement: A review." *Annual Review of Resource Economics* 6 (2014): 339-60.

Simonsohn, Uri, Leif D. Nelson, and Joseph P. Simmons. "P-curve: a key to the file-drawer." *Journal of Experimental Psychology: General* 143, no. 2 (2014): 534.

Storey, John D. "The positive false discovery rate: a Bayesian interpretation and the q-value." *The Annals of Statistics* 31, no. 6 (2003): 2013-2035.

Sullivan, Daniel M., and Alan Krupnick. "Using Satellite Data to Fill the Gaps in the US Air Pollution Monitoring Network." Resources for the Future Working Paper (2018): 18-21.

U.S. EPA. *Quality Assurance Handbook for Air Pollution Measurement Systems–Volume II–Ambient Air Quality Monitoring Program. Vol. 2.* EPA-454/B-13-003 (2013).

Walker, W. Reed. "The transitional costs of sectoral reallocation: Evidence from the clean air act and the workforce." *The Quarterly Journal of Economics* 128, no. 4 (2013): 1787-1835.

Yang, Lin. "Pollution Monitoring, Strategic Behavior, and Dynamic Representativeness." Working paper (2020).

Young, Alwyn. "Channeling fisher: Randomization tests and the statistical insignificance of seemingly significant experimental results." *The Quarterly Journal of Economics* 134, no. 2 (2019): 557-598.

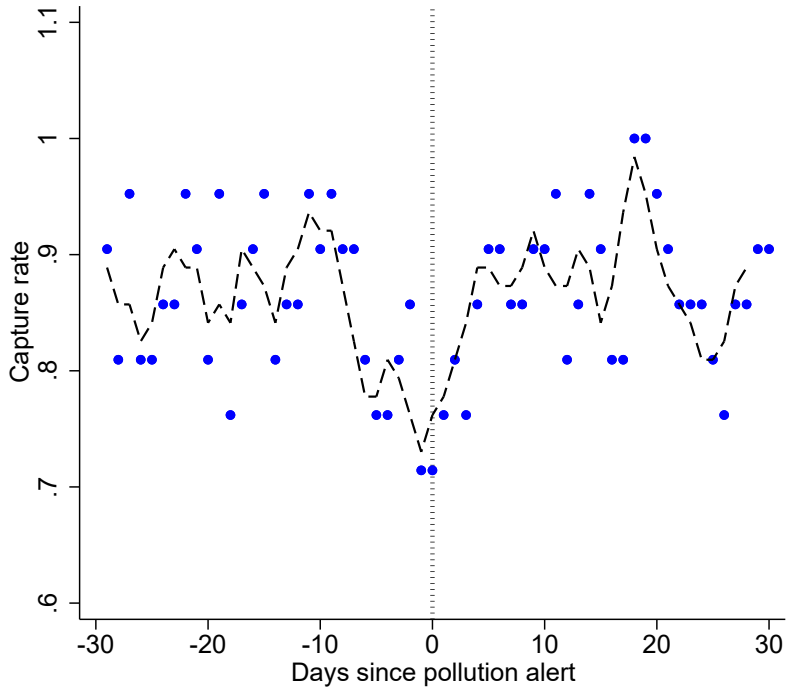
Zitzewitz, Eric. "Forensic economics." *Journal of Economic Literature* 50, no. 3 (2012): 731-69.

Zou, Eric. "Unwatched pollution: The effect of intermittent monitoring on air quality." *American Economic Review* 111, no. 7 (2021): 2101-26.

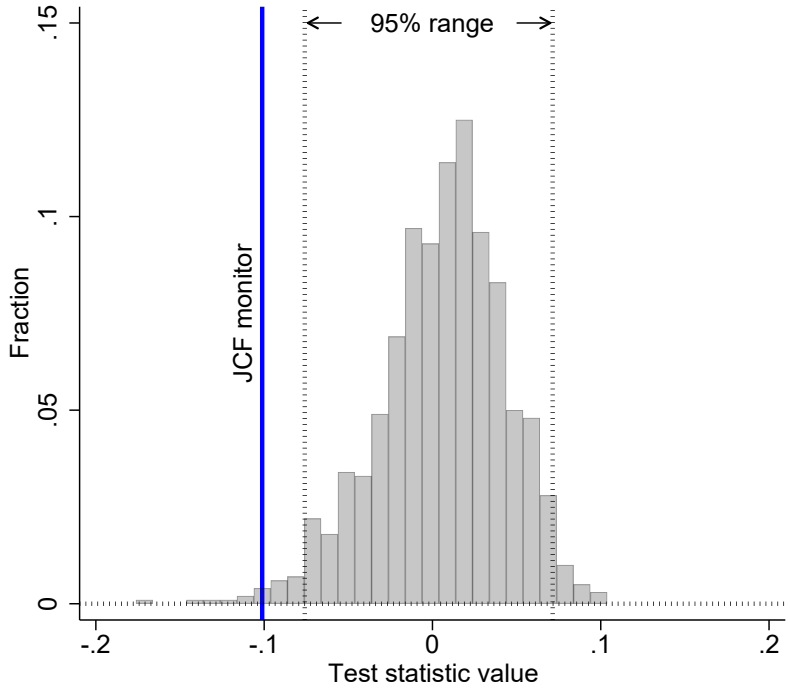
Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01477

Figure 1. Monitor's Sampling Behavior near Pollution Alerts: Jersey City Firehouse $PM_{2.5}$ Monitor

Panel A. Event study of monitor's capture rate



Panel B. Randomized inference with 5,000 placebo scenarios

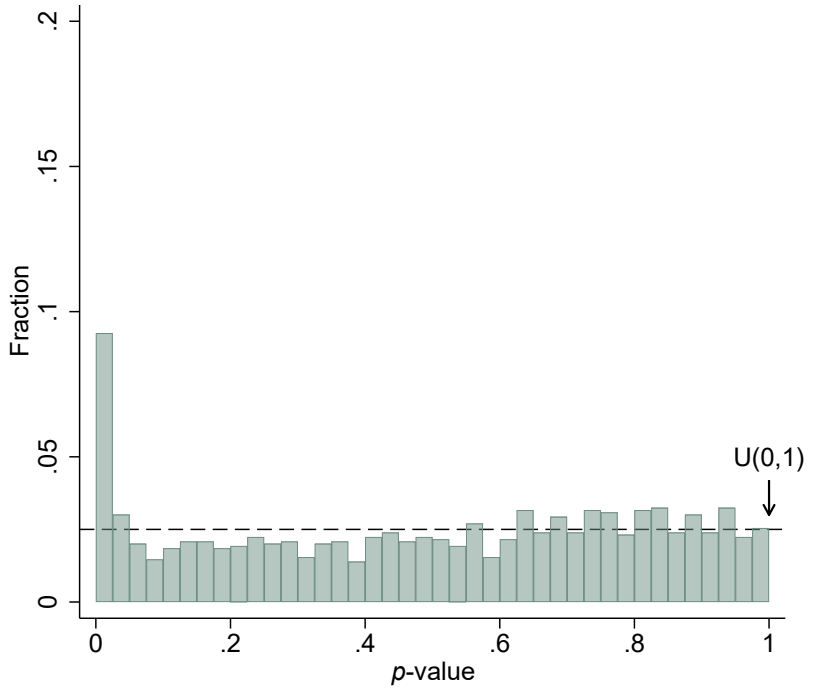


Notes: Panel A plots JCF monitor's average capture rate (i.e., one minus a dummy for missing data) as a function of days since pollution alerts issued by the Jersey City. Number of alerts = 21. Dashed line represents three-day moving average of point estimates. Panel B plots the distribution of the test statistics derived from 5,000 randomly assigned pollution alerts. Test statistic equals the difference between mean capture rates across event days [-3,3] and mean capture rates across event days [-30,-10]∪[10,30]. Solid vertical line is the observed (i.e., true) test statistic. Dashed vertical lines show 95% range of the randomized test statistics.

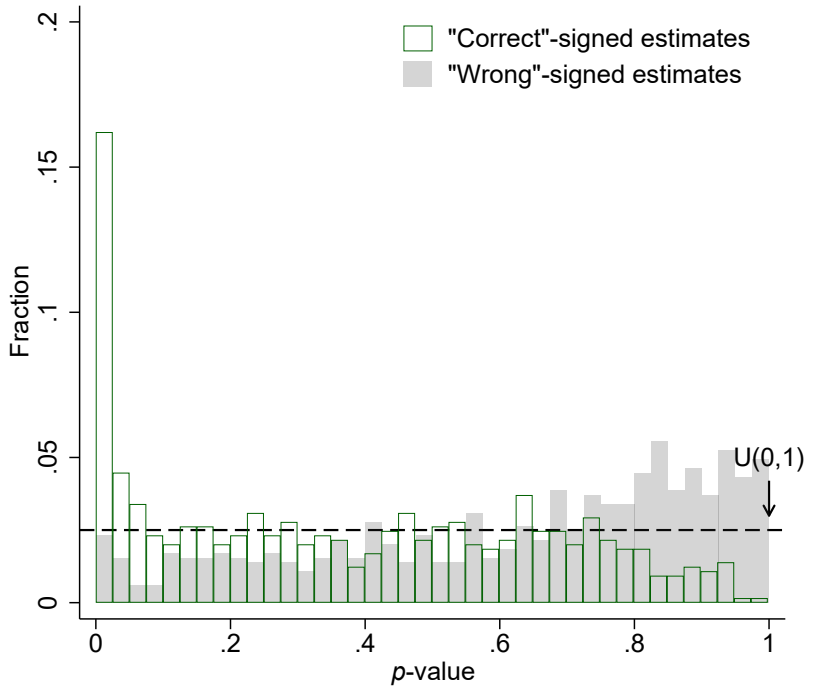
Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01477

Figure 2. Distribution of *p*-values, All Monitors

Panel A. Overall



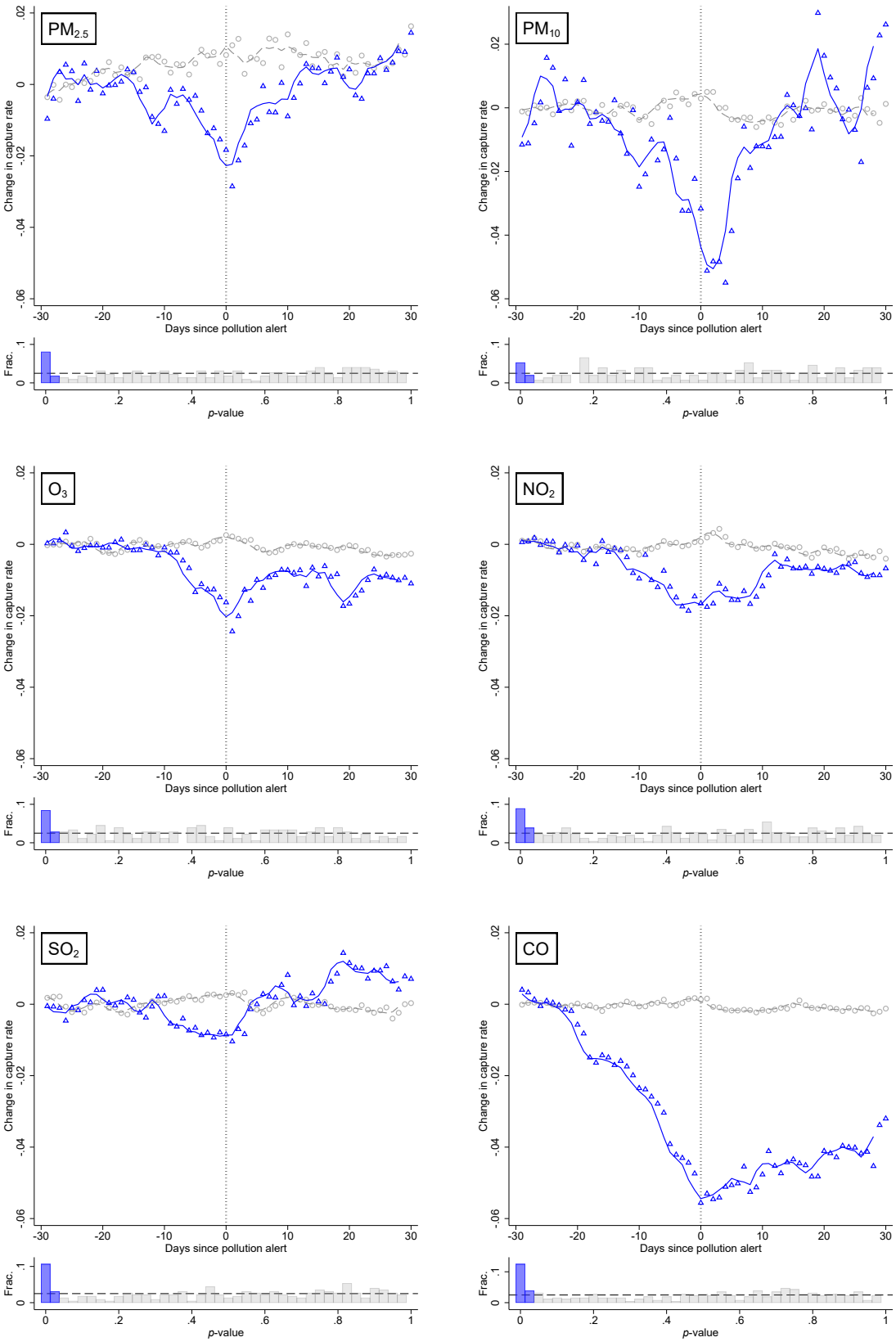
Panel B. By direction of the effect



Notes: Panel A shows the distribution of *p*-values for all monitors. Test statistic equals the difference between mean capture rates across event days [-3,3] and mean capture rates across event days [-30,-10]∪[10,30]. Panel B shows the breakdown by the sign of the estimated effect. Hollow bars show *p*-values for negative estimates (i.e., monitoring capture decreases around pollution alerts), and shaded bars show *p*-values for positive estimates (i.e., monitoring capture increases around pollution alerts). Horizontal dashed lines show the uniform distribution.

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01477

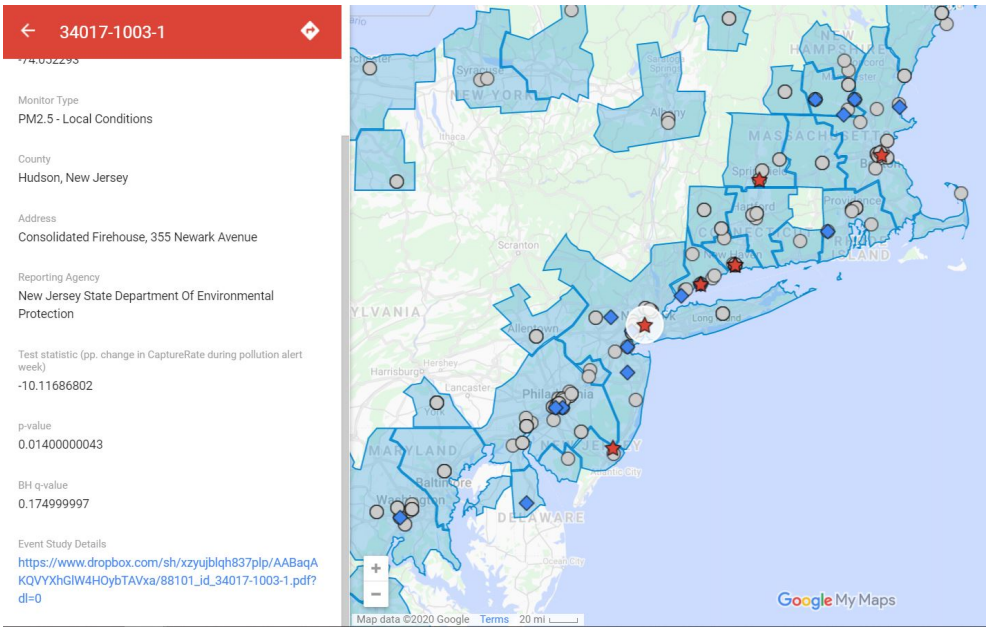
Figure 3. Capture Rate for "Interesting" Monitors (△) and Other Monitors (○)



Notes: This graph shows mean monitoring capture rate for "interesting" monitors (those with p–value<0.05) and other monitors. Data are demeaned by the average capture rate across the first ten event days. Fitted lines show three-day moving averages of point estimates. Each panel corresponds to one pollutant. Histograms show the distributions of p–values for the corresponding pollutant monitors.

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01477

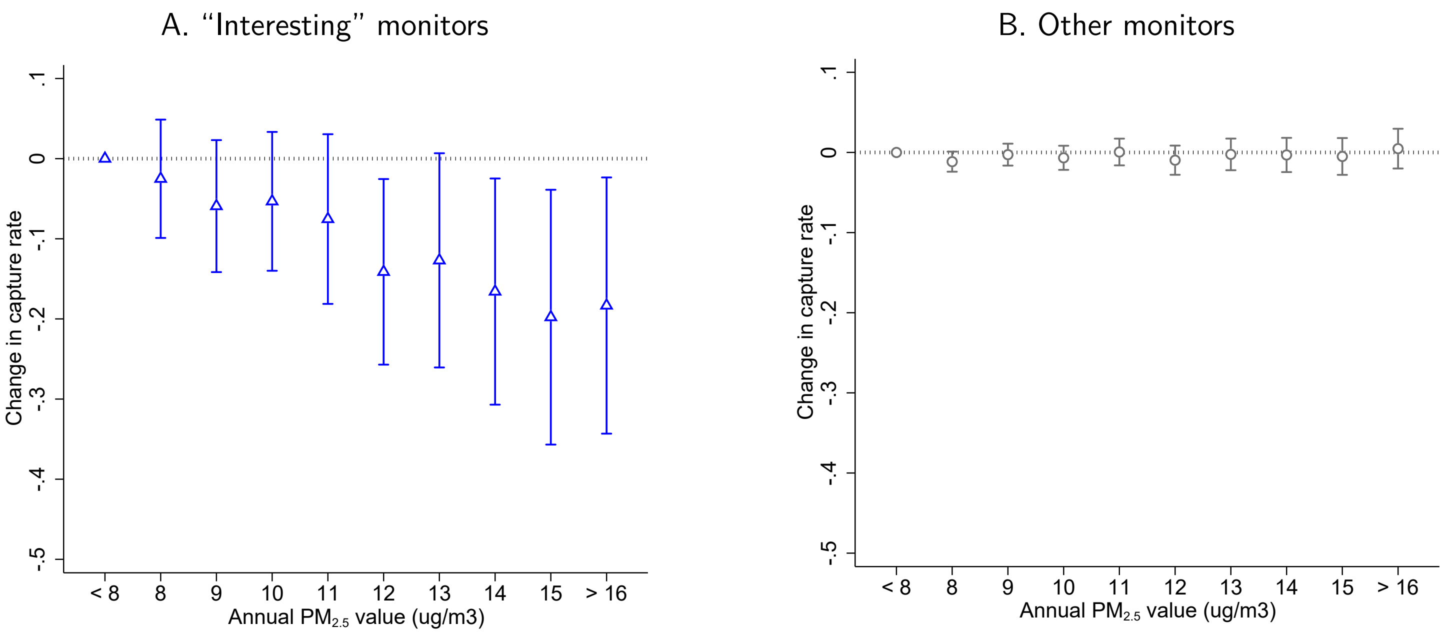
Figure 4. Study Website: Estimation Results for All Monitors



Notes: We store full estimation results at this website. Shaded areas highlight study regions. Click on each monitor to view estimation details.

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01477

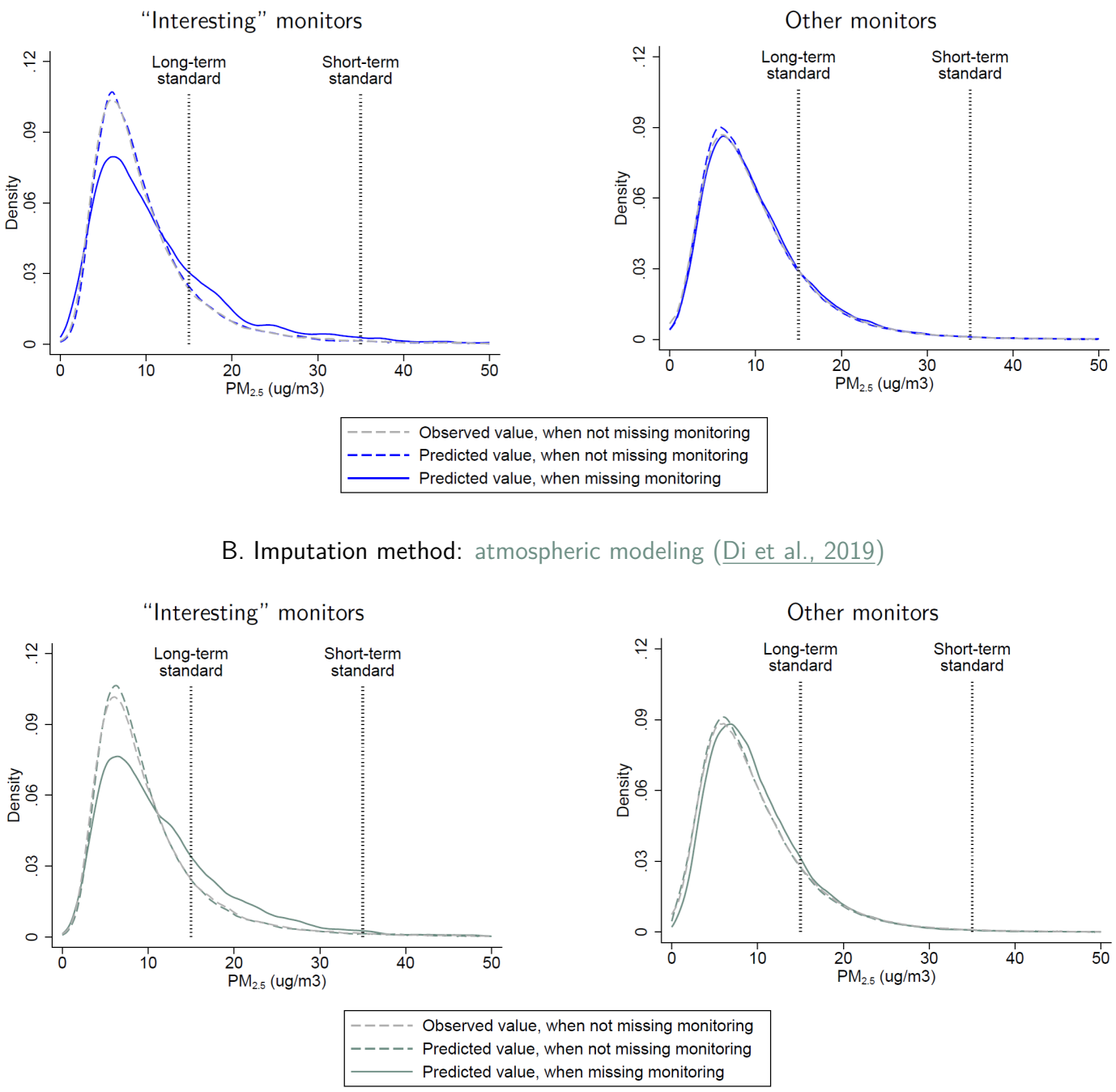
Figure 5. Quarterly Capture Rate vs. Annual $PM_{2.5}$ Design Value



Notes: This figure reports regression of a monitor's quarterly capture rate (valid sampling days divided by required sampling days in a quarter) on 1-ug/m3 bins of annual mean $PM_{2.5}$ concentration (i.e., design values for the $PM_{2.5}$ annual standard). The "< 8 ug/m3" bin is the omitted category. The regression controls for monitor fixed effects and year fixed effects. The regression is run separately for "interesting" monitors (panel A, number of observations = 300) and other monitors (panel B, number of observations = 7,688).

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01477

Figure 6. Distributions of Observed and Imputed $PM_{2.5}$ Concentration



Notes: Underlying data are monitor-daily level average $PM_{2.5}$ concentration. "Observed value" is concentration recorded on the monitor-day. In panel A, "Predicted value" is inverse distance-weighted concentration from all other operative $PM_{2.5}$ monitors within a 20-mile radius. In panel B, "Predicted value" is from 1 km × 1 km grid-daily prediction of $PM_{2.5}$ from atmospheric ensemble-based modeling (Di et al., 2019) which incorporates satellite observations. "Long-term standard" marks the 15 ug/m3 annual NAAQS standard. "Short-term standard" marks the 35 ug/m3 24-hr NAAQS standard.

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01477

Table 1. Correlates of "Interesting" Monitors

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Dep. var.:	1(p-value ≤ 0.05)				1(q-value ≤ 0.05)			
Non-attainment	0.066**				0.039*			
	(0.030)				(0.021)			
Non-attainment × 1("wrong" sign)		-0.014	0.011	-0.001		-0.002	0.012	0.022
		(0.033)	(0.034)	(0.041)		(0.024)	(0.024)	(0.030)
Non-attainment × 1("correct" sign)		0.203***	0.220***	0.223***		0.111***	0.124***	0.129***
		(0.055)	(0.055)	(0.061)		(0.039)	(0.039)	(0.044)
Above median Democrats			-0.022				-0.014	
			(0.027)				(0.019)	
Above median LCV score			-0.023				-0.021	
			(0.027)				(0.019)	
Above median government size			0.007				-0.001	
			(0.017)				(0.012)	
Above median corruption			0.035*				0.008	
			(0.018)				(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: Each column is a separate regression. Underlying data is a cross-section of monitors matched to parenting county's characteristics. Dependent variable is an indicator for whether the monitor's p–value is less than 0.05 (columns 1-4), or an indicator for whether the monitor's FDR-adjusted significance q–value is less than 0.05 (columns 5-8) where the family of tests is all 1,359 monitors. "Non-attainment" is an indicator for whether the county has ever been in NAAQS non-attainment throughout the study period. "1("correct" sign)" indicates a negative effect sign, i.e., capture rate drops near pollution alerts. "Above median"'s indicate the county has an above-median level of share of Democrats affiliation (2006 Gallup Poll), League of Conservation Voters score, share of government-sector employees (Bureau of Economic Analysis, NAICS=92), and per-capita number of federal convictions among state and local public officials (Glaeser and Saks, 2006). *: $p < 0.10$; **: $p < 0.05$; ***: $p < 0.01$.