Filling in the Gaps: Using Consumer Products to Replace Missing Pollution Data

Aaron Watt

April 2022

Abstract

The Clean Air Act and follow-on amendments (CAAA) constitute one of the largest-scale pollution policies in US history. The National Ambient Air Quality Standards (NAAQS), established by the CAAA, provide thresholds for a basic minimum standard of air quality in the US. However, the EPA only requires 75% completeness of air quality measurements. This minimum standard of completeness provides a plausible mechanism for locations to under-report high pollution to avoid expensive NAAQS penalties, thus biasing the air quality measurements that are compared to the NAAQS. This paper explores this issue of manipulability from omitting data by bringing in new consumer-based outdoor air pollution data to fill in the gaps in reported data. I find that though most tested locations do not have a significant bias in their reported measurements due to missing data, one of 15 sites tested has a statistically and economically large bias due to missing data that would be reported under a higher completeness standard. Additionally, I show that the minimum completeness criteria leaves room for a large amount of manipulation. While this is a test on a relatively small number of NAAQS sites, this methodology can be employed in future work to examine most or all of US air quality monitors.

1 Introduction

A critical input to good air quality regulation is reliable air quality measurement. In the United States, air quality is assessed by the federal government using a network of monitors that measure levels of ambient air pollution at the location of the monitor to a high degree of accuracy. The Environmental Protection Agency (EPA) requires these monitors to measure average daily air quality at specific frequencies to ensure enough data is collected for effective regulation.[1] During the days that are required to be measured, the goal is to accurately measure the daily average pollution concentration at the site of the monitor. Statistics of these daily averages, called *design values*, are then used to decide if a region is in or out of compliance with the National Ambient Air Quality Standards (NAAQS). Ideally, we would be able to measure air quality at every point in the US, at every hour and every day, to create design values that reflect statistics of the realized pollution distribution. In practice, we have a fairly sparse network of monitors that do not report 100% of the hours they are assigned to measure. The rules for minimum completeness of these data provide some flexibility in reporting for staff on the ground, allowing those in charge of the pollution monitors the ability to maintain the equipment – and take it offline when necessary – and not worry too much about missing a few hours of air quality measurements.

Though the measurements of air quality at the site of the monitor are fairly accurate when the monitor is on, omitting some measurements by turning the monitor off or removing measurements from the data before uploading to the EPA database could bias compliance statistics calculated from reported measurements. In the minimum reporting criteria (completeness goal) for measuring particulate matter in the air, the EPA allows 25% of daily measurements to be missing or omitted from each quarter. Additionally, 25% of hourly measurements are allowed to be missing from each day. Combined, these completeness goals imply that up to 43% of hourly measurements each quarter can be missing, representing a large source of potential bias in the estimation of the distribution of air quality at each monitoring site. (EPA, 2017).[2]

[1] The three main measurement frequencies require measuring daily average air quality every 1, 3 or 6 days.

[2] Design values are used to decide compliance with NAAQS and are statistics of daily averages. In calculating daily averages, the daily average is valid if at least 75% of the hourly readings (18 of 24 hours) are reported and valid. In calculating the design values, the design value is valid if at least 75% of daily averages in each quarter are reported and valid. Combining the daily and quarterly completeness measures,

Historically, the air quality monitoring stations have been placed in areas that are thought to have problems with low air quality. Though these air quality monitoring stations are regulated by the EPA, the EPA delegates the task of managing the stations and reporting the data to the states and regions being monitored. Effectively, the entities facing regulation are in charge of recording and reporting their own level of ambient air pollution. Additionally, if a region is out of compliance with the standards set by the EPA, the region or state can potentially face large penalties and forced adoption of expensive abatement technology.

The combination of large penalties and the discretion that local officials have to miss or drop measurements implies the possibility of misaligned incentives between federal regulators and local officials in charge of monitoring air quality. In theory, under reporting could lead to biased air quality statistics. Indeed, previous research suggests that there is mismeasurement of air quality statistics occurring; Zou (2021), Mu et al. (2021), Grainger et al. (2019) and Grainger and Schreiber (2019) provide evidence of strategic behavior in pollution measurement on behalf of local pollution regulators. This paper extends questions similar to Mu et al. (2021) about how monitor shutdowns and missing data can effect reported air quality data.

Specifically, I explore the question: is there a bias in reported air quality data due to missing observations and how much room is left for additional manipulation of compliance statistics? To explore these issues, I utilize a new network of consumer-based air quality sensors (PurpleAir sensors). Adoption of PurpleAir sensors has accelerated, especially in the last few years, providing a relatively dense network of independent groundtruth comparisons to air quality reported to the EPA via regulatory air quality monitors. The most promising new data coming from this sensor network are PM2.5 measurements – the concentration of particles in air that are 2.5 micrometers and below. Aside from data availability, PM2.5 is an incredibly important type of pollution to study, impacting health and productivity at levels well below current national standards, contributing to large social costs of pollution

the minimum reporting standard is actually about 57% of all required hourly PM2.5 readings per quarter. This is slightly different for each site depending on their reporting frequency (every 1, 3, or 6 days).

(Feng et al., 2016). This new sensor network provides multiple opportunities to address questions about air quality. One possible application the this new network of PM monitors is to examine how vulnerable the current NAAQS are to manipulation and omitted data.

Specifically, I combine PM2.5 measurements from multiple PurpleAir sensors that are near to federally-regulated monitoring stations to estimate the PM2.5 value at the monitoring station; I use inverse distance weighting to create a weighted average of PurpleAir measurements.[3] This allows me to construct predicted values of PM2.5 at the station during times when the station's readings would be used to calculated NAAQS compliance but when the station was shut down. I first examine how these predicted missing PM2.5 values compare to the reported values – if the missing data is missing at random, I would expect the data to be similar in distribution. Then I use the predicted values from PurpleAir sensors to fill in the gaps in the NAAQS monitor's reported data and use this reconstructed dataset to generate counterfactual NAAQS compliance statistics.

The NAAQS compliance statistics for PM2.5, called *design values*, are functions of the daily averages reported by air quality monitors. There are two primary design values for PM2.5: the "annual" design value is a three-year average of the daily averages; and "24-hour" design value is a three-year average of the annual 98^{th} percentile of daily averages.[4] Each quarter (3-month period), these two design values are calculated and compared to the NAAQS for PM2.5. If a monitor's design value is above the standard, then the monitor (and associated region) is determined to be in *non-attainment* (non-compliance) with the standard for that quarter. Using the reconstructed dataset of PM2.5 (PM2.5 estimates for all hours that would be reported from a given NAAQS monitor), I construct counterfactual estimates of the design values that determine if a region is in or out of attainment. I use these counterfactuals to determine which regions would have changed compliance status if they

[3] Inverse distance weighting has drawbacks: it can apply very large weight to sensors very near to the NAAQS monitor and it does not take into account that some PurpleAir sensors will be better predictors for the NAAQS monitor. See Appendix section 6.3 for an alternative strategy that I plan to implement.

[4] these statistics are discussed more in the Data section. Specific formulas for these statistics are listed in the appendix.

reported 100% of their PM2.5 measurements – I call these “flipped regions”. I also examine how close these flipped regions were to the regulatory threshold and report a measure of the bias related to the station’s missing PM2.5 readings.

Though I am examining the effect of pollution data that are missing from a monitor’s record (data missing *in time*), there is also the issue of attempting to measure a region’s ambient air quality using spatially sparse locations of monitors (you could consider this and issue of data missing *in space*). Previous literature has examined the sparse distribution of regulation-grade monitors and the resulting sensitivity of CAA air quality regulation. Grainger et al. (2019) and Grainger and Schreiber (2019) identify a principle-agent problem with the initial spatial placement of sparse pollution monitors; they find evidence that local regulators may be strategically locating their air quality monitors based on pollution, and possibly socioeconomic characteristics. To address the issue of sparse data and fill in the gaps, several authors have used satellite data products to provide finer resolution pollution data (Sullivan and Krupnick 2018, Fowlie et al. 2019). Moving to more time-based issues, Zou (2021) also uses satellite estimates to discuss the issue of strategic behavior in reaction to the timing of pollution monitoring. He provides evidence that some areas have significantly worse air quality on unmonitored days. In related work, Mu et al. (2021) show potential for strategic monitor shutdowns on days of expected high pollution, contributing to air quality data that is missing *in time*.

This paper contributes to this last line of research, where Mu et al. (2021) examine plausibly strategic monitor shutdowns. They examine how full-day omissions in the regulatory monitor data are related to the local regulator’s belief that high pollution is coming, proxied by local pollution alerts. This paper extends Mu et al. (2021) in two main ways: enabling the analysis of hour-level omissions in addition to day-level omissions; and examining the omission of observations that might occur after absent the local regulator’s prior belief that high pollution was coming in the near future.

First, Mu et al. look at local pollution alerts, somewhat limiting the study to missing day-level observations. Because I am examining hourly measurements, I can additionally study how the omission of hourly measurements can effect the distribution of reported measurements. As Mu et al. mention, up to 25% of daily measurements can be omitted while still complying with the federal completeness standard. Of the at least 75% of days that must be reported, the completeness goals also allow up to 6 hours to be omitted for each reported day. This represents another possible 18.75 percentage points of hourly observations that could be omitted from the reported days while maintaining completeness, on top of up to 25% missing full days.

Second, Mu et al. use pollution alerts to proxy for local regulator's prediction of high pollution, allowing them to test if data are more likely to be missing when a locality has signaled their belief of higher future pollution. This examines the likelihood of missing data that coincide with predictions of high pollution, regardless if the realization of pollution measurement at the location of the regulatory monitor is high or not following the pollution alert. However, there is also the possibility of data omissions resulting from high realized pollution but where no pollution alert was given (either because the high pollution was not predicted or because the prediction was not high enough to warrant a pollution alert). The concurrent PurpleAir data from sensors near the regulator monitor allow me to explore data omission occurring in all cases of realized high pollution. So an additional contribution of this application is the ability to examine how missing data might be distributed, possibly omitted *after* data was recorded, but in events when the high pollution was not predicted. However, I do not have a measure of the local authority's *belief* of high pollution and the lack of any information of the belief's of local authorities prevents this study from ascribing any form of intent to the local air quality authorities.

Additionally, this paper borrows the idea of computing counterfactual design values from Fowlie et al. (2019). Fowlie et al. (2019) use PM2.5 estimates generated from satellite data to examine counterfactual design values if *spatial* gaps between sensors could be filled in. In contrast, I am using the PurpleAir network to examine counterfactual design values if gaps *in time* can be filled in. Fowlie et al. end their analysis noting that the satellite-based data commonly used in these applications has significant prediction error in some areas; this can cause result in incorrect conclusions about design values. While satellite-based PM2.5 estimates have potential for large prediction errors, PurpleAir sensors can be fairly accurate measures of their local air quality[5] and can be averaged over multiple nearby sensors. PurpleAir data also have drawbacks however – the sensors are highly non-uniform in coverage across the US and are sensitive to specific placement by the consumer, perhaps leading to hyper-local estimates of air quality.

For these reasons, this analysis should be seen as a compliment to previous works by leveraging a new network of air pollution sensors to explore questions about our national air quality standards. As consumer sensors become more widespread, we can augment reliable federal air quality measurements with a growing number of auxiliary data points to better understand the shape of mismeasurement in air quality. In this paper, I explore one way of leveraging these data to test for issues with biased reporting of air quality. After predicting missing PM2.5 observations for 15 California air quality monitors using the PurpleAir network, I find a single monitor in Fresno, CA, that shows signs of bias. I find differences between design values calculated on reported NAAQS data and design values calculated on PurpleAir-imputed data; these differences persist for five quarters in 2020 and 2021 and are statistically significant at more than the 95% level, after adjusting for the multiplicity of hypothesis tests (198 hypothesis tests are conducted in total for the 15 sites). The largest discrepancy for Fresno is in the 24-hour design values, where the difference is more than $2.5\mu\text{g/m}^3$ of PM2.5 on average between 2018 and 2021. Fresno has been out of compliance for some time, and these results mirror previous literature that suggests larger pollution measurement problems in nonattainment areas.

I also conduct a simple exercise to examine how the maximum allowable omission of

[5] PurpleAir sensors have specifically been shown to be less accurate than regulation-grade monitors at high levels of PM2.5 concentration. However, the EPA has developed a correction technique that result in PurpleAir readings within 5% of co-located EPA monitors. This correction technique is used here and explained in more detail in the appendix.

observations could bias design values. Depending on the location, I find a large amount of room left for additional manipulation, with many sites able to reduce either 24-hour design value by 15-30 μg/m^3. The Fresno site in particular has a large amount of room for additional manipulation. However, this simple exercise assumes perfect foresight of which hours will have the highest PM readings (or the ability to manipulate the data after it's been recorded). In practice, there is likely much less room for manipulation but this provides an upper bound on how much more each site could have historically manipulated their design values.

Due to the small sample size of this study, there is not much evidence of widespread biased pollution standards. However, with the framework now setup, it would be possible to expand this type of analysis to all NAAQS monitors in the United States and all PurpleAir monitors available. A possible future outcome of this line of research is estimating the possible gains to be made in changing data completeness rules, NAAQS thresholds, and inclusion of currently excluded data. Increasing completeness standards to decrease allowable omitted observations may result in more non-attainment areas and further increases regulatory efficiency. Decreasing the NAAQS thresholds without changing completeness rules (or at least monitoring completeness) could result in more non-attainment areas, but with

The remainder of this article is organized as follows. Section 2 briefly reviews the history of air quality standards in the US and some key details of current regulations. Section 3 then discusses the data used and section 4 describes the empirical framework that will be applied to estimate the missing pollution and effects on NAAQS design values. Section 5 reviews the results of the empirical study and concludes.

2 Background

Amid growing public concern about air quality and pollution, the United States Congress passed the Clean Air Act of 1963 (CAA). Later additions to the CAA, the Clean Air Amendments of 1970, granted the Environmental Protection Agency (EPA) the regulatory authority to create and enforce air quality standards in the US. One major way air quality is regulated is through the National Ambient Air Quality Standards (NAAQS), which set concentration thresholds for a list of different "criteria" pollutants (91st US Congress, 1970). The EPA has since been in charge of setting and updating the NAAQS and require states to submit plans to bring their air quality to within NAAQS limits. An important aspect of enforcing the NAAQS is measuring criteria pollutants across the US by requiring states to install pollution monitoring stations in areas of questionable air quality. Because these monitoring stations are used for potentially costly enforcement, the equipment within each station must abide by specific regulations and are relatively costly to install and run.

Over the last decade, commercially available scientific equipment in measuring various air pollutants has evolved. There is now relatively cheap[6] equipment available to measure particulate matter (one of the criteria pollutants that regulated by the NAAQS). Specifically, the PurpleAir company produces devices that can measure particulate matter that has a diameter of less than 2.5 micrometers (designated as PM2.5).[7] PurpleAir is of particular interest because they have built an opt-out mechanism for end-users to allow their ambient air quality data to be stored in the cloud. They also provide multiple ways for researchers and the general public to use this crowd-sourced air quality data.

This paper is primarily concerned with the minimum reporting requirement. As with many federal regulations, there are many ways that states or emitters can cleverly navigate the rules to emit more than they are meant to according to the spirit of the regulation. One way of navigating the CAA regulations is through the choice of what data to report. The EPA currently requires a minimum threshold of air quality data to be reported – for PM

[6] e.g., a PurpleAir outdoor air quality sensor is about $250 to purchase with little upkeep from the end user, compared to roughly $100,000-200,000 to install EPA regulation-grade criteria pollutant monitors and trained staff to upkeep and record measurements. The cost alone is not a good comparison because the EPA monitors use different technology that is known to be more accurate across a wider range of pollution concentrations, have a better sense the sensor error, and measure more pollutants than the PurpleAir monitors. For the purposes of this analysis, PurpleAir monitors should be seen as a compliment to EPA monitors, not a potential replacement.

[7] PurpleAir devices can measure a few other criteria pollutants (namely ozone and PM10) but the comparability of the PM2.5 measurements between PurpleAir and EPA monitors are currently better understood.

2.5, 75% of daily measurements need to be reported, and each day must have 75% of hours reported. That leaves many choices of which hours to turn the monitor off for cleaning, calibration, or other reasons. I wish to understand how these timing decisions are affecting the distribution of reported data – specifically how it might be affecting a statistic of that distribution: the design value.

3 Data

3.1 EPA Regulation-grade Monitors

There are currently 388 air quality monitoring stations around the US that are used for NAAQS determination for PM2.5; I will refer to these monitors as *NAAQS monitors*. Each of the NAAQS monitors uploads hourly and/or daily PM2.5 measurements to the Air Quality System (AQS) database. There are more regulation-grade monitors that meet or approach the regulatory accuracy standards set by the EPA, but these 388 are the monitors that are officially used to calculate the design values that decide NAAQS attainment status. Of the 388 NAAQS monitors in the US, I limit my preliminary analysis on the 15 monitors in California that take hourly readings every day. Future analysis will include the full set of NAAQS monitors, which include monitors that only report daily averages (as opposed to hourly averages) and monitors that only report every 2, 3, 6, or 12 days.

Design Values. PM2.5 *design values* are statistics of hourly PM2.5 concentrations reported by the NAAQS monitors. In reality, design value determination for a monitor begins by calculating the initial design value on the non-missing data and then includes a negotiation step between EPA and the local regulator to decide the final, publicly-reported design value. I could use final design values for each monitor that are listed in EPA reports. However, I am interested in directly comparing the design values calculated from only reported data to design values that include predictions of unreported data. Because I cannot replicate the final negotiation process, I replicate work done in (Fowlie et al., 2019) to create *pseudo* *design values* by calculating the statistic on the data and making comparisons based on this initial design value. There are two NAAQS design values for PM2.5 explained in Section 4.2: annual and 24-hour. These design value calculations only use valid daily and annual averages, where validity is determined based on the number of reported and non-excluded observations.

Excluded Readings. There are a number of events that create air quality measurements that cannot be used in NAAQS determination; wildfires or machine calibrations (for example) can cause hour- or day-long readings to be invalid for the purposes of NAAQS determination. These times, referred to as *exceptional events* (EE), are events that are "not expected to recur routinely at a given location, or that [are] possibly uncontrollable or unrealistic to control through the [NAAQS regulatory] process"(EPA, 1990). These events are identified in the NAAQS monitor data and removed from the analysis: hours that have been labeled as EE are removed from both the PurpleAir and NAAQS monitor data before calculating design values.[8] These are not considered "missing" or "unreported" data for the sake of predicting missing values, however these are considered invalid observations in the design value calculation. Removing EE provides more realistic pseudo design value estimates.

Missing PM2.5 Data Figure 1 shows how complete the raw PM2.5 data is, as reported to the EPA and pulled from the AQS database.

3.2 PurpleAir Consumer Sensors

The last ten years have seen a growing interest in consumer-based air quality measurement. PurpleAir air quality sensors are designed to mainly measure PM2.5, but also measure other pollutants (PM10, ozone) and environmental factors (humidity, temperature).[9]. In my analysis, PurpleAir PM2.5 data plays a ground-truth role – it gives me an alternative source of PM2.5 measurements to rely on when the NAAQS monitor is shut off.

[8] See Appendix Section 6.6 Table 20 for a detailed list of reasons that observations are excluded from NAAQS determinations.

[9] See Appendix section 6.7 for pictures of both a NAAQS monitoring station and a typical PurpleAir outdoor pollution sensor

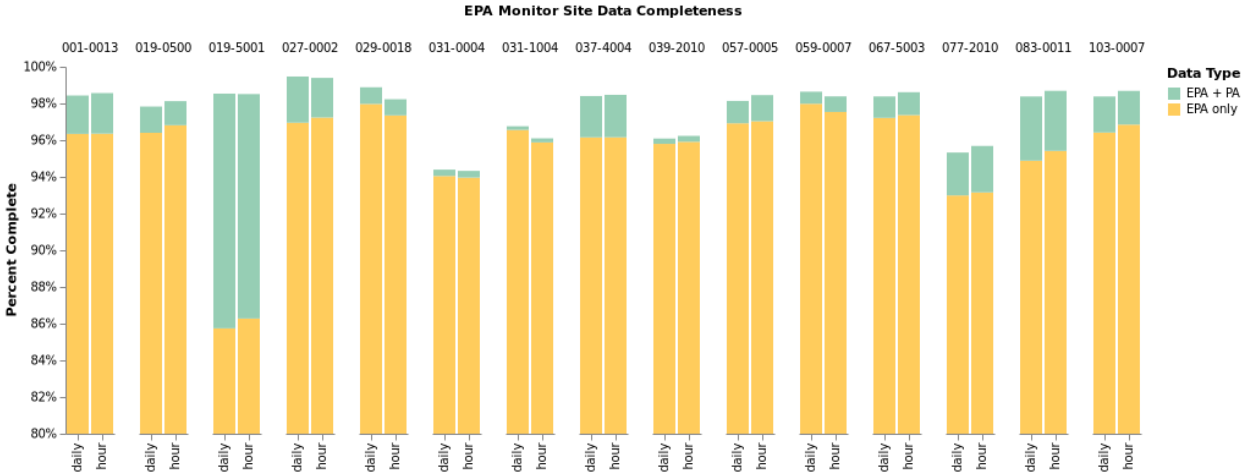

Figure 1: Completeness of PM2.5 observations for each site. Yellow: percentage of hourly and daily observations from each site that are not missing during that site's study period. Green: percentage of hourly and daily observations during study period that can be imputed using nearby PurpleAir PM2.5 measurements.

To examine how design values might be influenced by missing data, I predict missing PM2.5 hourly average concentrations from EPA NAAQS monitors using nearby PurpleAir PM2.5 sensors. For an initial analysis, I limit the sample to include PurpleAir sensors within 5 miles of each NAAQS monitor, or extending up to 25 miles to get 10 PurpleAir sensors minimum for each monitor.

This is a fairly new and rich dataset: there have been more than 16,000 public PurpleAir sensors brought online in the United States since 2015. When a consumer is setting up their sensor, they have the choice to make the sensor public or private. All sensors upload their PM2.5 readings to an online server, but only public sensors have data available for research use. The company asks consumers to make their data public if possible, attempting to contribute to more citizen science. Of the 16,000+ US sensors, there are 10,401 in California, Oregon, Nevada, and Arizona and I limit my sample to the 592 unique PurpleAir sensors within 5 miles of 15 NAAQS-primary monitors.

Correction of PurpleAir Readings. PurpleAir sensors are known to have worse readings at higher levels of pollution. I modify PurpleAir PM2.5 values using the EPA's correction equation for PurpleAir sensors. The calibrated this equation by studying co-located PurpleAir and NAAQS monitors.

$$\widetilde{PA}_{j,t} = \begin{cases} 0.52 * PA_{j,t} - 0.086 * H_{j,t} + 5.75, & \text{if } PA_{j,t} \leq 343\mu\text{g/m}^3 \\ 0.46 * PA_{j,t} + 0.(3.93e-4)PA_{j,t}^2 + 2.97, & \text{otherwise} \end{cases}$$

where $PA_{j,t}$ is the ambient PM2.5 measured by PurpleAir sensor j at time t and $H_{j,t}$ is the relative humidity (between 0 and 1) also measured by the PurpleAir device. This correction helps reduce concerns about heteroskedasticity due to larger errors in PurpleAir readings at high levels of PM2.5. Future work involves a more complex predictive model.

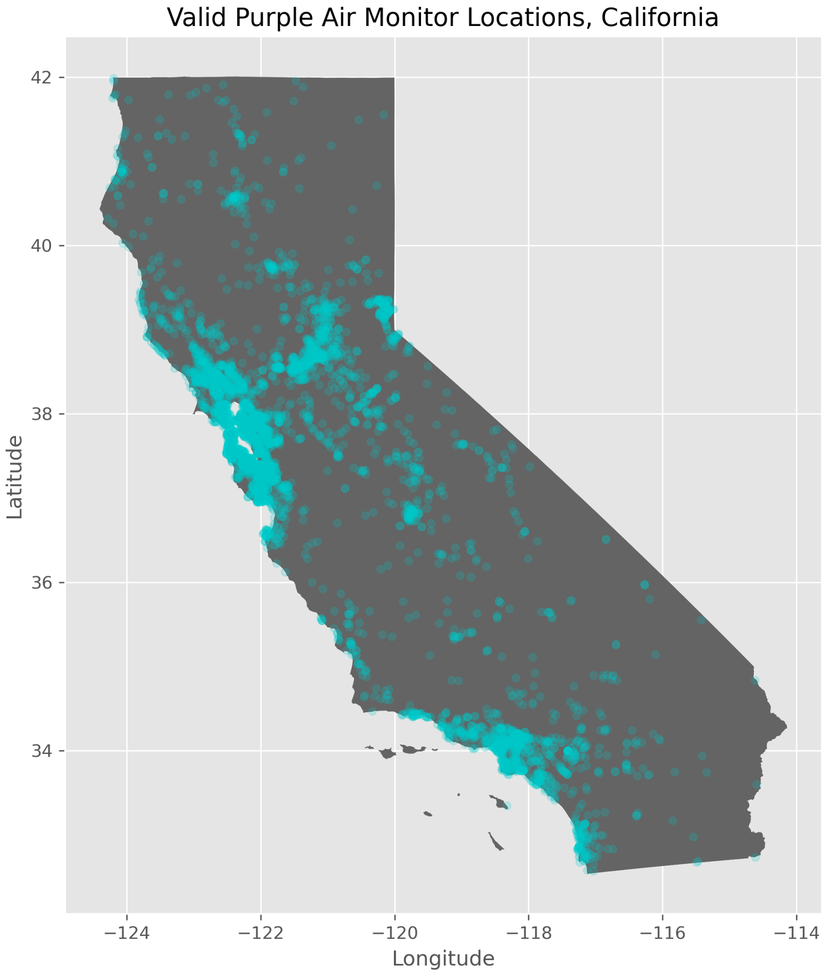

Figure 2: Map of PurpleAir sensors offering public, outdoor PM2.5 measurements. These are sensors that have offered any data in the past, so many are now inactive. The historical data is used in this analysis.

PurpleAir and a NAAQS Monitor. As an example, figure 3 depicts the high number of PurpleAir sensors in the vacinity of one of Los Angeles's NAAQS monitors. I select the PurpleAir sensors in pink, those within 5 miles of the monitor.

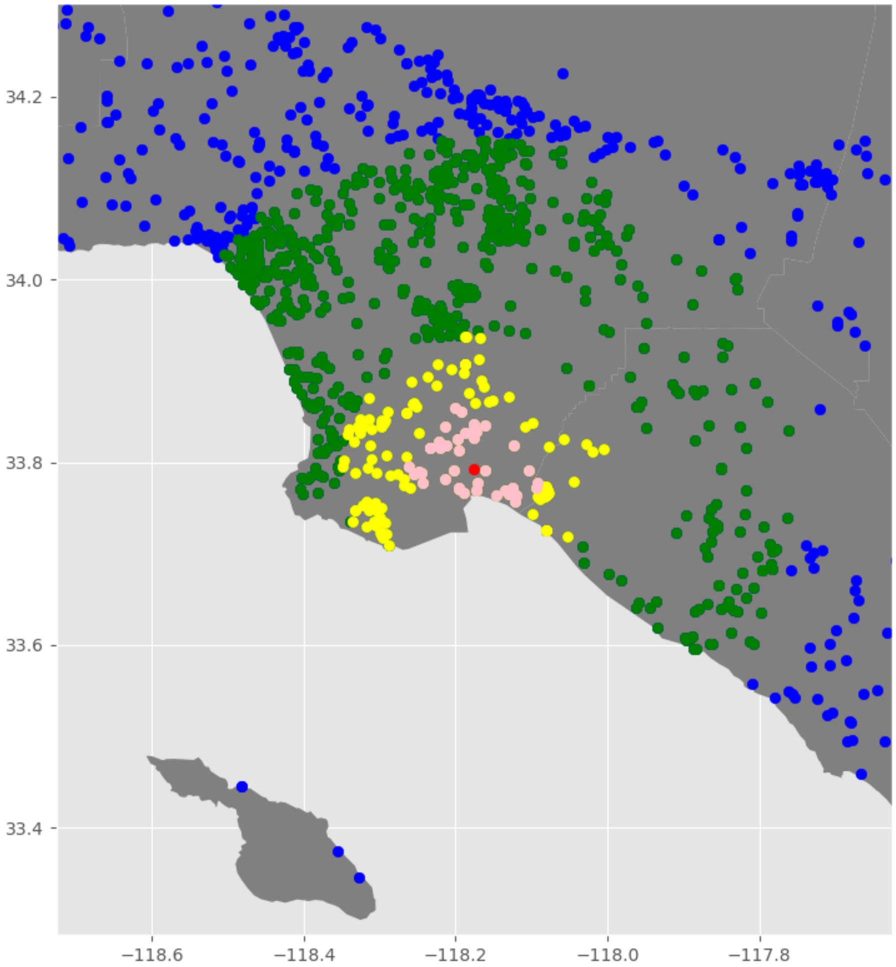

Figure 3: Map of an EPA NAAQS-primary monitoring station (red) surrounded by PurpleAir monitors within 5-mile (pink), 10-mile (yellow), and 25-mile (green) radii. This preliminary analysis uses the PurpleAir sensors within 5 miles (pink markers).

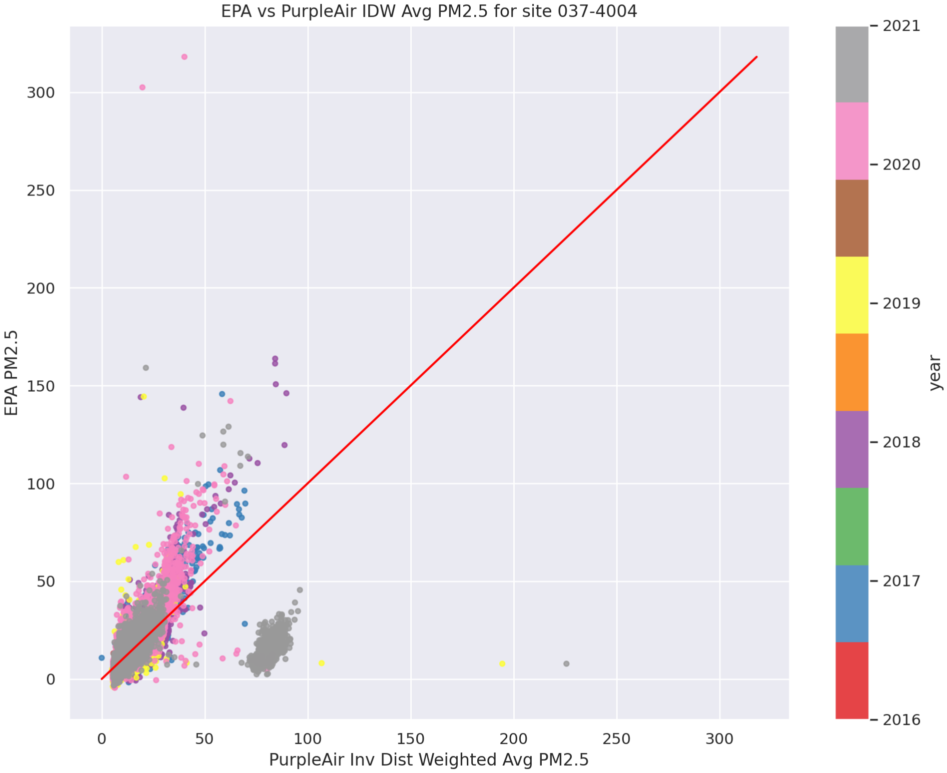

Figure 4: Scatter plot comparing reported hourly PM2.5 measurements: the x-axis represents the IDW-weighted average of PurpleAir measurements, the y-axis represents reported NAAQS-primary monitor measurements. The red line is a 45° line, representing perfect correlation between the PurpleAir average and the NAAQS-primary monitor. For this site, we can see the PurpleAir average is skewed to the right for readings from 2021. This is likely from a PurpleAir sensor coming online that was placed near a source of localized pollution that is not being picked up by the NAAQS-primary monitor.

3.3 Estimating Regulation-grade Readings with PurpleAir Sensors

I use inverse-distance weighting (IDW) with a power of 1 on the denominator. In their discussion of IDW in ambient pollution estimation, de Mesnard (2013) derives that a power between 1 and 3 is appropriate for diffuse particle distributions. I use a power of 1 here because I find evidence that some PurpleAir sensors that are very close to the NAAQS monitor are not very good predictors of the monitor's PM2.5 levels. These PurpleAir sensors seem to still have reliable estimates[10], and anecdotally seem to have very high PM2.5 readings when they disagree with the NAAQS monitor. This suggests they are measuring localized pollution that is out of the range of the NAAQS monitor (these sensors could be located next to a highway, for example).

I plan to fix this issue with a future implementation of a better prediction model[11]. In this iteration, I have implemented the sub-optimal IDW average to avoid excluding entire PurpleAir sensors and removing potentially useful sensor data.

4 Theoretical & Empirical Framework

4.1 Estimating PM2.5 at the NAAQS Monitor with PurpleAir Sensors

To examine the difference between reported pollution and that which is missing from the NAAQS monitors' dataset, I need some measure of ambient PM2.5 levels around the monitor. A good place to start is inverse distance weighting (IDW) to create a weighted average of PurpleAir sensors that can tell us about the PM2.5 levels near the NAAQS monitor.[12] To

[10] Each PurpleAir sensor has two internal sensors that measure PM2.5. Reliability of the PurpleAir sensors is determined by the agreement of the two sensors' hourly averages.

[11] See Appendix section 6.3

[12] This is not a good place to end, however. IDW produces fairly poor estimates of PM2.5 at the location of the NAAQS monitor (before OLS). I plan to implement a more rich prediction model using wind speed and direction – see the Appendix section 6.3 for model and data notes on this method. Ultimately, both satellite

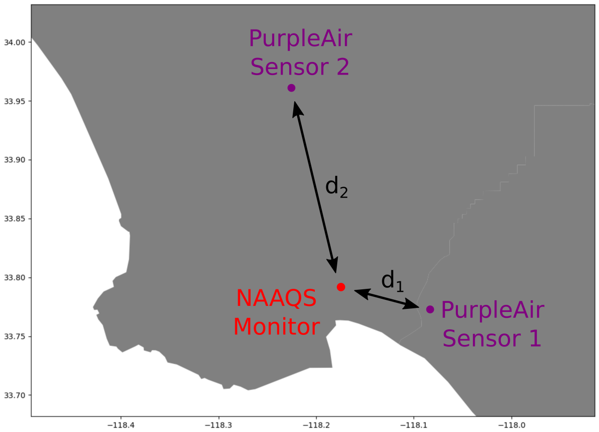

Figure 5: Example distances of two selected PurpleAir sensors near a NAAQS monitor in Los Angeles, CA.

help make the estimation process concrete, I will use a monitor in Los Angeles, CA as an example (see Fig. 5 for a visual representation).

Consider an EPA NAAQS monitor that measures PM2.5 concentration EPA_T at time t. Let $PA_{j,t}$ be PM2.5 concentration as measured by PurpleAir sensor j at time t at a distance d_j from the NAAQS monitor. Assume that there are J_t active PurpleAir sensors in the vicinity of the NAAQS monitor at some time t, each with their own PM measurements and distance. Then the Inverse-Distance Weighted average PurpleAir PM reading $PA_{j,t}^{IDW}$ is

$$PA_t^{IDW} = \sum_{j=1}^{J_t} \frac{\frac{1}{d_j} \cdot PA_{j,t}}{\sum_{j}^{J_t} \frac{1}{d_j}} = \sum_{j=1}^{J^t} w_{j,t} \cdot PA_{j,t} \qquad (1)$$

Note here that I am explicitly using the exponent of one on the distance to balance the desire to have closer sensors provide more weight but avoid having extraordinarily large weights on sensors that are relatively close to the monitor. Also note that the number of PurpleAir sensors J_t changes over time as sensors come online and exit. This means the weights of the weighted average need to be calculated separately for each period. This IDW average PurpleAir measurement provides me with a measure of ambient PM2.5 variation in the

and PurpleAir data could be combined with NAAQS monitor data to provide a more accurate depiction of pollution in the US.

vicinity of the NAAQS monitor at all the possible times there exists at least one PurpleAir monitor in the area. This helps provide good coverage and make implementing OLS in the next step easier.

OLS Prediction. Even if the PurpleAir sensors are highly accurate at measuring PM2.5 concentrations at their location, they may still biased as measures of PM2.5 near the NAAQS monitor. For example, someone might put up a PurpleAir sensor on a telephone pole right next to a highway, which might have high average PM compared to where the NAAQS monitor is placed. To help ensure an unbiased approximation of the missing monitor measurements *at the location of the NAAQS monitor*, I use OLS to regress the NAAQS monitor data on the weighted average PurpleAir data.

$$EPA_t = \beta_0 + \beta_1 PA_t^{IDW} + \varepsilon_t \tag{2}$$

I then predict missing NAAQS monitor data (out of sample) and combine the predicted PM estimate with the reported readings. Formally, suppose $\mathcal{M}$ is the set of **missing** times that we do not have a PM record from the NAAQS monitor (i.e., EPA_t does not exist for $t \in \mathcal{M}$). Let $\mathcal{N}$ be the **non-missing** (reported) times for the NAAQS monitor ($EPA_t \in \mathbb{R}\ \forall t \in \mathcal{N}$). For simplicity, assume that some PurpleAir data are available at all times ($PA_t^{IDW} \in \mathbb{R}_+\ \forall t \in \mathcal{M} \bigcup \mathcal{N}$). Using the PurpleAir data during the missing times, I predict the missing NAAQS data:

$$\widehat{EPA}_t = \hat{\beta}_0 + \hat{\beta}_1 PA_t^{IDW} \quad \forall t \in \mathcal{M} \tag{3}$$

I also estimate the 95% lower and upper bound on each predicted value ($\widehat{EPA}_t^L$ and $\widehat{EPA}_t^U$) to use later in calculating lower and upper bounds on the design values for each location.

4.2 Estimating Design Values

The NAAQS specify two primary statistics to determine if an area is in or out of attainment. These two statistics are referred to as the Annual and 24-hour Design Values. The annual design value is a 3-year average of annual averages of daily average PM2.5 levels. The 24-hour design value is a 3-year average of the annual 98th percentile of daily averages. There are also considerations about the proportion of allowed missing recordings (75%). See Appendix Section 6.4 for details on constructing these design values.

I first construct design values using the reported NAAQS monitor data. In the style of Fowlie et al. (2019), I will call these annual and 24-hour *pseudo design values*; DV_A and DV_H. I do not use the reported design values because there is a more complex negotiation that happens between the EPA and state regulators that can change some of the numbers. In order to compare design values between reported and reported + imputed datasets, I must calculate them on the actual reported PM2.5 readings from the NAAQS monitor.

I then create predicted NAAQS PM values using PurpleAir data as described in the previous section. I replace all missing NAAQS monitor values possible with the predicted values, leaving the original valid NAAQS readings. With this new imputed dataset, I calculate the new imputed design values: $\widetilde{\mathrm{DV}}_A$ and $\widetilde{\mathrm{DV}}_H$. Subtracting the original design value from the imputed design value gives an estimate of design value bias caused by missing data.

$$\mathrm{bias}^{miss}(\mathrm{DV}_A) \approx \widetilde{\mathrm{DV}}_A - \mathrm{DV}_A \tag{4}$$

$$\mathrm{bias}^{miss}(\mathrm{DV}_H) \approx \widetilde{\mathrm{DV}}_H - \mathrm{DV}_H \tag{5}$$

If this quantity is positive, then there is support that there is under-reporting of pollution via allowed missing data.

NAAQS. The standards set out in the National Ambient Air Quality Standards give specific thresholds to compare the design values to. Above the thresholds are considered nonattainment. The primary standard for the annual design value is 12 μg/m^3, and has a secondary standard of 12 μg/m^3. The 24-hour standard is 35 μg/m^3.

Exceptional Events. As mentioned in section 3, there are many events for which the EPA allows local regulators to exclude their readings from design value calculations. These readings were removed from the dataset before any design value calculations and do not get imputed.

4.3 Confidence Intervals & Inference

To gain an understanding of significance of the results, I also propagate the upper and lower bounds of the predicted observations through the design value calculation (similar to Fowlie et al. (2019), but they were estimating out of sample prediction errors). Propagating the bounds of the prediction intervals this way provides a confidence interval – upper and lower bounds on the imputed design values: $\widetilde{\mathrm{DV}}_A^{upper}$, $\widetilde{\mathrm{DV}}_A^{lower}$, $\widetilde{\mathrm{DV}_H}^{upper}$, and $\widetilde{\mathrm{DV}_H}^{lower}$.

Multiple Hypothesis Testing and Significance Levels. Suppose I am interested in confidence intervals at a level of significance α. If I simply propagate prediction intervals of level α through the design value calculation, I would end up with α-level confidence intervals that are valid for single-hypothesis inference – overstating the significance of the results and underestimating the true α-level confidence intervals. In total, I am conducting 198 different hypothesis tests, testing each of 15 sites for their two design values on up to 12 quarters. Following Benjamini et al. (2006), I use their two-stage procedure for selecting the correct significance level of the prediction intervals to propagate through the design value calculation, the correct level that will represent an α-level confidence interval given the multiplicity of hypotheses.[13]

For this paper, I present 95% confidence intervals (0.05 significance level), which correspond to a corrected significance level of approximately 0.002 applied to the prediction intervals and propagated through the design value calculation. See Appendix Section 6.5 for

[13] See Anderson (2008) for an helpful description of applying the two-stage procedure.

a full description of the iterative procedure used to attain the corrected significance level.

5 Results and Discussion

I conducted 15 design value tests on 15 different NAAQS monitors. To create the predicted design values for each NAAQS monitor, I regressed the NAAQS monitor PM2.5 readings on the weighted average PurpleAir readings. An example of this regression is below for one of the Los Angeles monitors. I ran regressions both with and without a constant, but because I wanted the prediction errors to have mean zero, I chose to use model (2) in the creation of the design values. Note that the R^2 value in model (1) is much higher, however R^2 values between regressions with constant and those without are not comparable due to the difference in denominators.

Table 1: 037-4004 NAAQS Monitor PM2.5 on Weighted Average PurpleAir PM2.5

	Reported NAAQS Monitor PM2.5	
	(1)	(2)
const		6.924^{***}
		(0.076)
PurpleAir IDW Average	0.741^{***}	0.444^{***}
	(0.003)	(0.004)
Preferred	No	Yes
Observations	36,813	36,813
R^2	0.658	0.240
Adjusted R^2	0.658	0.240
Residual Std. Error	9.870	8.920
F Statistic	70924.412^{***}	11642.169^{***}
Note:		*p<0.1; **p<0.05; ***p<0.01

The number of observations here are the number of hours between 2016 and 2021 where neither the NAAQS monitor or the weighted average PurpleAir readings were missing. The slope in both models is less than one, indicating that PurpleAir tends to measure higher PM2.5 concentrations. This could be a selection issue – consumers who choose to buy a

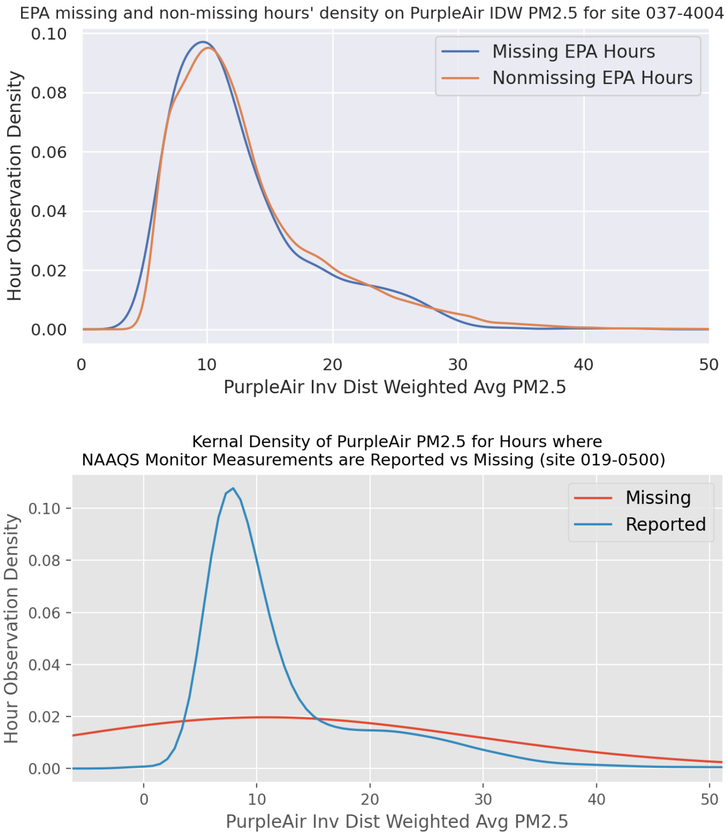

Figure 6: Comparison of estimated kernel densities of PM2.5 concentration for two sets of hours: reported (blue) and missing (red) hourly observations of the NAAQS monitor. Both densities use the hourly PurpleAir PM2.5 concentration estimates for this site, calculated using the IDW average of PurpleAir sensors within 5 miles of the NAAQS monitor location. The top image is for a monitor at a site in LA, and the bottom is at a site in Fresno, CA.

PurpleAir monitor and place it outside their house are probably concerned about pollution – or a measurement difference between the types of devices.

For all sites, I generated kernel density plots like those in Fig. 6. Only one of the 15 sites I tested had noticeable differences in the PM2.5 distributions for missing and reported hours (see the appendix for all sites' plots). However, the design values are the policy-relevant statistic of those distributions. Looking at the top and bottom figures below however, we can guess that the means and 98th percentile might be similar in the top case, and could plausibly be different in the bottom case.

Annual Design Values. For each site tested, Figure 7 graphically shows the difference between the pseudo design value (calculated using the reported data) and the imputed design value – what I believe is an estimate on the design value bias due to low reporting standards. We can see that not all sites have points for all quarters. This is common in the full results (Appendix Table 4) and this occurs because the site has too many invalid days in one of the 12 quarters used to calculate the design value for that quarter (the completeness criteria are violated).

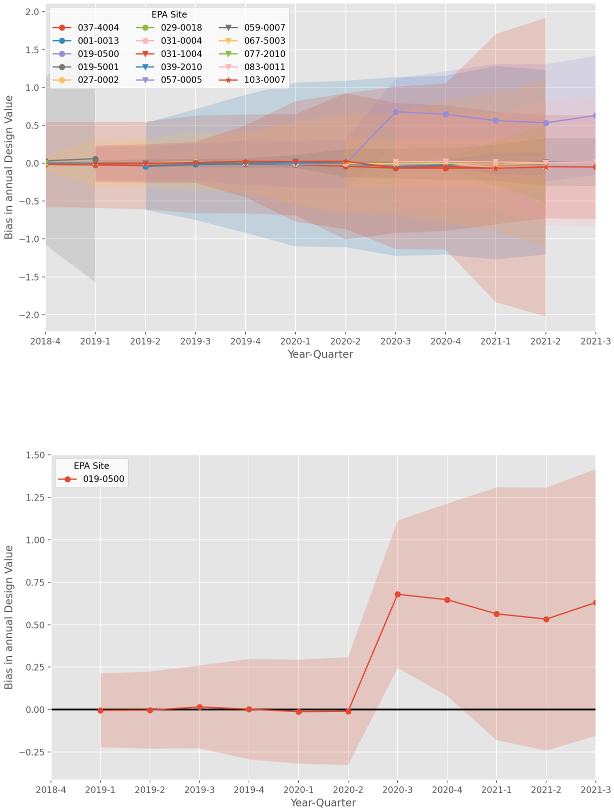

Figure 7: Difference between the annual pseudo design value from reported data and the predicted design value from filling in missing observations with predicted PM2.5 measurements. Fill represents the 95% confidence intervals after correcting for multiple hypothesis testing. Top: all 15 sites; Bottom: site 019-0500.

Note that design values are rounded to the nearest integer. So we can see that most of the differences in the graph are much smaller than possible rounding errors and have relatively wide confidence intervals containing zero difference. There is one site that jumps out (site 019-0500, Fresno, CA). Looking at that site in the bottom of Figure 7, we can see that the confidence intervals are significantly different from zero, but are not meaningfully large because the predicted difference in quarter 3 of 2020 is still less than 1 $\mu g/m^3$ and are not likely to flip the attainment status.

24-hour Design Values. Figure 8 depicts the difference in 24-hour design values. We can see in 2020 and 2021, there were fairly large and significant differences between the design values for site 019-0500 (Fresno). A positive value indicates that the imputed design value is larger than the design value based only on reported data. This means that having a higher data collection standard could have increased Fresno's design value significantly in these quarters. The increase in the Fresno 24-hour design value for the third quarter of 2020 is between 8.55 and 11.7 μg/m^3. This represents a meaningful difference because the 24-hour design value threshold is 35 μg/m^2 and if an area is near the threshold, a difference of 9 or 10 units could change the attainment status.

24-hour Design Value Confidence Intervals. The design value confidence intervals for the annual design values are nearly perfectly symmetric, while the 24-hour design value confidence intervals are highly asymmetric. This is expected because the annual design value is very close to a simply average of the daily observations, whereas the 24-hour design value is an order statistic that is sensitive to the placement of just a few predicted values that are higher in PM2.5 than the reported data. See Appendix Section 6.4 for a full discussion and example of why we can expect the lower bound of the 24-hour design value confidence interval to be much closer to the predicted value while the upper bound can be further away.

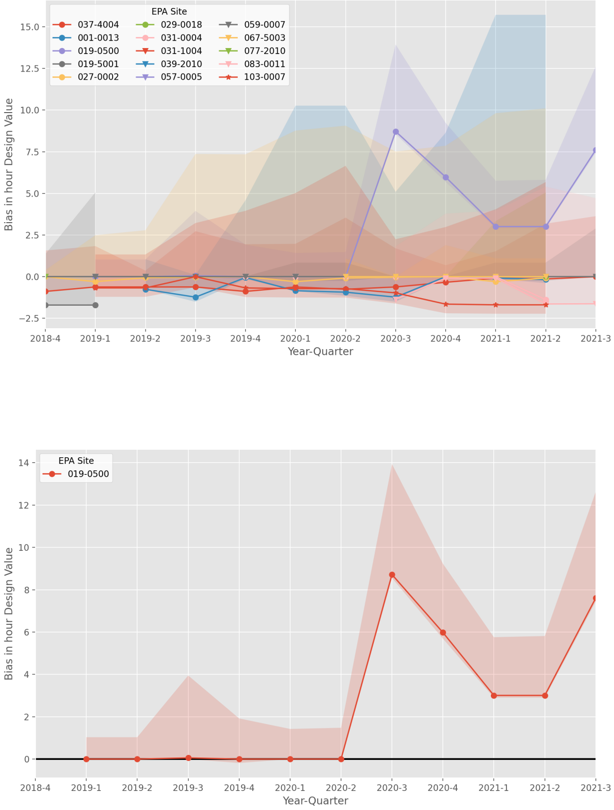

Figure 8: Difference between the 24-hour (98th percentile) pseudo design value from reported data and the predicted design value from filling in missing observations with predicted PM2.5 measurements. Fill represents the 95% confidence intervals after correcting for multiple hypothesis testing. Top: all 15 sites; Bottom: site 019-0500.

As noted in Section 4.3, to 95% construct confidence intervals that are robust to testing many hypotheses, I use an adjusted significance level to produce the OLS prediction intervals. In the specific case of this paper, the significance level ends up being about 0.002.

Magnitude of the Design Value Statistic. To help get a sense of the size of gap between the raw design values and the design values with imputed missing values, I have estimated the maximum amount of manipulation left in the raw, reported EPA data. For a moment, assume all PM2.5 data must be uploaded to the EPA database for each site at the end of each quarter. Also assume that each site has a data manager who can choose which observations to remove from the data before uploading. Under the NAAQS rules, all sites can omit up to 25% of days from each quarter, and up to 25% of hours from each day that is reported. To my knowledge, there are no rules about which observations to choose.

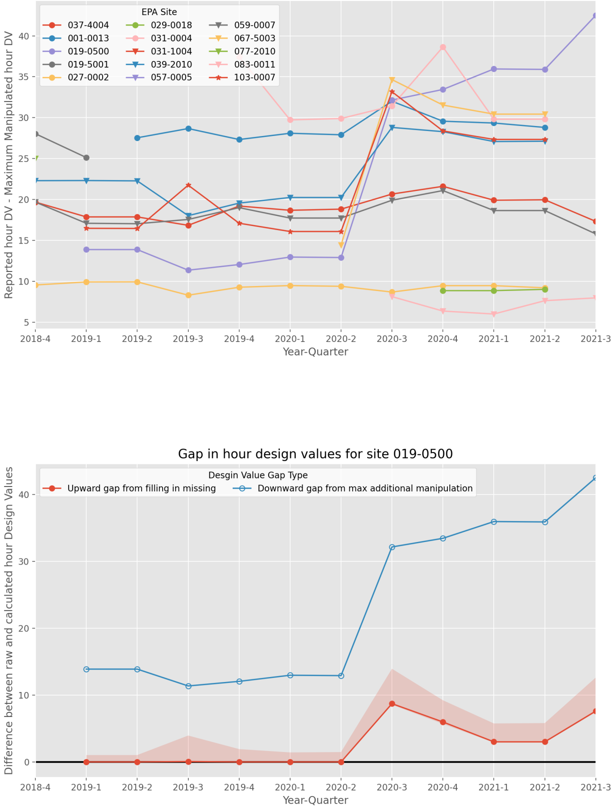

Figure 9: Room left for manipulation in raw EPA data – the number of pollution units (μg/m^3) that the 24-hour design value can be reduced by omitting more hourly observations, down the minimum completeness limit set by the NAAQS rules. This is the gap between the pseudo design value and a new design value caluclated from fewer observations. Top: room left for manipulation for all sites. Bottom: room left for manipulation (blue) compared to the estimated gap due to actual missing data (red).

Further assume that this fictitious data manager desires to reduce the annual and 24-hour design values as much as possible for their site. An easy way to reduce both the average and the 98th percentile of days is to remove the top PM2.5 hourly observations from each day until 75% of hours remain, create new daily averages from the hours left, and drop the top PM2.5 days from the quarter until the quarter only has 75% of days left. I have assumed the role of this fictitious data manager – the top of figure 9 shows how much the 24-hour design value can be reduced by dropping hourly and daily observations this way. The numbers are quite large with most sites having values above 15 μg/m^3 in all available quarters. Keeping in mind that the 24-hour NAAQS threshold is currently 35 μg/m^3, this shows that the 75% completeness rule leaves a meaningful amount of manipulation on the table, even for sites that have missing values already.

The bottom of figure 9 shows the Fresno site, comparing how the room left for manipulation in the reported data compares to the estimated gap due to actual missing observations. For Fresno specifically, we can see that there is a lot of room left for intentional manipulation if a bad actor had the ability to drop observations after recording the data. However, because this exercise used perfect hindsight (I can see the measured observations), it is likely an upper bound on the amount of possible manipulation. To my knowledge, there is no evidence of anyone deleting or intentionally omitting measurements that have already been recorded. Still, the blue line in the bottom panel of figure 9 gives us a sense of scale for how the missing observations affected the design value in 2020 and 2021.

Using OLS to predict the missing EPA hourly observations using PurpleAir data results in prediction errors that are mean 0. Since we are taking 3-year averages, it is plausible that the errors from the imputation method would average out. These combined with the lower bounds being well above zero suggest there was indeed under-reporting of pollution in Fresno in 2020 and 2021. I cannot say whether or not that it happened intentionally or by mere chance of turning off the monitor at times that would likely have recorded higher pollution.

Is this a meaningful difference in the design value? Is the magnitude of the bias from missing data meaningful for Fresno? Fresno has had nonattainment status for many years and is mentioned within the literature as having notable pollution measurement issues. So though Fresno's recorded design value would have been higher, they were already well over the 24-hour design threshold of 35 μg/m^3, having pseudo design values in the 50s and 60s in 2020 and 2021. However, there is evidence from previous research that pollution in non-attainment areas has been decreasing at significantly faster rates since the introduction of the CAA (Currie et al., 2020). So if Fresno were to fall below the threshold and into attainment based on reported values, a downward bias of 8-11 μg/m^3 in the 24-hour design value could be enough to flip the attainment status. So while this bias does not play a pivotal role in Fresno at this time, it may be important in the future and may be important for other areas not currently in this study that are closer to the nonattainment threshold.

It should also be noted that the EPA has been recently discussing both the lowering of the NAAQS (more areas would be nonattainment) and decreasing the exclusion of extreme events like wildfires (which could significantly increase some areas' design values). So large biases due to legally allowed omitted data could play an important role in determining attainment status in the near future.

It is also of note that there was only 1 site of 15 that showed significant bias due to omitted data. However, each site and each state acting as both the regulated and the monitor might have slightly different incentives depending on the local political and administrative climate. One might be concerned with individual actors or regions trying to get around the regulations. Since the Clean Air Act and the NAAQS are designed to ensure that all US residents can share in a minimum standard, it seems relevant to be able to diagnose the occurrence and size of mismeasurement, even if it is occurring at only a few sites in the NAAQS network. Especially considering how sparse regulatory air quality monitors are in the US, a single systematic mismeasurement of pollution can have impacts on millions of people.

References

91st US Congress (1970, December). Clean Air Amendments of 1970.

Anderson, M. L. (2008, December). 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*(484), 1481–1495. Publisher: Taylor & Francis _eprint: https://doi.org/10.1198/016214508000000841.

Benjamini, Y., A. M. Krieger, and D. Yekutieli (2006). Adaptive Linear Step-up Procedures That Control the False Discovery Rate. *Biometrika 93*(3), 491–507. Publisher: [Oxford University Press, Biometrika Trust].

Currie, J., J. Voorheis, and R. Walker (2020, January). What Caused Racial Disparities in Particulate Exposure to Fall? New Evidence from the Clean Air Act and Satellite-Based Measures of Air Quality. Working Paper 26659, National Bureau of Economic Research.

de Mesnard, L. (2013, March). Pollution models and inverse distance weighting: Some critical remarks. *Computers & Geosciences 52*, 459–469.

EPA, E. P. A. (1990). Event Qualifier. https://aqs.epa.gov/aqsweb/helpfiles/event_qualifier.htm.

EPA, E. P. A. (2017). Appendix N to Part 50, Title 40 – Interpretation of the National Ambient Air Quality Standards for PM2.5.

Feng, S., D. Gao, F. Liao, F. Zhou, and X. Wang (2016, June). The health effects of ambient PM2.5 and potential mechanisms. *Ecotoxicology and Environmental Safety 128*, 67–74.

Fowlie, M., E. Rubin, and R. Walker (2019, May). Bringing Satellite-Based Air Quality Estimates Down to Earth. *AEA Papers and Proceedings 109*, 283–288.

Grainger, C. and A. Schreiber (2019, May). Discrimination in Ambient Air Pollution Monitoring? *AEA Papers and Proceedings 109*, 277–282.

Grainger, C., A. Schreiber, and W. Chang (2019). Do Regulators Strategically Avoid Pollution Hotspots when Siting Monitors? Evidence from Remote Sensing of Air Pollution. Working Paper.

Mu, Y., E. A. Rubin, and E. Zou (2021, April). What's Missing in Environmental (Self-)Monitoring: Evidence from Strategic Shutdowns of Pollution Monitors. Working Paper 28735, National Bureau of Economic Research.

Sullivan, D. M. and A. Krupnick (2018). Using satellite data to fill the gaps in the US air pollution monitoring network. *Resources for the Future Working Paper*, 18–21.

Zou, E. Y. (2021, July). Unwatched Pollution: The Effect of Intermittent Monitoring on Air Quality. *American Economic Review 111*(7), 2101–2126.

6 Appendix

6.1 Faculty Feedback

Comment 1. The economic story behind an empirical observation is most often (though not always) a story of incentives. You describe data anomalies, but without setting the context in terms of incentives. Critically, we would like to know more if possible about who has the discretion to decide what data to omit, and what their incentives are. The introduction of the paper presumes a familiarity with local actor incentives to stay in attainment, but this may be assuming too much for a general economic audience, as opposed to someone specializing in air pollution. So, this comment pertains both to explaining the importance of attainment status and who the key stakeholders are to the broader audience, and offering additional insight on the details of who the "data managers" are, who hires them, and whether they are plausibly subject to pressure.

Response to Comment 1.

Comment 2. Another exercise would be to ask how much room is left for additional manipulation. That is, suppose a data controller cared about boosting numbers and wanted to drop as many data points as would be allowed. Given the data that are being reported and the slack in the number dropped, how much of a difference would additional manipulation make? This is sort of a scale exercise for asking how useful manipulation might be.

Response to Comment 2.

Comment 3. A related big picture comment is to ask why the objective is to infer an estimate for the missing data, rather than simply look for evidence of manipulation. (To be sure, they are related, but we think they are not identical.) You might have begun, for example, by predicting whether an EPA data point was missing (binary outcome), and asking whether high purple air readings make it more likely an observation is missing.

More generally, one might argue that this paper is a very interesting measurement exercise in search of an economic application. The possibility of flipping attainment status is one way to turn this into an economics/policy narrative, but you might also consider something more broad, like stating that "effective regulation requires reliable measurement," and then asking if the measure here is manipulable. Obviously we know that there are other types of manipulation (from prior literature), so your contribution is to look for the possibility of manipulation in this particular mechanism.

Comment 4. Isn't a kernel regression the right approach here? Asymptotic performance of current approach (inverse distance) may not work?

Comment 5. First table seems to involve regression to the mean (Purple Air a rhs variable, with measurement error).

Comment 6. Multiple testing (however many sites...)? If drawing inference that a *particular* site is engaged in manipulation, current version of inference seems problematic.

Comment 7. Confidence intervals for non-compliant site highly asymmetric—not clear why?

Response to Comment 7. A brief explanation of the asymmetry of the confidence intervals has been added to section 5 and a detailed discussion has been added to appendix section 6.4.

6.2 Research Project To-do's

Future Work

- I have only tested 15 of 388 possible sites. Now that I have written the bulk of the code to manage the data, most of the future work lies in acquiring the rest of the data for the NAAQS monitors and PurpleAir monitors.
- I am in the process of negotiating a data usage agreement with PurpleAir staff to get the entire historical dataset of all US sensors. This is the main bottleneck.
- Some of the NAAQS monitors report on a less frequent basis than hourly, so the analysis code will need to be generalized for other reporting frequencies.
- I have written code to download wind velocity data, but parsing the files to get wind velocity at a particular location and time requires more work. This will be used in my predictive model for missing PM2.5 data at the NAAQS monitor (see next section).

6.3 Improving Prediction of PM2.5 at the NAAQS Monitor Location

A more realistic model of predicting pollution at the EPA monitor could be used. Using wind direction (and possibly wind speed), we can intuitively put more weight on PurpleAir sensors that are North of the EPA monitor when the wind is blowing South. This model allows for mroe flexibility in dropping some sensors that may be measuring hyper-local pollution and are not good predictors of the EPA monitor's PM2.5 readings.

$$EPA_{i,t} = \gamma_{i,0} + \sum_{j \in J_i} \sum_{k=1}^{7} \gamma_{i,j,k} PA_{j,t} \cdot Winddir_{i,t,k} + u_{i,t}$$

- Each EPA monitor i has it's own set of weights for the PA sensors around it.
- Analysis is done at the quarter level; suppressing quarter subscript.

- t is a unique hour within a given quarter.
- EPA monitor i at time t reads PM2.5 pollution $EPA_{i,t}$.
- For each EPA monitor i, there are J_i Purple Air monitors within a 5-mile radius.
- Purple Air monitor $j \in J_i$ at time t reads PM2.5 pollution $PA_{j,t}$.
- $Winddir_{i,t,k}$ is a wind direction indicator; 1 if the prevailing wind near station i at time t is in the k^{th} bucket (of 8 buckets).
- I will also estimate a version with wind speed interacted in the sum. This could allow for sensors further away to have more predictive power when the winds are strong.
- This regression could be run as a LASSO first to determine which of the interactions for each PurpleAir sensor have the most predictive power.

6.4 Design Values

Notes about design values:

- Valid days: a day that has 18 or more valid hours in it.
- Valid quarters: a 3-month period that has at least 67. 68, or 69 valid days in it.
- Valid 3-year period: a 12-quarter period that has 12 valid quarters in it.
- A design value is only valid if its 3-year period is valid.
- All averages below assume there might be some data missing.
- Design values are based on quarters, so each quarter has a rolling average over the last 3 years before.

To construct the annual design value for a given quarter:

- Construct an average for every day.
- Construct an annual average of daily averages for every previous 12-month period before this quarter.
- Construct the 3-year average of those annual averages. Because some years have a different number of real or valid days in them, you cannot take a simple 3-year average of all available days or hours.

To construct the 24-hour design value:

- Construct an average for every day.
- Construct a 98th percentile of all daily averages in the 4-quarter period ending in this quarter. To avoid ambiguous design value construction, the EPA provides a lookup table for which n^{th} maximum daily value to take when constructing the 98th percentile.
- Take a 3-year average of the 98th percentiles.

Asymmetry of the design value confidence intervals:

The confidence intervals for the annual design value are fairly symmetric (in fact, not perfectly symmetric). This is expected because the the confidence interval for each OLS prediction is symmetric and the average design value is nearly a simple average (very close to the first moment of the hourly observations). "Nearly a simple average" because the daily design value is not a simple average of all hourly observations; it takes daily averages over 18 to 24 possibly valid hourly observations, and is a simple average of all valid daily averages. Since some daily averages are calculated on fewer than 24 hourly observations and therefore have different denominators, it is close to, but not exactly, a simple average.

The confidence intervals for the 24-hour design value are highly asymmetric. This is also expected as the 24-hour design value is an order statistic of the distribution of daily averages, not a moment of the distribution. The 24-hour design value takes the 98th percentile of daily averages, thus it is a function of a single daily observation of the right tail of the distribution and is highly sensitive to inserting new values further to the right of it (inserting higher PM2.5 daily observations).

For clarity, consider the following simplied example: assume that we only need one year's worth of daily PM2.5 observations to calculate a 24-hour design value for the fictional Albatross County. Further assume that we start with 100 valid daily PM2.5 observations in a single year, used to calculate the 24-hour design value. The design value will be the 98th percentile – **the third highest** – of the daily observations.

Say we can fill in 100 missing days with predicted values from PurpleAir measurements. The total is now 200 daily observations, of which, we would choose the sixth highest value for the 98th percentile. Assume Table 2 display the top three daily observations of the original 100 EPA daily observations and top six daily observations of the 100 filled in missing days. The table also includes symmetric upper and lower bounds for the prediction intervals of the daily

Table 2: Hypothetical PM2.5 daily averages from EPA reported measurements and PurpleAir-imputed daily averages that were missing from the EPA data. Column 1: top three original EPA PM2.5 measurements. Columns 2-4: The six added daily averages that were greater than the EPA top 3 daily averages, where added daily averages were filled-in from PurpleAir.

EPA original	PA Upper Bound	PA Predicted	PA Lower Bound
30	50	40	30
29	50	40	30
15	50	40	30
	50	40	30
	45	35	25
	40	30	20

Now we can examine the 98th percentile in each of the following four situations:

1. the original reported EPA data (column 1) which gives us the original design value
2. the EPA data with the upper bound of predicted data (columns 1 and 2) – the upper bound of the predicted design value
3. the EPA data with the predicted data (columns 1 and 3) – the predicted design value
4. the EPA data with the lower bound of predicted data (columns 1 and 4) – the lower bound of the predicted design value

In Table 3, the data is combined and sorted for each of the four cases. We can see that the 98th percentile (bolded) for all three combined datasets (columns 2-4) is larger than the 98th percentile of the original data (column 1). We can also see that both the predicted 98th percentile and the lower bound of the prediction are equal to 30, while the upper bound of the prediction is 40. This hypothetical example details a simple property of the order statistics for this paper's use case: when adding new predicted data points to the original data, the confidence interval of the 98th percentile (any order statistics in fact) can be highly asymmetric, especially when the predicted values have relatively large prediction intervals.

Table 3: Hypothetical top PM2.5 daily observations, combined and sorted from Table 2. Column 1: Sorted top three original EPA daily observations (no new data, 100 days). Columns 2-4: Sorted top observations, combining original EPA data with predicted data (the upper bound of the predictions, the prediction, and the lower bound of the predictions, all with 200 total days). The original EPA data is in red and the 98th percentile of each dataset is bolded.

EPA original	Upper Bound	Predicted	Lower Bound
30	50	40	30
30	50	40	30
15	50	40	30
	50	40	30
	45	35	30
	40	**30**	**30**
	30	30	25
	30	30	20
	15	15	15

6.5 Adjusting Confidence Intervals for Multiple Hypotheses

Benjamini et al. (2006) show that we can use a two-stage procedure for correcting our p-value threshold for the issue of multiple hypothesis testing. In this procedure, you first estimate your statistics and test all hypotheses at the uncorrected level of significance. You can then use the number of total and rejected hypotheses to correct the significance threshold and conduct the hypothesis tests again. Given that the practitioner has p-values of the estimated statistics, this allows one to only run the estimation procedure once. Due to the complexity of the design values and the mixture of non-random data with random data,[14] it is significantly less cumbersome to produce confidence intervals at a given level of significance during the estimation process than to produce p-values. The original two-stage method is as follows:

[14] In this application, the raw data reported to the EPA are treated as fixed and the missing data are estimated and have uncertainty.

Let α be the desired level of significance, m be the total number of hypotheses being tested, and m_0 be the true number of correct null hypotheses. Define

$$\alpha' = \frac{\alpha}{1+\alpha}$$

Stage 1: Conduct all m single-hypothesis tests at a level of significance α. Denote r_1 as the number of rejected null hypotheses.

Stage 2: Let $\hat{m}_0 = m - r_1$ be the estimated number of correct null hypotheses. Define

$$\alpha^* = \alpha' \frac{m}{\hat{m}_0}$$

Denote p_i as the p-value from the i^{th} statistic, and let the ordered p-values be $p_{(1)} \leq ... \leq p_{(m)}$. Then define the number of multiple-hypothesis-corrected rejected hypotheses as

$$k = \max\{i : p_{(i)} \leq \alpha^* \frac{i}{m}\}$$

This k can also be found using the step-down procedure, starting at the largest p-value and iterating to lower p-values until $p_{(i)} \leq \alpha^* \frac{i}{m} = (\alpha' \frac{m}{\hat{m}_0}) \frac{i}{m} = \alpha' \frac{i}{\hat{m}_0}$.

In the case of this paper, I did not attain p-values during estimation. But it is relatively simple to re-estimate confidence intervals. So I have created a slightly modified version of the above procedure in order to iterate through successively larger confidence intervals until I converge on the correct estimate of the number of correct null hypotheses. The modified procedure is as follows:

1. Estimate the design value confidence intervals for all m design values using naive α-level prediction upper and lower bounds. (I have $m = 198$ separate design values and use an $\alpha = 0.05$ level of confidence.)

2. Denote r_1 as the number of design values with naive confidence intervals outside of zero. Estimate the number of correct null hypotheses in iteration 1 as $\hat{m}_0^1 = m - r_1$. (I had $r_1 = 24$ confidence intervals that did not include zero, and $\hat{m}_0^1 = 198 - 24 = 174$ estimated correct null hypotheses.)

3. I now need to know if the number of rejected hypotheses r_1 change if I adjust the significance level of the prediction intervals. But I need to know what level of significance to compare to, since the original procedure makes the comparison $p_{(i)} \leq \alpha' \frac{i}{\hat{m}_0}$. I know the r_1^{th} statistic is the least-significant statistic and is closest to becoming insignificant. So testing if r_1 (the number of rejected null hypotheses) changes is the same as testing if the r_1^{th} statistics stays significant. For the r_1^{th} statistic to be significant, we want $p_{(r_1)} \leq \alpha' \frac{r_1}{\hat{m}_0^1}$. This holds if the r_1^{th} single-hypothesis confidence interval of level $\alpha' \frac{r_1}{\hat{m}_0^1}$ does not contain zero. So I rerun the estimation process, creating prediction intervals at the $\alpha' \frac{r_1}{\hat{m}_0^1}$ level. (Here, $\alpha' \frac{r_1}{\hat{m}_0^1} \approx 0.048 \frac{24}{198} \approx 0.0066$)

4. Denote r_2 as the new number of rejected null hypotheses (confidence intervals not containing zero). If $r_2 = r_1$, we have applied the correct confidence intervals and the r_1 statistics are still significant. If $r_2 = 0$, then we can stop and none of the null hypotheses are rejected. If $r_2 < r_1$, then we must re-estimate confidence intervals at the new corrected level using r_2 and $\hat{m}_0^2 = m - r_2$, resulting in r_3 rejected null hypotheses. We then need to check if $r_3 = r_2$.

5. Repeat this until $r_k = r_{k-1}$. We now have a consistent estimate of the true number of rejected null hypotheses. Then, to get the α-level multiple hypothesis confidence intervals, the appropriate single-hypothesis confidence interval to apply is of level $\alpha' \frac{r_{k-1}}{\hat{m}_0^{k-1}}$. (Here, this converges after 2 iterations, going from 24 rejected null hy-

potheses, to 10, then to 8. The final level of significance used on the prediction intervals is $\alpha' \frac{r_3}{\hat{m}_0^3} \approx 0.048 \frac{8}{190} \approx 0.002$. This is a $\frac{0.05}{0.002} = 25$ factor reduction in the significance level applied to the prediction intervals, which are then propagated through the design value calculation to get 95% confidence intervals.)

6.6 Tables

Table 4: Design Value differences and bounds: NAAQS Monitor DV subtracted from the imputed DV. The imputed DV was created by imputing the missing NAAQS Monitor observations with OLS predictions from a regression of the NAAQS PM2.5 values on the weighted average of nearby PurpleAir sensors' PM2.5 values. Positive values represent an upward bias in the missing PM2.5 measurements. The lower (upper) bound of the DV difference is calculated by finding the lower (upper) bound of the imputed DV and holding the "pure" DV constant. The imputed DV lower bound is found by imputing missing values with the 95% lower bound of each predicted observation (the prediction confidence interval), then re-computing the imputed DV with the lower bound data.

County	Site	Year-Quarter	County In Attainment ?	In Attainment based on Psedo DV?	Annual DV Difference	Lower Bound	Upper Bound	Hour DV Difference	Lower Bound	Upper Bound
37	4004	2018-4	N	N	-0.016	0.340	-0.373	-0.885	-0.885	-0.885
37	4004	2019-1	N	N	-0.023	0.337	-0.383	-0.617	-0.114	-0.617
37	4004	2019-2	N	N	-0.031	0.336	-0.397	-0.617	-0.617	-0.617
37	4004	2019-3	N	N	-0.016	0.392	-0.424	-0.617	1.570	-0.617
37	4004	2019-4	N	N	-0.010	0.404	-0.424	-0.885	-0.792	-0.885
37	4004	2020-1	N	N	-0.021	0.403	-0.445	-0.617	0.023	-0.617
37	4004	2020-2	N	N	-0.040	0.573	-0.653	-0.754	-0.356	-0.754
37	4004	2020-3	N	N	-0.065	0.480	-0.609	-0.617	0.113	-0.617

37	4004	2020-4	N	N	-0.065	0.461	-0.591	-0.336	-0.243	-0.336
37	4004	2021-1	N	N	-0.065	0.406	-0.536	-0.083	0.557	-0.083
37	4004	2021-2	N	N	-0.046	0.387	-0.478	-0.138	0.260	-0.138
37	4004	2021-3	N	N	-0.050	0.385	-0.485	0.000	0.730	0.000
31	1004	2018-4	N	–	Invalid DV	–	–	Invalid DV	–	–
31	1004	2019-1	N	–	Invalid DV	–	–	Invalid DV	–	–
31	1004	2019-2	N	–	Invalid DV	–	–	Invalid DV	–	–
31	1004	2019-3	N	–	Invalid DV	–	–	Invalid DV	–	–
31	1004	2019-4	N	–	Invalid DV	–	–	Invalid DV	–	–
31	1004	2020-1	N	–	Invalid DV	–	–	Invalid DV	–	–
31	1004	2020-2	N	–	Invalid DV	–	–	Invalid DV	–	–
31	1004	2020-3	N	–	Invalid DV	–	–	Invalid DV	–	–
31	1004	2020-4	N	–	Invalid DV	–	–	Invalid DV	–	–
31	1004	2021-1	N	–	Invalid DV	–	–	Invalid DV	–	–
57	5	2020-3	Y	–	Invalid DV	–	–	Invalid DV	–	–

57	5	2020-4	Y	–	Invalid DV	–	–	Invalid DV	–	–
57	5	2021-1	Y	–	Invalid DV	–	–	Invalid DV	–	–
57	5	2021-2	Y	–	Invalid DV	–	–	Invalid DV	–	–
57	5	2021-3	Y	–	Invalid DV	–	–	Invalid DV	–	–
1	13	2019-2	Y	N	-0.044	0.320	-0.408	-0.758	0.263	-0.758
1	13	2019-3	Y	N	-0.018	0.446	-0.482	-1.228	-0.684	-1.506
1	13	2019-4	Y	N	-0.008	0.569	-0.586	-0.049	3.469	-0.049
1	13	2020-1	Y	N	-0.018	0.668	-0.703	-0.842	4.791	-0.842
1	13	2020-2	Y	N	-0.010	0.687	-0.708	-0.929	4.791	-1.147
1	13	2020-3	Y	N	-0.044	0.705	-0.793	-1.228	-0.337	-1.506
1	13	2020-4	Y	N	-0.028	0.721	-0.777	0.000	3.740	0.000
1	13	2021-1	Y	N	0.007	0.817	-0.803	-0.083	6.474	-0.083
1	13	2021-2	Y	N	0.012	0.785	-0.760	-0.171	6.474	-0.389
19	500	2018-4	N	–	Invalid DV			Invalid DV		
19	500	2019-1	N	Y	-0.005	0.133	-0.143	0.000	0.252	0.000

19	500	2019-2	N	Y	-0.003	0.141	-0.147	0.000	0.252	0.000
19	500	2019-3	N	Y	0.015	0.170	-0.139	0.058	2.202	0.000
19	500	2019-4	N	Y	0.002	0.190	-0.185	0.000	1.460	-0.024
19	500	2020-1	N	Y	-0.012	0.182	-0.207	0.000	0.335	0.000
19	500	2020-2	N	Y	-0.010	0.191	-0.211	0.000	0.376	0.000
19	500	2020-3	N	N	0.679	0.954	0.403	8.718	11.704	8.556
19	500	2020-4	N	N	0.647	1.006	0.288	5.979	8.281	5.851
19	500	2021-1	N	N	0.564	1.036	0.091	3.007	4.184	2.903
19	500	2021-2	N	N	0.533	1.024	0.042	3.007	4.225	2.903
19	500	2021-3	N	N	0.630	1.129	0.132	7.607	10.557	7.444
19	5001	2018-4	N	N	0.031	0.733	-0.671	-1.703	-1.703	-1.703
19	5001	2019-1	N	N	0.060	1.095	-0.975	-1.702	0.118	-1.702
19	5001	2019-2	N	–	Invalid DV	–	–	Invalid DV	–	–
19	5001	2019-3	N	–	Invalid DV	–	–	Invalid DV	–	–
19	5001	2019-4	N	–	Invalid DV	–	–	Invalid DV	–	–

19	5001	2020-1	N	–	Invalid DV	–	–	Invalid DV	–	–
19	5001	2020-2	N	–	Invalid DV	–	–	Invalid DV	–	–
19	5001	2020-3	N	–	Invalid DV	–	–	Invalid DV	–	–
19	5001	2020-4	N	–	Invalid DV	–	–	Invalid DV	–	–
19	5001	2021-1	N	–	Invalid DV	–	–	Invalid DV	–	–
27	2	2018-4	Y	Y	-0.001	0.058	-0.060	0.000	0.000	0.000
27	2	2019-1	Y	Y	0.002	0.206	-0.203	-0.299	-0.299	-0.299
27	2	2019-2	Y	Y	0.005	0.211	-0.202	-0.061	-0.061	-0.061
27	2	2019-3	Y	Y	0.018	0.257	-0.220	-0.039	4.018	-0.039
27	2	2019-4	Y	Y	0.006	0.250	-0.239	0.000	3.794	0.000
27	2	2020-1	Y	Y	-0.006	0.346	-0.358	-0.299	2.637	-0.299
27	2	2020-2	Y	Y	-0.007	0.406	-0.420	-0.082	2.869	-0.082
27	2	2020-3	Y	Y	-0.009	0.420	-0.438	-0.039	4.018	-0.039
27	2	2020-4	Y	Y	0.021	0.531	-0.490	0.000	4.292	0.000
27	2	2021-1	Y	Y	0.017	0.595	-0.562	-0.299	3.134	-0.299

27	2	2021-2	Y	Y	0.001	0.695	-0.693	-0.082	3.367	-0.082
29	18	2020-3	N	–	Invalid DV	–	–	Invalid DV	–	–
29	18	2020-4	N	Y	0.002	0.047	-0.043	0.000	0.000	0.000
29	18	2021-1	N	Y	-0.008	0.174	-0.190	0.000	0.000	0.000
29	18	2021-2	N	Y	-0.006	0.318	-0.331	0.000	0.000	0.000
31	4	2019-4	N	N	0.000	0.000	0.000	0.000	0.000	0.000
31	4	2020-1	N	N	0.000	0.000	0.000	0.000	0.000	0.000
31	4	2020-2	N	N	0.000	0.000	0.000	0.000	0.000	0.000
31	4	2020-3	N	N	0.000	0.000	0.000	0.000	0.000	0.000
31	4	2020-4	N	N	-0.002	0.012	-0.015	0.000	0.000	0.000
31	4	2021-1	N	N	-0.057	0.035	-0.150	0.000	0.000	0.000
31	4	2021-2	N	N	-0.052	0.047	-0.150	-1.376	-1.376	-1.376
39	2010	2018-4	N	N	0.000	0.000	0.000	0.000	0.000	0.000
39	2010	2019-1	N	N	0.000	0.000	0.000	0.000	0.000	0.000
39	2010	2019-2	N	N	0.000	0.000	0.000	0.000	0.000	0.000

39	2010	2019-3	N	N	0.000	0.000	0.000	0.000	0.000	0.000
39	2010	2019-4	N	N	0.000	0.001	-0.002	0.000	0.000	0.000
39	2010	2020-1	N	N	0.001	0.009	-0.007	0.000	0.000	0.000
39	2010	2020-2	N	N	0.000	0.016	-0.015	0.000	0.000	0.000
39	2010	2020-3	N	N	-0.001	0.020	-0.021	0.000	0.000	0.000
39	2010	2020-4	N	N	-0.001	0.024	-0.025	0.000	0.000	0.000
39	2010	2021-1	N	N	-0.007	0.085	-0.098	0.000	0.000	0.000
39	2010	2021-2	N	N	-0.007	0.089	-0.103	0.000	0.000	0.000
59	7	2018-4	N	N	0.002	0.018	-0.014	0.000	0.000	0.000
59	7	2019-1	N	N	0.002	0.026	-0.022	0.000	0.000	0.000
59	7	2019-2	N	Y	0.002	0.031	-0.027	0.000	0.000	0.000
59	7	2019-3	N	Y	0.002	0.037	-0.033	0.000	0.000	0.000
59	7	2019-4	N	Y	0.002	0.041	-0.038	0.000	0.000	0.000
59	7	2020-1	N	Y	0.023	0.078	-0.031	0.000	0.492	-0.254
59	7	2020-2	N	Y	0.001	0.118	-0.117	0.000	0.492	-0.254

59	7	2020-3	N	Y	-0.003	0.121	-0.127	0.000	0.000	0.000
59	7	2020-4	N	Y	-0.011	0.120	-0.143	0.000	0.000	0.000
59	7	2021-1	N	Y	0.006	0.155	-0.143	0.000	0.492	-0.254
59	7	2021-2	N	Y	0.016	0.214	-0.183	0.000	0.492	-0.254
59	7	2021-3	N	Y	0.015	0.215	-0.186	0.000	0.771	0.000
67	5003	2020-1	Y	–	Invalid DV	–	–	Invalid DV	–	–
67	5003	2020-2	Y	Y	-0.002	0.202	-0.206	0.000	0.833	0.000
67	5003	2020-3	Y	N	-0.005	0.201	-0.210	0.000	0.000	0.000
67	5003	2020-4	Y	N	-0.003	0.204	-0.210	0.000	1.601	-0.029
67	5003	2021-1	Y	N	-0.011	0.199	-0.220	0.000	0.833	0.000
67	5003	2021-2	Y	N	-0.001	0.216	-0.219	0.000	0.833	0.000
77	2010	2018-4	N	N	0.005	0.039	-0.028	0.000	0.000	0.000
77	2010	2019-1	N	–	Invalid DV	–	–	Invalid DV	–	–
77	2010	2019-2	N	–	Invalid DV	–	–	Invalid DV	–	–
77	2010	2019-3	N	–	Invalid DV	–	–	Invalid DV	–	–

77	2010	2019-4	N	–	Invalid DV	–	–	Invalid DV	–	–
77	2010	2020-1	N	–	Invalid DV	–	–	Invalid DV	–	–
77	2010	2020-2	N	–	Invalid DV	–	–	Invalid DV	–	–
77	2010	2020-3	N	–	Invalid DV	–	–	Invalid DV	–	–
77	2010	2020-4	N	–	Invalid DV	–	–	Invalid DV	–	–
77	2010	2021-1	N	–	Invalid DV	–	–	Invalid DV	–	–
77	2010	2021-2	N	–	Invalid DV	–	–	Invalid DV	–	–
83	11	2020-2	Y	–	Invalid DV	–	–	Invalid DV	–	–
83	11	2020-3	Y	Y	0.020	0.423	-0.382	-1.361	-0.560	-1.361
83	11	2020-4	Y	Y	0.024	0.429	-0.380	-0.083	1.398	-0.083
83	11	2021-1	Y	Y	0.016	0.419	-0.388	-0.083	1.398	-0.083
83	11	2021-2	Y	Y	0.008	0.537	-0.522	-1.639	0.641	-1.639
83	11	2021-3	Y	Y	0.017	0.559	-0.526	-1.625	-0.052	-1.625
103	7	2018-4	Y	–	Invalid DV	–	–	Invalid DV	–	–
103	7	2019-1	Y	Y	-0.007	0.146	-0.160	-0.675	0.000	-1.208

103	7	2019-2	Y	Y	-0.005	0.155	-0.164	-0.675	0.000	-1.208
103	7	2019-3	Y	N	0.010	0.182	-0.162	0.000	2.597	-0.323
103	7	2019-4	Y	Y	0.024	0.325	-0.277	-0.675	1.855	-1.222
103	7	2020-1	Y	Y	0.023	0.528	-0.481	-0.716	3.444	-1.250
103	7	2020-2	Y	Y	0.026	0.594	-0.542	-0.716	3.444	-1.250
103	7	2020-3	Y	N	-0.062	0.618	-0.741	-0.972	1.625	-1.295
103	7	2020-4	Y	N	-0.043	0.652	-0.739	-1.647	0.883	-2.194
103	7	2021-1	Y	N	-0.065	1.058	-1.188	-1.689	2.472	-2.222
103	7	2021-2	Y	N	-0.052	1.199	-1.303	-1.689	2.472	-2.222

Table 5: Current NAAQS Nonattainment Counties

State	**Area Name**	**EPA Designated Nonattainment Counties**	**FIPS**	**# of monitors in sample**
CA	Imperial County, CA	Imperial, CA (p)	025	
	San Joaquin Valley Air Basin, CA	Fresno, CA	019	2
		Kern, CA (p)	029	1
		Kings, CA	031	2
		Madera, CA	039	1
		Merced, CA	047	
		San Joaquin, CA	077	1
		Stanislaus, CA	099	
		Tulare, CA	107	
	Los Angeles-South Coast Air Basin, CA	Los Angeles, CA (p)	037	1
		Orange, CA	059	1
		Riverside, CA (p)	065	
		San Bernardino, CA (p)	071	
	Plumas County, CA	Plumas, CA (p)	063	
ID	West Silver Valley, ID	Shoshone, ID (p)	079	
OH	Cleveland, OH	Cuyahoga, OH	035	
		Lorain, OH	093	
PA	Delaware County, PA	Delaware, PA	035	

	Lebanon County, PA	Lebanon, PA	075	
	Allegheny, PA	Allegheny, PA	005	

Table 6: 001-0013 NAAQS Monitor PM2.5 on Weighted Average PurpleAir PM2.5

	Reported NAAQS Monitor PM2.5	
	(1)	(2)
const		6.766***
		(0.093)
PurpleAir IDW Average	0.338***	0.192***
	(0.002)	(0.003)
Preferred	No	Yes
Observations	34,619	34,619
R^2	0.399	0.111
Adjusted R^2	0.398	0.111
Residual Std. Error	13.364	12.454
F Statistic	22935.628***	4340.693***
Note:		*p<0.1; **p<0.05; ***p<0.01

Table 7: 019-0500 NAAQS Monitor PM2.5 on Weighted Average PurpleAir PM2.5

	Reported NAAQS Monitor PM2.5	
	(1)	(2)
const		-7.724***
		(0.081)
PurpleAir IDW Average	1.125***	1.401***
	(0.003)	(0.004)
Preferred	No	Yes
Observations	32,016	32,016
R^2	0.784	0.784
Adjusted R^2	0.784	0.784
Residual Std. Error	11.640	10.269
F Statistic	116065.720***	116450.963***
Note:		*p<0.1; **p<0.05; ***p<0.01

Table 8: 019-5001 NAAQS Monitor PM2.5 on Weighted Average PurpleAir PM2.5

	Reported NAAQS Monitor PM2.5	
	(1)	(2)
const		13.919***
		(0.115)
PurpleAir IDW Average	0.177***	0.075***
	(0.002)	(0.002)
Preferred	No	Yes
Observations	26,838	26,838
R^2	0.171	0.047
Adjusted R^2	0.171	0.047
Residual Std. Error	21.444	17.244
F Statistic	5549.054***	1311.050***
Note:		*p<0.1; **p<0.05; ***p<0.01

Table 9: 027-0002 NAAQS Monitor PM2.5 on Weighted Average PurpleAir PM2.5

	Reported NAAQS Monitor PM2.5	
	(1)	(2)
const		-4.507***
		(0.081)
PurpleAir IDW Average	1.046***	1.329***
	(0.005)	(0.007)
Preferred	No	Yes
Observations	25,671	25,671
R^2	0.624	0.586
Adjusted R^2	0.624	0.586
Residual Std. Error	9.404	8.882
F Statistic	42589.056***	36294.550***
Note:		*p<0.1; **p<0.05; ***p<0.01

Table 10: 029-0018 NAAQS Monitor PM2.5 on Weighted Average PurpleAir PM2.5

	Reported NAAQS Monitor PM2.5	
	(1)	(2)
const		10.655***
		(0.259)
PurpleAir IDW Average	0.042***	-0.020***
	(0.002)	(0.002)
Preferred	No	Yes
Observations	8,520	8,520
R^2	0.052	0.008
Adjusted R^2	0.052	0.008
Residual Std. Error	19.940	18.209
F Statistic	471.879***	72.168***
Note:	*p<0.1; **p<0.05; ***p<0.01	

Table 11: 031-0004 NAAQS Monitor PM2.5 on Weighted Average PurpleAir PM2.5

	Reported NAAQS Monitor PM2.5	
	(1)	(2)
const		-2.227***
		(0.159)
PurpleAir IDW Average	1.109***	1.198***
	(0.005)	(0.008)
Preferred	No	Yes
Observations	9,132	9,132
R^2	0.861	0.719
Adjusted R^2	0.861	0.719
Residual Std. Error	9.036	8.941
F Statistic	56771.178***	23331.302***
Note:	*p<0.1; **p<0.05; ***p<0.01	

Table 12: 031-1004 NAAQS Monitor PM2.5 on Weighted Average PurpleAir PM2.5

	Reported NAAQS Monitor PM2.5	
	(1)	(2)
const		-1.542***
		(0.196)
PurpleAir IDW Average	1.198***	1.255***
	(0.006)	(0.009)
Preferred	No	Yes
Observations	6,981	6,981
R^2	0.868	0.728
Adjusted R^2	0.868	0.728
Residual Std. Error	9.941	9.898
F Statistic	45870.404***	18650.990***
Note:		*p<0.1; **p<0.05; ***p<0.01

Table 13: 039-2010 NAAQS Monitor PM2.5 on Weighted Average PurpleAir PM2.5

	Reported NAAQS Monitor PM2.5	
	(1)	(2)
const		-8.235***
		(0.105)
PurpleAir IDW Average	1.075***	1.383***
	(0.004)	(0.005)
Preferred	No	Yes
Observations	15,227	15,227
R^2	0.843	0.831
Adjusted R^2	0.843	0.831
Residual Std. Error	9.642	8.137
F Statistic	81867.760***	75091.538***
Note:		*p<0.1; **p<0.05; ***p<0.01

Table 14: 057-0005 NAAQS Monitor PM2.5 on Weighted Average PurpleAir PM2.5

	Reported NAAQS Monitor PM2.5	
	(1)	(2)
const		9.149***
		(0.146)
PurpleAir IDW Average	0.020***	0.005***
	(0.001)	(0.001)
Preferred	No	Yes
Observations	24,423	24,423
R^2	0.021	0.001
Adjusted R^2	0.021	0.001
Residual Std. Error	23.532	21.836
F Statistic	530.156***	35.924***
Note:		*p<0.1; **p<0.05; ***p<0.01

Table 15: 059-0007 NAAQS Monitor PM2.5 on Weighted Average PurpleAir PM2.5

	Reported NAAQS Monitor PM2.5	
	(1)	(2)
const		-3.169***
		(0.082)
PurpleAir IDW Average	0.992***	1.213***
	(0.003)	(0.006)
Preferred	No	Yes
Observations	29,577	29,577
R^2	0.781	0.545
Adjusted R^2	0.781	0.545
Residual Std. Error	6.705	6.542
F Statistic	105282.862***	35369.085***
Note:		*p<0.1; **p<0.05; ***p<0.01

Table 16: 067-5003 NAAQS Monitor PM2.5 on Weighted Average PurpleAir PM2.5

	Reported NAAQS Monitor PM2.5	
	(1)	(2)
const		-2.488***
		(0.078)
PurpleAir IDW Average	1.047***	1.205***
	(0.005)	(0.007)
Preferred	No	Yes
Observations	20,504	20,504
R^2	0.683	0.594
Adjusted R^2	0.683	0.594
Residual Std. Error	8.016	7.825
F Statistic	44083.188***	29970.462***
Note:	*p<0.1; **p<0.05; ***p<0.01	

Table 17: 077-2010 NAAQS Monitor PM2.5 on Weighted Average PurpleAir PM2.5

	Reported NAAQS Monitor PM2.5	
	(1)	(2)
const		-4.593***
		(0.096)
PurpleAir IDW Average	0.951***	1.147***
	(0.003)	(0.005)
Preferred	No	Yes
Observations	18,026	18,026
R^2	0.825	0.736
Adjusted R^2	0.825	0.736
Residual Std. Error	8.238	7.760
F Statistic	84882.128***	50316.689***
Note:	*p<0.1; **p<0.05; ***p<0.01	

Table 18: 083-0011 NAAQS Monitor PM2.5 on Weighted Average PurpleAir PM2.5

	Reported NAAQS Monitor PM2.5	
	(1)	(2)
const		7.949^{***}
		(0.035)
PurpleAir IDW Average	0.009^{***}	0.003^{***}
	(0.000)	(0.000)
Preferred	No	Yes
Observations	33,171	33,171
R^2	0.050	0.015
Adjusted R^2	0.050	0.015
Residual Std. Error	10.052	6.334
F Statistic	1762.982^{***}	519.691^{***}
Note:		*p<0.1; **p<0.05; ***p<0.01

Table 19: 103-0007 NAAQS Monitor PM2.5 on Weighted Average PurpleAir PM2.5

	Reported NAAQS Monitor PM2.5	
	(1)	(2)
const		9.376^{***}
		(0.128)
PurpleAir IDW Average	0.150^{***}	0.097^{***}
	(0.003)	(0.003)
Preferred	No	Yes
Observations	25,260	25,260
R^2	0.110	0.054
Adjusted R^2	0.110	0.054
Residual Std. Error	21.525	19.560
F Statistic	3126.880^{***}	1453.660^{***}
Note:		*p<0.1; **p<0.05; ***p<0.01

The EPA AQS data system used to get the NAAQS monitor PM2.5 hourly readings uses data qualifier flags to describe what readings need to be excluded from calculations. Below are all the "NULL" and "REQEXC" type qualifiers, representing null/canceled data and requests for exceptional event designation. All data with REQEXC flags were removed from the analysis, and all data with NULL type flags have empty readings in the dataset and were treated as missing for replacement by PurpleAir PM2.5 measurement.

Table 20: Null and excluded EPA AQS data qualifier flags.

Qualifier Code	Qualifier Description	Qualifier Type Code
AA	Sample Pressure out of Limits.	NULL
AB	Technician Unavaliable.	NULL
AC	Construction/Repairs in Area.	NULL
AD	Shelter Storm Damage.	NULL
AE	Shelter Temperature Outside Limits.	NULL
AF	Scheduled but not Collected.	NULL
AG	Sample Time out of Limits.	NULL
AH	Sample Flow Rate or CV out of Limits.	NULL
AI	Insufficient Data (cannot calculate).	NULL
AJ	Filter Damage.	NULL
AK	Filter Leak.	NULL
AL	Voided by Operator.	NULL
AM	Miscellaneous Void.	NULL
AN	Machine Malfunction.	NULL
AO	Bad Weather.	NULL
AP	Vandalism.	NULL
AQ	Collection Error.	NULL
AR	Lab Error.	NULL
AS	Poor Quality Assurance Results.	NULL

AT	Calibration.	NULL
AU	Monitoring Waived.	NULL
AV	Power Failure.	NULL
AW	Wildlife Damage.	NULL
AX	Precision Check.	NULL
AY	Q C Control Points (zero/span).	NULL
AZ	Q C Audit.	NULL
BA	Maintenance/Routine Repairs.	NULL
BB	Unable to Reach Site.	NULL
BC	Multi-point Calibration.	NULL
BD	Auto Calibration.	NULL
BE	Building/Site Repair.	NULL
BF	Precision/Zero/Span.	NULL
BG	Missing ozone data not likely to exceed level of standard.	NULL
BH	Interference/co-elution/misidentification.	NULL
BI	Lost or damaged in transit.	NULL
BJ	Operator Error.	NULL
BK	Site computer/data logger down.	NULL
BL	QA Audit.	NULL
BM	Accuracy check.	NULL
BN	Sample Value Exceeds Media Limit.	NULL
BR	Sample Value Below Acceptable Range.	NULL
CS	Laboratory Calibration Standard.	NULL
DA	Aberrant Data (Corrupt Files, Aberrant Chromatography, Spikes, Shifts).	NULL
DL	Detection Limit Analyses.	NULL

EC	Exceeds Critical Criteria.	NULL
FI	Filter Inspection Flag.	NULL
MB	Method Blank (Analytical).	NULL
MC	Module End Cap Missing.	NULL
QV	Quality Control Multi-point Verification.	NULL
SA	Storm Approaching.	NULL
SC	Sampler Contamination.	NULL
ST	Calibration Verification Standard.	NULL
SV	Sample Volume out of limits.	NULL
TC	Component Check & Retention Time Standard.	NULL
TS	Holding Time Or Transport Temperature Is Out Of Specs.	NULL
XX	Experimental Data.	NULL
1C	A 1-Point QC check exceeds acceptance criteria but there is compelling evidence that the analyzer data is valid.	NULL QC
1F	No 1 Point QC but need to count for completeness	NULL QC
E	Forest Fire.	REQEXC
RA	African Dust.	REQEXC
RB	Asian Dust.	REQEXC
RC	Chemical Spills & Industrial Accidents.	REQEXC
RD	Cleanup After a Major Disaster.	REQEXC
RE	Demolition.	REQEXC
RF	Fire - Canadian.	REQEXC
RG	Fire - Mexico/Central America.	REQEXC
RH	Fireworks.	REQEXC
RI	High Pollen Count.	REQEXC
RJ	High Winds.	REQEXC

RK	Infrequent Large Gatherings.	REQEXC
RL	Other.	REQEXC
RM	Prescribed Fire.	REQEXC
RN	Seismic Activity.	REQEXC
RO	Stratospheric Ozone Intrusion.	REQEXC
RP	Structural Fire.	REQEXC
RQ	Terrorist Act.	REQEXC
RR	Unique Traffic Disruption.	REQEXC
RS	Volcanic Eruptions.	REQEXC
RT	Wildfire-U. S.	REQEXC
RU	Wildland Fire Use Fire-U. S.	REQEXC

6.7 Pictures of PM2.5 monitors

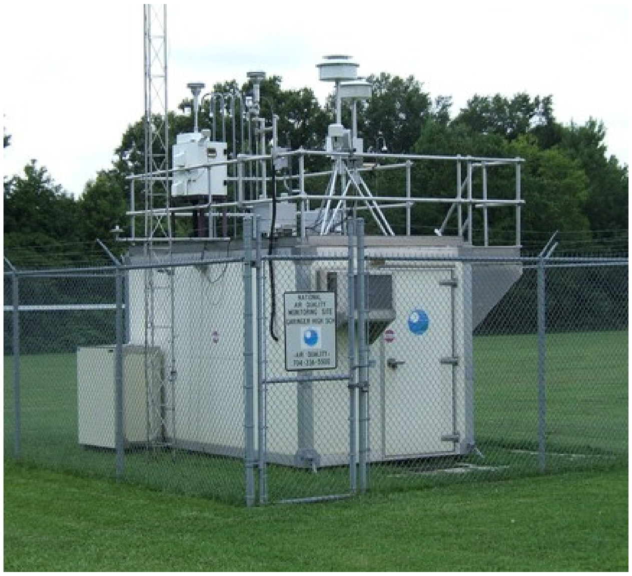

Figure 10: A typical NAAQS-primary grade air quality monitoring station.

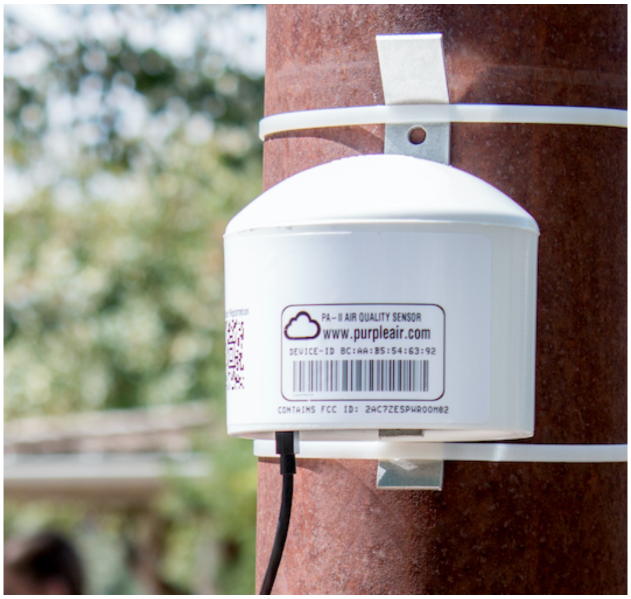

Figure 11: One of PurpleAir's two main outdoor air pollution monitors.

6.8 Plots for Other California Hourly NAAQS Monitors