# Measuring bias in (strategically) missing EPA pollution data

Aaron C Watt

December 2, 2021

# Motivation

## Clean Air

# Motivation

## Clean Air

- The Clean Air Act (1970) established National Ambient Air Quality Standards (NAAQS) for US counties.

# Motivation

## Clean Air

- The Clean Air Act (1970) established National Ambient Air Quality Standards (NAAQS) for US counties.
- Binary standards: either "attainment" or "non-attainment"

# Motivation

## Clean Air

- The Clean Air Act (1970) established National Ambient Air Quality Standards (NAAQS) for US counties.
- Binary standards: either “attainment” or “non-attainment”
- Air quality can vary significantly throughout a day, week, quarter.

# Motivation

## Clean Air

- The Clean Air Act (1970) established National Ambient Air Quality Standards (NAAQS) for US counties.
- Binary standards: either "attainment" or "non-attainment"
- Air quality can vary significantly throughout a day, week, quarter.
- NAAQS only requires each monitor to report 75% of their readings, per quarter.

# Motivation

## Clean Air

- The Clean Air Act (1970) established National Ambient Air Quality Standards (NAAQS) for US counties.
- Binary standards: either “attainment” or “non-attainment”
- Air quality can vary significantly throughout a day, week, quarter.
- NAAQS only requires each monitor to report 75% of their readings, per quarter.
- EPA pollution monitors can be shut off for unreported reasons.

# Motivation

## Clean Air

- The Clean Air Act (1970) established National Ambient Air Quality Standards (NAAQS) for US counties.
- Binary standards: either "attainment" or "non-attainment"
- Air quality can vary significantly throughout a day, week, quarter.
- NAAQS only requires each monitor to report 75% of their readings, per quarter.
- EPA pollution monitors can be shut off for unreported reasons.

## Research Questions

# Motivation

## Clean Air

- The Clean Air Act (1970) established National Ambient Air Quality Standards (NAAQS) for US counties.
- Binary standards: either "attainment" or "non-attainment"
- Air quality can vary significantly throughout a day, week, quarter.
- NAAQS only requires each monitor to report 75% of their readings, per quarter.
- EPA pollution monitors can be shut off for unreported reasons.

## Research Questions

- How biased are EPA monitor-based measures of local air quality?

# Motivation

## Clean Air

- The Clean Air Act (1970) established National Ambient Air Quality Standards (NAAQS) for US counties.
- Binary standards: either “attainment” or “non-attainment”
- Air quality can vary significantly throughout a day, week, quarter.
- NAAQS only requires each monitor to report 75% of their readings, per quarter.
- EPA pollution monitors can be shut off for unreported reasons.

## Research Questions

- How biased are EPA monitor-based measures of local air quality?
- Does this bias significantly change NAAQS attainment status?

# Context

## Previous Works

# Context

## Previous Works

- Spotting Strategic Behavior in EPA Monitor Shutoffs: Mu et al. 2021

# Context

## Previous Works

- Spotting Strategic Behavior in EPA Monitor Shutoffs: Mu et al. 2021
- Strategic decisions of monitor locations: Grainger et al. 2017

# Context

## Previous Works

- Spotting Strategic Behavior in EPA Monitor Shutoffs: Mu et al. 2021
- Strategic decisions of monitor locations: Grainger et al. 2017
- Using satellite data to fill gaps in air pollution monitoring: Sullivan, Krupnick 2018

# Context

## Previous Works

- Spotting Strategic Behavior in EPA Monitor Shutoffs: Mu et al. 2021
- Strategic decisions of monitor locations: Grainger et al. 2017
- Using satellite data to fill gaps in air pollution monitoring: Sullivan, Krupnick 2018
- Prediction error is significant in satellite-based estimates of air pollution: Fowlie, Rubin, Walker 2019

# Context

## Previous Works

- Spotting Strategic Behavior in EPA Monitor Shutoffs: Mu et al. 2021
- Strategic decisions of monitor locations: Grainger et al. 2017
- Using satellite data to fill gaps in air pollution monitoring: Sullivan, Krupnick 2018
- Prediction error is significant in satellite-based estimates of air pollution: Fowlie, Rubin, Walker 2019

## This project

# Context

## Previous Works

- Spotting Strategic Behavior in EPA Monitor Shutoffs: Mu et al. 2021
- Strategic decisions of monitor locations: Grainger et al. 2017
- Using satellite data to fill gaps in air pollution monitoring: Sullivan, Krupnick 2018
- Prediction error is significant in satellite-based estimates of air pollution: Fowlie, Rubin, Walker 2019

## This project

- Using new consumer-based pollution monitors (PurpleAir).

# Context

## Previous Works

- Spotting Strategic Behavior in EPA Monitor Shutoffs: Mu et al. 2021
- Strategic decisions of monitor locations: Grainger et al. 2017
- Using satellite data to fill gaps in air pollution monitoring: Sullivan, Krupnick 2018
- Prediction error is significant in satellite-based estimates of air pollution: Fowlie, Rubin, Walker 2019

## This project

- Using new consumer-based pollution monitors (PurpleAir).
- Focus on EPA pollution data that is missing *in time*.

# Context

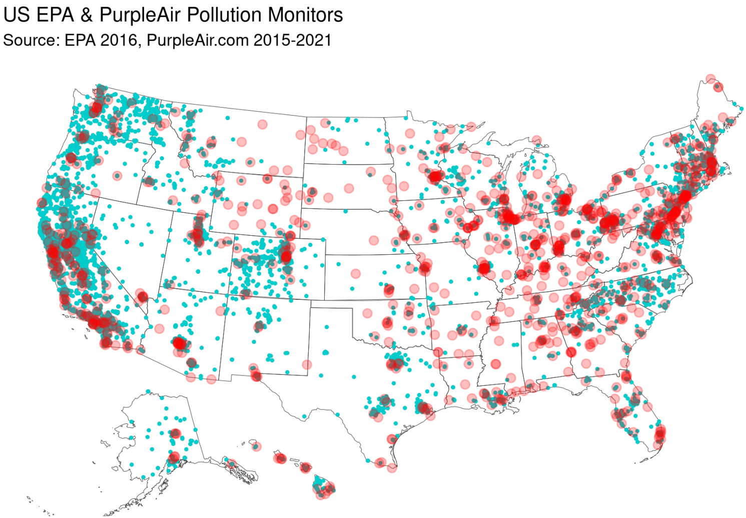
## Previous Works

- Spotting Strategic Behavior in EPA Monitor Shutoffs: Mu et al. 2021
- Strategic decisions of monitor locations: Grainger et al. 2017
- Using satellite data to fill gaps in air pollution monitoring: Sullivan, Krupnick 2018
- Prediction error is significant in satellite-based estimates of air pollution: Fowlie, Rubin, Walker 2019

## This project

- Using new consumer-based pollution monitors (PurpleAir).
- Focus on EPA pollution data that is missing *in time*.
- Avoids using satellite estimates (has been shown to have significant error).

# PM2.5 Air Pollution Monitors





# Models

- **Estimate missing pollution observations:** Use PurpleAir data to predict EPA pollution at missing times

# Models

- **Estimate missing pollution observations:** Use PurpleAir data to predict EPA pollution at missing times
- **Estimate bias of reported EPA pollution:** difference between predicted pollution at missing times and reported pollution at nonmissing times.

# Models

- **Estimate missing pollution observations:** Use PurpleAir data to predict EPA pollution at missing times
- **Estimate bias of reported EPA pollution:** difference between predicted pollution at missing times and reported pollution at nonmissing times.
- **Estimate counties' counterfactual attainment status:** Include estimated missing pollution data.

# Models: Predictive model of each EPA monitor PM2.5 pollution

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

- Analysis done at the month and quarter level; suppressing that subscript.

# Models: Predictive model of each EPA monitor PM2.5 pollution

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

- Analysis done at the month and quarter level; suppressing that subscript.
- $t$ is a unique hour within a given month or quarter.

# Models: Predictive model of each EPA monitor PM2.5 pollution

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

- Analysis done at the month and quarter level; suppressing that subscript.
- $t$ is a unique hour within a given month or quarter.
- EPA monitor $i$ at time $t$ reads PM2.5 pollution $EPA_{i,t}$.

# Models: Predictive model of each EPA monitor PM2.5 pollution

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

- Analysis done at the month and quarter level; suppressing that subscript.
- $t$ is a unique hour within a given month or 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 10-mile radius.

# Models: Predictive model of each EPA monitor PM2.5 pollution

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

- Analysis done at the month and quarter level; suppressing that subscript.
- $t$ is a unique hour within a given month or 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 10-mile radius.
- Purple Air monitor $j \in J_i$ at time $t$ reads PM2.5 pollution $PA_{j,t}$.

# Models: Predictive model of each EPA monitor PM2.5 pollution

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

- Analysis done at the month and quarter level; suppressing that subscript.
- $t$ is a unique hour within a given month or 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 10-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).

# Models: Hour-by-Day-of-week Bias of Missing EPA Monitor Pollution Data

**Missingness Bias:**

$$Bias_{i,h,d} = \frac{1}{|\mathcal{N}_{i,h,d}|} \sum_{t \in \mathcal{N}_{i,h,d}} EPA_{i,t} - \frac{1}{|\mathcal{M}_{h,d}|} \sum_{t \in \mathcal{M}_{h,d}} \widehat{EPA}_{i,t}$$

$$\text{where } \mathcal{M}_{i,h,d} = \{t : t \text{ is at hour } h \text{ and day } d \text{ and } EPA_{i,t} \text{ is Missing}\};$$
$$\mathcal{N}_{i,h,d} = \{t : t \text{ is at hour } h \text{ and day } d \text{ and } EPA_{i,t} \text{ is Non-missing}\}$$

# Models: Hour-by-Day-of-week Bias of Missing EPA Monitor Pollution Data

**Missingness Bias:**

$$Bias_{i,h,d} = \frac{1}{|\mathcal{N}_{i,h,d}|} \sum_{t \in \mathcal{N}_{i,h,d}} EPA_{i,t} - \frac{1}{|\mathcal{M}_{h,d}|} \sum_{t \in \mathcal{M}_{h,d}} \widehat{EPA}_{i,t}$$

$$\text{where } \mathcal{M}_{i,h,d} = \{t : t \text{ is at hour } h \text{ and day } d \text{ and } EPA_{i,t} \text{ is Missing}\};$$

$$\mathcal{N}_{i,h,d} = \{t : t \text{ is at hour } h \text{ and day } d \text{ and } EPA_{i,t} \text{ is Non-missing}\}$$

We can also define the **algorithm bias** as the Hour-by-Day-of-week prediction error

$$\widetilde{Bias}_{i,h,d} = \frac{1}{|\mathcal{N}_{i,h,d}|} \sum_{t \in \mathcal{N}_{i,h,d}} EPA_{i,t} - \frac{1}{|\mathcal{N}_{h,d}|} \sum_{t \in \mathcal{N}_{h,d}} \widehat{EPA}_{i,t}$$

# Models: Hour-by-Day-of-week Bias of Missing EPA Monitor Pollution Data

**Missingness Bias:**

$$Bias_{i,h,d} = \frac{1}{|\mathcal{N}_{i,h,d}|} \sum_{t \in \mathcal{N}_{i,h,d}} EPA_{i,t} - \frac{1}{|\mathcal{M}_{h,d}|} \sum_{t \in \mathcal{M}_{h,d}} \widehat{EPA}_{i,t}$$

where $\mathcal{M}_{i,h,d} = \{t : t \text{ is at hour } h \text{ and day } d \text{ and } EPA_{i,t} \text{ is Missing}\};$

$\mathcal{N}_{i,h,d} = \{t : t \text{ is at hour } h \text{ and day } d \text{ and } EPA_{i,t} \text{ is Non-missing}\}$

We can also define the **algorithm bias** as the Hour-by-Day-of-week prediction error

$$\widetilde{Bias}_{i,h,d} = \frac{1}{|\mathcal{N}_{i,h,d}|} \sum_{t \in \mathcal{N}_{i,h,d}} EPA_{i,t} - \frac{1}{|\mathcal{N}_{h,d}|} \sum_{t \in \mathcal{N}_{h,d}} \widehat{EPA}_{i,t}$$

We can also define $Bias_{j,h,d}$ and $\widetilde{Bias}_{j,h,d}$ for PA monitor $j$ (we'll come back to this).

# Models: County Attainment Status

$$Attain_c^{annual} = 1 \text{ if } \textbf{reported} \text{ annual average PM2.5 below } 15.0\ \mu\text{g/m}^{3*}$$

$$Attain_c^{daily} = 1 \text{ if } 98^{th} \text{ percentile of } \textbf{reported} \text{ daily average PM2.5 below } 35\ \mu\text{g/m}^{3*}$$

$$\widehat{Attain}_c^{annual} = 1 \text{ if } \textbf{predicted} \text{ annual average PM2.5 below threshold}^*$$

$$\widehat{Attain}_c^{daily} = 1 \text{ if } 98^{th} \text{ percentile of } \textbf{predicted} \text{ daily average PM2.5 below threshold}^*$$

*averaged over 3 years in NAAQS standard. [fill in equations and thresholds]

# Identification Strategy

- Direct causal link between reported pollution levels, attainment status, and regulatory penalties / attainment requirements

# Identification Strategy

- Direct causal link between reported pollution levels, attainment status, and regulatory penalties / attainment requirements
- Assumption: nearby PurpleAir monitors that are good predictors for EPA monitors during non-missing times will also be good predictors during missing times.

# Identification Strategy

- Direct causal link between reported pollution levels, attainment status, and regulatory penalties / attainment requirements
- Assumption: nearby PurpleAir monitors that are good predictors for EPA monitors during non-missing times will also be good predictors during missing times.
  - Specifically, reasons for EPA data missingness are not correlated with missingness or measurement error in PurpleAir data

# Proposed Statistical Test

- The $J_i$ group of PurpleAir sensors is (in a sense) a synthetic control for the EPA sensor $i$.

# Proposed Statistical Test

- The $J_i$ group of PurpleAir sensors is (in a sense) a synthetic control for the EPA sensor $i$.
- **The question of bias can be stated**: are the data observed during the times when the EPA monitor is turned off significantly different from the data observed when the monitor is turned on? Is it more different than by random chance?

# Proposed Statistical Test

- The $J_i$ group of PurpleAir sensors is (in a sense) a synthetic control for the EPA sensor $i$.
- **The question of bias can be stated**: are the data observed during the times when the EPA monitor is turned off significantly different from the data observed when the monitor is turned on? Is it more different than by random chance?
- Implies an Abadie et al. 2011 style permutation inference test for each EPA monitor $i$.

# Proposed Statistical Test (Is the bias larger than by random chance?)

- Calculate missingness bias and algorithm bias for the EPA monitor

# Proposed Statistical Test (Is the bias larger than by random chance?)

- Calculate missingness bias and algorithm bias for the EPA monitor
- Pick PA monitor $j \in J_i$, temporarily remove data for original hours missing from EPA monitor ($\mathcal{M}_{i,h,d}$).

# Proposed Statistical Test (Is the bias larger than by random chance?)

- Calculate missingness bias and algorithm bias for the EPA monitor
- Pick PA monitor $j \in J_i$, temporarily remove data for original hours missing from EPA monitor ($\mathcal{M}_{i,h,d}$).
- Construct similar placebo synthetic control for PA monitor $j$ (predict $\widehat{PA}_{j,t}$).

# Proposed Statistical Test (Is the bias larger than by random chance?)

- Calculate missingness bias and algorithm bias for the EPA monitor
- Pick PA monitor $j \in J_i$, temporarily remove data for original hours missing from EPA monitor ($\mathcal{M}_{i,h,d}$).
- Construct similar placebo synthetic control for PA monitor $j$ (predict $\widehat{PA}_{j,t}$).
- Calculate missingness bias and algorithm bias for the PA monitor: $Bias_{j,h,d}$

# Proposed Statistical Test (Is the bias larger than by random chance?)

- Calculate missingness bias and algorithm bias for the EPA monitor
- Pick PA monitor $j \in J_i$, temporarily remove data for original hours missing from EPA monitor ($\mathcal{M}_{i,h,d}$).
- Construct similar placebo synthetic control for PA monitor $j$ (predict $\widehat{PA}_{j,t}$).
- Calculate missingness bias and algorithm bias for the PA monitor: $Bias_{j,h,d}$
- Repeat for all PA sensors.

# Proposed Statistical Test (Is the bias larger than by random chance?)

- **Graphical test**: For EPA sensor $i$, compare graph of $Bias_{i,h,d}$ to placebo $Bias_{j,h,d}$ for $j \in J_i$.

# Proposed Statistical Test (Is the bias larger than by random chance?)

- **Graphical test**: For EPA sensor $i$, compare graph of $Bias_{i,h,d}$ to placebo $Bias_{j,h,d}$ for $j \in J_i$.
- **Permutation inference p-value**:

# Proposed Statistical Test (Is the bias larger than by random chance?)

- **Graphical test**: For EPA sensor $i$, compare graph of $Bias_{i,h,d}$ to placebo $Bias_{j,h,d}$ for $j \in J_i$.
- **Permutation inference p-value**:
  - Calculate sum of squared missingness bias and sum of squared algorithm bias for EPA sensor $i$ and PA sensors $j \in J_i$.

# Proposed Statistical Test (Is the bias larger than by random chance?)

- **Graphical test**: For EPA sensor $i$, compare graph of $Bias_{i,h,d}$ to placebo $Bias_{j,h,d}$ for $j \in J_i$.
- **Permutation inference p-value**:
  - Calculate sum of squared missingness bias and sum of squared algorithm bias for EPA sensor $i$ and PA sensors $j \in J_i$.
  - $\text{Ratio}_k$ = sum of squared missingness bias / sum of squared algorithm bias

# Proposed Statistical Test (Is the bias larger than by random chance?)

- **Graphical test**: For EPA sensor $i$, compare graph of $Bias_{i,h,d}$ to placebo $Bias_{j,h,d}$ for $j \in J_i$.
- **Permutation inference p-value**:
  - Calculate sum of squared missingness bias and sum of squared algorithm bias for EPA sensor $i$ and PA sensors $j \in J_i$.
  - $\text{Ratio}_k$ = sum of squared missingness bias / sum of squared algorithm bias
  - p-value $= \frac{\text{\# of PA sensors in } i\text{'s radius with Ratio}_j \text{ larger than Ratio}_i}{\text{\# of PA sensors in } i\text{'s radius}}$

# Extensions

- Welfare analysis based on attainment status changes and required reductions in pollution.

# Extensions

- Welfare analysis based on attainment status changes and required reductions in pollution.
- Comparing county population-weighted PM2.5 pollution to EPA sensors to estimate location-based bias.

# Appendix A: PurpleAir monitor correction factor

| | |
|---|---|
| Low Concentration $PA_{cf\_1} \leq 343\ \mu g\ m^{-3}$ ~176-185 $\mu g\ m^{-3}$ as measured by the corrected sensor | $PM_{2.5} = 0.52 \times PA_{cf\_1} - 0.086 \times RH + 5.75$ |
| High Concentration $PA_{cf\_1} > 343\ \mu g\ m^{-3}$ ~207 $\mu g\ m^{-3}$ as measured by the corrected sensor | $PM_{2.5} = 0.46 \times PA_{cf\_1} + 3.93 \times 10^{-4} \times PA_{cf\_1}^{2} + 2.97$ |

$PA_{cf\_1}$ = PurpleAir $PM_{2.5}$ from the higher of the 2 correction factors (cf) currently labeled as cf_1 32

Figure 2: PurpleAir correction equation for EPA monitor PM2.5 (RH = relative humidity, also measured by PA monitor)

Source: https://www.epa.gov/air-sensor-toolbox/technical-approaches-sensor-data-airnow-fire-and-smoke-map

# Appendix B: Data Plan

## Datasets

# Appendix B: Data Plan

## Datasets

- Hourly PM2.5 Pollution data from California EPA pollution monitors (2015-2020)

# Appendix B: Data Plan

## Datasets

- Hourly PM2.5 Pollution data from California EPA pollution monitors (2015-2020)
- 2-minute PM2.5 Pollution data from California PurpleAir sensors, hourly averages taken

# Appendix B: Data Plan

## Datasets

- Hourly PM2.5 Pollution data from California EPA pollution monitors (2015-2020)
- 2-minute PM2.5 Pollution data from California PurpleAir sensors, hourly averages taken

## Deadlines

# Appendix B: Data Plan

## Datasets

- Hourly PM2.5 Pollution data from California EPA pollution monitors (2015-2020)
- 2-minute PM2.5 Pollution data from California PurpleAir sensors, hourly averages taken

## Deadlines

- Dec. 5: PurpleAir is downloading/averaging on 4 AWS tiny linux instances, sending CSVs to S3 bucket

# Appendix B: Data Plan

## Datasets

- Hourly PM2.5 Pollution data from California EPA pollution monitors (2015-2020)
- 2-minute PM2.5 Pollution data from California PurpleAir sensors, hourly averages taken

## Deadlines

- Dec. 5: PurpleAir is downloading/averaging on 4 AWS tiny linux instances, sending CSVs to S3 bucket
- Dec. 12: Proof of concept for 2 EPA sensors (Fresno, and [need to pick another low on Mu's list])

# Appendix B: Data Plan

## Datasets

- Hourly PM2.5 Pollution data from California EPA pollution monitors (2015-2020)
- 2-minute PM2.5 Pollution data from California PurpleAir sensors, hourly averages taken

## Deadlines

- Dec. 5: PurpleAir is downloading/averaging on 4 AWS tiny linux instances, sending CSVs to S3 bucket
- Dec. 12: Proof of concept for 2 EPA sensors (Fresno, and [need to pick another low on Mu's list])
- Dec. 19: Data warehouse setup and transfer of existing Purple Air data

# Appendix B: Data Plan

Data Wearhouse

# Appendix B: Data Plan

## Data Wearhouse

- AWS Linux Cassandra database (noSQL columnar, designed for large queries of columns)

# Appendix B: Data Plan

## Data Wearhouse

- AWS Linux Cassandra database (noSQL columnar, designed for large queries of columns)
- Python pushes and pulls

# Appendix B: Data Plan

## Data Wearhouse

- AWS Linux Cassandra database (noSQL columnar, designed for large queries of columns)
- Python pushes and pulls

## Storage costs

# Appendix B: Data Plan

## Data Wearhouse

- AWS Linux Cassandra database (noSQL columnar, designed for large queries of columns)
- Python pushes and pulls

## Storage costs

- $\sim$ 30,000 sensors, 50 variables, 2 minute intervals, 5 years of data $=$ 107.55 Terabytes

# Appendix B: Data Plan

## Data Wearhouse

- AWS Linux Cassandra database (noSQL columnar, designed for large queries of columns)
- Python pushes and pulls

## Storage costs

- $\sim$ 30,000 sensors, 50 variables, 2 minute intervals, 5 years of data = 107.55 Terabytes
- Depending on method of storage: $2,900 - $13,300 per month

# Appendix B: Data Plan

## Data Wearhouse

- AWS Linux Cassandra database (noSQL columnar, designed for large queries of columns)
- Python pushes and pulls

## Storage costs

- $\sim$ 30,000 sensors, 50 variables, 2 minute intervals, 5 years of data = 107.55 Terabytes
- Depending on method of storage: $2,900 - $13,300 per month
- Only storing hourly means and SD: $4 - $15 per month

# Appendix C: PurpleAir Takeup

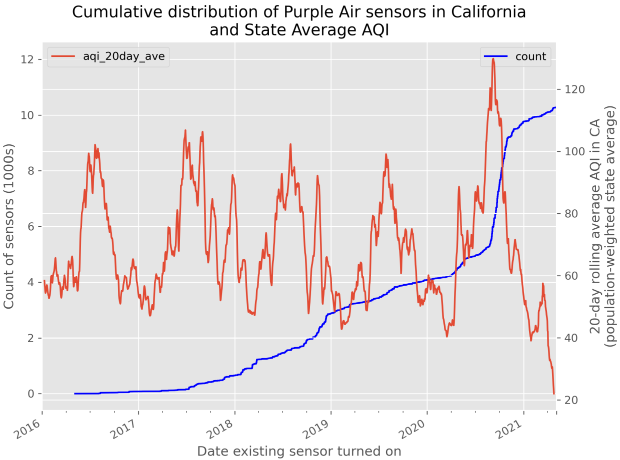



Figure 3: Valid Purple Air Monitor Locations, Contiguous United States

Source: api.purpleair.com