Estimation and Implications of Bias in EPA Pollution Measurement

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
- Either “attainment” or “non-attainment”, large penalties and costly restrictions

Motivation

Clean Air

- The Clean Air Act (1970) established National Ambient Air Quality Standards (NAAQS) for US counties
- Either "attainment" or "non-attainment", large penalties and costly restrictions
- Minimum requirement of 75% of readings, per quarter

Motivation

Clean Air

- The Clean Air Act (1970) established National Ambient Air Quality Standards (NAAQS) for US counties
- Either "attainment" or "non-attainment", large penalties and costly restrictions
- Minimum requirement of 75% of readings, per quarter
- Air quality can change quickly

Motivation

Clean Air

- The Clean Air Act (1970) established National Ambient Air Quality Standards (NAAQS) for US counties
- Either "attainment" or "non-attainment", large penalties and costly restrictions
- Minimum requirement of 75% of readings, per quarter
- Air quality can change quickly
- Monitor shutoffs are common

Motivation

Clean Air

- The Clean Air Act (1970) established National Ambient Air Quality Standards (NAAQS) for US counties
- Either “attainment” or “non-attainment”, large penalties and costly restrictions
- Minimum requirement of 75% of readings, per quarter
- Air quality can change quickly
- Monitor shutoffs are common

Research Questions

Motivation

Clean Air

- The Clean Air Act (1970) established National Ambient Air Quality Standards (NAAQS) for US counties
- Either "attainment" or "non-attainment", large penalties and costly restrictions
- Minimum requirement of 75% of readings, per quarter
- Air quality can change quickly
- Monitor shutoffs are common

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
- Either "attainment" or "non-attainment", large penalties and costly restrictions
- Minimum requirement of 75% of readings, per quarter
- Air quality can change quickly
- Monitor shutoffs are common

Research Questions

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

Context

Previous Works in EPA Pollution Monitors

Satellite Data:

Location Pollution Alerts:

Context

Previous Works in EPA Pollution Monitors

Satellite Data:

- Grainger et al. 2017

Location Pollution Alerts:

Context

Previous Works in EPA Pollution Monitors

Satellite Data:

- Grainger et al. 2017
- Sullivan, Krupnick 2018

Location Pollution Alerts:

Context

Previous Works in EPA Pollution Monitors

Satellite Data:

- Grainger et al. 2017
- Sullivan, Krupnick 2018
- Fowlie, Rubin, Walker 2019

Location Pollution Alerts:

Context

Previous Works in EPA Pollution Monitors

Satellite Data:

- Grainger et al. 2017
- Sullivan, Krupnick 2018
- Fowlie, Rubin, Walker 2019
- Zou 2021

Location Pollution Alerts:

Context

Previous Works in EPA Pollution Monitors

Satellite Data:

- Grainger et al. 2017
- Sullivan, Krupnick 2018
- Fowlie, Rubin, Walker 2019
- Zou 2021

Location Pollution Alerts:

- Mu, Rubin, Zou 2021

Context

Previous Works in EPA Pollution Monitors

Satellite Data:

- Grainger et al. 2017
- Sullivan, Krupnick 2018
- Fowlie, Rubin, Walker 2019
- Zou 2021

Location Pollution Alerts:

- Mu, Rubin, Zou 2021

This project

Context

Previous Works in EPA Pollution Monitors

Satellite Data:

- Grainger et al. 2017
- Sullivan, Krupnick 2018
- Fowlie, Rubin, Walker 2019
- Zou 2021

Location Pollution Alerts:

- Mu, Rubin, Zou 2021

This project

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

Context

Previous Works in EPA Pollution Monitors

Satellite Data:

- Grainger et al. 2017
- Sullivan, Krupnick 2018
- Fowlie, Rubin, Walker 2019
- Zou 2021

Location Pollution Alerts:

- Mu, Rubin, Zou 2021

This project

- Using new consumer-based pollution monitors (PurpleAir).
- Avoids using satellite estimates (has been shown to have significant error).

Context

Previous Works in EPA Pollution Monitors

Satellite Data:

- Grainger et al. 2017
- Sullivan, Krupnick 2018
- Fowlie, Rubin, Walker 2019
- Zou 2021

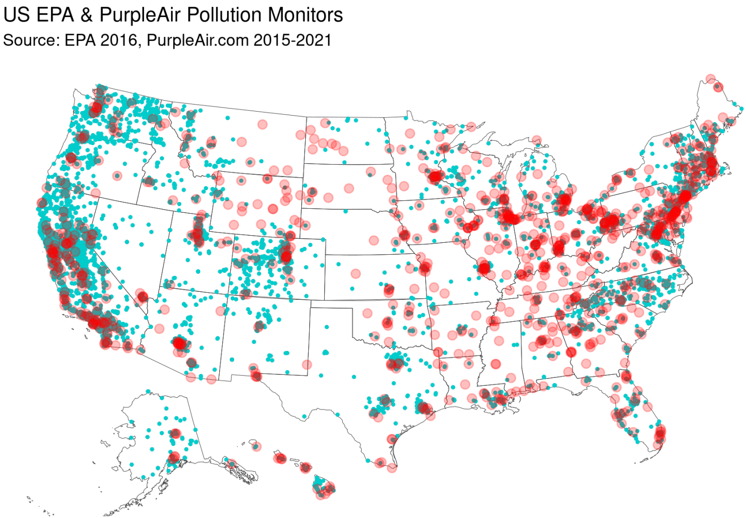
Location Pollution Alerts:

- Mu, Rubin, Zou 2021

This project

- Using new consumer-based pollution monitors (PurpleAir).
- Avoids using satellite estimates (has been shown to have significant error).
- Focus on EPA pollution data that is missing *in time*, specifically PM2.5

Use PM2.5 Air Pollution Monitors to Predict Missing EPA Data



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.
- 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: 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: 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: Predictive model of each EPA monitor PM2.5 pollution

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

- 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 $PA_{j,t}$.
- $\color{red}{Winddir_{i,t,k}}$ 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:

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

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

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

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

Missingness Bias:

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

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

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

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

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

Missingness Bias:

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

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

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

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

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

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

$Attain_c^{annual} = 1$ if **reported** annual average PM2.5 below 15.0 μg/m^{3*}

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

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

Models: County Attainment Status

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

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

$$\widehat{Attain}_c^{annual} = 1 \text{ if } \textbf{predicted} \text{ annual average PM2.5 below 15.0 } \mu\text{g/m}^{3*}$$

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

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

Identification

Identification

Is the missing pollution data identified?

Identification

Is the missing pollution data identified?

- 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

Identification

Is the missing pollution data identified?

- 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

Are attainment status changes identified?

Identification

Is the missing pollution data identified?

- 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

Are attainment status changes identified?

- Yes: simple rule based on pollution concentrations.

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 in EPA monitor data can be stated**: is pollution concentration at the monitor during reported times significantly different from the concentration at the monitor at missing times? Is the difference greater than random variations?

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 in EPA monitor data can be stated**: is pollution concentration at the monitor during reported times significantly different from the concentration at the monitor at missing times? Is the difference greater than random variations?
- Implies an 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 i^{th} EPA monitor using J_i PA monitors with n_i missing hour observations.

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

- Calculate missingness bias and algorithm bias for the i^{th} EPA monitor using J_i PA monitors with n_i missing hour observations.
- Pick PA monitor $j \in J_i$, temporarily remove random n_i hour observations.

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

- Calculate missingness bias and algorithm bias for the i^{th} EPA monitor using J_i PA monitors with n_i missing hour observations.
- Pick PA monitor $j \in J_i$, temporarily remove random n_i hour observations.
- Construct placebo synthetic control for PA monitor j and predict $\widehat{PA}_{j,t}$.

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

- Calculate missingness bias and algorithm bias for the i^{th} EPA monitor using J_i PA monitors with n_i missing hour observations.
- Pick PA monitor $j \in J_i$, temporarily remove random n_i hour observations.
- Construct placebo synthetic control for PA monitor j and predict $\widehat{PA}_{j,t}$.
- Calculate missingness bias and algorithm bias for PA monitor j: $Bias_{j,h,d}$

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

- Calculate missingness bias and algorithm bias for the i^{th} EPA monitor using J_i PA monitors with n_i missing hour observations.
- Pick PA monitor $j \in J_i$, temporarily remove random n_i hour observations.
- Construct placebo synthetic control for PA monitor j and predict $\widehat{PA}_{j,t}$.
- Calculate missingness bias and algorithm bias for PA monitor j: $Bias_{j,h,d}$
- 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 $j \in J_i$.
  - 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

- Use PurpleAir data to create population-weighted pollution measure $\implies$ counterfactual attainment.

Extensions

- Use PurpleAir data to create population-weighted pollution measure $\implies$ counterfactual attainment.
- Welfare analysis based on attainment status changes and required reductions in pollution.

Extensions

- Use PurpleAir data to create population-weighted pollution measure $\Longrightarrow$ counterfactual attainment.
- 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

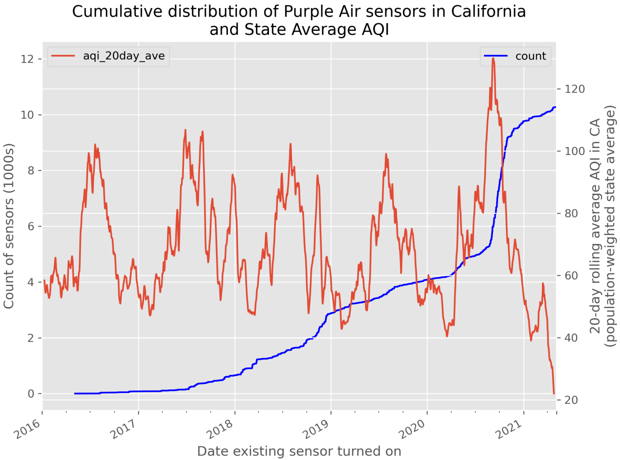


Figure 3: Valid Purple Air Monitor Locations, Contiguous United States