

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

A. C. Watt¹

¹Agricultural & Resource Economics
University of California, Berkeley

Second Year Paper, March 2022

State of project

Data (2016-2021)

- have: 15 CA EPA sites
- have: all hourly data for PA sites within 50 miles of all 15 EPA sites
- want: all US PA data
- can before April 11: get 10-50 more EPA sites

Analysis

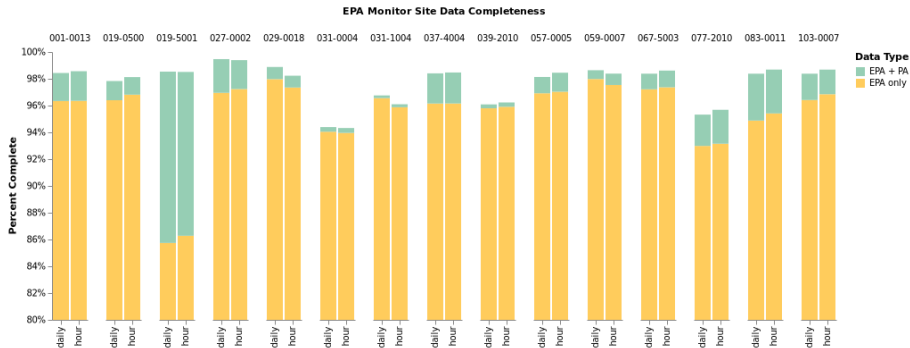
- have: Inverse distance weighted mean of PA
- have: OLS to predict EPA with PA, give prediction intervals
$$EPA_t = \beta_0 + \beta_1 PA_t^{IDW} + \varepsilon_t$$
- want: prediction (1): more complicated OLS with windspeed and direction
- prediction (2): machine learning model
- can before April 11: probably windspeed and direction model

Data Use Agreement

- Gave Adrian Dybwad draft person DUA
- Adrian said legal folks ask if we could do this at the University level
- I asked if we could do a temporary DUA while we get the university level one figured out

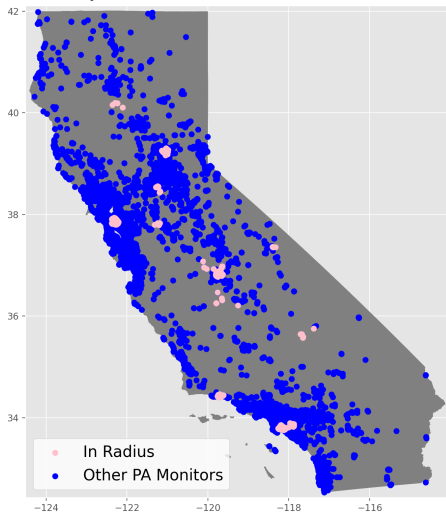
What are the first steps to getting a university level DUA?

Observed Completeness of NAAQS Monitors in sample

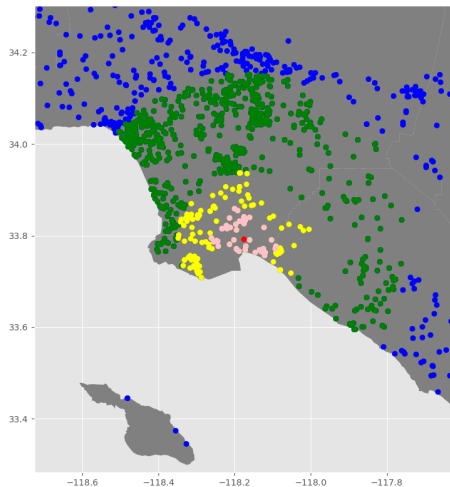


PurpleAir Outdoor Monitors

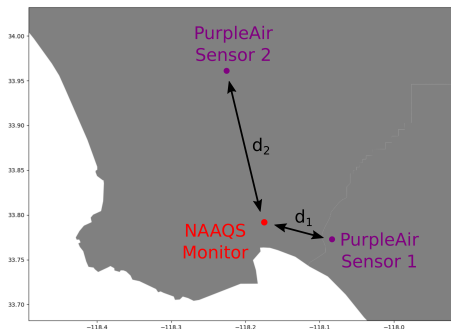
PurpleAir Sensors in California



LA Site Example



Alternate measure of ambient PM2.5 Concentration



Inverse-distance Weighted Average Ambient PM2.5

$$PA_t^{IDW} = \sum_{j=1}^{J_t} \frac{\frac{1}{d_j} \cdot PA_{j,t}}{\sum_j \frac{1}{d_j}} = \sum_{j=1}^{J_t} w_{j,t} \cdot PA_{j,t}$$

- J_t = active PurpleAir sensors around the NAAQS monitor at time t

Predicting Missing EPA Data

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

Table: Reported NAAQS Monitor PM2.5 (site 037-4004)

	(1)	(2)
intercept		6.924*** (0.076)
PurpleAir IDW Average	0.741*** (0.003)	0.444*** (0.004)
Observations	36,813	36,813
R^2	0.658	0.240
F Statistic	70924.412***	11642.169***

*p<0.1; **p<0.05; ***p<0.01

Calculating Design Values for an EPA Site

Pseudo Design Values

DV_p = p design value for quarter, calculated using **reported** PM2.5 data from EPA monitor, $\forall p \in \{\text{Daily, 24-Hour}\}$

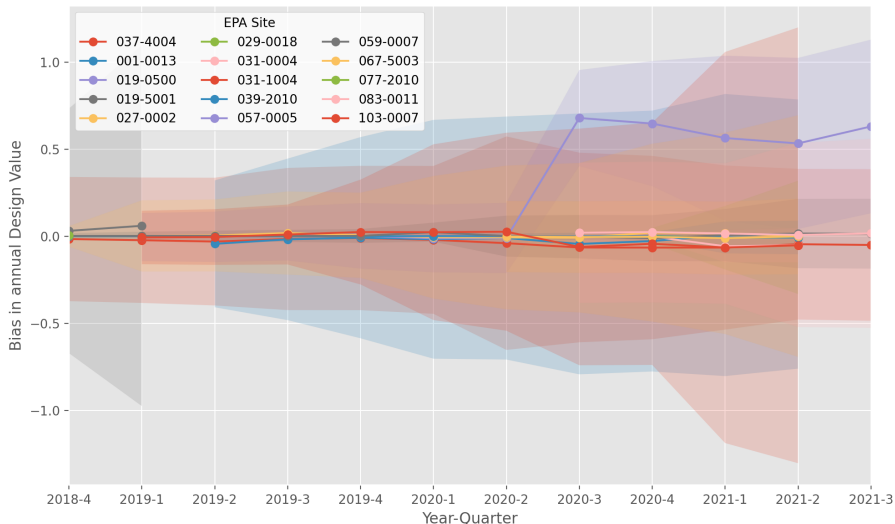
Imputed Design Values

\widetilde{DV}_p = p design value for quarter, calculated using **imputed** PM2.5 data from EPA monitor and PA sensors

Bias in Design Values

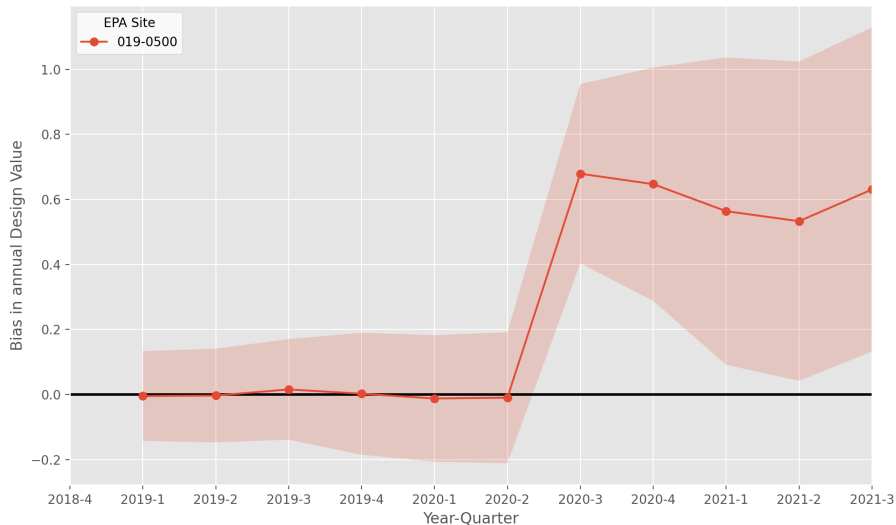
$$\text{bias from missing data in } DV_p \approx \widetilde{DV}_p - DV_p$$

Results for Daily Design Value: Sample EPA Sites



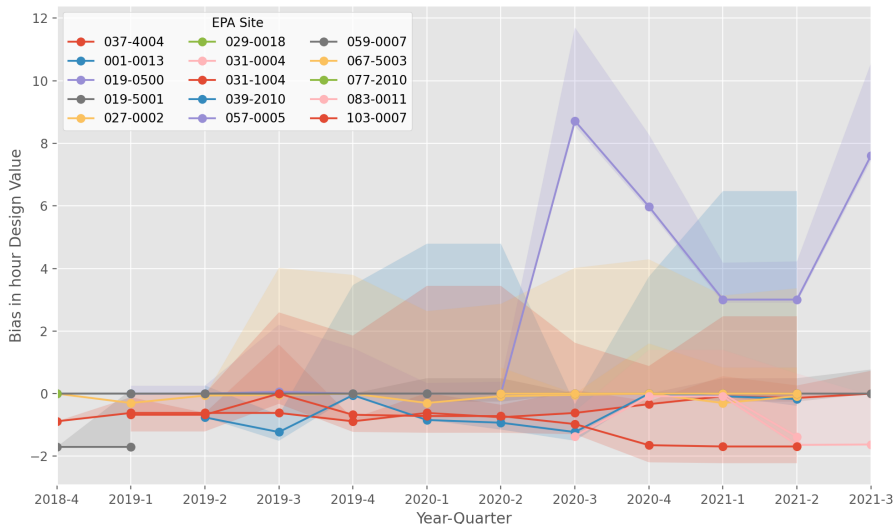
Shaded regions are 95% confidence intervals from interpolating the data.

Results for Daily Design Value: Fresno



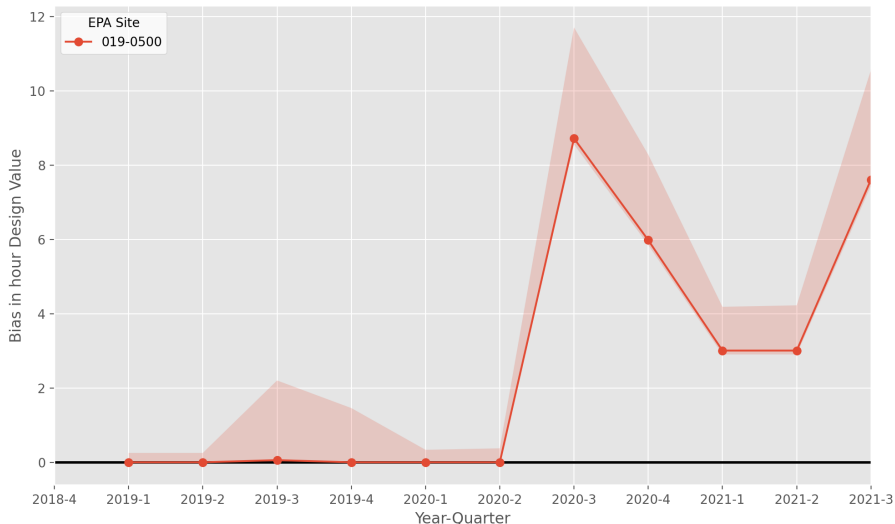
Shaded regions are 95% confidence intervals from interpolating the data.

Results for 24-hour Design Value: Sample EPA Sites



Shaded regions are 95% confidence intervals from interpolating the data.

Results for 24-hour Design Value: Fresno



Shaded regions are 95% confidence intervals from interpolating the data.

Conclusions & Discussion

- 1 Most tested sites show little evidence of bias from missing data, but one has large, meaningful bias
- 2 Even one site can affect millions of people due to the sparsity of monitoring sites
- 3 As high-pollution locations continue to reduce pollution, this bias may play an important role in determining NAAQS compliance
- 4 Underlines importance of expanding the monitor network or exploring alternative measures of ambient air quality

- Optimal regulation of ambient pollution under monitor expense-accuracy tradeoff
- Expand test to rest of US monitors
- Explore spatial distribution of air quality in unmonitored locations

Appendix: Correction of PurpleAir Readings

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

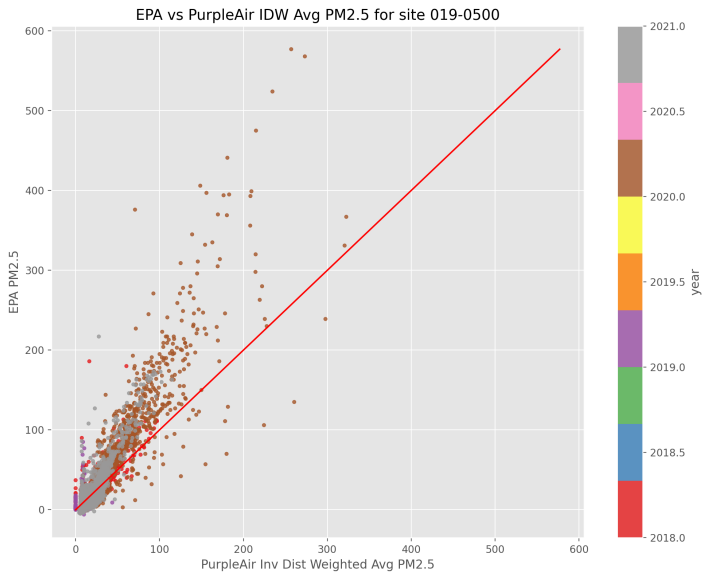
- $PA_{j,t}$ = ambient PM2.5 measured by PurpleAir sensor j at time t
- $H_{j,t} \in [0, 1]$ is the relative humidity

Appendix: Better Prediction of EPA PM2.5

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

- $PA_{j,t}$ = ambient PM2.5 measured by PurpleAir sensor j at time t
- $H_{j,t} \in [0, 1]$ is the relative humidity

Appendix: Fresno Site Data



Appendix: Fresno, California DV Bias Table

Table: Design Value Comparison for Fresno, CA. (95% CI Bounds)

Year-Quarter	Annual DV Difference	Upper Bound	Lower Bound	Hour DV Difference	Upper Bound	Lower Bound
2018-4	Invalid DV			Invalid DV		
2019-1	-0.005	0.133	-0.143	0.000	0.252	0.000
2019-2	-0.003	0.141	-0.147	0.000	0.252	0.000
2019-3	0.015	0.170	-0.139	0.058	2.202	0.000
2019-4	0.002	0.190	-0.185	0.000	1.460	-0.024
2020-1	-0.012	0.182	-0.207	0.000	0.335	0.000
2020-2	-0.010	0.191	-0.211	0.000	0.376	0.000
2020-3	0.679	0.954	0.403	8.718	11.704	8.556
2020-4	0.647	1.006	0.288	5.979	8.281	5.851
2021-1	0.564	1.036	0.091	3.007	4.184	2.903
2021-2	0.533	1.024	0.042	3.007	4.225	2.903
2021-3	0.630	1.129	0.132	7.607	10.557	7.444