Estimation and Implications of Bias in EPA Pollution Measurement

Aaron C Watt

December 2, 2021

Contents

Agenda

- communication over break and next term
- status of project
- winter break next steps
- discuss statistical analysis

Communication

- How far in advance of our meetings
  - send updated materials 2 days in advance
- How often should/can we meet
  - 2 weeks as baseline, scheduling more if things come up
- Email vs. zoom calls
  - zoom preferred

Status of project

- Have tested downloading full date range of PurpleAir data (python code ready)
- Have tested uploading files to S3 bucket (python code ready)

Winter break plans

Data

- Download hourly PA PM2.5 data from all US sensors (just in range of EPA monitors?)
- Download hours EPA PM2.5 data from all 88101 sensors (if more granular than spreadsheets)
- Add to S3 bucket

Analysis

- Develop understanding of regulatory process (going from data to attainment status)
    - Make “pseudo design values” using my data (just using thresholds, don’t worry about perfectly replicating the chosen attainment status/design values, but could have a map/stats comparing real to pseudo)
- Write code to implement attainment status given hourly data and check against actual attainment status

Datasets

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

Deadlines

- Dec. 24: PurpleAir is downloading hourly averages on 8 AWS tiny linux instances, sending CSVs to S3 bucket
- Jan. 5: Proof of concept of statistical analysis for 2 EPA sensors
    - Send Meredith update
- Jan. 14: Draft paper sections (esp. intro and data section)
- Jan. 21: Draft of all paper sections
- Feb 18: Paper due
- Feb 18: Presentation + other paper reviews
- Apr 11: Final paper (incorporating feedback)

Research Questions

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

Particulars of this project

- Using PurpleAir data on PM2.5
- Avoids using satellite estimates (but has drawbacks of it’s own – can document the extent of these).
- Focus on EPA pollution data that is missing *in time* (starting with every day, hourly sensors).

- Extensions to summer:
    - Missingness in space (placement of EPA monitors)
    - How biased are current measures of PM2.5 pollution faced by residents (population-weighted pollution)
    - Optimal regulation design: Given imperfect/incomplete reporting, spatial nature of weather, and the range of quality and price of different pollution monitors, what is the optimal regulation mechanism? At least, what is the optimal distribution of cheap vs. expensive monitors given budget constraints?

Models

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

Model 1: Predictive model of each EPA monitor PM2.5 pollution

Two models: - one with sensor-specific weights/regression ($\gamma_{i,j,k}$) - one with pooled sensor weights/regression ($\gamma_{j,k}$)

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

- 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).

Extreme Event from EPA data: indicator for imputed data when there was some nearby extreme event (eg, car fire), only use real data not imputed to create *gamma* weights.

Model 2: 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

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

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

- ML correction: include region, humidity, temperature, etc of PA monitors within radius.
- Come up with my own correction equation? How does this compare to ML correction?
- Compare both types of bias from uncorrected predictions, EPA corrected predictions, ML corrected predictions'

Model 3: 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.

Proposed Statistical Test

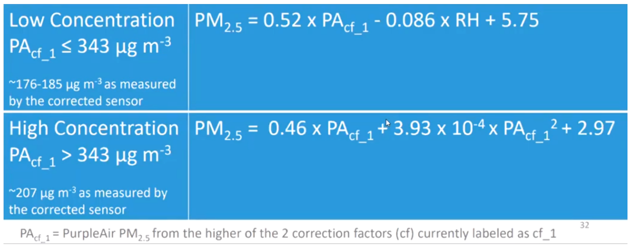
- The J_i group of PurpleAir sensors is (in a sense) a synthetic control for the EPA sensor i.
- **The question of bias in EPA monitor data can be stated**: is pollution concentration at the monitor during reported times significantly different from the concentration at the monitor at missing times? Is the difference greater than random variations?
- Implies an permutation inference test for each EPA monitor i.
- 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.
- **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 = $\frac{\text{\# of PA sensors in } i\text{'s radius with Ratio}_j \text{ larger than Ratio}_i}{\text{\# of PA sensors in } i\text{'s radius}}$

Welfare implications: use cheap comparison from Fowlie et al. 2019: look at counties that would be non-attainment if missing data were included, calibrate model (mortality, etc) on those specific counties and predict # of lives that would be saved, etc

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 1: 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

Data Wearhouse

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

Storage costs

- ~ 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