Second Year Paper Summary Statistics

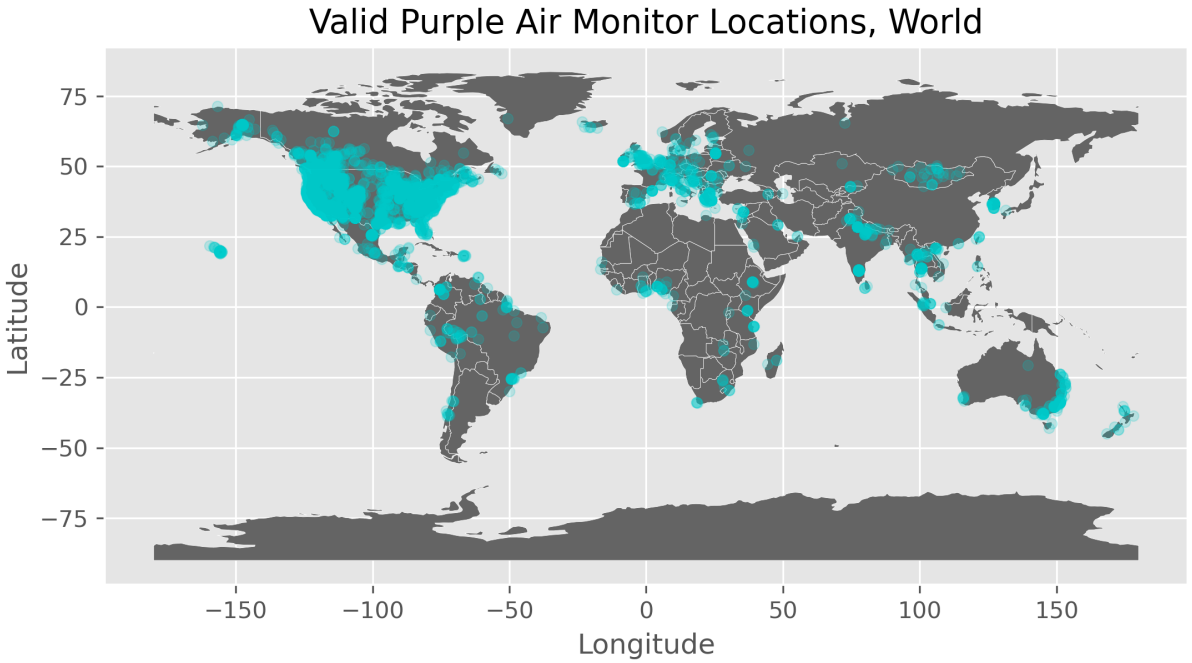
Aaron Watt

2021-11-08

Contents

Geographical Distribution of Purple Air Sensors

We can see from the figure below that the majority of Purple Air sensors are currently concentrated in the United States and Europe.



Note: These location data were collected in 2021, and I am unsure if there is a way to see the historical location of existing sensors or information on sensors that dropped out of the network.

In the US, the sensors are concentrated on the coastlines, with the majority in California.

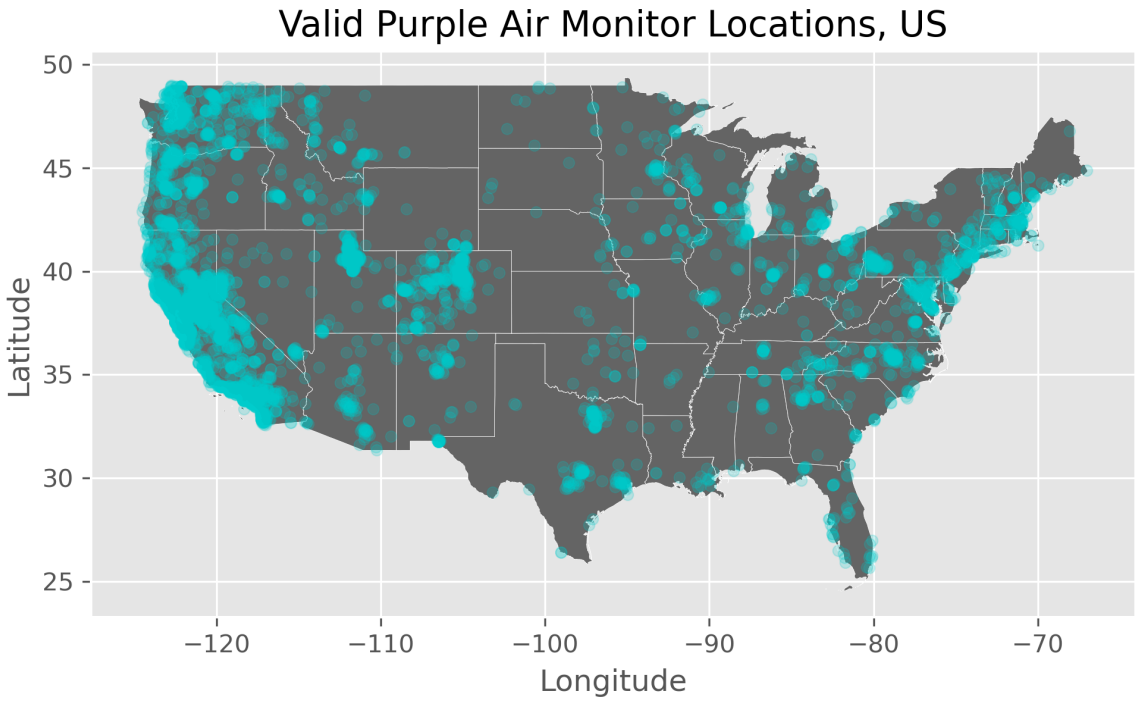


Figure 1: Valid Purple Air Monitor Locations, Contiguous United States

Within California, we can see the sensors are clustered around the Bay area and LA.

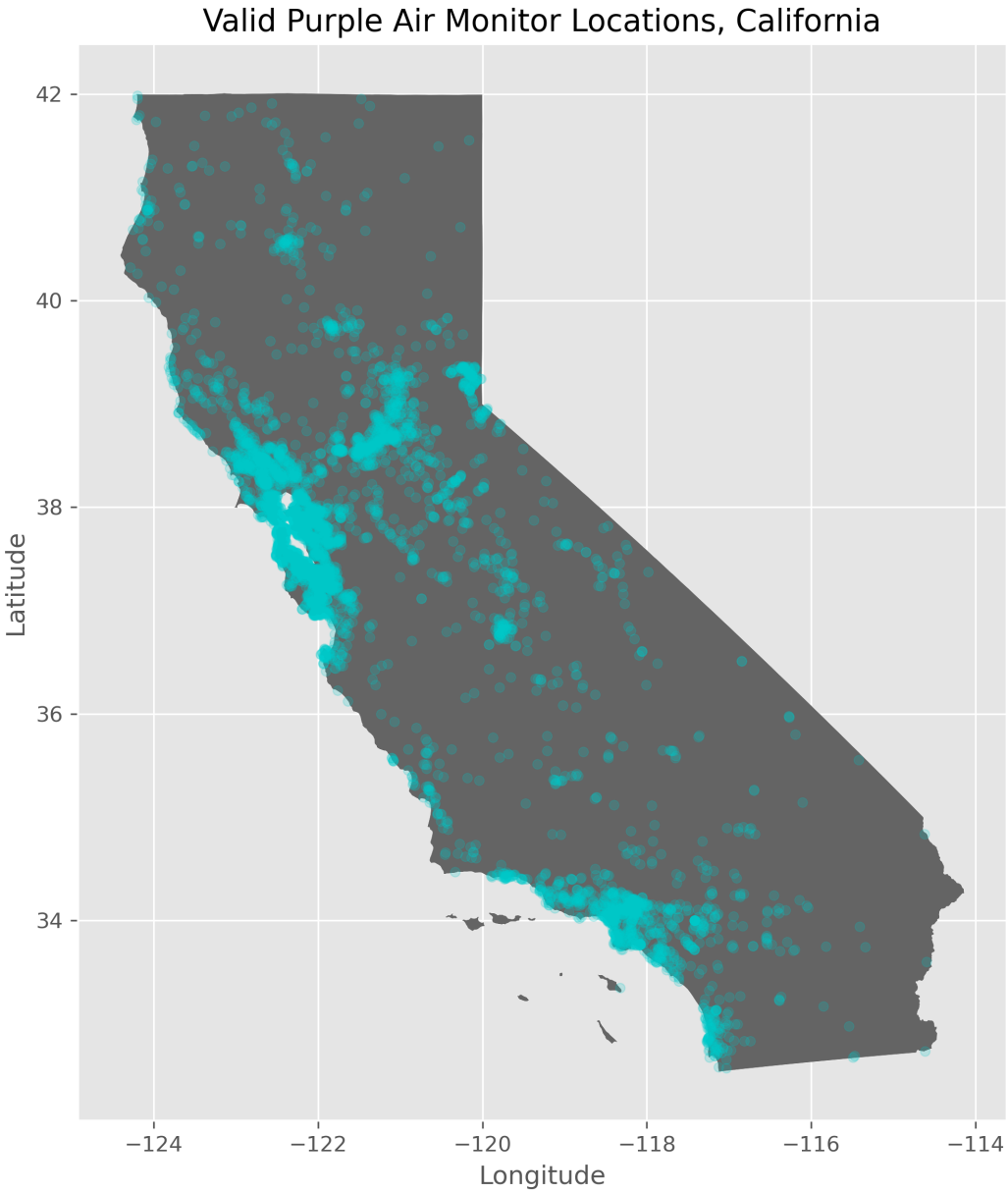


Figure 2: Valid Purple Air Monitor Locations, California

Distribution of adoption over time

For California, we can see from the below graph that adoption of Purple Air pollution monitors (in blue) increased steadily until summer of 2020, when adoption went up dramatically. There were particularly bad smoke days in populated areas of California that summer (e.g., "Mars Day"). However, these data are biased toward recent increases in adoption because I only have access to the sensors that are currently in the network – I have no data on the sensors that dropped out, therefore the increases pre-2020 are likely understated.

A more interesting red graph would probably be population-distance-weighted fires per-month or a population-weighted measure of visibility.

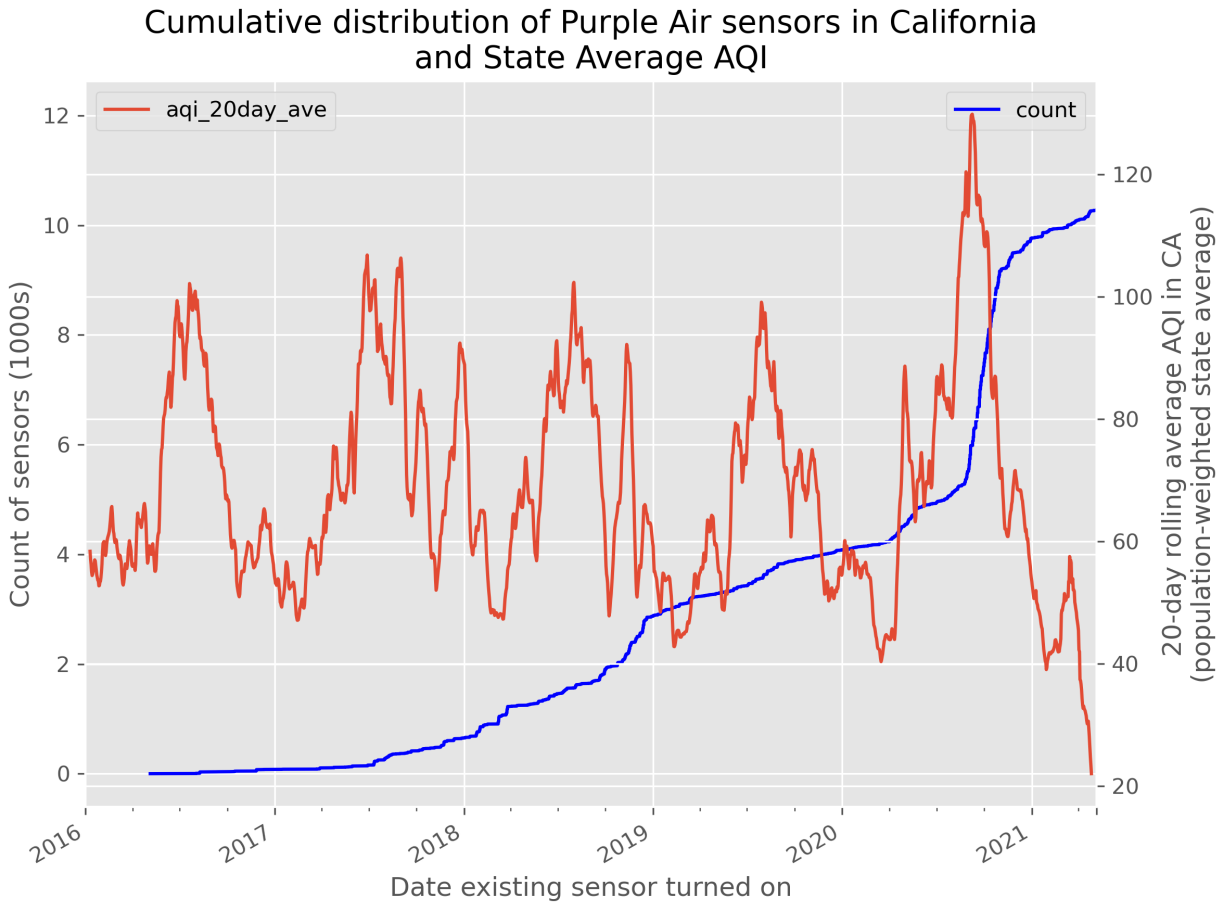


Figure 3: Purple Air Adoption over time vs. Air Quality in California

Table 1: Statistics of Purple Air Pollution Sensors Joining the Network

Year Sensor Joined the Network	Mean PM2.5	Sensors joined in CA	Max # of Sensors joining in any CA county
2016	19.599	78	31
2017	53.835	559	179
2018	15.069	2108	362
2019	17.650	1138	289
2020	7.003	5539	659
2021	7.936	2043	221

In the Figure 3 and Table 1, we see that many Purple Air pollution sensors joined the network in years of high pollution, or just after years of high pollution. One research question I hope to address is what might explain the spatial and temporal missingness of pollution data from county-level sensors that the EPA requires non-attainment counties to keep. Because Purple Air monitors create a network spread over the state, I can estimate pollution in areas that don't have EPA monitors or in places that have EPA monitors but at times when they are shut off. However, from Figure 2, we can see the distribution of Purple Air monitors is not distributed evenly or randomly.

Data plan

I am planning to limit my second year project to California and focus on comparing Purple Air PM2.5 data to EPA monitor data. I have not yet finished collecting PM2.5 data from Purple Air monitors but plan to collect years 2016-2021 for all California sensors via the Purple Air API. I am also in the process of collecting EPA PM2.5 data via the EPA API and have a small sample from Meredith.

In order to calibrate a selection model into Purple Air adoption, I plan to collect Census Block Group level demographic data and calibrate the model on the census block group level over 2018-2020 (inclusive). I also plan to collect weather data from PRISM.

To model measurement error of sensors (EPA sensors vs. Purple Air sensors), I plan to use pollution distribution data generated for IPCC climate models. I've been told these data are available to download and are probably the best measure of "ground truth" that I have access to.

Issues that will affect my estimates of pollution

Measurement error from the Purple Air monitors will need to be addressed. I plan to explore a combination of averaging over multiple sensors and regression-based correction using EPA monitors that are near some Purple Air monitors as ground truth pollution for that location.

I also plan to estimate a selection model into Purple Air adoption so omitted variable bias will be a large concern. I will have census data at the census block group level that I can match to the location of the PA monitors. But that is not unit-level data so also suffers from measurement error if I am estimating an adoption model at the unit level. I am also interested in modeling this as a binary discrete choice model with block- or county-level adoption shares (percentage of the PA monitors adopted in that month, in that county). With either of these models of adoption, I will need to instrument PA sensor prices, but since it is a homogeneous good sold by only one firm and the prices have remained nearly the same for 4 years, I would predict I will have a weak instrument problem.