Time series modeling for air pollution monitoring with a focus on the correcting low-cost PM2.5 sensor measurements

Oladimeji Mudele, Ph.D

Post-Doc Research Fellow
Harvard University, Boston, MA, USA

January 21, 2023

Bits of my background

- B.Eng, Elect/Elect. Engineering, FUTA, Nigeria
- M.Sc Electronic Engineering (summa cum laude), UNIPV, Italy.
- Ph.D., Electronics, Computer Science and Electrical Engineering (2021)
- Experience as a Research Engineer in a deep tech. company.
- Currently at Harvard as a Post-Doc Research Fellow working on AI, causal inference, and public health problems.

Air pollution

Definition

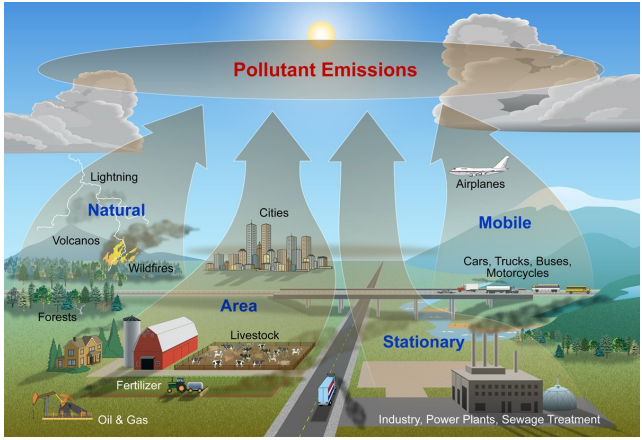
Air pollution is the contamination of the indoor or outdoor environment by any chemical, physical or biological agent that modifies the natural characteristics of the atmosphere.

Example

Particulate matter, carbon monoxide, ozone, nitrogen dioxide and sulfur dioxide.

Sources of air pollution

1



[1] (https://www.nps.gov/subjects/air/sources.htm)

Particulate matter - PM2.5

Definition

Atmospheric particulate matter with aerodynamic diameter of 2.5 microns (μm) or less.

Health outcomes

- Exposure to PM2.5 has been associated with negative health outcomes including **asthma, chronic obstructive pulmonary diseases**, and others
- Short-term PM2.5 exposure is linked to coughing, sneezing, and shortness of breath.

Outdoor sources

- Automobile exhausts, burning of fuels such as wood, oil or coal, power plants, and forest fires.

Air pollution - the need for low-cost sensing

- In the U.S., traditional air quality measurements follow the metrics established by the United States Environmental Protection Agency (EPA) using equipment that implement the federal reference method (FRM) or federal equivalence method (FEM).
- FEM/FRM-based monitors cost tens of thousands of dollars and require significant infrastructure and trained personnel to operate.
- In 2019, the global mean population distance to the nearest PM2.5 monitor was 220 km.
- In the U.S., more than 70% of counties do not have regulatory PM2.5 monitoring[2].

[2] (`https://www.denvergov.org/`)

Low-cost air quality sensors - correction methods

While low-cost sensors cannot replace traditional systems, they can expand access to air quality monitoring.

Correction methods

1. Nominal correction (laboratory).
2. Model-based correction using collocated FEM/FRM data.

Training a correction model (statistical/ML) is generally more accurate because it is based on data from realistic meteorological and air pollution conditions.

Modeling

Correction model

$$y_r = f(y_l, \mathbf{x})$$

- y_r: amount of PM2.5 in the atmosphere measured by FEM/FRM reference monitor.
- y_l amount of PM2.5 measured by the low-cost sensor we want to correct.
- $\mathbf{x}$: other features including spatial, temporal, and meteorological terms (e.g temperature and humidity)
- In reality, the model output will be $\hat{y}_r$ which will be much closer to y_r than y_l is.

Case study: Love My Air (LMA) program - Denver

- Denver experiences significant construction and traffic congestion.
- Denver is the 14th-worst among major US cities in air quality
- Goal: To empower community with local air quality data using a network of low-cost and traditional sensing infrastructure.
- Project: Low-cost PM2.5 sensors installed in a network of Denver Public Schools (DPS)[3].

[3](Source: `https://www.denvergov.org/`)

Reference study using LMA data

- Considine, Ellen M., et al. "Improving accuracy of air pollution exposure measurements: Statistical correction of a municipal low-cost airborne particulate matter sensor network." Environmental Pollution 268 (2021): 115833.
- The study applied random forest and linear model to correct air quality data measured using five LMA (DDPHE) low-cost sensors across the city of Denver, US.
- Results show that in addition to y_l, **temperature, and humidity, plus a near-highway indicator** produced the best model (random forest)

Study data: sensing

- Love My Air sensors are Canary-S (CS) models
- Records PM2.5, temperature ,and humidity; and uploads minute-resolution measurements via cellular data.
- "True" PM2.5 amount data are obtained from FEM reference sensors are part of the AirNow[4] network.
- FEM data are provided hourly data.

[4] https://www.airnow.gov/

Study data: collocation sites

- Aug. 2018 to May 2019 - National Jewish Hospital (NJH), La Casa, and I25-Globeville. Three CS sensors were collocated with the I25-Globeville reference FEM.
- Sept. 2019 to mid Dec. 2019 - CAMP, I25-Denver.
- Studies have shown that thoughtful placement of at least three collocation sites is preferable.

Study data: bringing it all together

PM2.5

Hourly measurements and hourly averages from AirNow and CS sensors, respectively.

Other features to account for in our model

- Temperature
- Humidity
- Distance to highway - near highway I-25-Globeville and I-25 Denver.
- Road lengths within buffer radius (City of Denver Open Data Catalog)
- Cyclical values of month and hour of the day to account for daily and seasonal variations

Study data: training and test

Goal

Obtain a model that generalise into all seasons in the year and to new locations (unseen) during training

- **Training**: Aug. 2018 to May 2019 - National Jewish Hospital (NJH), La Casa, and I25-Globeville.
- **Test**: Sept. 2019 to mid Dec. 2019 - CAMP, I25-Denver.

Useful questions to guide our lab sessions

- Can we correct the PM2.5 readings from low-cost sensors using data from synchronous collocated FEM sensors and a model?
- Above is Yes. But, what kind of model would be useful and which model can produce the best results for our purpose?
- What variables do we need to account for in our model(s)?
- What are the contributions of each variables/features account for in the model to model performance?
- How do we evaluate this model? (Error evaluation and cross-validation)

Now, let's dive into code