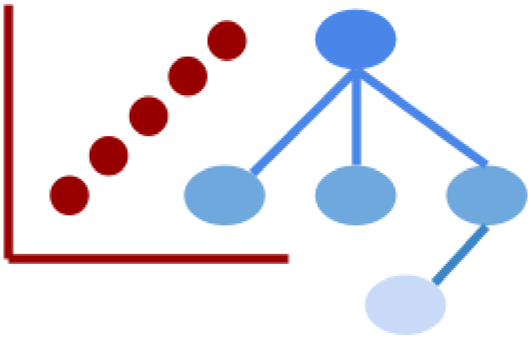
An Exploration of Machine Learning Techniques for Predicting Air Quality Using Low-Cost Sensors



by

Garima Raheja

PhD Student, Department of Earth and Environmental Sciences
Lamont-Doherty Earth Observatory
Of Columbia University

EAEEE4000 Machine Learning for
Environmental Engineering
Final Project

December 24, 2021

garima.raheja@columbia.edu

1. Introduction

a. Background

Air pollution is a burgeoning global health crisis. In 2019, it rose from the 5th to the 4th leading risk factor for premature death, associated with 6.67 million premature deaths and up to 6 million premature births. $PM_{2.5}$, air pollution composed of inhalable particulate matter smaller than 2.5µm in aerodynamic diameter, is linked to asthma, ischemic heart disease, type II diabetes, lung cancer and other deleterious health effects. These particles are emitted from vehicles, coal-burning power plants, waste incineration and other anthropogenic and natural sources. Thus far, most academic research, monitoring and media attention regarding $PM_{2.5}$ exposure has been largely focused on the United States, Europe and recently, China. Additional research is vital and urgent for other regions where air pollution levels might be even higher: in particular, countries in Asia, Africa and the Middle East face the highest levels of ambient $PM_{2.5}$ yet remain scarcely monitored.

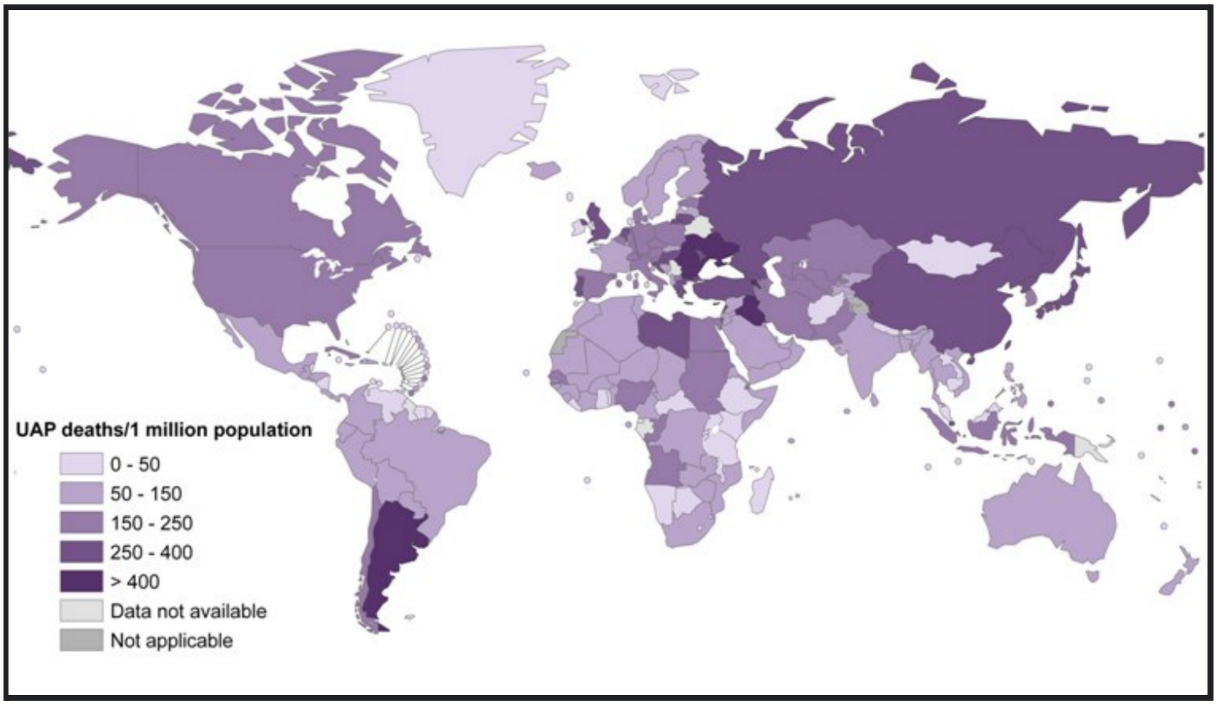


Figure 1. Deaths attributable to air pollution. Source: World Health Organization

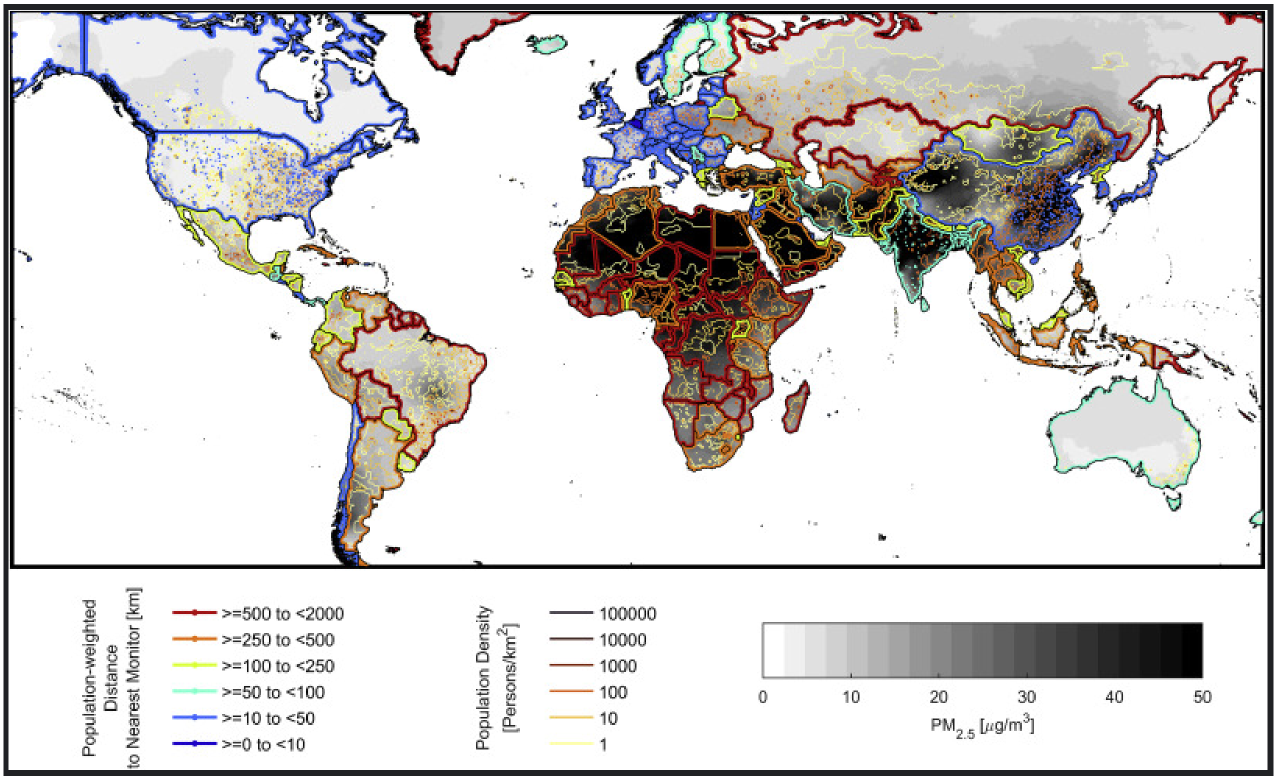

Population-weighted Distance to Nearest Monitor [km]
>=500 to <2000
>=250 to <500
>=100 to <250
>=50 to <100
>=10 to <50
>=0 to <10
Population Density [Persons/km^2]
100000
10000
1000
100
10
1
0
10
20
30
40
50
PM$_{2.5}$ [μg/m^3]

Figure 2. Population-weighted distance to nearest air pollution monitor. From Martin et al 2019.

Sparsity in air pollution monitoring creates high uncertainty in exposure and health impact assessments. The global mean population distance of a person to a $PM_{2.5}$ monitor is about 220 km. This includes large portions of low- and middle-income countries (LMICs). Often, this sparsity is engendered by the high cost of equipment; Federal Equivalent Method or Federal Reference Method equipment (referred to as reference monitors) which measure ambient $PM_{2.5}$ concentrations, such as the frequently used MetOne Beta Attenuation Monitor (BAM) 1020 or Teledyne T640, can cost several orders of magnitude higher than low-cost (LCS) sensors when accounting for sensor climate control and associated maintenance costs. Even when a city or regional agency accrues the funds to purchase a monitor, the singular sensor cannot capture the vast heterogeneity that has been shown to impact neighborhood-scale exposure**.**

Recent advances in LCS measurement technology, and calibration of LCS sensor data using data science techniques, allow for high-density, real-time monitoring, and automated web-based archiving of ambient $PM_{2.5}$ concentrations. LCS for measuring air pollution and identifying sources offer a possible path forward to remedy the lack of data in resource-limited locations such as sub-Saharan Africa. For LCS to provide useful, actionable information, understanding local conditions is vital. Calibration factors and sensor technical performance vary strongly with environmental conditions, such as temperature, humidity, and air pollution loading. The PurpleAir sensor, when calibrated and corrected correctly, has demonstrated high accuracy in comparison to reference-grade monitors.

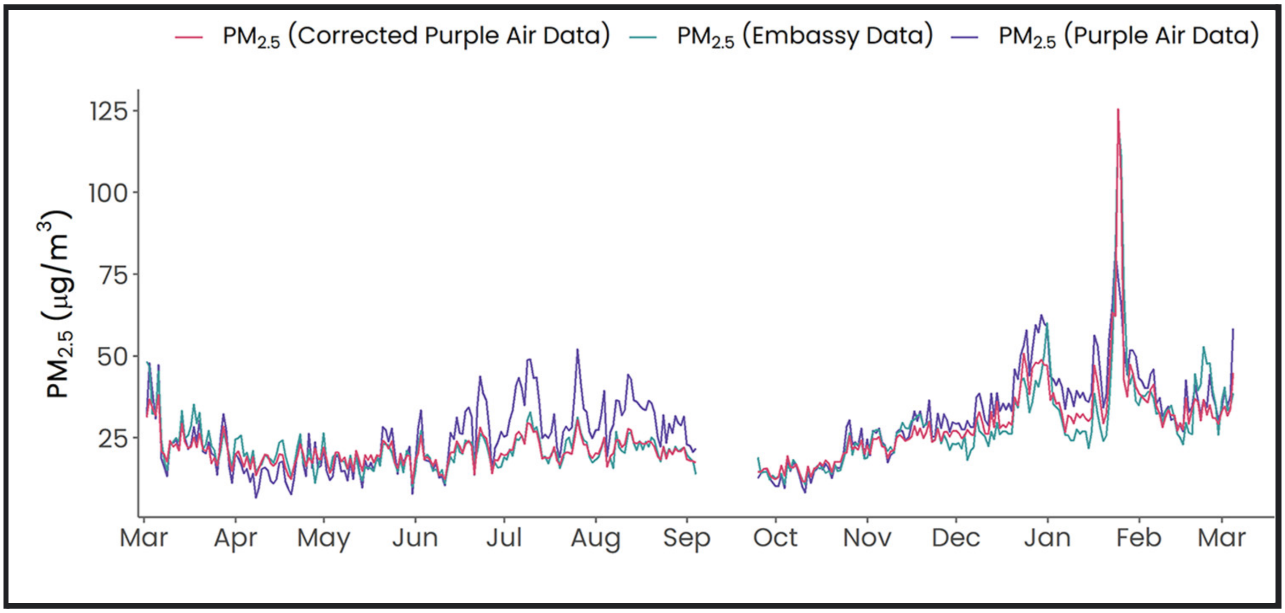


Figure 3. Corrected Purple Air compared to reference-grade monitor at the US Embassy in Accra, Ghana from March 2020 to March 2021. Soure: McFarlane et al, 2021.

b. Problem Statement

Traditionally, LCS corrections have largely been limited to estimates, or in some advanced cases, simple linear regressions. Though this is due to the novelty of the research area, it is also because many communities or groups using LCS often lack the technical knowledge to attempt other methods. Consequently, there are many available avenues of exploration of techniques that use advanced data science methods to enhance the LCS corrections, and I am uniquely positioned to further the field.

2. Dataset

a. Acquiring the Dataset

In order to use machine learning to create better LCS correction models, I must first acquire a large, diverse dataset from LCS that are co-located with reference grade monitors. To do this, I used a two-pronged approach:

1. Publicly Available Open Source Data

Using the python requests package, I queried the OpenAQ API to scrape lists of all the LCS and reference-grade monitors available in the archives. I then used a nearest neighbor methodology to cross-reference both lists to find 12,062 co-locations of LCS and reference-grade monitors within 200 m of each other. From this, I selected co-locations where both sensors were monitoring outdoor conditions, had sufficient metadata, and had 80% completeness for one year of data. Using the python requests package and parallelization, I downloaded one year of data from each co-located LCS and reference monitor, and then cleaned the dataset to remove rows with missing values or values above physical reality (humidity > 100%, $PM_{2.5} < 0$, etc). While this proved to be an immense resource, the OpenAQ data download does not include a field for the sensor name – there is no way to differentiate between a Purple Air measurement versus a Clarity or MODULAIR sensor measurement. Though many different LCS use the Plantower optical particle counter inside, each sensor brand uses a unique and secret methodology to convert raw Plantower data into $PM_{2.5}$ numbers. I used co-located data from Accra, Ghana to investigate this.

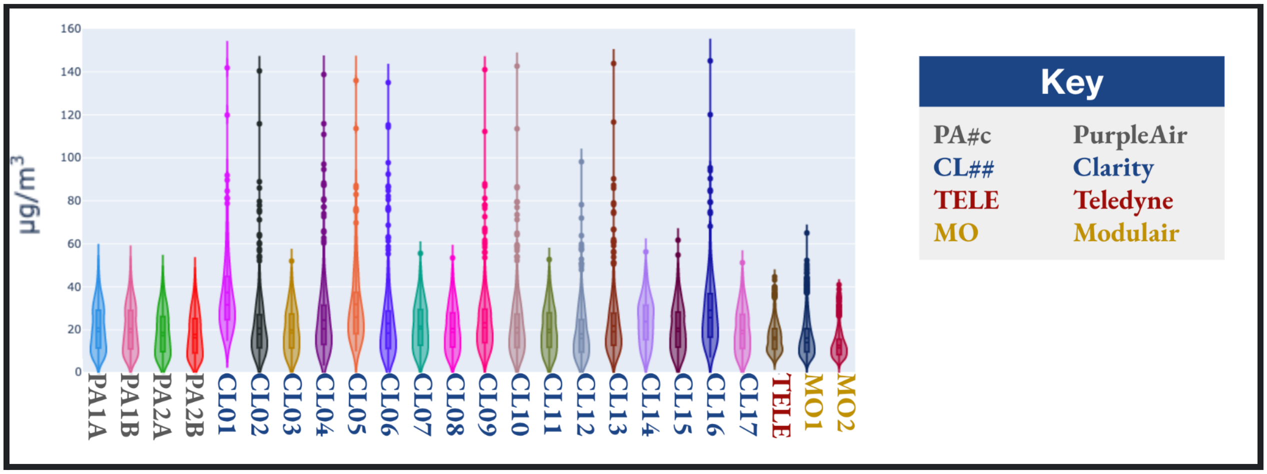


Figure 4. Co-located raw data from 2 Purple Air, 17 Clarity, 2 MODULAIR (all low-cost) and 1 Teledyne (reference grade monitor) from two months in a laboratory setting in Accra, Ghana.

Since the colocation here shows that the low-cost sensors generally perform similarly under laboratory conditions, for now it is acceptable to use all the OpenAQ data as one set. However, for future work, it will be important to understand how the different LCS behave when deployed outside the laboratory. For this, I contacted OpenAQ and suggested adding a field for the “type of LCS” when data is retrieved.

2. CAMS-Net

Working with my advisor, Prof. Dan Westervelt, and a fellow researcher on our team, Celeste McFarlane, we reached out to members of the Clean Air Monitoring and Solutions Network (CAMS-Net). CAMS-Net is an NSF-funded global collaboration of air pollution researchers and policymakers. We requested colocated LCS / reference monitor data from teams around the world, allowing us access to a diverse dataset showing LCS in differing ambient conditions from around the world. So far, we have received and cleaned data from 8 locations, with more teams continuing to send their data access information to us.

The most common data feature used to predict actual air quality is the LCS measurement. Further features of the dataset can come from co-measurements of temperature, humidity, pressure and other meteorological parameters measured by the devices. Additionally, features can be created using inherent properties, such as converting the timestamp to individual day, month, week, and year columns. The “true” value is the number measured by the reference-grade monitor, which is considered the best possible estimate of ambient air quality.

b. Exploratory Data Analysis - Comparing LCS

First, I wanted to investigate the correlation between LCS and reference monitors in a laboratory setting, and understand how they differ between sensor types. This will inform which machine learning techniques I attempt for this project. Using the laboratory co-location in Accra, Ghana, I plotted each LCS against the Teledyne reference monitor.

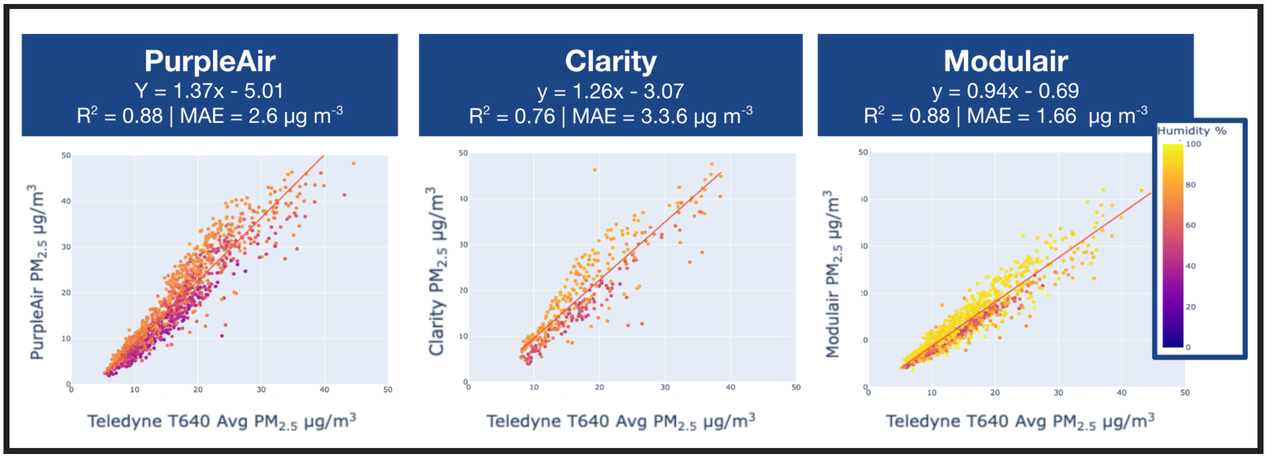


Figure 5. Simple linear regression of each LCS type against reference monitor, with ordinary least squares line drawn in red. The color of the dot represents humidity at each measurement.

This shows that generally, there is a linear relationship between LCS and reference monitors, but the linearity of the relationship is slightly different for each monitor type. Further, the linearity changes with the ambient meteorology regime. For example, in Figure 5, the Purple Air plot seems like the error would decrease if separate linear regressions were used for data < 50% humidity and > 50% humidity. In the Clarity plot, the higher humidity regime shows a slightly nonlinear relationship, but the lower humidity regime shows a strong linear relationship – neither is captured perfectly by the aggregated ordinary least squares line. This means that to develop a truly universal correction model, we need to create a method that allows for differentiation of sensor types and further allows for linear and nonlinear relationships within the correlations.

Since the OpenAQ dataset does not contain this vital “type of LCS” feature, I decided not to use the OpenAQ dataset at this time, and decided to focus on just the CAMS-Net collection which only contains data from Purple Air sensors.

c. Exploratory Data Analysis - Understanding the CAMS-Net Dataset

I used the dataprep.eda module for conducting an exploratory analysis of the CAMS-Net Dataset, which contained 1157 rows of measurements.

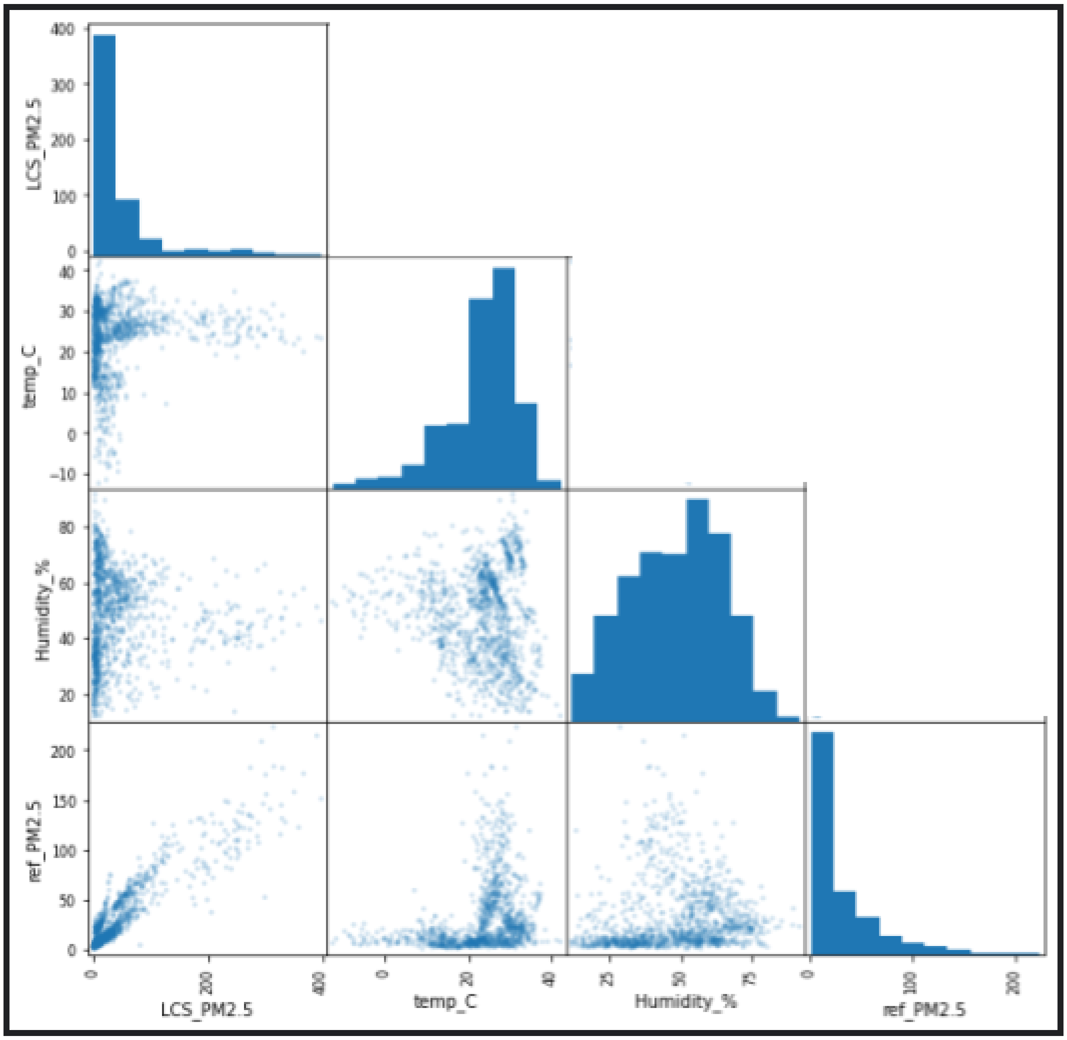


Figure 6. Correlation plots between the 4 measured variables in the CAMS-Net dataset. Diagonals show the histogram of each feature. The humidity and temperature show a vaguely normal distribution; the LCS and reference $PM_{2.5}$ are skewed similarly to the right.

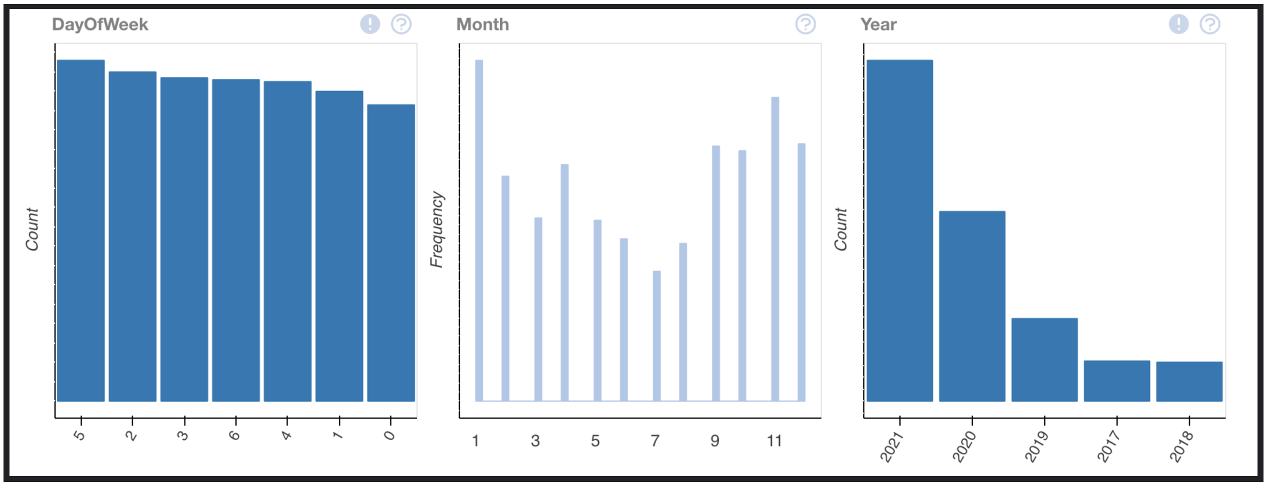


Figure 7. Histograms of the features derived from the timestamp of the CAMS-Net dataset. This shows an even spread across days of the week and months of the year. Most of the data is from the last two years.

d. Key Takeaways

The exploratory data analysis using the figures 6-7 and others not shown indicates that this dataset covers diverse spatiotemporal regimes and contains only rows within physical realities. It also shows that the LCS to reference monitor correlations change based on a number of underlying factors. This means that our universal correction model will need access to all the features but might choose that some are redundant. It also means that we will need to choose a model that allows for differential modeling within regimes.

3. Methods

a. Establishing a Baseline

To establish a baseline, I conducted a naive linear regression of all LCS and reference data. The raw data has a mean absolute error of 19.94 $\mu g/m^3$. An ordinary least squares of all the LCS and reference data within the dataset gives an R^2 value of 0.82. We will use this to assess the quality of future models.

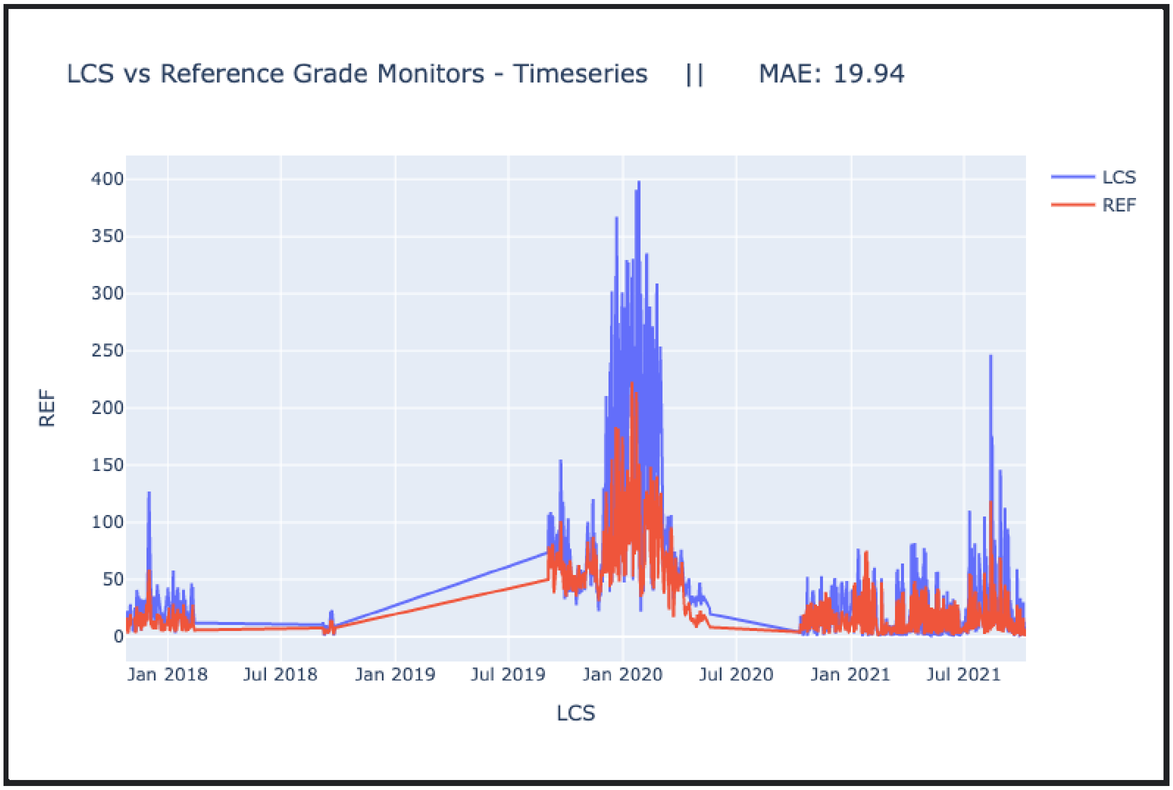


Figure 8. Raw LCS and reference data from all sites plotted as a collective timeseries. The blue line is LCS and the red line is reference monitor data.

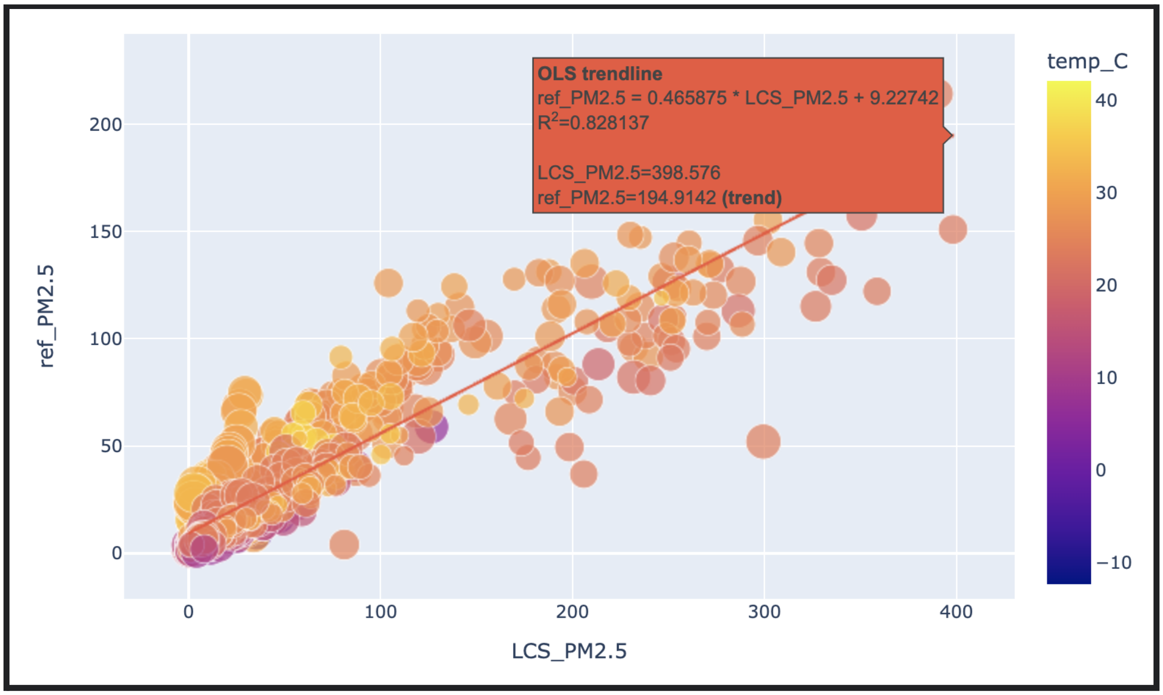


Figure 9. Comparing all LCS to reference $PM_{2.5}$. Points are shaded by temperature and sized by humidity values.

Similarly, I separated data by site and found MAE values for all individual sites, to use for comparison. The starting MAE at the Kampala site is 14.79 µg/m³.

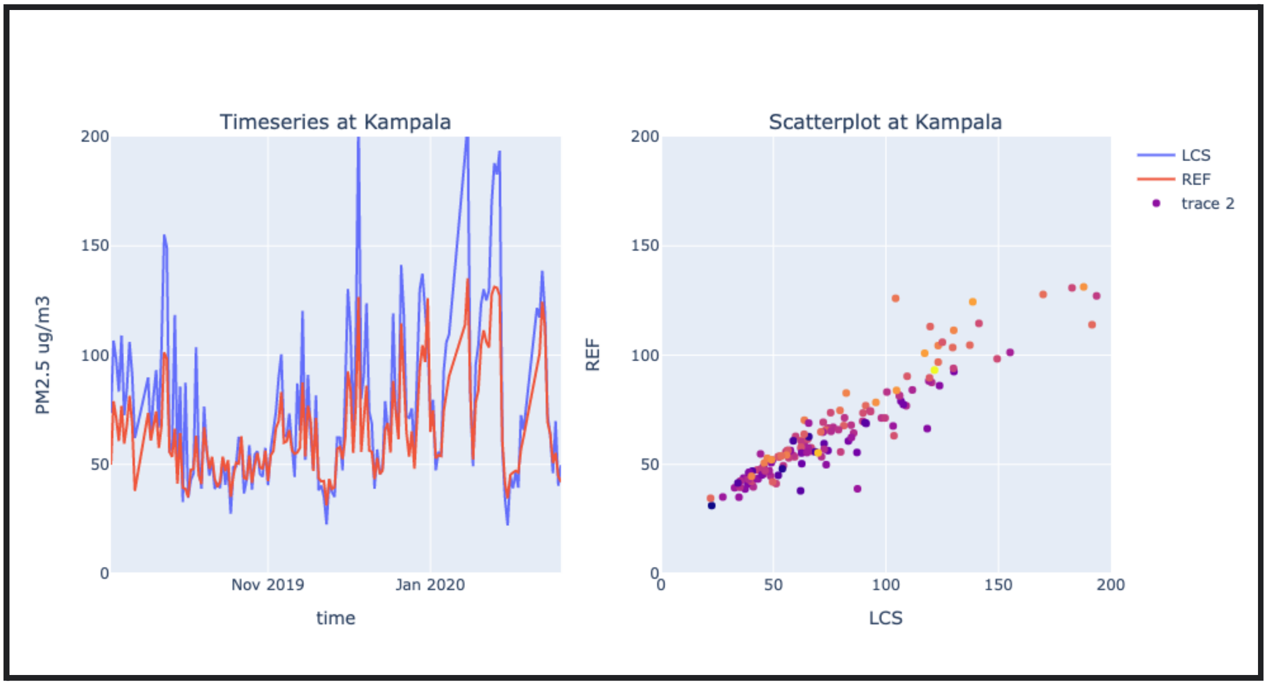
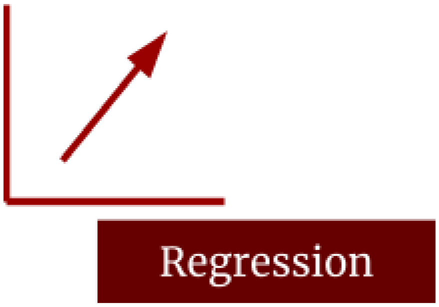


Figure 10. Timeseries and scatterplot at the Kampala site.

b. Linear Regression

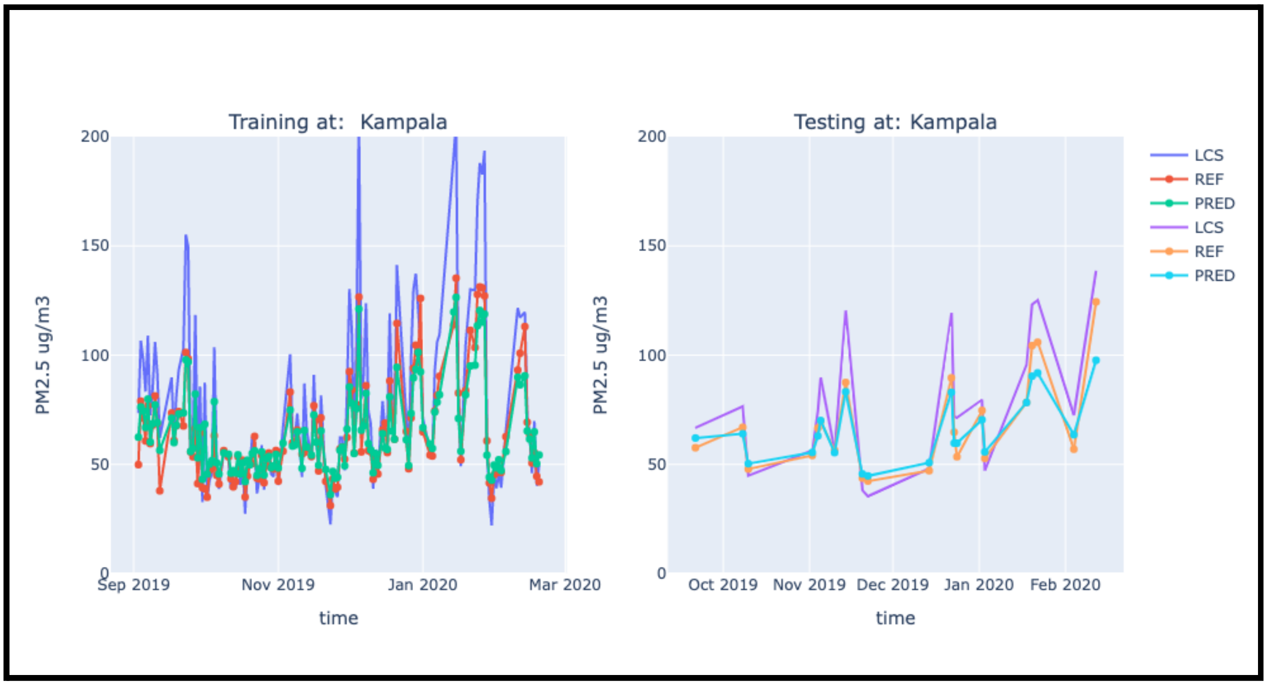


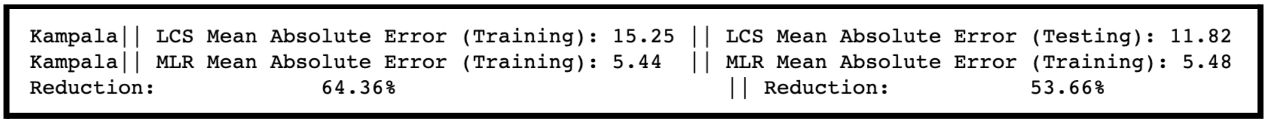
First, I used sklearn to create a 80-20 train-test data split and then train amultiple linear regressions for the entire dataset, and then applied the MLR to each LCS dataset and compared the corrected LCS to the reference datasets in both the training and testing sets. Removing all the coefficients < 0.001, the MLR equation was derived to be this:

$$Y_{pred} = 0.48\,LCS + 2.84\,Temperature - 0.19\,Humidity - 0.016\,LCS_{lat} + 0.006\,Distance_{CDI} + 0.002\,Distance_{GC} - 0.02\,Distance_{Woodland} - 0.02\,Distance_{Kampala} - 0.01\,Distance_{Kolkata} + 0.01Distance_{PGHPkwE} - 0.01\,Distance_{PGHLin} + 0.02\,DayOfWk + 0.08\,Month - 2\,Year - 0.02\,Week$$

This indicates that the most important features are the year of measurement and the LCS measurement. The LCS measurement is obviously important; the year could be important since climate change is causing concentrations to rise in many regions.

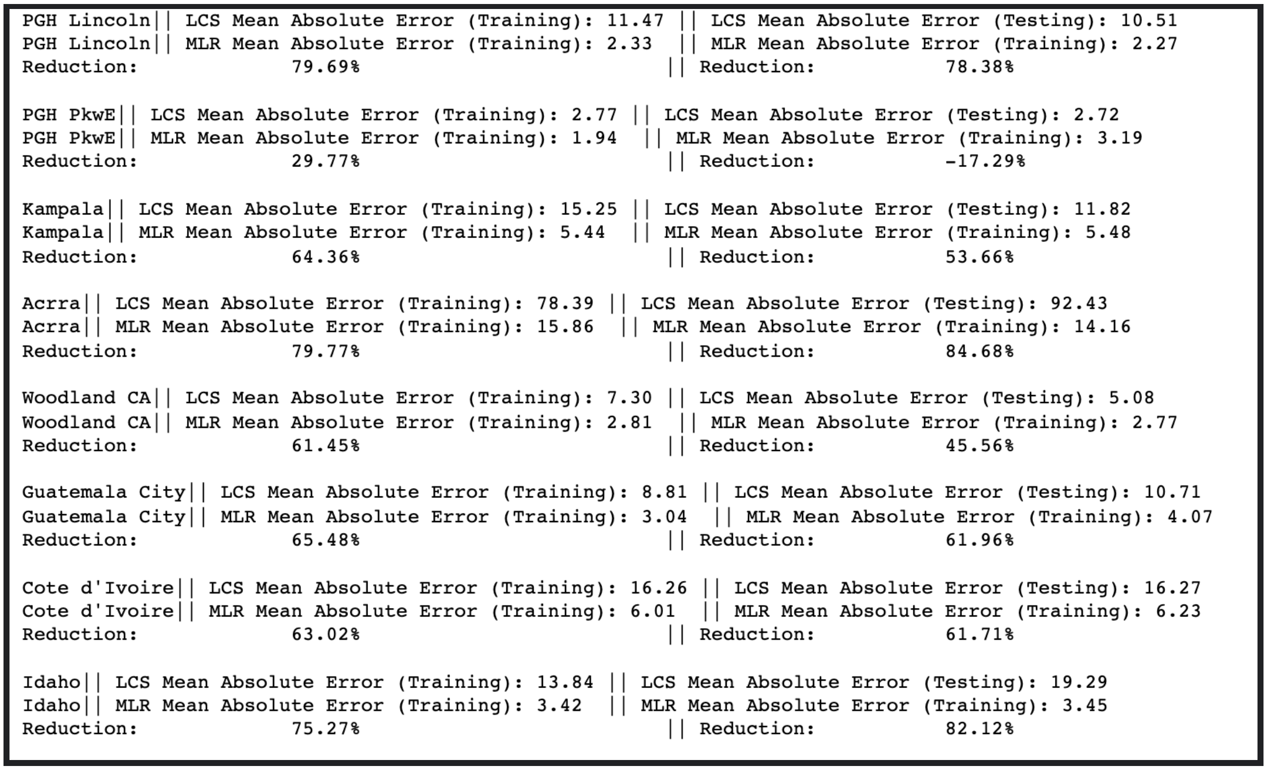
This MLR drastically reduced MAE at all sites, as demonstrated in Table 1.





```
Kampala|| LCS Mean Absolute Error (Training): 15.25  || LCS Mean Absolute Error (Testing): 11.82
Kampala|| MLR Mean Absolute Error (Training): 5.44   || MLR Mean Absolute Error (Training): 5.48
Reduction:              64.36%                         || Reduction:            53.66%
```

Figure 11. MLR results from Kampala. The LCS data is shown in blue and purple. The reference monitor data is shown in the dotted red and orange. The MLR-predicted values are shown in the dotted green and teal.



```
PGH Lincoln|| LCS Mean Absolute Error (Training): 11.47 || LCS Mean Absolute Error (Testing): 10.51
PGH Lincoln|| MLR Mean Absolute Error (Training): 2.33  || MLR Mean Absolute Error (Training): 2.27
Reduction:              79.69%                         || Reduction:            78.38%

PGH PkwE|| LCS Mean Absolute Error (Training): 2.77 || LCS Mean Absolute Error (Testing): 2.72
PGH PkwE|| MLR Mean Absolute Error (Training): 1.94  || MLR Mean Absolute Error (Training): 3.19
Reduction:              29.77%                         || Reduction:            -17.29%

Kampala|| LCS Mean Absolute Error (Training): 15.25 || LCS Mean Absolute Error (Testing): 11.82
Kampala|| MLR Mean Absolute Error (Training): 5.44  || MLR Mean Absolute Error (Training): 5.48
Reduction:              64.36%                         || Reduction:            53.66%

Acrra|| LCS Mean Absolute Error (Training): 78.39 || LCS Mean Absolute Error (Testing): 92.43
Acrra|| MLR Mean Absolute Error (Training): 15.86  || MLR Mean Absolute Error (Training): 14.16
Reduction:              79.77%                         || Reduction:            84.68%

Woodland CA|| LCS Mean Absolute Error (Training): 7.30 || LCS Mean Absolute Error (Testing): 5.08
Woodland CA|| MLR Mean Absolute Error (Training): 2.81  || MLR Mean Absolute Error (Training): 2.77
Reduction:              61.45%                         || Reduction:            45.56%

Guatemala City|| LCS Mean Absolute Error (Training): 8.81 || LCS Mean Absolute Error (Testing): 10.71
Guatemala City|| MLR Mean Absolute Error (Training): 3.04  || MLR Mean Absolute Error (Training): 4.07
Reduction:              65.48%                         || Reduction:            61.96%

Cote d'Ivoire|| LCS Mean Absolute Error (Training): 16.26 || LCS Mean Absolute Error (Testing): 16.27
Cote d'Ivoire|| MLR Mean Absolute Error (Training): 6.01  || MLR Mean Absolute Error (Training): 6.23
Reduction:              63.02%                         || Reduction:            61.71%

Idaho|| LCS Mean Absolute Error (Training): 13.84 || LCS Mean Absolute Error (Testing): 19.29
Idaho|| MLR Mean Absolute Error (Training): 3.42  || MLR Mean Absolute Error (Training): 3.45
Reduction:              75.27%                         || Reduction:            82.12%
```

Table 1. MLR-based MAE reductions across training and testing datasets at each site.

Table 1 shows that MLR is most successful in reducing MAE when the starting MAE is very high. In Accra, the starting MAE was > 75 µg/m^3 and the MLR was able to bring MAE to < 16 µg/m^3– reductions of 80-85%. However, the MLR is less effective when the initial MAE is very low. In the PGH PkwE site, the initial MAE is approximately 2.7 µg/m^3 but the MLR MAE is 1.9 - 3.1 µg/m^3 . In the testing case, the MAE is actually higher than baseline.

c. Decision Trees

I decided to try decision trees because they replicate how many in this field generally attack the problem of correcting LCS data. First, they usually decide based on sensor type, and then based on ambient conditions (meteorology and average $PM_{2.5}$ loading), and then based on humidity and temperature. I was curious to see if the machine learning result would be similar to the method used in the field.

First, to decide the depth of the decision tree, I analyzed a few different options.

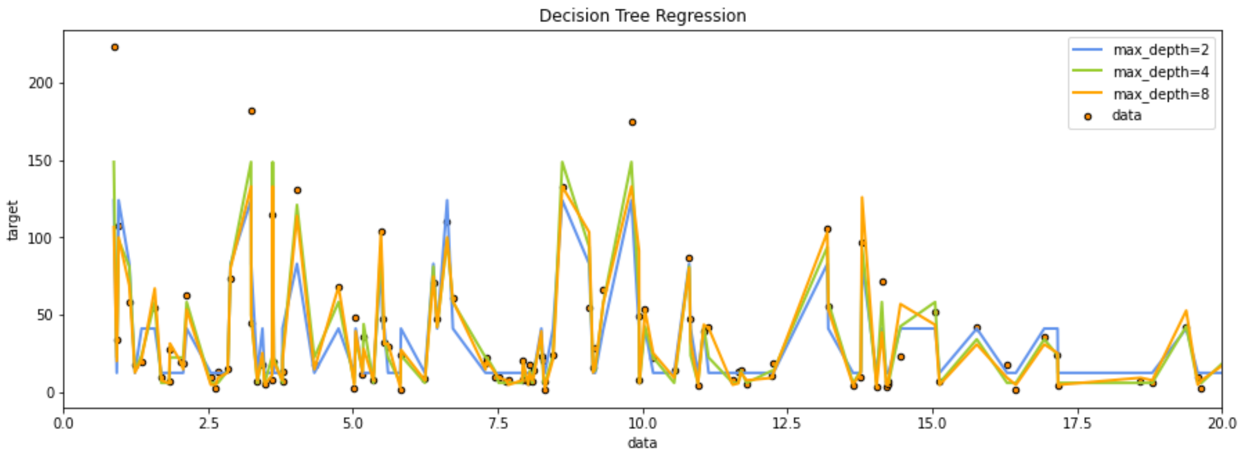


Figure 12. The x axis shows the data and the y axis represents the "target" to be hit by the ML correction. Each line shows the corrected LCS value based on the max depth of the decision tree. The depth-2 tree captures most of the variation; the depth-4 tree makes some progress but the depth-8 tree is more computational intensive yet still not perfect.

Since the depth-4 tree gave a decent performance for the trade-off of needing smaller storage and computational intensity, I decided to apply a depth-4 decision tree to all of the data.

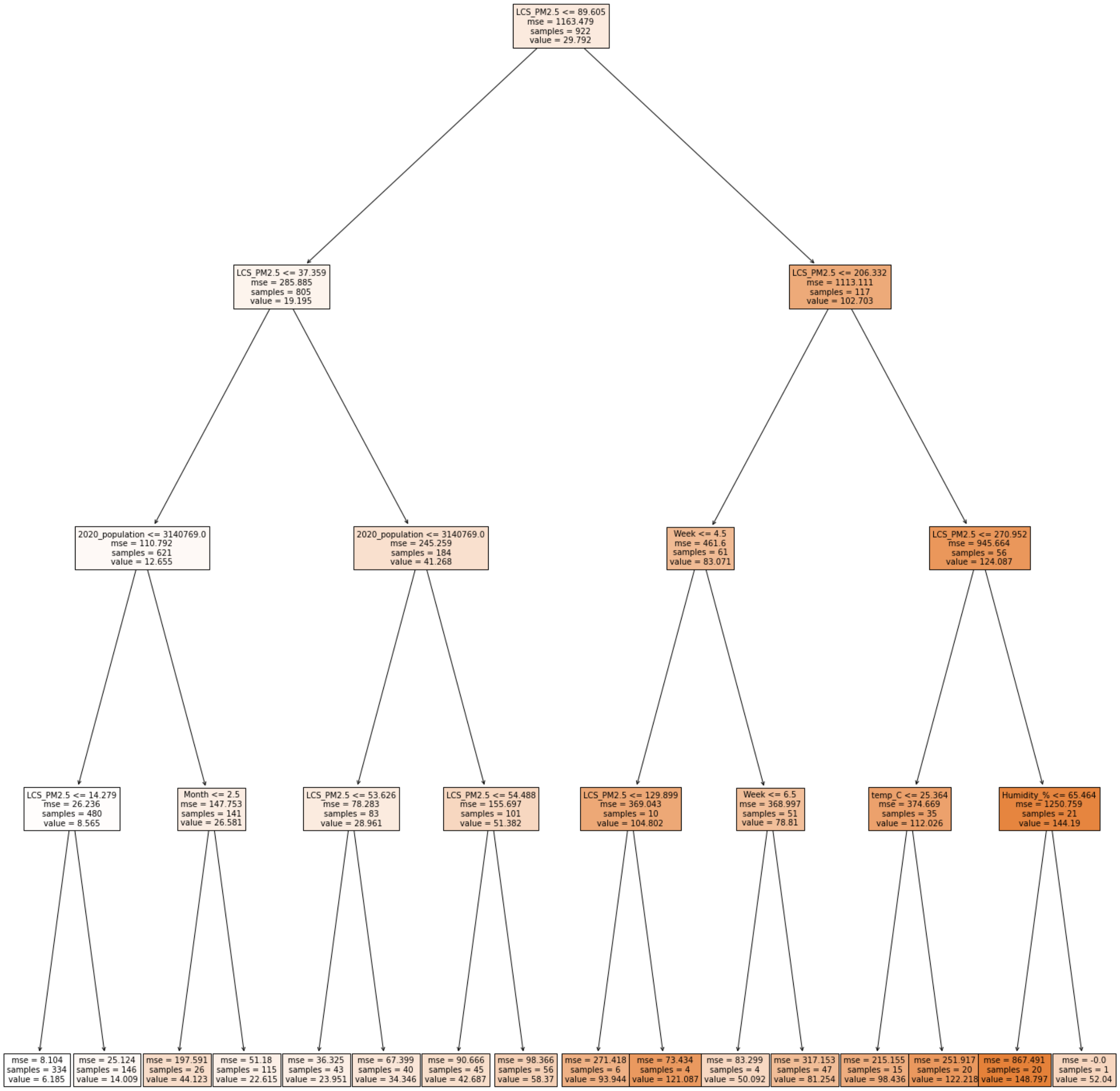
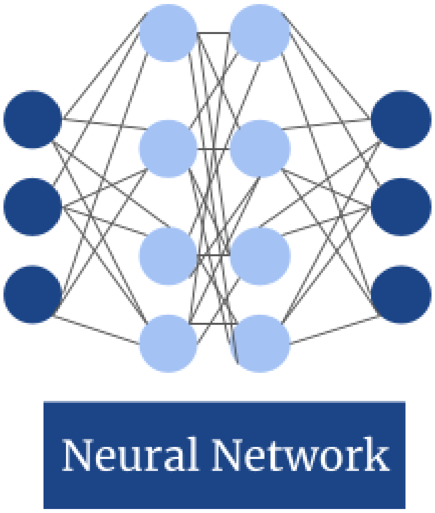


Figure 13. Depth-4 Decision Tree for the CAMS-Net dataset. Each branch represents a decision indicated by the top row of each box. Shaded boxes represent the majority of data split. Boxes include decision, mse, number of samples and final estimated value. The initial MSE is 1163. After depth-4, the maximum mse is 867, with the final mse largely falling in the 25-200 range. This indicates a significant dropoff in mean squared error.

d. Neural Network



Finally, I tried a neural network regressor method, using a ReLu activation function, α=0.001, the adam solver, 100 hidden layers, and 5000 maximum iterations. The neural network seems to create appropriate prediction regimes for the diverse sections of the CAMS-Net dataset, by creating two offset linear regressions with approximately the same coefficient but different intercepts. This is different from the naive methods I was trying, which all used 0-intercepts. I tried a lot of different inputs for the neural net, but the best resulting MAE I could achieve with a neural network is 26.5 µg/m^3 which is slightly worse than our starting MAE. Thus, our usage of neural networks is not over – while the implementation currently is not objectively perfect, it provides a new way of attempting the problem which I hope to tackle in my research work next semester.

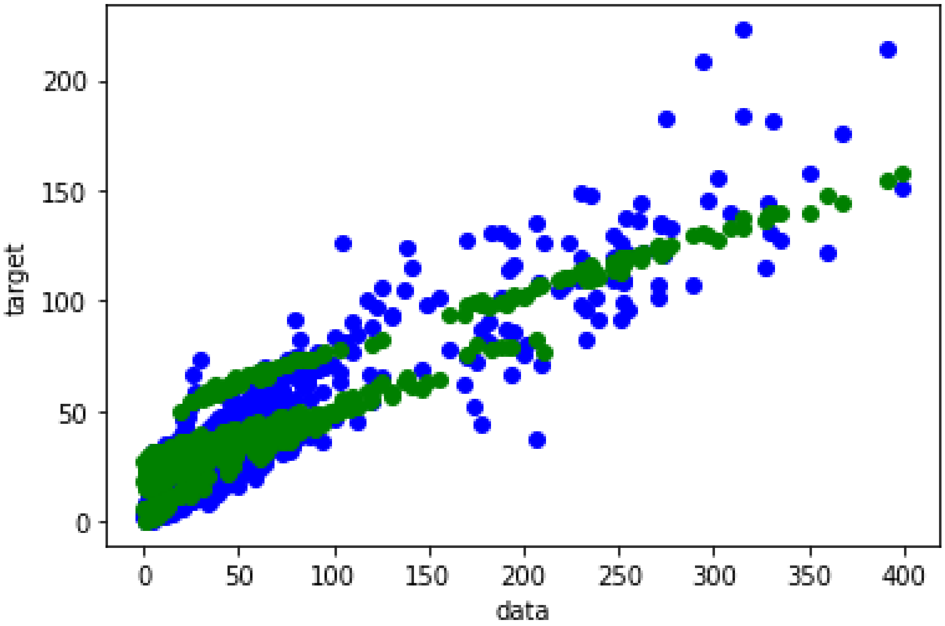


Figure 14. Neural network results. The blue dots represent the measured LCS data versus the reference monitor target. The green dots represent the neural network predicted values.

4. Conclusion

In this project, I used a dataset of low-cost and reference sensor measurements from around the world to create a universal machine learning model to correct low-cost data, in the hopes of trying to replicate reference-grade data from instruments that are more accessible to a wider array of community groups, researchers, analysts, policymakers and government agencies. I used three main types of machine learning models: multiple linear regression, decision trees, and neural networks. The multiple linear regression proved to be effective in reducing mean absolute error by up to 75% in sites that had high disagreement between low-cost and reference monitor data, but the MLR was less useful in reducing error when the raw data error was already low. The decision tree provided interesting visualizable insight into the dataset, indicating that correlation regimes shifted based on the average LCS measurement, 2020 population and temperature, which were the features used for the depth-4 tree. Trees deeper than 4 did not provide significant improvements in accuracy. Finally, the neural network proved challenging to fine tune to this dataset and will require further feature engineering for full optimization; however, it provided insight into choosing the best naive methods for correction low-cost sensor data. From all the modelling activities, it is clear that more specific feature creating to ascertain the type of sensor, its location relative to important emissions sites, the local time of measurement and the seasonality of the measurement region, and most importantly, the brand/type of sensor will be important features to add into the dataset. Finally, as we collect data from more colocations around the world and find a way to utilize the very large OpenAQ dataset, the ML models piloted in this study will have more “big data” for training. Results from this study will be built upon next semester, but even in their current state, can be useful for government agencies in resource-limited regions who do not have a reference-grade monitor to establish air quality levels using corrected low-cost sensor data.

5. Acknowledgements

Thank you to Professor Pierre Gentine and Teaching Assistant Olya Skulovich for their excellent teaching and help in the EAEEE 4000 class. Thank you to Professor Dan Westervelt and Research Assistant Celeste McFarlane for help with the CAMS-Net dataset collection.