Improving Accuracy of Air Pollution Exposure Measurements: Statistical Correction of a Municipal Low-Cost Airborne Particulate Matter Sensor Network

Authors: Ellen M. Considine[1], Colleen E. Reid[2*], Michael R. Ogletree[3], Timothy Dye[4]
[1] Department of Applied Math, University of Colorado Boulder, USA
[2] Department of Geography, University of Colorado Boulder, USA
[3] Denver Department of Public Health and Environment, USA
[4] TD Environmental Services, LLC, USA

Correspondence: colleen.reid@colorado.edu

Abstract

Low-cost air quality sensors can help increase spatial and temporal resolution of air pollution exposure measurements. These sensors, however, most often produce data of lower accuracy than higher-end instruments. In this study, we investigated linear and random forest models to correct $PM_{2.5}$ measurements from the Denver Department of Public Health and Environment (DDPHE)'s network of low-cost sensors against measurements from co-located U.S. Environmental Protection Agency Federal Equivalence Method (FEM) monitors. Our training set included data from five DDPHE sensors from August 2018 through May 2019. Our testing set included data from two newly deployed DDPHE sensors from September 2019 through mid-December 2019. In addition to $PM_{2.5}$, temperature, and relative humidity from the low-cost sensors, we explored using additional temporal and spatial variables to capture unexplained variability in sensor measurements. We evaluated results using spatial and temporal cross-validation techniques. For the long-term dataset, a random forest model with all time-varying covariates and length of arterial roads within 500 meters was the most accurate (testing RMSE = 2.9 $\mu g/m^3$ and R^2 = 0.75; leave-one-location-out (LOLO)-validation metrics on the training set: RMSE = 2.2 $\mu g/m^3$ and R^2 = 0.93). For on-the-fly correction, we found that a multiple linear regression model using the past eight weeks of low-cost sensor $PM_{2.5}$, temperature, and humidity data plus a near-highway indicator predicted each new week of data best (testing RMSE = 3.1 $\mu g/m^3$ and R^2 = 0.78; LOLO-validation metrics on the training set: RMSE = 2.3 $\mu g/m^3$ and R^2 = 0.90). The statistical methods detailed here will be used to correct low-cost sensor measurements to better understand $PM_{2.5}$ pollution within the city of Denver. This work can also guide similar implementations in other municipalities by highlighting the improved accuracy from inclusion of variables other than temperature and relative humidity to improve accuracy of low-cost sensor $PM_{2.5}$ data.

Keywords:

air pollution; low-cost sensor; machine learning; on-the-fly calibration; Plantower sensor; cross-validation

1. Introduction

Low spatial coverage of air pollution monitors is a major barrier to quantifying the air pollution to which people are exposed and investigating the health impacts of this exposure. In 2019, the global mean population distance to the nearest $PM_{2.5}$ (atmospheric particulate matter with aerodynamic diameter of less than 2.5 μm) monitor was 220 km (Martin et al., 2019). In the U.S., more than 70% of counties do not have regulatory $PM_{2.5}$ monitoring (Bi et al., 2020). This shortage of air quality measurements prevents accurate exposure assessment for epidemiological studies of the health impacts of air pollution.

Low-cost sensors allow for a higher density network of air quality monitors to be deployed across a city, assuming the same municipal air quality monitoring budget. In addition to community education and hazard warning systems (Kumar et al., 2015), deploying such a network creates opportunities for detection of air pollution hotspots or high-pollution sources, reactive ("smart city") systems (such as dynamic traffic controls based on pollution levels), and improved environmental health research (Budde et al., 2014). The downside of low-cost sensors is that they most often produce data of lower accuracy (in terms of bias, noise, etc.) than federal reference method (FRM) or federal equivalence method (FEM) monitors (Cromar et al., 2019; Bi et al., 2020).

One remedy for low-cost sensors' inaccuracy is the development of statistical models to correct measurements from low-cost sensors to measurements from a collocated FRM or FEM monitor. Many commercial sensors are nominally corrected (calibrated) in laboratory settings but training the correction model on field data is generally more accurate because then the sensor experiences more realistic meteorological and air pollution conditions (Kumar et al., 2015; Castell et al., 2017). Correction or calibration models for air pollution sensors can be characterized by the extent to which they are based on known physical properties of the atmosphere and sensors and/or based on empirical observations from the sensors. In this paper, we focus on the latter type of correction model, which Malings et al. (2019b) showed tend to be as accurate as correction models based on physical properties. For low-cost particulate matter sensors, recent studies have used linear regression (Holstius et al., 2014; Magi et al., 2020; Zusman et al., 2020) and higher-order polynomial regression (Gao et al., 2015; Malings et al., 2019b) and machine learning algorithms such as extreme gradient boosting (Si et al., 2020) and artificial neural networks (Badura et al., 2019; Si et al., 2020). Researchers have also found that a blend of statistical models, for example linear regression with different coefficients above a threshold (Malings et al., 2019b) and gaussian process regression (kriging) combined with linear regression (Zheng et al., 2019) can help to capture nonlinear sensor response.

Because many air quality sensors' readings are influenced by temperature and humidity, measurements of these variables are often taken on site and can be used in correction models (Holstius et al., 2014; Malings et al., 2019b; Zusman et al., 2020). Otherwise, low-cost air pollution sensor correction studies tend to avoid incorporating external parameters into their models. As Hagler et al. (2018) argue, it is critical that corrections of sensor data are transparent and do not pull too far away from the original ("ground truth") data by using needlessly complex algorithms. However, large seasonal variations in accuracy have been reported in studies which do not take time into account (Malings et al., 2019b; Sahayi et al., 2019). Some researchers attempt to address this issue by calculating different regression coefficients for different seasons (Zheng et al., 2018;

Malings et al., 2019b), however, it is possible that use of temporal terms in the model could achieve similar adjustment for seasonal or other temporal variation in correction accuracy.

One challenge in accurately correcting a low-cost air pollution sensor network is that the accuracy (at least the bias) of many low-cost sensors (for both airborne particulate matter and gases) has been shown to degrade (or "drift") over time (Kumar et al., 2015; Budde et al., 2014; Malings et al., 2019b; Sayahi et al., 2019; Delaine et al., 2019) -- regularly updating the correction model is recommended. For low-cost particulate matter sensors, several different techniques have been proposed to counter the effects of sensor degradation. One approach is to estimate the bias of a low-cost sensor compared to a reference monitor and then simply adjust the constant term (the bias) in the correction equation over time (Malings et al. 2019b). Another approach is to regularly re-run the whole regression for the correction model. A benefit of the latter approach is that it can address the possibility that aspects of the correction other than the bias (constant) change over time. However, while the latter approach has been shown to help maintain low-cost air pollution sensor correction accuracy over time (Zheng et al., 2019; Zimmerman et al., 2018), it also introduces the added complexity of needing to decide how much data (or how long a "lookback") to use to train the correction model each time it is run.

Another major challenge in low-cost sensor correction is that it is necessary to develop a generalized model that works without having to collocate every low-cost sensor with an FRM or FEM monitor, but it is unknown how many collocations are needed within an urban area. Because statistical models are likely to perform worse on new data than on data used to train the models, many studies have utilized cross-validation methods to evaluate the accuracy of their correction strategies on new data (Badura et al., 2019; Zheng et al., 2019; Magi et al., 2020). Recent studies have highlighted the importance of spatial and temporal cross-validation (Malings et al., 2019b; Zusman et al. 2020). Specifically, Zusman et al. (2020) concluded that leave-one-location-out (LOLO) cross-validation is more accurate when three or more collocation sites are in use, while 10-fold cross-validation by week is more accurate when only one or two sites are in use.

Denver, Colorado was one of nine cities across the U.S. to win the 2018 Bloomberg Philanthropies' Mayors Challenge. The Mayors Challenge encourages cities to develop innovative programs which increase sustainability and equity, and which ultimately can be scaled to other cities after proof of concept. Denver is using its $1 million award to install a system of low-cost air quality monitors at public schools across the city (targeting schools with high asthma rates and in lower-income neighborhoods), build an online platform for real-time reporting of air quality, and engage in community education about air quality and environmental health. This program, managed by the Denver Department of Public Health & Environment (DDPHE), is called the Love My Air program (formerly the Air Quality Community Action Network, or AQ-CAN).

In this study, we develop statistical correction for the Denver Love My Air sensors. Our study is novel in several ways. First, we develop two different models to correct data from low-cost particulate matter sensors: a long-term model to correct archived data and an on-the-fly model to correct data in real time. Second, we employ robust spatial and temporal cross-validation techniques to test the performance of our models on data from new locations and time periods. Third, we explore the inclusion of temporal and landcover variables. Finally, this was a direct partnership between academics and the DDPHE, ensuring that our models will be incorporated into the Denver system, helping to correct air quality data and inform public warning systems.

2. Methods

2.1 Data Sources

Between August 2018 and May 2019 (one academic year), Denver Love My Air collected data from five low-cost $PM_{2.5}$ sensors in stable locations, collocated with U.S. EPA FEM monitors. There were three different sites (National Jewish Hospital, La Casa, and I25-Globeville); three sensors were collocated at the I25-Globeville location. In fall 2019, two additional Love My Air sensors were stationed at the CAMP and I25-Denver FEM locations (see Figure 1 for a map of these locations). This work is in line with the conclusion of Zusman et al. (2020), that thoughtful placement of at least three collocation sites is preferable for this kind of correction. More Love My Air sensors have been deployed across the city.

The Love My Air sensors are Canary-S models equipped with a Plantower 5003, made by Lunar Outpost. The Canary-S sensors detect $PM_{2.5}$, temperature, and humidity, and upload minute-resolution measurements to an online platform via cellular data. We obtained hourly $PM_{2.5}$ measurements from the three FEM monitors and hourly averages from the five Canary-S sensors between August 20, 2018 and May 30, 2019. After removing missing values in the $PM_{2.5}$, temperature and humidity data (coded as either NA or -1) and $PM_{2.5}$ values above 1,500 $\mu g/m^3$ (unrealistically high concentrations) from the Canary-S sensors ($N_{missing}$ = 4,313, N_{high} = 2), we were left with 29,770 hourly observations. Time series of the measurements from each sensor are shown in Figure S1. These time series plots illustrate that there is reasonable overall agreement between the measurements from the reference monitors and low-cost sensors, but that the low-cost sensors tend to overestimate $PM_{2.5}$, especially at high concentrations.

Because of daily, weekly, and seasonal variation in $PM_{2.5}$ that may be due to factors beyond temperature and relative humidity, we extracted hour, weekend, and month variables from the Canary-S sensors and converted hour and month into cyclic values by taking the cosine and sine of hour*2π/24 and month*2π/12. Sinusoidal correction for season has been shown to improve accuracy of $PM_{2.5}$ measurements (Eberly et al., 2002).

Along with adjusting for variability in time, we investigated variability in space. The position of an air quality sensor within a city, especially relative to known sources of pollution such as highways, is likely to affect the characteristics of the air pollution in that area: the type and size of particulates, timing of fluctuations in air pollution, etc. We investigated including two different kinds of landcover variables: a binary variable indicating whether a monitor was near or far from a highway (based on local knowledge, I-25-Globeville and I-25 Denver were classified as near-highway and NJH, La Casa, and CAMP were not) and the lengths of different sizes of roads within a certain distance from a monitor. To derive the latter, we used a road dataset from the City of Denver Open Data Catalog (see Figure 1) and calculated the lengths of arterial, collector and local (large, medium, and small) roads within circular buffers surrounding each monitor location. We considered buffers of radius 50, 100, 250 and 500 meters. Preliminary testing showed that five of the road variables – arterial roads within 500 and 50 meters and local roads within 250, 100, and 50 meters - were the most important. We used these in the rest of the analyses. The values of these road length variables are shown in Table S1.

2.2 Statistical Modeling

We developed two correction models: one for archived data and one for on-the-fly data. Archived data can be used for long-term evaluations including environmental public health research, while real-time data can be used to warn people about hazardous air quality conditions. The reason for doing two different types of correction is that while long-term models tend to be more accurate over the entire spatiotemporal data set, it is inefficient to re-run large models frequently (incorporating new data). Also, on-the-fly correction can help characterize short-term variation in air pollution and sensor characteristics, improving public health warnings. Both types of correction allow for use of low-cost sensors to inform air quality monitoring at finer spatial and temporal scales than is possible using only FRM or FEM monitors, given the few FRM and FEM monitoring sites in the U.S., particularly in the western states (Martin et al., 2019).

2.2.1 Modeling: Long-Term Correction

The goal of this correction is to predict, as accurately as possible, the "true" $PM_{2.5}$ concentration at a location given the $PM_{2.5}$ measurement from a Canary-S sensor at that location. Thus, the EPA FEM $PM_{2.5}$ measurements, which we take to be the "true" concentration of $PM_{2.5}$ at that location, are the dependent variable in the correction models that will then be predicted by the correction model at locations without an FEM monitor.

We tested simple and multiple linear models, mixed effects linear models (otherwise known as random effects models or hierarchical linear models), and random forest models. Mixed effects models can help account for the violation of independence between repeated measurements from each monitor by specifying a random effect term in the model to account for variation in the correction at different measurement locations. Unlike including a near-highway indicator or a road-length variable in the model, however, using a random effect for the monitoring location in the model does not allow us to account for location-dependent variability in the prediction/correction step, only in the training step. Random forest is a decision-tree-based machine learning algorithm that can capture more complicated nonlinear effects (for instance, unknown relationships between additional spatial and temporal variables) and tends to perform well in air quality prediction (Malings et al., 2019a; Zimmerman et al., 2018; Xu et al., 2018). We used a random forest algorithm called *ranger* using the R package caret (Kuhn, 2008).

When selecting and evaluating our models, we used root-mean squared error (RMSE) and the correlation coefficient R^2 as performance metrics. Lower RMSE values and higher R^2 values indicate more accurate models. With such a large sample size, we found that our R^2 values were numerically equivalent to adjusted R^2 values. In terms of variable selection, we only kept terms that appeared to improve the results in the validation step. For the linear models, this included a preliminary investigation of using higher-order polynomial terms and transformations such as logarithms, but none of these significantly improved the predictions. Before training the random forest models, we tuned the hyperparameters for the ranger algorithm using a random subset of the training data. The first random forest model we trained used all available data from the 2018-2019 academic year (our entire training/validation data set from the original five collocated sensors, including all the time-varying and road length covariates).

During model development, we used a LOLO cross-validation strategy (as explained in Zusman et al., 2020) to validate the model results. For further evaluation, we tested our final models on completely held-out data from the CAMP and I-25 Denver reference monitors (deployed in early fall 2019) for testing to obtain our final performance metrics. Having the completely held-out data from the CAMP and I-25 Denver monitors in the testing set is especially helpful because CAMP is in the middle of downtown Denver and I-25 Denver is next to an Interstate highway, providing us with test metrics reflecting different environments. These test set data spanned September 2019 through mid-December 2019. However, the EPA FEM monitor at CAMP shut off during mid-October, leaving much less test set data for that monitor than for the I-25 Denver monitor. After removing missing values and values where the reference monitor reported exactly zero, we were left with 3,011 hourly observations in the test set.

2.2.2 Modeling: On-the-Fly Correction

The analysis described above was backward correction: we used all the data, including the most recent, to correct all the data, which is the best choice for correcting long-term archived data. Hasenfrantz et al. (2012) found that backward correction reduced measurement error from forward correction by a factor of two. However, due to data availability, Love My Air's real-time air quality reporting must rely on forward correction: using past data to correct new data which was not included in the correction model.

An important question is how many days/weeks of past data are needed to get an accurate on-the-fly correction model to predict forward and how far into the future such a model can accurately predict. In addition to accuracy, however, we must consider practical constraints, such as how often an on-the-fly correction model can be run because of computational limitations. With too little training data (such as weeks when there are a lot of missing observations), some linear regressions will not converge, and random forest models with too little data are likely to overfit. We assessed the performance of all possible combinations of 1-8 weeks of training data (lookbacks) with 1 or 2 weeks of testing data (predictions) for several linear models, mixed effects models, and random forest models. Each model was tested on held-out data from La Casa because, of the original five low-cost sensors in the training set, its data displayed average performance in the data summary statistics and long-term data correction models.

Here is a repository with the R code used in these analyses: https://github.com/EllenConsidine/Love_My_Air/tree/master/R

To facilitate discussion about models tested in both the archived and on-the-fly analyses, we use the following model-naming conventions: A = archived, O = on-the-fly; LR = linear regression, ME = mixed effects linear regression, and RF = random forest.

3. Results

3.1 Data Summary

The summary statistics in Table 1 provide context for the performance of the training/validation and testing set monitors. In the training/validation set, we observe that both the FEM (AirNow) monitors and the Canary-S sensors measure lower $PM_{2.5}$ at the National Jewish Hospital monitor and higher $PM_{2.5}$ at the I-25 Globeville monitor. This is expected given that the National Jewish Hospital monitor is not directly next to a highway, while the I-25 Globeville monitor is. Also, the National Jewish Hospital FEM monitor is a Teledyne T640 while all the other

FEM sites use GRIMM EDM 180 monitors. The La Casa monitor $PM_{2.5}$ levels were in the middle for these monitors, with an average of 10.4 μg/m^3.

In the test set (CAMP and I-25 Denver), we observe lower $PM_{2.5}$ at the CAMP monitor than at the I-25 Denver location, which again is expected given CAMP's location far from a highway and I-25 Denver's location next to an Interstate highway. We also note that the measurements from the CAMP monitor have much lower variance than the other monitors, likely due to its much shorter period of reporting data before shutting down.

For comparison, prior to correction, the raw low-cost sensor measurements in the training/validation set had RMSE = 5.5 μg/m^3 and R^2 = 0.81 compared to the reference measurements. The raw testing set had RMSE = 7.1 μg/m^3 and R^2 = 0.73.

Table S2 provides descriptive statistics for the environmental variables (temperature and relative humidity). In general, the temperatures in the testing set are higher than those in the training/validation set. Specifically, the CAMP sensor reported high temperatures, in part because it shut off in mid-fall. By contrast, both testing set sensors measured much lower values of relative humidity, while the third low-cost sensor at the I-25 Globeville location reported much higher values of relative humidity.

3.2 Long-Term Correction

Table 2 displays the training/validation and testing set RMSE values of the linear, linear mixed effects, and random forest models (R^2 values are in Table S3). In general, the more complex models tend to do better in the LOLO cross-validation (training). However, there is not such a clear pattern for the test set. The CAMP results from linear models including Aroad_500 illustrate the danger of using a continuous variable like road length with relatively few observations to extrapolate to new locations: clearly whatever linear relationship is specified in the training does not apply to CAMP. Interestingly, the random forest models with Aroad_500 do not have this problem when testing on CAMP, indicating that the relationship is likely nonlinear.

Based on both the training/validation and the testing results, the best models were A.RF.4 and A.RF.5, the random forest models with $PM_{2.5}$, temperature, humidity, month, time, weekend, and one or more road length variables. We observed an improvement from the inclusion of multiple road variables (A.RF.5), but it was sufficiently small that it may be overlooked in the interests of model simplicity. Figure 2 illustrates the relationship between the reference data and the corrected low-cost sensor data. Based on only the training/validation results, we would have selected A.RF.3, the random forest model with $PM_{2.5}$, temperature, humidity, month, time, weekend, and the near-highway indicator. However, the testing results for I-25 Denver were much worse for this model. Thus, A.RF.4 (a random forest model with $PM_{2.5}$, temperature, humidity, month, time, weekend, and the length of arterial roads within 500 meters of the monitor location) is our final selection.

When we calculated variable importance in the random forest models using the permutation method, we found that all of the temporally-dependent variables ($PM_{2.5}$ from the low-cost sensors, temperature, relative humidity, and time) were more important than the stationary variables. We note that while multicollinearity between the predictors does not impair the predictive accuracy of the random forest models, it does make the variable importance scores inexact (Gregorutti et al., 2017).

3.3 On-the-Fly Correction

Table 3 displays the on-the-fly correction results from the best model for each algorithm regarding which training and testing timespans yielded the lowest RMSE value when tested on the data from the La Casa monitor, which was left out of the trainings for these models.

In this table, we see that O.LR.3, the multiple linear regression model with temperature, humidity, and the near-highway indicator, had the lowest RMSE values compared to the other model types (algorithm plus subsets of covariates). In general, random forest models perform better on larger datasets than the on-the-fly corrections and thus in this analysis yielded less accurate results than the linear models.

4. Discussion

We found that using a random forest model accounting for temperature, humidity, month, hour, and road lengths within 500 meters was the most accurate in correcting long-term (archived) $PM_{2.5}$ measurements from the Canary-S sensors to the EPA FEM monitor measurements, using data from five monitors from the 2018-2019 academic year and two additional monitors from fall 2019. We note that using a time-invariant land cover variable in this machine learning model is akin to using a random effect in mixed effects linear models in terms of capturing sensor- or location-specific characteristics that could influence the correction. The average LOLO performance metrics for the validation set were RMSE = 2.2 μg/m^3 and R^2 = 0.93. The average performance metrics for the testing set were RMSE = 2.6 μg/m^3 and R^2 = 0.76. Weighting the test set performance metrics to account for the number of observations from each test monitor (CAMP = 25%, I-25 Denver = 75%) yielded RMSE = 2.9 μg/m^3 and R^2 = 0.75.

We found the higher computational cost of random forest (in exchange for higher accuracy compared to linear regression models) to be worthwhile for applications which require the correction of archived data sets, such as long-term environmental health research studies. Other nonlinear models, such as generalized additive models (GAMs), might also be employed for this purpose. However, the improvement from random forest over linear regression for the archived data was modest. Compared to the best multiple linear regression model, the best random forest model reduced the RMSE by about 1 μg/m^3. For ease of comparison, Table 2 details the accuracy of all our linear regression, linear mixed effects regression, and random forest models.

For on-the-fly correction, we found that the most accurate approach was using a multiple linear regression with the past eight weeks of training data to correct each new week of data with the following predictor variables: Canary-S $PM_{2.5}$, temperature, humidity, and a near-highway indicator. The performance metrics for the validation set (data from the La Casa monitor) were RMSE = 2.3 μg/m^3 and R^2 = 0.90. The performance metrics for the testing set (just I-25 Denver due to lack of data from CAMP) were RMSE = 3.5 μg/m^3 and R^2 = 0.77. For comparison's sake: if we were to use a lookback of 3 weeks with this model, the CAMP testing results would be RMSE = 1.8 μg/m^3 and R^2 = 0.79. Weighting the test set performance metrics to account for the number of observations from each test monitor would yield RMSE = 3.1 μg/m^3 and R^2 = 0.78.

Of the five comparable studies to ours that we found, which used statistical techniques to correct hourly data from low-cost $PM_{2.5}$ sensors in regions with relatively low ambient air pollution (and which reported the magnitudes of their error as opposed to just R^2), four achieved RMSEs between 3.4 and 4.2 μg/m^3 (Holstius et al., 2014; Badura et al., 2019; Magi et al., 2020; Si et al., 2020) and one achieved an average (across testing sites) MAE (mean absolute error) of 2.3 μg/m^3 (Malings et al., 2019b). While these last results are impressive, it is important to keep in mind that RMSE is always greater than or equal to MAE; squaring the errors before averaging penalizes variance (Chai and Draxler, 2014). Also, when we consider only Malings et al.'s (2019b) results that used Plantower sensors like ours, their MAE was 2.7 μg/m^3.

Another factor frustrating direct comparison between these studies and ours is different pre-processing. Some studies removed values for which the low-cost sensors measured beyond certain thresholds, for instance over 50 μg/m^3 (Magi et al., 2020) or under 1 μg/m^3 (Sayahi et al. 2019). Malings et al. (2019b) averaged the values from the two sensors within the Plantower device. Zusman et al. (2020) removed unusually high values from time periods with fireworks and wildfires and then averaged the values from the two sensors. Compared to these previous studies, our study differs by correcting both archived and on-the-fly data, investigating inclusion of variables to capture variation in time and space beyond temperature and relative humidity, and using spatiotemporal cross-validation strategies for model evaluation, which can cause worse performance metrics than plain cross-validation (Zusman et al., 2020).

To contextualize our results, we refer to low-cost $PM_{2.5}$ sensor accuracy standards proposed by multiple groups. Malings et al. (2019b) assert that determining whether regulatory standards are being met necessitates accuracy around ±10% of the average air pollution levels in an area; mapping spatial gradients and monitoring microenvironments (e.g. for environmental health studies) could be done with ±25% accuracy, while ±50% accuracy is still useful for tracking large sources of air pollution and informing the public about which areas of a city are more polluted or less polluted. Williams et al. (2018) reviewed standards from multiple countries and concluded that for decision support applications, including regulatory monitoring, ±25% accuracy in 24h averages or $R^2 \geq 0.72$ is acceptable. All our training and testing R^2 values were ≥ 0.75. For our archived model, the ratio of RMSE to average $PM_{2.5}$ for our validation set was 23% and for our (weighted) testing set was 30%. For our on-the-fly model, the ratio of RMSE to average $PM_{2.5}$ for our validation set was 22% and for our (weighted) testing set was 32%. Given that our testing set measurements were taken nearly half a year after our training set measurements and at new locations, we interpret these results to mean that our models are in line with these proposed standards. We also note that these standards or accuracy percentages or R^2 thresholds that were all made for 24h-average measurements of air pollution may not be the right standards to use for hourly-average measurements, as we have used in this study. Averaging across 24 hours likely increases accuracy, therefore we would expect to get worse accuracy metrics using hourly data.

Another way to evaluate our model performance is to view the plots of the corrected measurements versus reference measurements (Figure 2). In addition to the general shape around the one-to-one line, an eye-catching feature of these plots is the set of roughly half a dozen outlier points. Early in this project, we experimented with creating an outlier detection algorithm to identify the combination of large jumps between sequential measurements and large discrepancies between the two sensors within each Plantower device. Further investigation revealed that these points were all on days with low temperature and high humidity, specifically days right around when it snowed in Denver. However, some of the snow day points (especially in the test set) went undetected by this algorithm. Several papers have reviewed outlier detection algorithms for this kind of application (Zhang et al., 2010; van Zoest et al., 2018; Ottosen and Kumar, 2019; Delaine et al., 2019), however more work needs to be done to ensure that measurements from true high air pollution events, which are extremely important for health impact studies, are not being classified as low-cost sensor malfunctioning. This assertion is in line with the findings of Williams et al. (2018), that more studies using non-regulatory air pollution sensors need to explicitly address treatment of erroneous data. We decided against removing the suspected outlier points for the analysis, even though removing them would slightly improve our RMSE and R^2 values.

Overall, the instances of discrepancy between temperature and relative humidity measurements within the training and testing sets indicates a potential limitation of using measurements of environmental variables from low-cost sensors. For instance, there is reason to suspect that the highest humidity measurements in our training set indicate sensor malfunction because 100% humidity in Colorado is quite rare. If the temperature and relative humidity sensors are inaccurate, this will interfere with statistical corrections which use these variables. Even if the environmental measurements were accurate in our study, the fact that they were noticeably different overall between the training/validation and testing sets means that our testing set results may show the correction models to be worse than they actually are. In general, these kinds of correction models are likely to perform worse on domains they were not trained on, including extreme meteorological conditions, new peak air pollution events, and different geographic regions (Zusman et al., 2020). This highlights the importance of having a large training set and checking the accuracy of the correction model(s) over the domain to which they are being applied.

Another limitation of this study is that the National Jewish Hospital FEM monitor is a Teledyne T640 while all the other FEM sites use GRIMM EDM180 monitors. We observed that the $PM_{2.5}$ measurements from the National Jewish Hospital reference monitor had lower variance than those from the other reference monitors. If this was in part due to the instrumentation as opposed to only the location by National Jewish Hospital, then this may have interfered with our exploration of including additional spatial/landcover terms in the models. For reference, a GRIMM and a T640 monitor were collocated for two weeks in September 2019 in Denver. The R^2 between the measurements from these two monitors was 0.82. A time series of the measurements of these two monitors, along with the measurements from a collocated BAM monitor, is shown in Figure S2.

Regarding our accounting for additional spatiotemporal variation in the models: for the archived-data correction, we found that including additional temporal variables (a weekend indicator and cyclic versions of time and month) was generally unhelpful when using linear or mixed linear models. For the random forest models, including additional temporal variables was most helpful when paired with additional spatial variables; the two different kinds of spatial variables performed roughly the same in the validation, but the road length variables performed better in the testing. For the linear models, including an additional spatial variable often appeared to help in the validation but not in the testing. In general, the mixed effect models did not outperform their plain linear counterparts. For the on-the-fly correction, including additional temporal variables did not appear to be helpful, but including an additional spatial variable did. Here, the near-highway variable slightly outperformed the arterial road length variable. For comparison: when we ran a random forest regression on our archived training / validation set (without cross-validation) not including low-cost sensor $PM_{2.5}$ but including temperature, relative humidity, month, time, a weekend indicator, and the length of arterial roads within 500 meters, we got an RMSE of 5.3 $\mu g/m^3$ and an R^2 of 0.52; under the same conditions (without cross-validation) but including low-cost sensor $PM_{2.5}$, we got an RMSE of 2.1 $\mu g/m^3$ and an R^2 of 0.93. This indicates that, at least with a "greedy" algorithm such as random forest which can capture nonlinear effects, a lot of the variation in $PM_{2.5}$ can be explained by these spatiotemporal factors, but the low-cost $PM_{2.5}$ measurements are still very important. The results of our exploration suggest that future low-cost air pollution sensor correction studies may want to investigate including additional temporal and spatial variables in their correction models, for correction of both archived and on-the-fly data. A couple of limitations of the land cover variables in this study are that we are assuming any variability in sensor performance due to location can be explained by proximity to roadways, and that creating something like the near-highway indicator relies on local knowledge. There may be location-dependent variability that could be explained, at least in part, by other land cover variables. Future studies might also consider incorporating traffic count data if such data are available.

We have also identified several other directions for future study: (1) working more on outlier detection; (2) determining whether imputing missing data points from low-cost airborne particulate matter sensors is useful, and if so, how it should be done; (3) optimizing the number and relative placement of collocation sites within a city or region (Zheng et al., 2019 investigated the optimal number for a large air pollution monitoring network in Delhi via simulation, but similar work remains to be done for smaller-scale municipalities with lower ambient air pollution); (4) determining whether and how to adjust for different types of FEM monitors when doing similar corrections (along the lines of work by Zheng et al., 2018); (5) investigating how effectively low-cost sensor correction models can be transferred between networks, cities, or regions (Zusman et al. 2020); (6) optimizing the timespan after which a long-term correction model should be updated, which is likely dependent on the monitoring network (e.g. sensor type and environmental characteristics of the city).

5. Conclusion

In this study, we investigated both on-the-fly and archived data correction, exploring the use of additional temporal and spatial variables to capture variation not explained by temperature and relative humidity, and employing extensive cross-validation to evaluate our correction models' performance in space and time. For the long-term dataset, a random forest model with all the time-varying covariates and the length of arterial roads within 500 meters was the most accurate. For the on-the-fly correction for each new week of data, we found that a multiple linear regression using the past eight weeks of low-cost sensor $PM_{2.5}$, temperature, and humidity data plus a near-highway indicator performed best. This work was the result of a direct partnership between academics and the DDPHE. Our correction models will be incorporated into the Love My Air platform for all sensors in this network, ultimately helping to communicate $PM_{2.5}$ levels to the public in Denver and inform future environmental health studies at local schools. Key directions for future study include developing methods for dealing with outliers and missing data, informing best practices in the deployment of collocated low-cost sensor and reference monitor pairs at the municipal level, and further exploring the inclusion of covariates to explicitly capture variability over time and space, as this study suggests these can help to improve low-cost sensor correction.

6. Figures and Tables

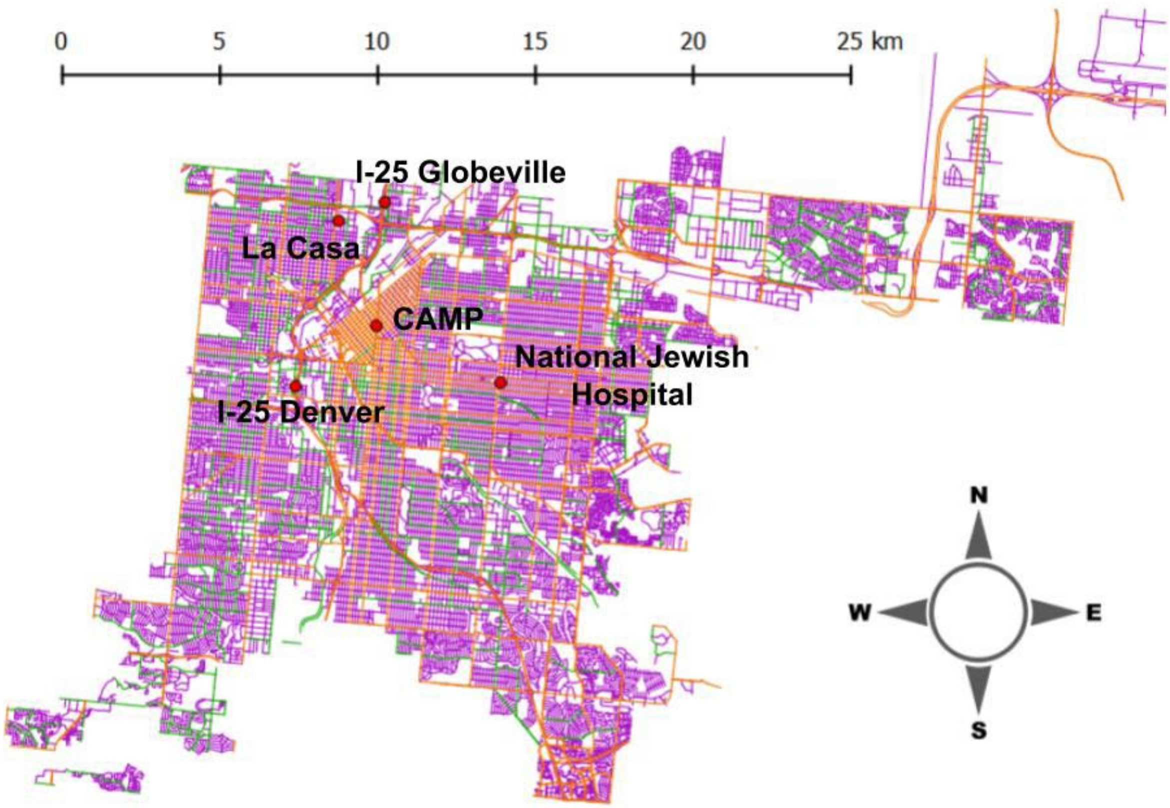

Figure 1: Map of collocated monitor locations and roads. Map of Denver County's U.S. EPA $PM_{2.5}$ FEM monitors at which Canary-S sensors have been collocated (red points), as well as arterial roads (orange), collector roads (green), and local roads (purple) in Denver (truncated to exclude the airport area in which there were no monitors). Note: I-25 Globeville has three collocated Canary-S sensors.

Table 1: Summary statistics of observations from the training/validation and testing sets

	Canary-S					AirNow				
Monitor	**Mean**	**Median**	**IQR**	**SD**	**Max.**	**Mean**	**Median**	**IQR**	**SD**	**Max.**

NJH	7.7	4.0	(1.2, 9.8)	10.1	91.8	7.7	5.8	(3.8, 8.9)	6.7	74.2
La Casa	10.4	6.4	(2.4, 13.5)	11.9	104.0	8.2	6.2	(4.0, 10.1)	7.1	76.5
I-25 Globeville 1	12.2	8.1	(3.5, 16.1)	12.7	170.7	11.0	8.8	(5.3, 14.1)	8.5	72.8
I-25 Globeville 2	9.1	6.4	(2.7, 12.3)	9.1	75.1	10.4	8.6	(5.3, 13.6)	7.0	54.1
I-25 Globeville 3	10.9	7.1	(3.0, 14.0)	11.7	99.0	11.0	8.8	(5.3, 14.1)	8.4	72.8
CAMP	5.5	4.1	(2.1, 7.3)	4.9	30.9	6.3	5.5	(3.8, 7.9)	3.6	27.2
I-25 Denver	11.2	7.3	(3.5, 14.1)	11.6	68.9	7.8	6.4	(3.9, 9.9)	5.7	56.2

Table 2: Root Mean Square Error (RMSE) values in μg/m³ for the training/validation set monitors for specific models using LOLO cross-validation where the metric provided is for when that monitor is the left out monitor, and RMSE in μg/m³ for the test set monitors by comparing the prediction value from the training model on the testing data that was completely held out of the training.

Statistical Model	Variables and CV folds (if applicable)	LOLO Training/Validation RMSE (μg/m³)					Testing RMSE (μg/m³)	
		NJH	La Casa	I25.1	I25.2	I25.3	CAMP	I25 Denver
A.LR.1	$PM_{2.5}$	2.3	3.2	4.0	3.7	3.7	1.6	4.5
A.LR.2	$PM_{2.5}$, Temperature, Humidity	2.5	3.1	3.9	3.7	3.7	1.8	4.9
A.LR.3	$PM_{2.5}$, Temperature, Humidity, Near_hwy	2.3	2.5	4.0	3.4	3.5	1.8	5.6
A.LR.4	$PM_{2.5}$, Temperature, Humidity, Aroad_500	3.0	2.7	3.9	3.4	3.5	17.3	3.8
A.LR.5	$PM_{2.5}$, Temperature, Humidity, Month, Time, Weekend	2.6	3.2	3.7	3.4	3.5	1.9	4.6
A.LR.6	$PM_{2.5}$, Temperature, Humidity, Month, Time, Weekend, Near_hwy	2.6	2.6	3.9	3.2	3.3	2.0	5.2
A.LR.7	$PM_{2.5}$, Temperature, Humidity, Month, Time, Weekend, Aroad_500	3.3	2.8	3.8	3.2	3.3	16.1	3.8
A.ME.1	Fixed = $PM_{2.5}$, Temperature, Humidity; Random = Intercept	2.4	2.8	3.8	3.5	3.6	1.8	5.0
A.ME.2	Fixed = $PM_{2.5}$, Temperature, Humidity; Random = Intercept, $PM_{2.5}$	2.4	2.8	3.9	3.5	3.6	1.8	5.0

A.ME.3	Fixed = $PM_{2.5}$, Temperature, Humidity, Month, Time, Weekend; Random = Intercept, $PM_{2.5}$	2.4	2.9	3.7	3.3	3.4	1.8	5.0
A.RF.1	$PM_{2.5}$, Temperature, Humidity	2.7	3.1	3.4	3.6	3.3	1.8	4.8
A.RF.2	$PM_{2.5}$, Temperature, Humidity, Month, Time, Weekend	2.3	2.9	2.5	2.8	2.3	1.7	3.9
A.RF.3	$PM_{2.5}$, Temperature, Humidity, Month, Time, Weekend, Near_hwy	2.2	2.2	2.5	2.2	1.9	1.7	4.5
A.RF.4	$PM_{2.5}$, Temperature, Humidity, Month, Time, Weekend, Aroad_500	2.2	2.3	2.5	2.2	1.9	1.8	3.3
A.RF.5	$PM_{2.5}$, Temperature, Humidity, Month, Time, Weekend, Aroad_500, Lroad_100, Aroad_50, Lroad_250, Lroad_50	2.2	2.2	2.6	2.2	1.9	1.7	3.4

In the statistical model column: A = archived data (as opposed to on-the-fly); LR = linear regression; ME = mixed effect linear regression; RF = random forest. In the variable column: Aroad = arterial road; Lroad = local road; the number following is the radial buffer size in meters within which the length of that type of road is being totaled. Near_hwy = near-highway indicator.

Note that "Time" and "Month" are the sinusoidal (cyclic) versions. Preliminary testing showed that including both sine and cosine of the hour of day did not improve performance in the linear models, and that including both sine and cosine of the month led to model non-convergence in the linear mixed effect models. Thus, for the linear and linear mixed effect models, "Time" refers only to cosine of hour of day; for the linear mixed effect models, "Month" refers only to cosine of month. All other references to "Time" and "Month" imply the inclusion of both sine and cosine.

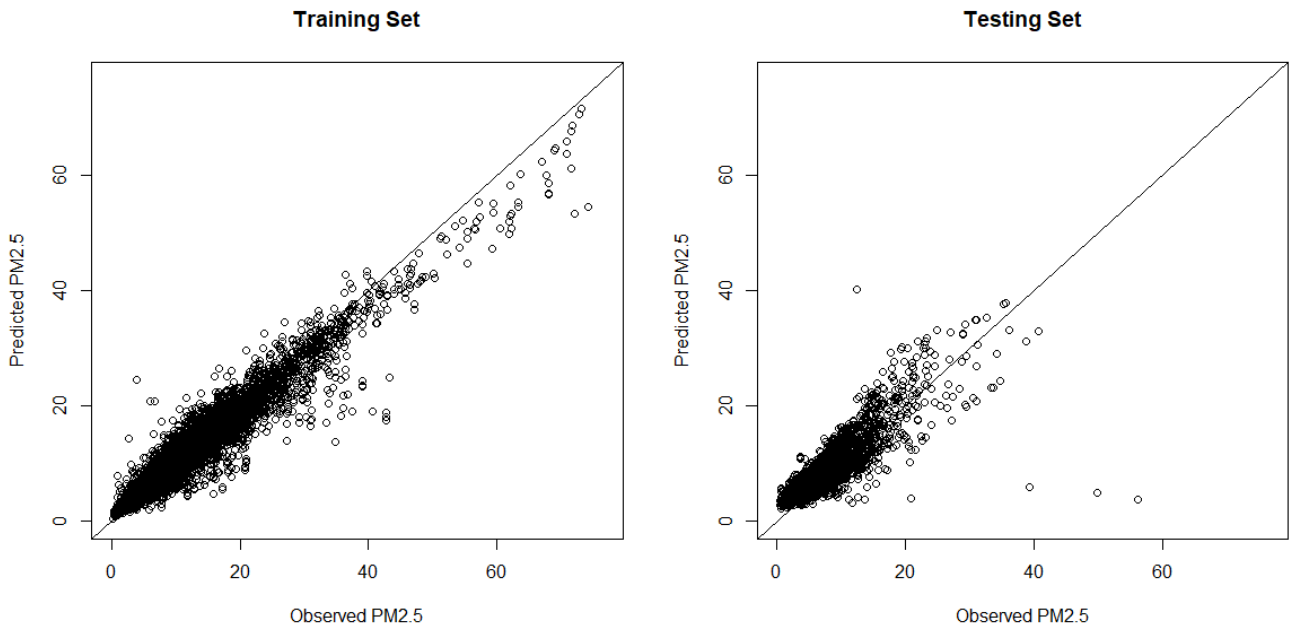

Figure 2: Visual representation of the performance of the model for correcting archived data. Fitted (predicted) versus observed $PM_{2.5}$ values ($\mu g/m^3$) using the A.RF.4 model.

Table 3: RMSE ($\mu g/m^3$) values for the best model of each type (optimal training set time span out of all tested (1 – 8 weeks) and optimal testing set time span out of all tested (1 or 2 weeks)). Grayed text indicates a rank-deficient fit reported in R for 11 out of the 41 weeks in the training set, where there was insufficient data. Blank cells indicate lack of sufficient training data from that monitor to train on the optimal time span (for example: the CAMP monitor shut off one week into October, thus we were unable to train a model on 8 weeks of data, as was selected to be optimal by the O.LR.3 model).

Statistical Model	**Variables and CV folds (if applicable)**	**Optimal Training Set Size (weeks prior to prediction)**	**Optimal Testing Set Size (prediction weeks)**	**La Casa Testing (RMSE in $\mu g/m^3$, R^2)**	**CAMP Testing (RMSE in $\mu g/m^3$, R^2)**	**I25-Denver Testing (RMSE in $\mu g/m^3$, R^2)**
O.LR.1	$PM_{2.5}$	3	1	3.1, 0.88	1.7, 0.83	3.7, 0.69
O.LR.2	$PM_{2.5}$, Temperature, Humidity	3	1	3.1, 0.89	1.8, 0.79	3.6, 0.69
O.LR.3	$PM_{2.5}$, Temperature, Humidity, Near_hwy	8	1	2.3, 0.90	-----	3.5, 0.77
O.LR.4	$PM_{2.5}$, Temperature, Humidity, Aroad_500	8	1	2.6, 0.91	-----	3.5, 0.77

O.LR.5	$PM_{2.5}$, Temperature, Humidity, Month, Time, Weekend	3	1	3.2, 0.88	1.8, 0.78	3.7, 0.68
O.LR.6	$PM_{2.5}$, Temperature, Humidity, Month, Time, Weekend, Near_hwy	8	1	2.5, 0.89	-----	3.7, 0.74
O.LR.7	$PM_{2.5}$, Temperature, Humidity, Month, Time, Weekend, Aroad_500	8	1	2.7, 0.89	-----	3.7, 0.74
O.ME.1	Fixed = $PM_{2.5}$, Temperature, Humidity, Time, Weekend; Random = Intercept, $PM_{2.5}$	3	1	3.1, 0.89	1.8, 0.79	3.6, 0.70
O.ME.2	Fixed = $PM_{2.5}$, Temperature, Humidity, Time, Weekend, Near_hwy; Random = Intercept	8	1	2.4, 0.90	-----	3.5, 0.77
O.RF.1	$PM_{2.5}$, Temperature, Humidity	3	1	3.5, 0.80	2.0, 0.75	3.7, 0.64
O.RF.2	$PM_{2.5}$, Temperature, Humidity, Month, Time, Weekend	3	2	4.0, 0.72	2.1, 0.77	4.0, 0.61
O.RF.3	$PM_{2.5}$, Temperature, Humidity, Month, Time, Weekend, Near_hwy	7	2	3.3, 0.80	-----	4.1, 0.66
O.RF.4	$PM_{2.5}$, Temperature, Humidity, Month, Time, Weekend, Aroad_500	7	2	3.5, 0.80	-----	4.1, 0.66

In the statistical model column: O =on-the-fly data (as opposed to archived); LR = linear regression; ME = mixed effect linear regression; RF = random forest. In the variable column: Aroad = arterial road; Lroad = local road; the number following is the radial buffer size in meters within which the length of that type of road is being totaled. Near_hwy = near-highway indicator.

Note that "Time" and "Month" are the sinusoidal (cyclic) versions. Preliminary testing showed that including both sine and cosine of the hour of day did not improve performance in the linear models, and that including both sine and cosine of the month led to model non-convergence in the linear mixed effect models. Thus, for the linear and linear mixed effect models, "Time" refers only to cosine of hour of day; for the linear mixed effect models, "Month" *refers only to cosine of month. All other references to "Time" and "Month" imply the inclusion of both sine and cosine.*

Author Contributions:

Ellen Considine: conceptualization, formal analysis, writing (original draft); Colleen Reid: supervision, conceptualization, and writing (review and editing); Michael Ogletree: data curation and conceptualization; Tim Dye: data curation

Acknowledgements:

The data for this project were collected through funds from the 2018 Bloomberg Mayors Challenge. Funding for this analysis was provided in part by the University of Colorado Boulder's Undergraduate Research Opportunities Program. This work was supported by Earth Lab through the University of Colorado Boulder's Grand Challenge Initiative.

Conflicts of Interest

The authors declare no conflict of interest.

References

Badura, M., P. Batog, A. Drzeniecka-Osiadacz, and P. Modzel. "Regression Methods in the Calibration of Low-Cost Sensors for Ambient Particulate Matter Measurements." *SN Applied Sciences* 1, no. 6 (June 2019): 622. https://doi.org/10.1007/s42452-019-0630-1.

Bi, J. *et al.*, "Contribution of low-cost sensor measurements to the prediction of PM2.5 levels: A case study in Imperial County, California, USA," *Environ. Res.*, vol. 180, p. 108810, Jan. 2020, doi: 10.1016/j.envres.2019.108810.

Budde, M., L. Zhang, and M. Beigl, "Distributed, Low-cost Particulate Matter Sensing: Scenarios, Challenges, Approaches," p. 7, 2014.

Castell, N., *et al.*, "Can commercial low-cost sensor platforms contribute to air quality monitoring and exposure estimates?," *Environ. Int.*, vol. 99, pp. 293–302, Feb. 2017, doi: 10.1016/j.envint.2016.12.007.

Chai, T., and R. R. Draxler. "Root Mean Square Error (RMSE) or Mean Absolute Error (MAE)? – Arguments against Avoiding RMSE in the Literature." *Geoscientific Model Development* 7, no. 3 (June 30, 2014): 1247–50. https://doi.org/10.5194/gmd-7-1247-2014.

City of Denver Open Data Catalog, "Street Centerline," *https://www.denvergov.org/opendata*. https://www.denvergov.org/opendata/dataset/city-and-county-of-denver-street-centerline (accessed May 15, 2019).

Cromar, K. R., *et al.*, "Air Pollution Monitoring for Health Research and Patient Care. An Official American Thoracic Society Workshop Report," vol. 16, no. 10, p. 8, 2019.

Delaine, F., B. Lebental, and H. Rivano. "In Situ Calibration Algorithms for Environmental Sensor Networks: A Review." *IEEE Sensors Journal* 19, no. 15 (August 1, 2019): 5968–78. https://doi.org/10.1109/JSEN.2019.2910317.

Eberly, S., *et al.,* “Data Quality Objectives (DQOs) for relating Federal Reference Method (FRM) and Continuous PM2.5 Measurements to Report an Air Quality Objective (AQI).” US EPA, Office of Air Quality Planning and Standards. 2002.

Gao, M., J. Cao, and E. Seto, “A distributed network of low-cost continuous reading sensors to measure spatiotemporal variations of PM2.5 in Xi’an, China,” *Environ. Pollut.*, vol. 199, pp. 56–65, Apr. 2015, doi: 10.1016/j.envpol.2015.01.013.

Gregorutti, B., B. Michel, and P. Saint-Pierre. “Correlation and Variable Importance in Random Forests.” *Statistics and Computing* 27, no. 3 (May 2017): 659–78. https://doi.org/10.1007/s11222-016-9646-1.

Hagler, G. S. W., R. Williams, V. Papapostolou, and A. Polidori. “Air Quality Sensors and Data Adjustment Algorithms: When Is It No Longer a Measurement?” *Environmental Science & Technology* 52, no. 10 (May 15, 2018): 5530–31. https://doi.org/10.1021/acs.est.8b01826.

Hasenfratz, D., O. Saukh, and L. Thiele, “On-the-Fly Calibration of Low-Cost Gas Sensors,” in *Wireless Sensor Networks*, vol. 7158, G. P. Picco and W. Heinzelman, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 228–244.

Holstius, D. M., A. Pillarisetti, K. R. Smith, and E. Seto, “Field calibrations of a low-cost aerosol sensor at a regulatory monitoring site in California,” *Atmospheric Meas. Tech.*, vol. 7, no. 4, pp. 1121–1131, Apr. 2014, doi: 10.5194/amt-7-1121-2014.

Kuhn, M., *caret package (R)*. *Journal of Statistical Software*, 28(5), 2008.

Kumar, P. *et al*., “The rise of low-cost sensing for managing air pollution in cities,” *Environ. Int*., vol. 75, pp. 199–205, Feb. 2015, doi: 10.1016/j.envint.2014.11.019.

Magi, B. I., C. Cupini, J. Francis, M. Green, and C. Hauser. “Evaluation of PM2.5 Measured in an Urban Setting Using a Low-Cost Optical Particle Counter and a Federal Equivalent Method Beta Attenuation Monitor.” *Aerosol Science and Technology* 54, no. 2 (February 1, 2020): 147–59. https://doi.org/10.1080/02786826.2019.1619915.

Malings, C., *et al.* (2019a), “Development of a general calibration model and long-term performance evaluation of low-cost sensors for air pollutant gas monitoring,” *Atmospheric Meas. Tech.*, vol. 12, no. 2, pp. 903–920, Feb. 2019, doi: 10.5194/amt-12-903-2019.

Malings, C., *et al.* (2019b), “Fine Particle Mass Monitoring with Low-Cost Sensors: Corrections and Long-Term Performance Evaluation.” *Aerosol Science and Technology* 54, no. 2 (February 1, 2020): 160–74. https://doi.org/10.1080/02786826.2019.1623863.

Martin, R. V., M. Brauer, A. van Donkelaar, G. Shaddick, U. Narain, and S. Dey, “No one knows which city has the highest concentration of fine particulate matter,” *Atmospheric Environ. X*, vol. 3, p. 100040, Jul. 2019, doi: 10.1016/j.aeaoa.2019.100040.

Ottosen, T.B. and P. Kumar. “Outlier Detection and Gap Filling Methodologies for Low-Cost Air Quality Measurements.” *Environmental Science: Processes & Impacts*, 21 (February 2019): 701-713

Sayahi, T., A. Butterfield, and K.E. Kelly. "Long-Term Field Evaluation of the Plantower PMS Low-Cost Particulate Matter Sensors." *Environmental Pollution* 245 (February 2019): 932–40. https://doi.org/10.1016/j.envpol.2018.11.065.

Si, M., Y. Xiong, S. Du, and K. Du. "Evaluation and Calibration of a Low-Cost Particle Sensor in Ambient Conditions Using Machine-Learning Methods." *Atmospheric Measurement Techniques* 13, no. 4 (April 7, 2020): 1693–1707. https://doi.org/10.5194/amt-13-1693-2020.

Williams, R., *et al.,* "Peer Review and Supporting Literature Review of Air Sensor Technology Performance Targets," US Environmental Protection Agency, Office of Research and Development National Exposure Research Laboratory. September 2018.

Xu, Y., *et al.*, "Evaluation of machine learning techniques with multiple remote sensing datasets in estimating monthly concentrations of ground-level PM2.5," *Environ. Pollut.*, vol. 242, pp. 1417–1426, Nov. 2018, doi: 10.1016/j.envpol.2018.08.029.

Zhang, Y., N. Meratnia, and P. Havinga. "Outlier Detection Techniques for Wireless Sensor Networks: A Survey." IEEE Communications Surveys & Tutorials 12, no. 2 (2010): 159–70. https://doi.org/10.1109/SURV.2010.021510.00088.

Zheng, T., *et al.*, "Field Evaluation of Low-Cost Particulate Matter Sensors in High- and Low-Concentration Environments." *Atmospheric Measurement Techniques* 11, no. 8 (August 22, 2018): 4823–46. https://doi.org/10.5194/amt-11-4823-2018.

Zheng, T., *et al.*, "Gaussian Process Regression Model for Dynamically Calibrating a Wireless Low-Cost Particulate Matter Sensor Network in Delhi." Preprint. Aerosols/In Situ Measurement/Data Processing and Information Retrieval, March 1, 2019. https://doi.org/10.5194/amt-2019-55.

Zimmerman, N., *et al*., "A machine learning calibration model using random forests to improve sensor performance for lower-cost air quality monitoring," *Atmospheric Meas. Tech*. 11, no. 1 (Jan. 2018): 291–313, doi: 10.5194/amt-11-291-2018.

Zoest, V. M. van, A. Stein, and G. Hoek. "Outlier Detection in Urban Air Quality Sensor Networks." *Water, Air, & Soil Pollution* 229, no. 4 (April 2018): 111. https://doi.org/10.1007/s11270-018-3756-7.

Zusman, M., *et al.,* "Calibration of Low-Cost Particulate Matter Sensors: Model Development for a Multi-City Epidemiological Study." *Environment International* 134 (January 2020): 105329. https://doi.org/10.1016/j.envint.2019.105329.