2022

Supervised Machine Learning Techniques Applied to Low-Cost Air Quality Sensor Suites

Peter Wahman
Minnesota State University, Mankato

Supervised Machine Learning Techniques Applied to Low-Cost Air Quality Sensor Suites

By

Peter Wahman

A Senior Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of
Bachelor of Science
in
Engineering with an Electrical Focus

Minnesota State University, Mankato
Mankato, Minnesota
March 2022

June 3, 2022

Supervised Machine Learning Techniques Applied to Low-Cost Air Quality Sensor Suites

Peter Wahman

This Senior Thesis has been examined and approved by the following members of the student's committee.

Dr. Jacob Swanson

Dr. Rebecca Bates

Dr. Robert Sleezer

Abstract

Low-cost PM sensors have garnered interest for their ability to reduce the cost of investigating PM concentrations in both indoor and outdoor spaces. They perform well in high concentration lab testing with correlation coefficients greater than 0.9. In real-world applications, the correlation coefficients drop significantly because of sensing floors and adverse ambient conditions. There are plenty of supervised machine learning techniques that aim to correct the measurements ranging from linear regression to more advanced neural networks and random forests. This work aims to use those more complicated techniques to adjust the measurements using other data sets gathered by a sensor suite. The Minnesota Pollution Control Agency (MPCA) has deployed a network of 47 AQ-Mesh sensors around the Minneapolis-St. Paul Metro Area. The network was active for two years, with mass colocations at a regional federal sensing site before and after deployment. The sensor suite includes electrochemical sensors for nitric oxide (NO), nitrogen dioxide (NO2), carbon monoxide (CO), ozone (O3), and sulfur dioxide (SO2). The suite also has an Alphasense OPC-N2 particle counter for PM measurements, along with temperature, pressure, and relative humidity sensors. Using most of these sensors in combination with basic supervised machine learning regression and a large dataset spanning over two years predictors are trained, applied, and examined for stability.

Acknowledgments

I would like to acknowledge all the people who helped make this project possible. From the University of Minnesota, Mankato, Dr. Jacob Swanson managed this project with great patience. Monika Vidali and Margaret McCourtney from the Minnesota Pollution Control Agency managed the physical sensors and the data access. Additionally, I would like to thank the professors and administrative staff at Twin Cities Engineering. Finally, I would like to thank my family and friends for supporting me through this process.

Funding for this project was provided by the Minnesota Environment and Natural Resources Trust Fund as recommended by the Legislative-Citizen Commission on Minnesota Resources (LCCMR). The Trust Fund is a permanent fund constitutionally established by the citizens of Minnesota to assist in the protection, conservation, preservation, and enhancement of the state's air, water, land, fish, wildlife, and other natural resources.

ENVIRONMENT
AND NATURAL RESOURCES
TRUST FUND

MINNESOTA STATE

Minnesota State University, Mankato
A member of Minnesota State

Table of Contents

Introduction

Air quality has been a public health concern in the past, but more recent studies have suggested that poor air quality rivals smoking and eclipses violence as a measure of reduction in lifespan. (Jos Lelieveld, 2020) Considering the major health effects caused by high concentrations of gasses and particles, it follows that many regulatory agencies are investigating new approaches to monitoring.

In the US, the Environmental Protection Agency (EPA) operates a network of continuous monitors using sensors defined as federal equivalent methods (FEM). These sensors are quite expensive, along with the need for staff to maintain and operate them, so there are inherent financial limitations leading to relatively few numbers of monitoring stations in the state and urban areas. The low number of high-quality monitors inherently limits the spatial resolution of air quality data.

Recent advances in electrochemical sensing of concentrations have moved the detection limit of sensors from the parts per million level to the parts per billion level necessary for the detection of the prominent gasses that the regulatory agencies are interested in. There have also been advances in optical measurements for particulate matter (PM) that have been integrated into sensors that can be produced at a much lower cost than the more traditional PM sensing techniques. Several companies have used these advances to produce low-cost sensing suites that cover a wide range of pollutants that regulatory agencies are interested in. The Minnesota Pollution Control Agency (MPCA) had an interest in using these sensors to look at air quality in the Minneapolis-St. Paul metro area.

The low-cost sensing sites offer a chance at improving the spatial resolution of air quality data; however, the wide range of environmental conditions that the low-cost sensors are exposed to, along with low concentrations of pollutants in some places, poses a challenge for low-cost sensors. For example, electrochemical sensors are affected by temperature and cross-gas sensitivity, and optical sensors have a sensitivity to humidity and temperature. In laboratory testing, many of these sensors show a strong correlation to a reference sensor. (Andrea Polidoi, 2016) When moving from the laboratory into the field, various corrections are applied to calibrate the sensor. They include firmware corrections along with database corrections applied. These corrections are focused on cross-gas sensitivity and temperature changes for the electrochemical sensors along with multipliers and offsets based on laboratory calibration for both electrochemical and optical sensors. (Air Quality Sensor Performance Evaluation Center, 2021)

Initially, these corrections are applied to the sensors at the factory and sometimes updated in a calibration colocation before deployment. Over time the sensors drift due to chemical changes in an electrochemical sensor, or the harsh environment affects both types of sensors. The changes, along with a generally lower level of pollutant concentration, lead to the correlation, sensitivity, and accuracy of the sensors being reduced. Attempts have been made to incorporate some environmental factors to improve the quality of the measurements. There have been some corrections that draw heavily on aerosol science (Carl Malings, 2020). The idea was to consider that the FEM monitors were using a standard temperature and humidity and the low cost sensors were not. They employed methods to correct for the activation of water at different temperatures and humidities and then used that information along with information about they composition of the particles to calculate a correction factor for water condensing on the particles.

The other major technique that has been attempted is algorithmic corrections. There are two major branches of study, one based on linear regression and another based on random forests. These both try to correct the low cost air quality sensor data using multiple inputs, but used empirical data to arrive at those corrections instead of using a theoretical model. Past studies have applied both of these corrections in (N. Zimmerman, 2018) and (E. Esposito, 2016) but generally they have data sets that are weeks to a few months in length. The work detailed in this report has a dataset that approaches two years in length.

This work focuses on those algorithmic corrections; these corrections are all a form of supervised machine learning (SUPML). In SUPML, a predictor is trained on a set of data with various input features and a known target output. After building the predictor, test data is fed to the predictor with a known target to analyze how accurate the predicted value is. There are various forms of SUPML applied, but the ones specifically used in this project are Random Forests (RF) and Neural Networks (NN). Random Forests are an ensemble of decision trees; a decision tree effectively takes an input value and adjusts it at various decision points based on the features that the predictor requires. This is repeated hundreds of times as many decision trees make a random forest; the final number is averaged from the outputs of all the trees. (Brieman, 2001) A neural network is different from the random forest in that it is not an ensemble; there is one predictor that outputs the final number. A neural network is also built from linear regressions, meaning that, unlike a random forest, it can forecast predictions outside of the range it was trained on. Both of these techniques are quite common in machine learning and have implementations in various open-source libraries.

There are a few exciting factors in this application of SUPML to low-cost air quality sensor data; the first is the length of time that the sensors were deployed. Two years of data were collected, meaning that any annual variations or drift would be captured by the predictors. The second interesting factor is the environment that the sensors were put in, with a range of average temperatures of greater than 50°C. The third challenging factor for the sensors was the low concentration environment inherent in Minnesota as a sensing location. Some of the concentrations involved were below the manufacturer's recommended levels. All of these factors combine to pose an interesting challenge; in this environment, can quality data be extracted from the low-cost sensor suite, and if the uncorrected data is poor, can SUPML methods improve the data quality?

Background

AQ-Mesh Suite

The AQ-Mesh sensing suite integrates three main sensor types, logs data from them, and sends that data to the MPCA through AQ-Mesh. The three types of sensor include an optical particle counter, electrochemical gas sensors for various gasses, and basic environmental sensors. All of these sensors have different sample rates and data delivery types. The data is integrated and logged as one 15-minute measurement and sent to the MPCA servers. On the MPCA side, the data is averaged to 1-hour increments to be comparable to Federal Equivalent Measurement(FEM) methods.

Alphasense Optical Particle Counter

The optical PM sensor onboard the AQ-Mesh sensor is an Alphasense OPC N2. This sensor is a fan-driven optical particle counter with a measurement range of 0.38 µm to 17 µm split into 16 bin sizes. The sensor uses an onboard algorithm to calculate the mass concentration of the particles and give the mass concentration of PM1.0, PM2.5, and PM10. The sensor has a preset density of the particles it uses to shift from number concentration to mass concentration; the preset value is 1.65 g/cm^3. Mass concentration out of the sensor can be reported at 1.4-20 second intervals; AQ-Mesh processes this into 15-minute averages, and to align with the federal monitors, those 15-minute averages are averaged into a 1-hour average. The sensor does not preprocess any of the aerosols; there is no dry air sheath flow, or temperature modifications like the federal sites have. This means that there is a variation in how much water may be condensed on a particle as it passes through the optical particle counter. When humidity is high, the particles will appear bigger to the particle counter; their density will also be reduced because water has a density of 1.0 g/cm^3. The composition of particles is also not a constant, and the bulk aerosol hygroscopicity is related to the composition of the particles. The composition of the particles changes, generally seasonally, corrections based on κ-Kohler theory will adjust these over the seasons. (Carl Malings, 2020). Using the Neural Network and Random Forest corrections to adjust for the multiple sources of variation in the final output seemed to be the best way to achieve improvements across the full suite of sensors.

Alphasense B4 Electrochemical Sensors

The electrochemical sensors in the AQ-Mesh suite are all types of Alphasense B4 electrochemical sensors. They all contain an electrolyte behind a gas permeable membrane that varies its resistance based on the concentration of the target gas. The variable resistance is measured by the voltage differential across the terminals. For known cross-gas effects, AQ-Mesh suites preprocess the data before sending it to the MPCA for analysis; this generally involves subtracting the NO2 concentration from the O3 concentration. (Aphasense ltd., 2019) Sample rates of the B4 sensors are close to 2s per sample. To match the federal equivalent sites, the data is processed by AQ-Mesh using a proprietary algorithm into 15-minute averages, and those 15-minute averages are averaged by the MPCA into hourly averages.

An electrochemical sensor is a sensitive piece of equipment; rapid environmental changes have a negative effect on the seal and the electrolytes in the electrode. Humidity and temperature changes are the main factors that the sensors are responsive to. These environmental factors affecting the sensor gave an insight into what features may be interesting to try to add into a corrective function.

The limit of detection (LOD) for most of these sensors is 1ppb; CO has a higher LOD of 10ppb. AQ-Mesh has tested these devices and has its own limits. They have a special confidence measure called a limit of confidence (LOC). This confidence measure has to do with the standard deviation of their measurements and the concentration being measured. They have found that for NO, NO2, and O3, their confidence level is 10ppb. For the CO sensor, their confidence level is 50ppb. (Air Quality Sensor Performance Evaluation Center, 2021)

Specifics about environmental sensors

The three main environmental sensors are pressure (P), temperature (T), and relative humidity (RH). These sensors are common on most environmental sensing packages and are well characterized and well understood. As inputs into predictive models, general trends are more valuable than absolute accuracy.

NN/RF Specifics

The data used to train the networks contained matched pairs of measurements from both a federal equivalent method source and an AQ-Mesh source. There were only two gas sites but four and five sites for PM10 and PM2.5, respectively. This data imbalance meant that to get a complete snapshot of a moment in time; there was really only one site that could give a full picture of both the federal measurements and the AQ-Mesh measurements at the same time. To increase the amount of data available to train the models, the PM predictors did not use the gasses as features; conversely, the gas predictors did not use the PM data to make predictions.

The PM predictors used AQ-Mesh data for RH, Temperature, Pressure, PM2.5, and PM10 as features to achieve a correction matching the FEM outputs. The gaseous predictors used RH, Temperature, Pressure, CO, NO, NO2, and O3 as features to correct for their respective FEM outputs. The data used to build the predictors was the first half of the data from March 2019 to August 2020. The predicted or test data goes until May 2021.

The predictors were not built to be machine-specific corrections. Because they learned on data from multiple sources and some sensor changes during the project life, the corrections are of a more generalized form. The corrections may apply to sensors across the network.

There are two different predictor networks used to improve the final data output of the AQ-Mesh suite. The first is a random forest type predictor; the algorithm is Breiman's original algorithm implemented in the randomforest package for R (Brieman, 2001). The number of decision points at each node was five, the number of trees that were grown was 2000 for PM predictors and 1000 for gaseous predictors. The number of variables sampled at a decision point was 4 for both gaseous and PM predictors. The predictors used mean standard error (MSE) as a regressor to learn on.

The random forest correction has one major drawback: non-parametric correction. Non-parametric, in this case, means that it can only predict in the range it has been trained on; if inputs are given that exceed the range of training data, it will output the maximum concentration that it has seen before. This lack of range can be an issue when there is a short period of training data available; the full range of conditions the sensor is expected to sense will probably happen in a short amount of time. There are annual variations, and there could be spikes in pollutant concentration during a short-term event. This means that for this correction to work well, an extensive range of concentrations needs to be seen by the sensor to accurately train the predictor.

The second type of predictor built was a neural network type predictor. The neuralnet package for r was used to create this type of predictor. The neural network predictor used the same data and features as the random forest type predictor but needed different network variables. The variables of importance are how many hidden layers to include and how many steps the network is allowed to learn on. For the gaseous predictor, there were four hidden layers with 10^7 steps allowed before convergence.

There is a major benefit to the neural network type predictor when it comes to regression predictions. The predictor is parametric, so it can adjust to conditions outside of what it has been trained on. The parametric nature of neural networks may allow more options for training with reduced ranges.

Methods

Physical Layout of Sensors

The MPCA AQ-Mesh network is made up of around 45 sensor suites in every zip code in the Minneapolis-St. Paul city limits. Most of the network is mounted about 3m high on various poles the MPCA negotiated access to. In Minneapolis, they were mounted on wooden light poles owned by Xcel energy; in St. Paul, they were mounted on light poles in school parking lots owned by the St. Paul school district; a select few are collocated within 3m of a FEM site. The two gaseous FEM sites are located at a regional airport outside the city and next to a large interstate highway interchange in the core of the city. These two sites are the source of the synchronized data for the gaseous training and test data.

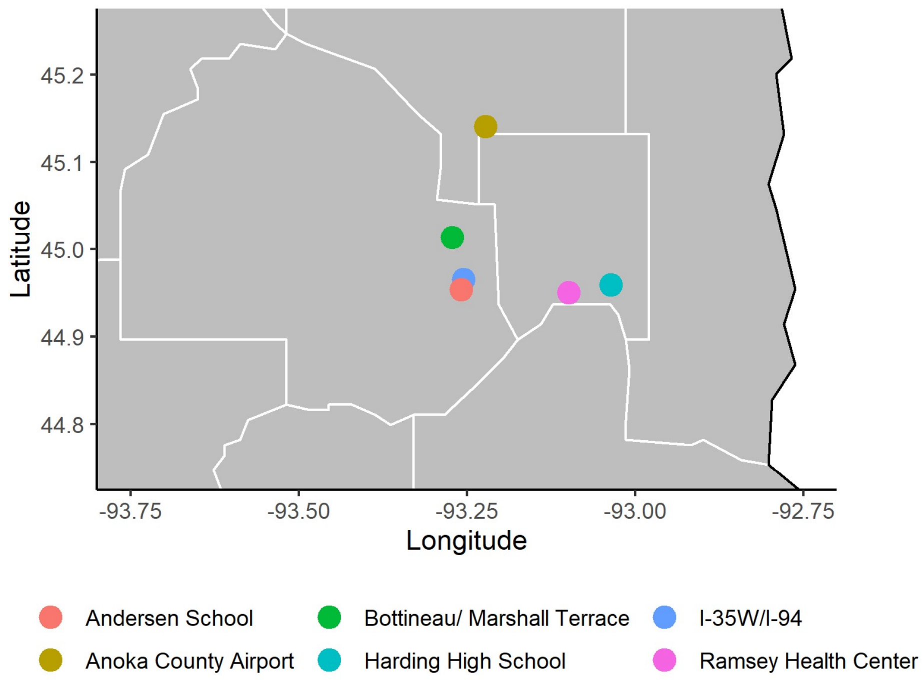

Figure 1 Locations of Co-located sensor suites in the Minneapolis St. Paul metro area.

There are also several sites that have a FEM for PM data; these are at both the regional airport site and the interstate highway site. There are also a few more at other locations around the city; in (Figure 1), the sites are shown.

Data Cleaning

There are three major steps in the data cleaning process, the first is checking for major outliers, the second step is duplicate data checks, and the third is using data points within AQ-Mesh confidence levels.

The first cleaning goal was to reduce major outliers; the univariate mean and standard deviation technique was used. This technique is commonly applied to many datasets and is less restrictive than other techniques (Wada, 2020). Both the AQ-Mesh data and the AQM data have their averages and standard deviations calculated, and any data points further than three standard deviations from the mean are considered outliers and removed from the dataset.

The second challenge is when there are multiple types of FEM sensors at a collocated site. This can add duplicate data points into a dataset. For example, some gaseous sensors have a trace sensor and a base sensor; this would lead to duplicate measurements for the federal monitors at that hour. To address duplicate points created by multiple senor types, only one type of FEM sensor is considered for each site and pollutant.

The final data filter is for data that is outside of AQ-Mesh's loss of confidence (LOC) range. This is a limit that is higher than the limit of detection (LOD) that the manufacturer recommends (AQ-Mesh, 2021). It is a limit where the variation in AQ-Mesh's testing they found that the standard deviation in their data is greater than the concentration of the pollutant measured. For most of the gas sensors, this limit is 15ppb; for the CO sensor, it is 50ppb. Most of the sites have a very low concentration of pollutants; for CO, our average was 40ppb over the life of the project. For sites where the average concentration was above the LOC limit, data below the LOC limit was excluded. For CO, all of the data was included because 90% of the data was below the LOC limit of 50ppb.

The data is then collected into a records table for all measurements. In the case of the gaseous measurements, they had columns joined on sample time of the CO, NO, NO2, and O3 measurements, both federal and AQMesh values, and the environmental measurements of T, RH, and P. The PM sensors underwent a similar joining, the PM2.5 data was a record of the site, sample time, PM2.5 Federal, PM2.5 AQMesh, T, RH, and P. The PM10 was similar except the for looking at PM10 data instead of PM2.5. When the measurements were collected, a final tally of 9190 measurements over the life of the project was collected for the gaseous set; 37270 were collected for the PM10 set, and 61960 were collected for the PM10 set.

Metrics

The metrics chosen to examine how well the sensors performed were the slope between collocated sensors, the R^2 between the sensors, the relative mean, standard error (RMSE), and the relative standard error (RSE). These metrics give a good look at how closely the AQ-Mesh sensors match the FEM sensors.

The slope created by a linear regression of the AQ-Mesh data compared with the FEM data shows how the response of the AQ-Mesh sensor compares to the FEM. Ideally, the slope has a value of one; the EPA defines a method to be equivalent if the slope is within 0.1 of 1.

The R^2 of the linear regression between the AQ-Mesh sensors and the FEM gives a good indication of what percentage of the variability of the AQ-Mesh sensor is caused by the change in the FEM sensor. This metric isn't actually related to how accurate the measurement is it is a precision metric. It is a metric that can be communicated quickly that shows the precision of the sensor in question. In the lab, most of these sensors are showing R^2 values of <0.9 (AQ-Mesh, 2021).

RMSE gives the standard deviation of the residuals in a regression. This is useful because it gives an indication of the confidence levels of a sensor. This measure is also in the same domain as the measurement. This gives access to another measure, the relative standard error (RSE); this measurement divides the RMSE by the average concentration so that a fraction of variability in the measurement can be compared. Ideally, the RSE is below 0.25 for a sensor; RSEs greater than that mean that many of the measurements have a lower confidence than the measurement.

These metrics were calculated monthly to gauge the quality of the sensors over time. Not all hours in a month had full pictures of the data; some months had very few hours where all sensors from both suites were operating. To reduce the impact of months where there was little data, a lower limit of 100hrs of data per month was used to analyze these metrics.

Data Splits

Data splits and testing are something to consider when implementing a machine learning correction. For many machine learning applications, the dataset doesn't care about time. For correcting sensor data, the timing of the data is essential. Generally, there is a correction/calibration phase and then a measurement phase of a project. Because of this time sensitivity, random sampling of data is not warranted. In this application, annual fluctuations were noticed in the uncorrected data, so the data split tries to capture some of that fluctuation.

The training data for all sensors starts in March 2019 and goes through August 2020. The test data that has not been seen by the predictor happens after August 2020 and runs through May 2021. The final testing metrics are month-to-month values over the life of the project so there will be months where the predictor has seen all the data available and months where the predictor has seen none of the data from those months.

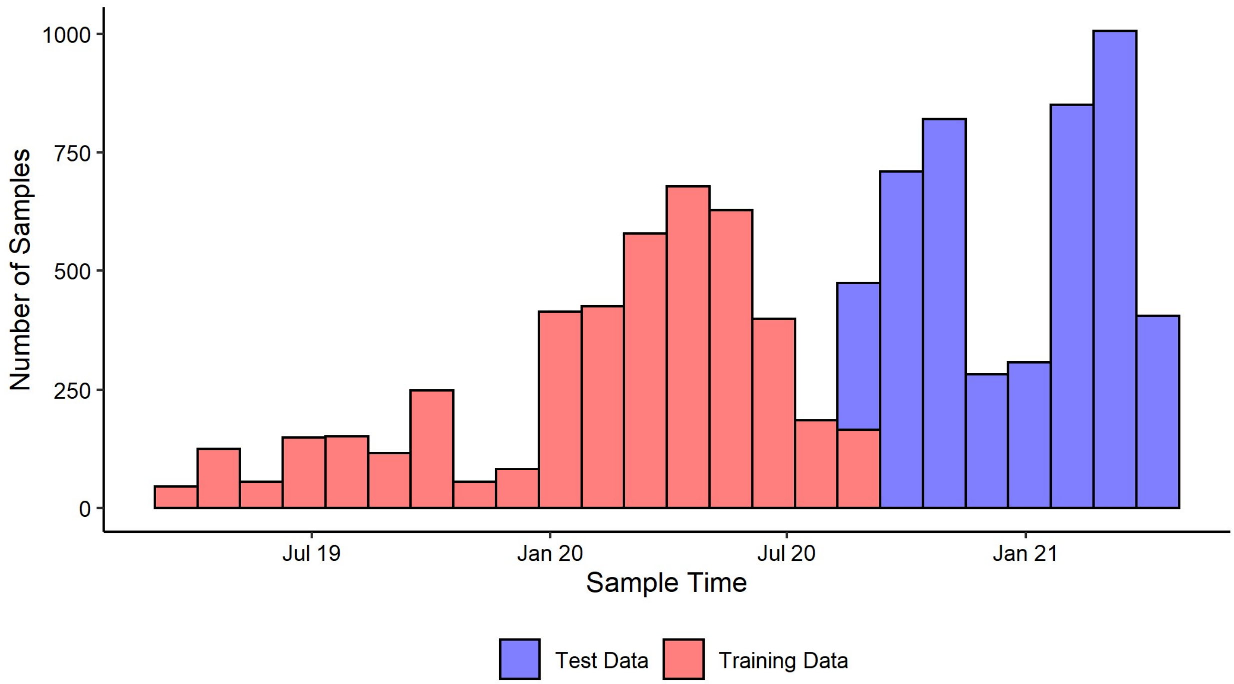

Figure 2 Data split for gaseous data set. Each bar represents a month of meaurments collected at collocated gaseous sites. There are 17 months that are included in the training set and 8 months that are included in the test set.

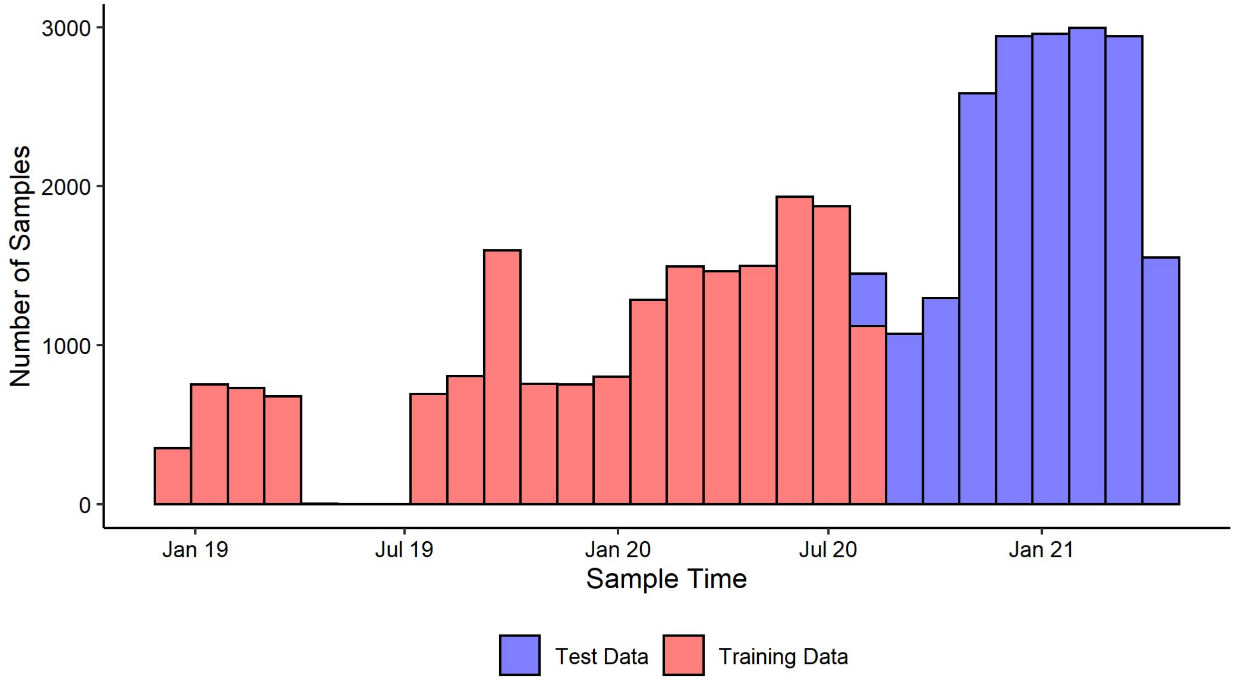

Figure 3 Data split for PM10 data set. Each bar represents a month of meaurments collected at collocated PM10 sites. There are 17 months that are included in the training set and 11 months that are included in the test set. There is a gap of three months in the training set.

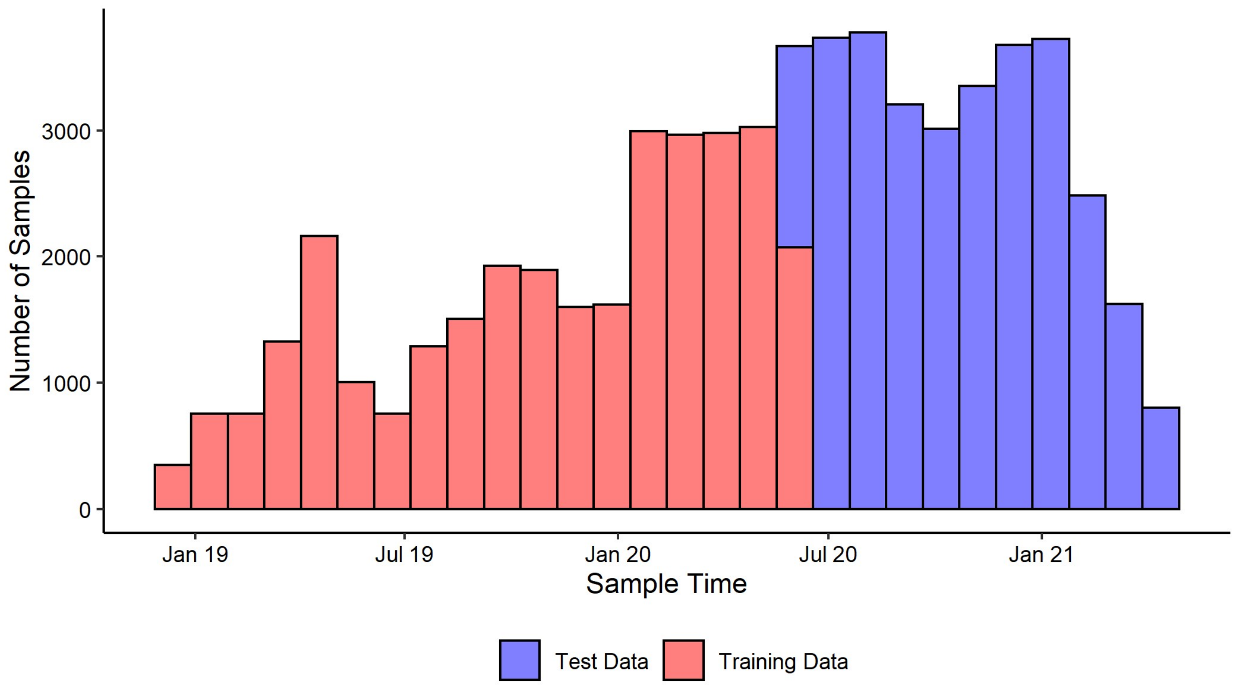

Figure 4 Data split for the PM2.5 data set. Each bar represents a month of meaurments collected at collocated PM2.5 sites. There are 17 months that are included in the training set and 11 months that are included in the test set.

Results and Discussion

Sensor Results After Random Forest Correction

The results shown are the effect of the Random Forest predictor on the data using the metrics established. The data shows the 8 months at the the two gaseous sites and 11 months at the various PM sites. First is a series of three graphs showing the effect of the predictor on all of the target sensors. Next is a time series with two selected sensors showing slope and R^2 and the effect of the predictor over time.

Slope Results For all Sensors and Sites

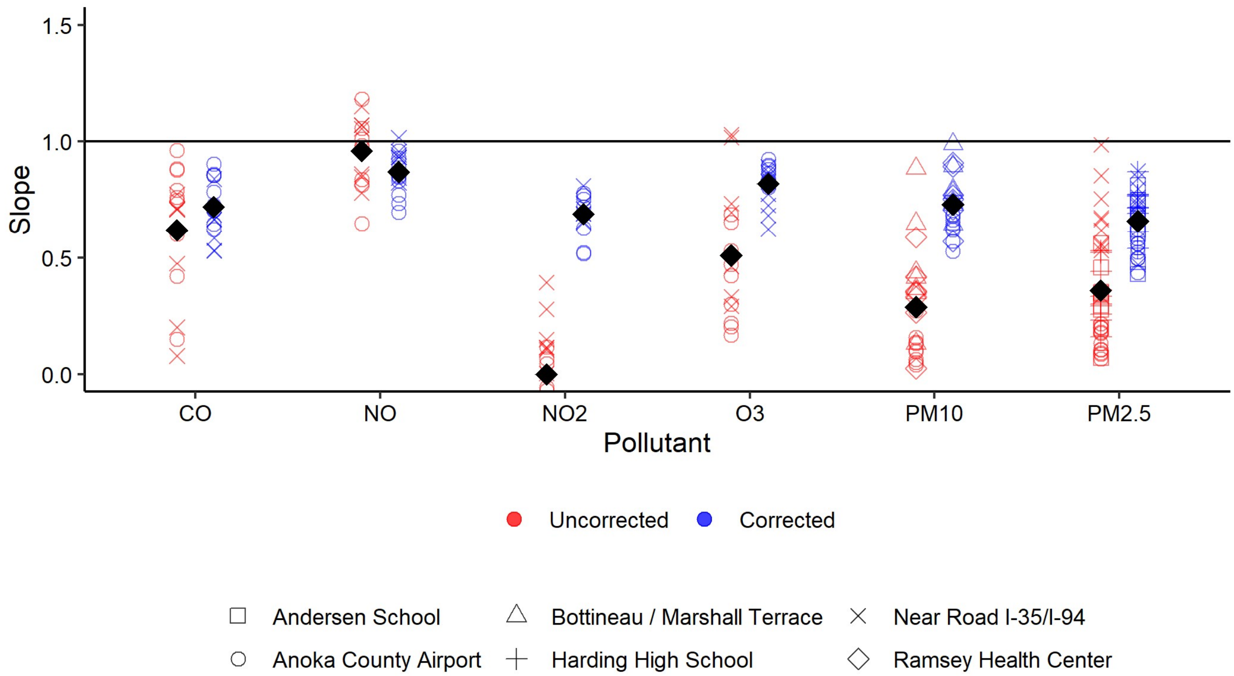

Figure 5 Slope of the line between federal reference sensor and AQMesh sensor before and after Random Forest Correction applied. Each point is one month at a specific site, and the black squares are the mean of the measurements.

In general, there was an improvement in response for the sensors, the black points showing the mean are mostly closer to the ideal of a 1:1 ratio between the federal and the AQM sensor. The sensor that showed the most improvement was the nitrogen dioxide (NO2) sensor, with an average of 0.03 to 0.69. The second biggest improvement comes from the PM10 sensor moving from 0.28 to 0.72. The nitric oxide (NO) sensor that didn't make any numerical improvements on average did make some improvements on the variance of the points. The standard deviation of the spread went from 0.25 to 0.08.

R^2 For All Sensors and Sites

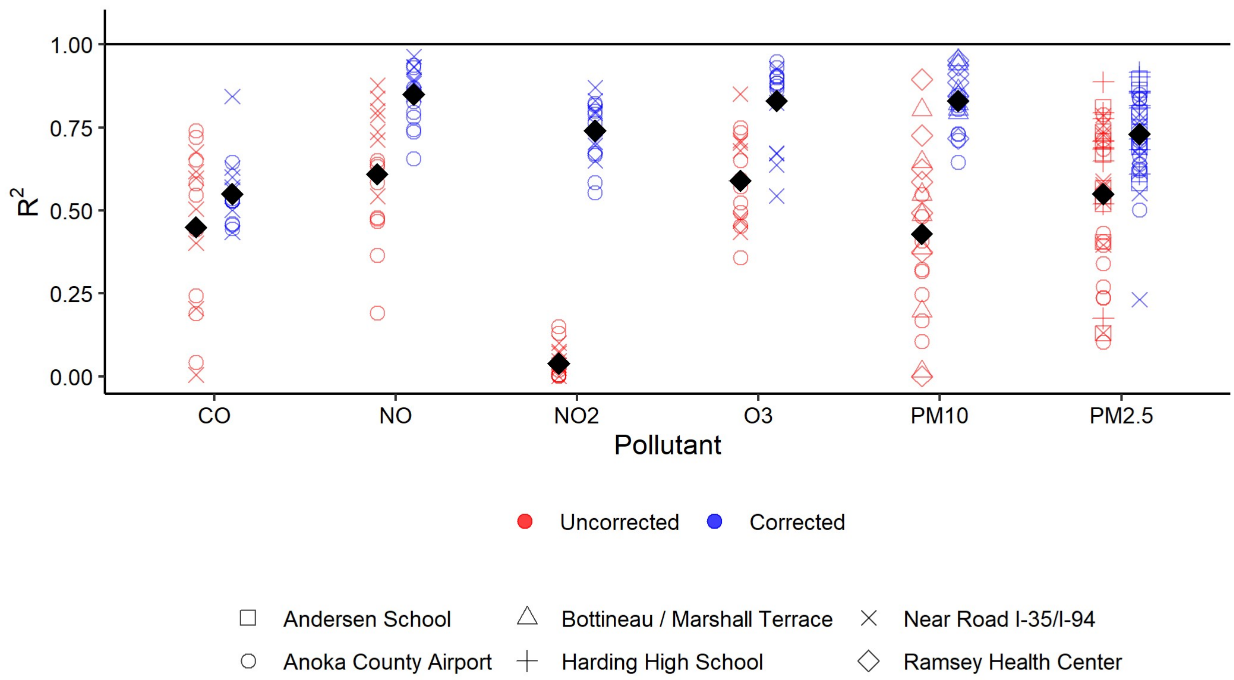

Figure 6 R-Squared of Random Forest correction for sensors compared to a federal equivalent. Black dots represent the mean of the measurements in the column.

The R^2 is improved across all sensors and sites with the random forest correction. All sensors except the carbon monoxide (CO) sensor improve to levels greater than 0.7. The sensors that show the greatest improvement are again the nitrogen dioxide (NO2) sensor moving from 0.04 to 0.73, and the PM10 sensor which improved from 0.43 to 0.83. Similar to the slope measurement, even the worst-performing sensor in this metric made an improvement in the variance of the measurements. The CO sensor improved its average from 0.45 to 0.55 and reduced the standard deviation of the correlation from 0.24 to 0.1.

Relative Standard Error for All Sensors and Sites

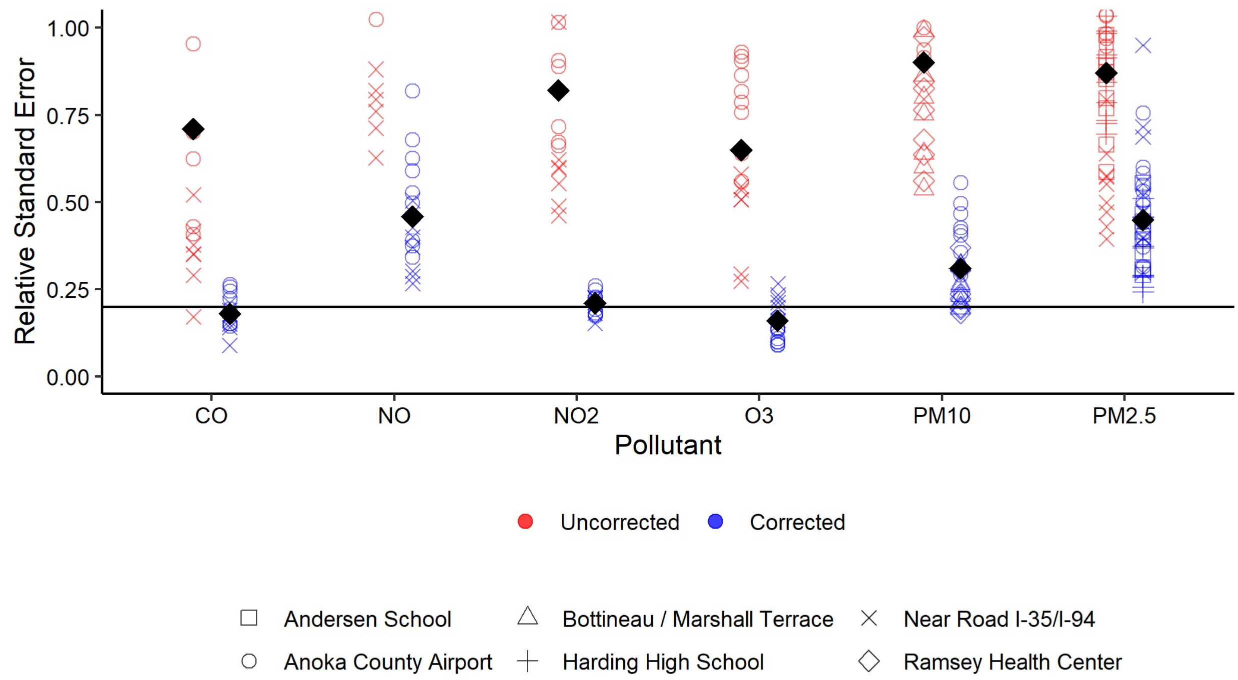

Figure 7 Relative Standard Error of sensors after a Random Forest correction is applied, with black markers for the average. The target to reach is zero, but an RSE of under 0.2 is a goal for many low-cost sensors.

Across all sensors and sites, there was an improvement in the RSE of the sensors after the correction was applied. The highest performing sensors in this metric were the O3 sensors after correction showing a reduction of the mean for 0.65 to 0.15. CO and NO2 also showed improvement CO had a mean of 0.18, passing the 0.2 threshold and NO2 got close with 0.21. CO NO2 and O3 also had a standard deviation of 0.05, 0.03, and 0.06, respectively.

Random Forest Correction Staying Power

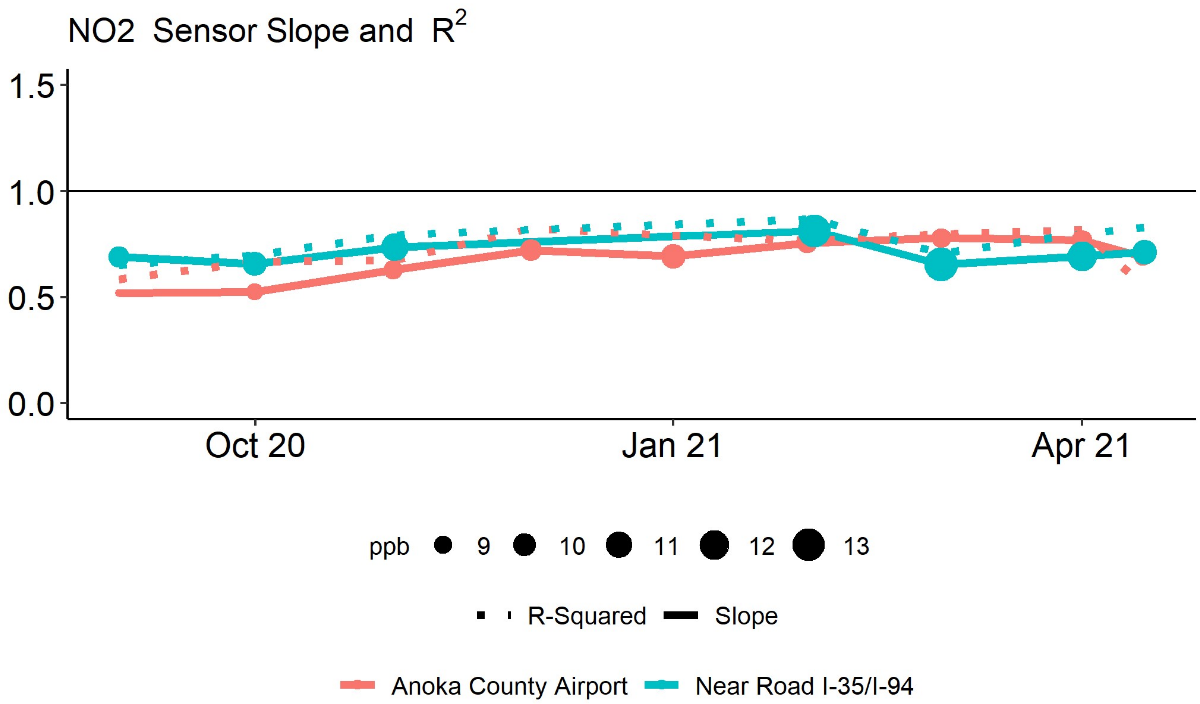

Figure 8 NO2 sensor slope, R^2, and average concentration over time at the 35W site and the Anoka Site using RF correction.

The electrochemical gas sensor that showed the most improvement after the correction was the NO2 sensor; the correction shows consistent improvements throughout the rest of the 9 months of data available. Both slope and R^2 were improved across two different sites in a variety of average concentration levels for the entire time period. The corrections also give similar sensor responses and R^2 for both sites.

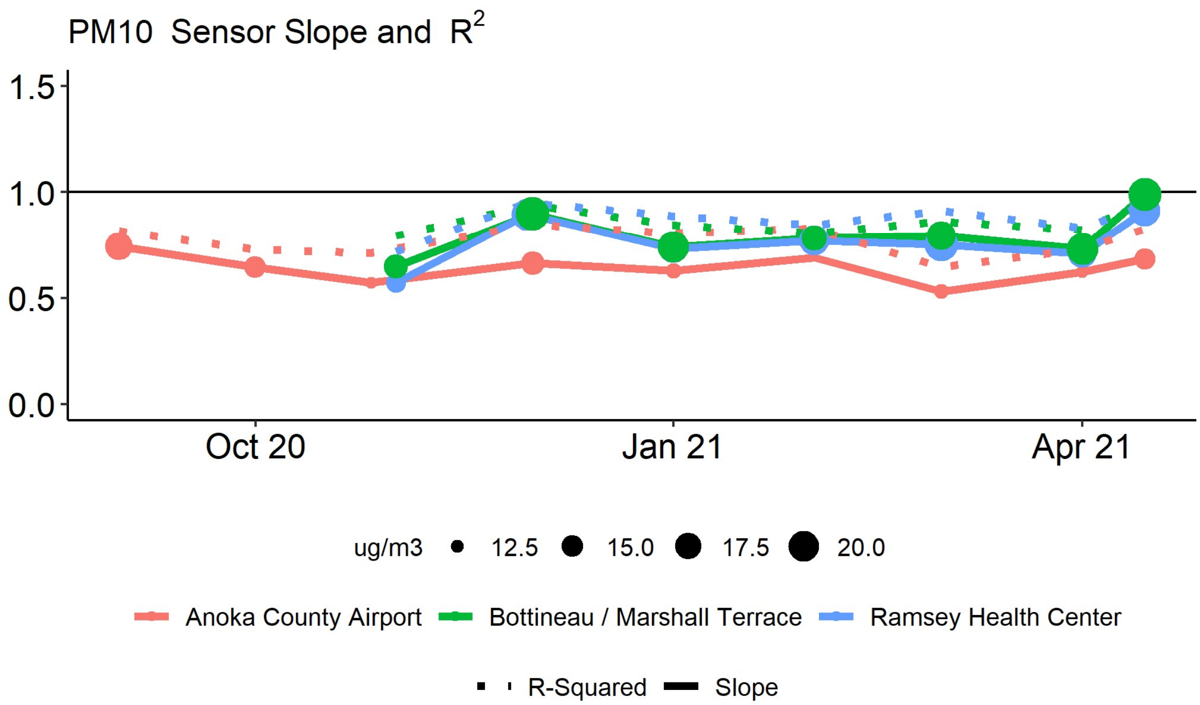

Figure 9 PM10 sensor slope, R^2, and average concentration over time at the Ramsey Health Center (RHC) site, the Bottineau Marshall Terrace site, and the Anoka Site using RF correction.

The optical sensor correction showed a similar staying power to the electrochemical sensor correction. Slope and R^2 were improved across all sites in many different average concentrations for months after the correction was applied.

Sensor Results After Neural Network Correction

The results shown are the effect of the Neural Network predictor on the data using the metrics established. The data shows the 8 months at the the two gaseous sites and 11 months at the various PM sites. First is a series of three graphs showing the effect of the predictor on all of the target sensors. Next is a time series with two selected sensors showing slope and R^2 and the effect of the predictor over time.

Slope For All Sensors and Sites

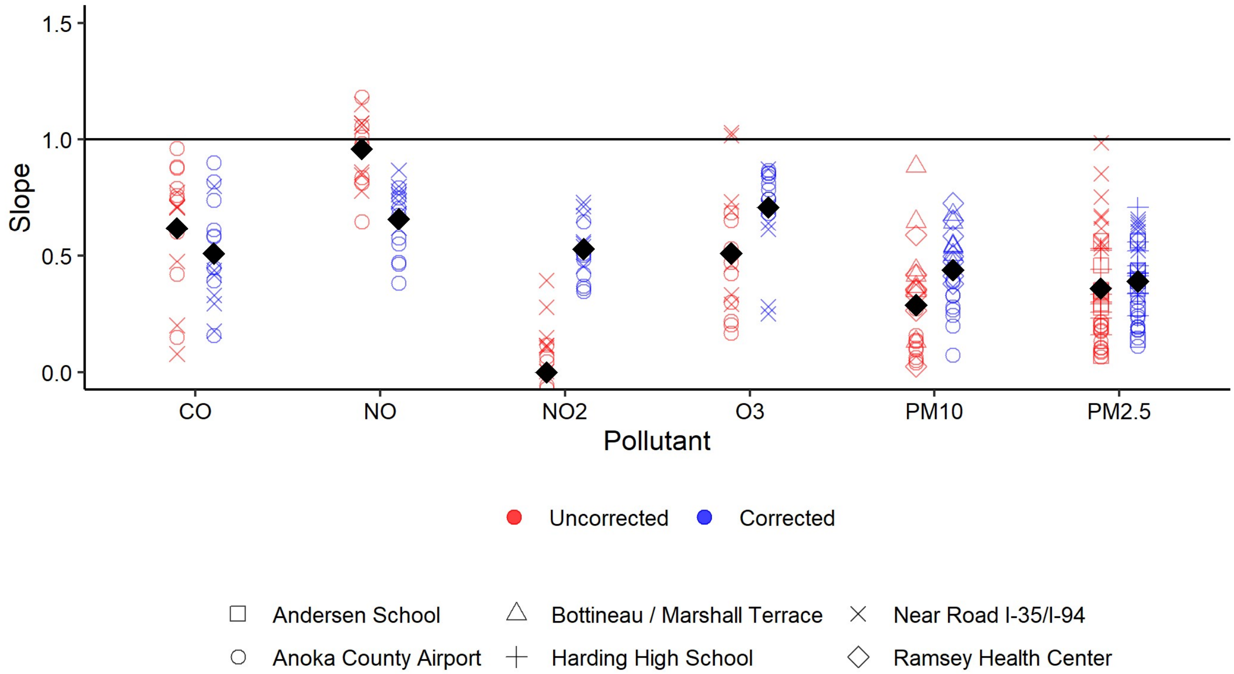

Figure 10 Slope of line between federal reference sensor and AQ Mesh sensor before and after Neural Network Correction applied. Each point is one month at a specific site and the black squares are the mean of the measurements.

The Neural Network (NN) correction had a mixed record for correcting the slope in the sensor suite. The average of the slopes declined for CO and NO but improved for all other sensors. The most improved is NO2 moving from 0.003 to 0.52. All sensors had reduced variability in the measurements. Sensors had a standard deviation in slope measurements of 0.2 to 0.27 before correction. After correction, the standard deviations were 0.14 to 0.22, an improvement in consistency across the suite. In general, this indicates mixed news on sensor responsiveness but increased predictability of that sensor responsiveness after correction.

R^2 For All Sensors and Sites

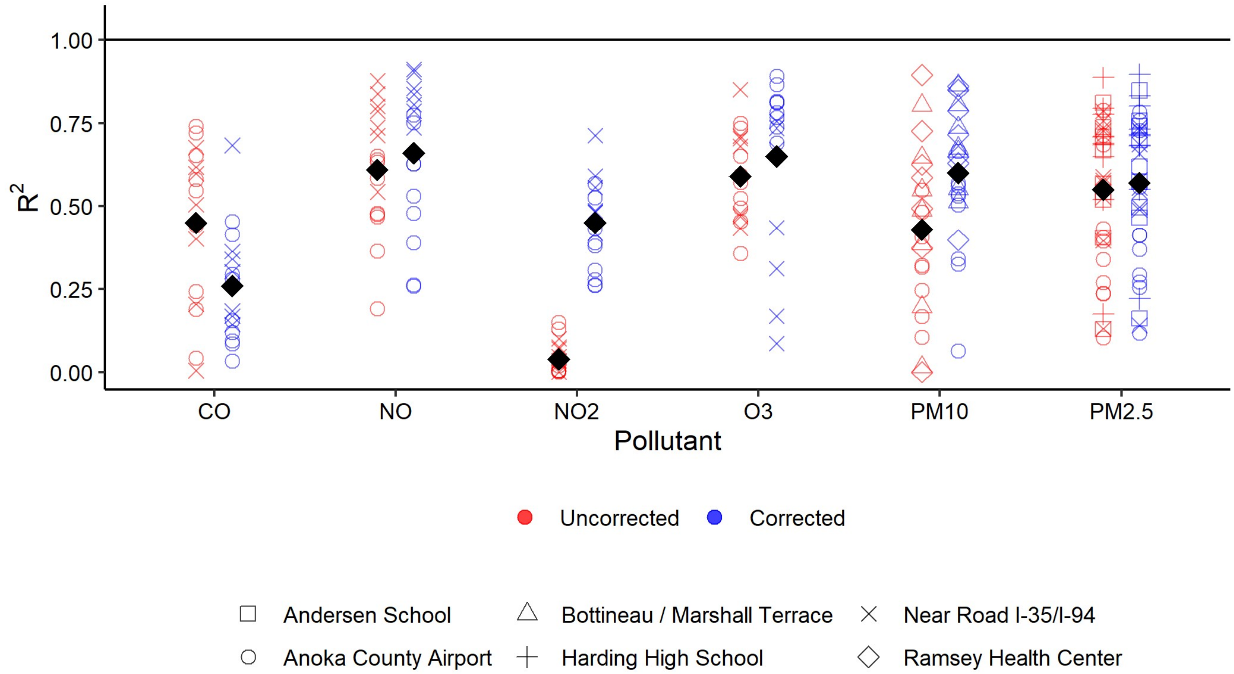

Figure 11 R-Squared of Neural Network correction for sensors compared to a federal equivalent. Black dots represent the mean of the measurements in the column.

The Neural Network regression correction made improvements to the mean R^2 across all types of sensors except CO, which showed a decrease in the average. The increases were relatively small for NO, O3, and PM2.5. There were larger changes for NO2 and PM10; both saw an increased correlation with the FEM after the correction was applied.

RSE For All Sensors and Sites

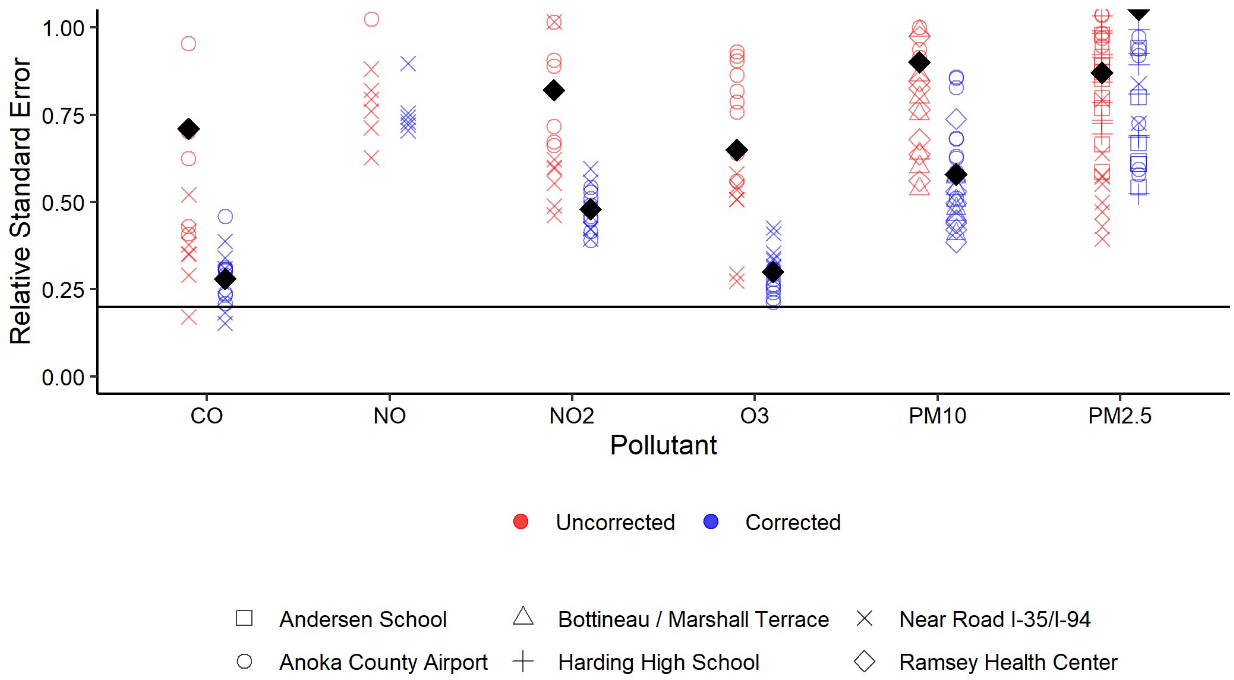

Figure 12 Relative Standard Error of sensors after a Neural Network correction is applied, with black markers for the average. The target to reach is zero, but an RSE of under 0.2 is a goal for many low-cost sensors.

The Neural Network corrections effect on the Relative Standard Error of the sensors in the suite is generally positive; however, only CO and O3 approach the target of 0.2. Many of the sensors have a more significant proportion of error in their monthly measurement than the proportion of measurement. Specifically, PM2.5 and NO datapoints are leaving the boundary of the chart.

Neural Network Correction Staying Power

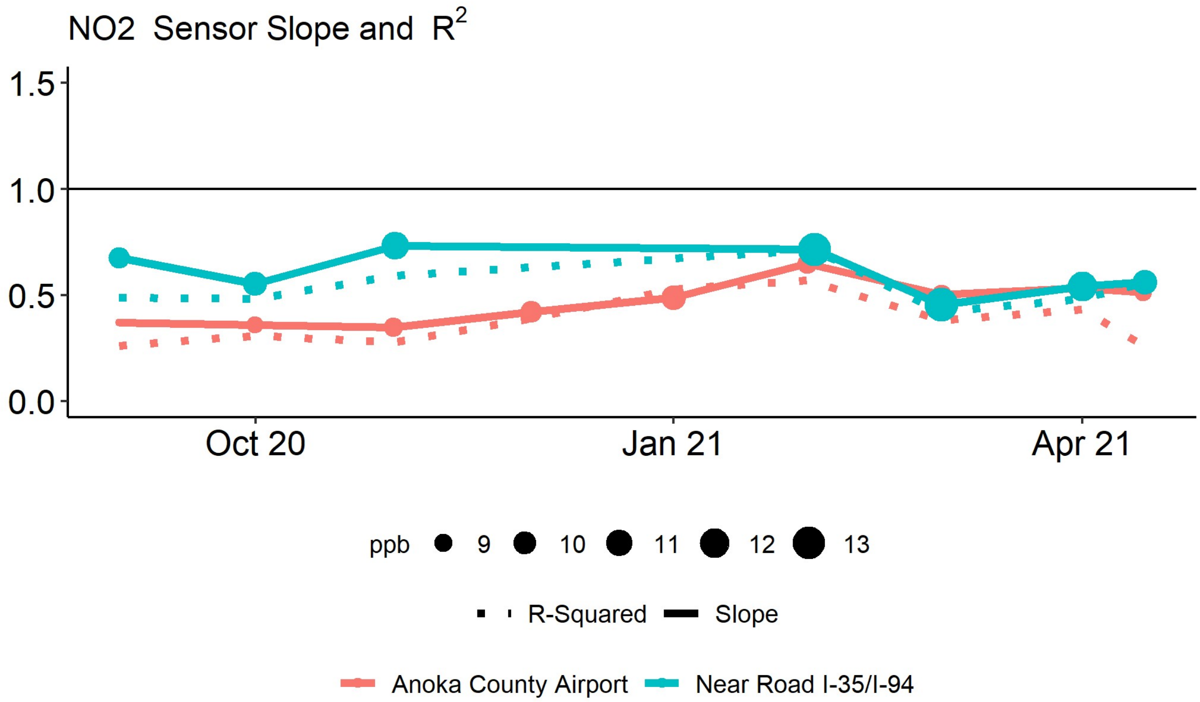

Figure 13 NO2 sensor slope, R^2, and average concentration over time at the 35W site and the Anoka Site using NN correction.

The Neural Network correction for the NO2 electrochemical sensor shows signs of being stable after implementation. The slope and R^2 are similar for both sites and stay at a relatively consistent level throughout the nine months that the correction is active. This happens while the average monthly concentration changes.

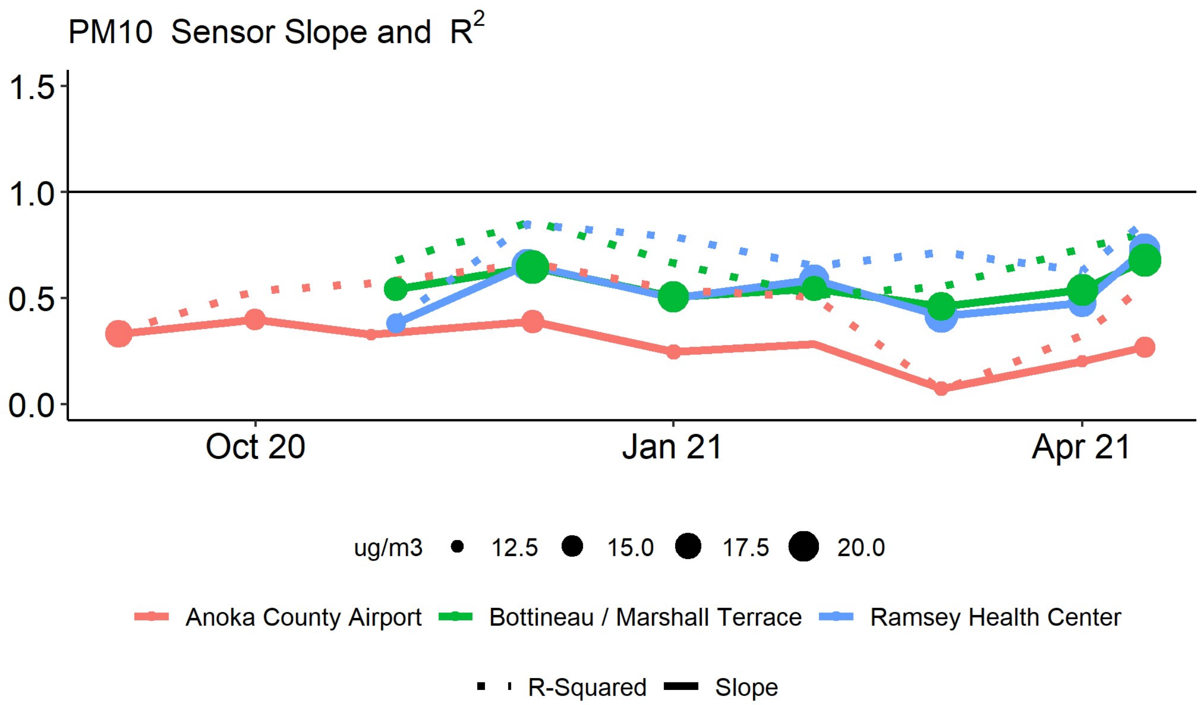

Figure 14 PM10 sensor slope, R^2, and average concentration over time at the Ramsey Health Center (RHC) site, the Bottineau Marshall Terrace site, and the Anoka Site using NN correction.

The optical sensor correction using the Neural Network predictor shows signs of being stable in places where the average concentration is higher. Both the RHC site (Green) and the Bottineau site (Blue), where there is a higher concentration of PM10, perform better than the Anoka site (Pink). The Anoka site also seems to perform better when there is a higher concentration of PM10.

Generality of Predictors

The predictors were trained on data from multiple different sensors at multiple different sites. The predictors performed well regardless of which site they were at; this hints at a possible generality of the predictors to be used across the network, regardless of where they are placed. Implementing this would be a massive improvement over the current state of the network. Uncorrected, it leaves questions about the accuracy of the data coming from the network using the standard linear regression corrections.

Future Work

Most future work involves taking the lessons learned about how these predictors work in Minnesota and expanding on their uses in the low-cost air quality network. Correcting the pods at the FEM sites is excellent, but expanding those corrections out across the whole network would be a massive benefit for the regulatory agencies and the general public.

Gold Pod technique RF

The gold pod technique is a technique developed by AQ-Mesh to provide corrections around the network. In this technique, a pod is collocated next to a FEM site for a period of time; corrections are made based on that data; this establishes the pod as a 'gold' pod. Then that gold pod is moved to be collocated with another low-cost pod, not near a FEM site. The pods are compared, and the low-cost pod is corrected to match the gold pod.

This technique is suited for a random forest type correction; the 'gold' pod needs to be as accurate as possible; the RF correction was the most promising out of the corrections attempted. Future deployments of the MPCA network would probably benefit from a system of gold pod corrections.

The gold pod technique could also help prove the applicability of using the predictors across the network, giving hints about the generality of the predictors.

Long Distance Scaling NN

Long-distance scaling is another technique that many are interested in that could be served by one of these corrections. The basic idea is to take the lowest quartile of data and to set that as a regional background level. When the FEM sites have a similar reading in their lowest quartile, then those two readings are assumed to be measuring the same concentration. This data filtering is artificially synchronizing low readings of data. The range of this data is much below and local spikes or readings so the full range of measurements can't be compared. The neural network correction, being parametric, can predict higher readings, though. So, a future project to look at may be working the neural network correction into a long-distance scaling technique, reducing the maintenance and movement required compared to the current state of the art, the gold pod technique.

Conclusion

The two supervised machine learning techniques used improve the accuracy of many of the low-cost sensors used on the AQ-Mesh suite. The Random Forest technique seems the most promising of the techniques to implement in a way that is familiar to the field. The Neural Network technique shows some promise, but it is not as accurate as the Random Forest predictor. Groups thinking of deploying sensor networks should consider applying supervised machine learning corrections to data they have gathered to improve the accuracy of the data gathered.

Works Cited

Air Quality Sensor Performance Evaluation Center. (2021). *Field Evaluation AQ-Mesh v5.1-Gas.* Diamond Bar: Air Quality Sensor Performance Evaluation Center. Retrieved from http://www.aqmd.gov/docs/default-source/aq-spec/field-evaluations/aqmesh-v5-1_gas---field-evaluation.pdf?sfvrsn=20

Andrea Polidoi, V. P. (2016). *Laboratory evaluation of low-cost air quality sensors.* Diamondbar, CA, USA: South Coast Air Quality Management District.

Aphasense ltd. (2019, July). OX-A431 Oxidising Gas Sensor Ozone + Nitrogen Dioxide 4 Electrode. Essex, United Kingdom. Retrieved from https://www.alphasense.com/wp-content/uploads/2019/09/OX-A431.pdf

AQ-Mesh. (2021, December 13). AQ-Mesh Technical Specification Gas algorithim V5.3 PM algorithm V3.0h.

Brieman, L. (2001). Random Forests. *Machine Learning*(45), 5-32. doi:10.1023/A:1010933404324

Carl Malings, R. T. (2020). Fine particle mass monitoring with low-cost sensors: Corrections and long-term performance evaluation. *Aerosol Sience and Technology, 54*(2), 160-174. doi:10.1080/02786826.2019.1623863

E. Esposito, S. D. (2016). Dynamic neural network architectures for on field stochastic calibration of indicative low cost air quality sensing systems. *Sensors and Actuators B: Chemical, 231*, 701-713. doi:10.1016/j.snb.2016.03.038

Jos Lelieveld, A. P. (2020, September 1). Loss of life expectancy from air pollution compared to other risk factors: a worldwide perspective. *Cardiovascular Research, 116*(11), 1910-1917. doi:10.1093/cvr/cvaa025

N. Zimmerman, A. A. (2018). A machine learning calibration model using random forests to improve sensor performance for lower-cost air quality monitoring. *Atmospheric Measurment Techniques, 11*(1), 291-313. doi:10.5194/amt-11-291-2018

Wada, K. (2020, December). Outliers in official statistics. *Jpn J Stat Data Sci, 3*, 669-691. doi:https://doi.org/10.1007/s42081-020-00091-y