Atmospheric Environment 184 (2018) 9–16

Contents lists available at ScienceDirect

Atmospheric Environment

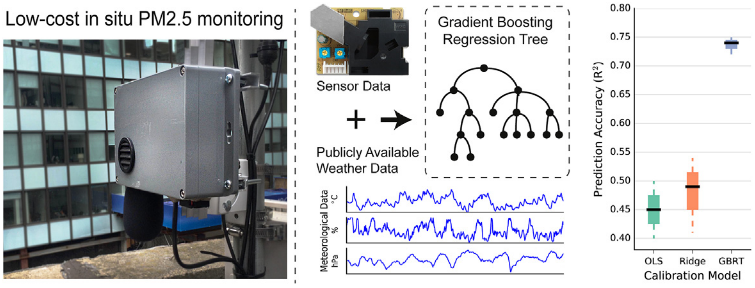
journal homepage: www.elsevier.com/locate/atmosenv

Using a gradient boosting model to improve the performance of low-cost aerosol monitors in a dense, heterogeneous urban environment

Nicholas E. Johnson[a,b], Bartosz Bonczak[b], Constantine E. Kontokosta[b,c,*]

[a] University of Warwick, United Kingdom
[b] Center for Urban Science and Progress, New York University, United States
[c] Department of Civil and Urban Engineering, New York University, United States

GRAPHICAL ABSTRACT



ARTICLE INFO

Keywords:
Machine learning
Low-cost sensing
Air quality
Urban
Calibration

ABSTRACT

The increased availability and improved quality of new sensing technologies have catalyzed a growing body of research to evaluate and leverage these tools in order to quantify and describe urban environments. Air quality, in particular, has received greater attention because of the well-established links to serious respiratory illnesses and the unprecedented levels of air pollution in developed and developing countries and cities around the world. Though numerous laboratory and field evaluation studies have begun to explore the use and potential of low-cost air quality monitoring devices, the performance and stability of these tools has not been adequately evaluated in complex urban environments, and further research is needed. In this study, we present the design of a low-cost air quality monitoring platform based on the Shinyei PPD42 aerosol monitor and examine the suitability of the sensor for deployment in a dense heterogeneous urban environment. We assess the sensor's performance during a field calibration campaign from February 7th to March 25th 2017 with a reference instrument in New York City, and present a novel calibration approach using a machine learning method that incorporates publicly available meteorological data in order to improve overall sensor performance. We find that while the PPD42 performs well in relation to the reference instrument using linear regression (R^2 = 0.36–0.51), a gradient boosting regression tree model can significantly improve device calibration (R^2 = 0.68–0.76). We discuss the sensor's performance and reliability when deployed in a dense, heterogeneous urban environment during a period of significant variation in weather conditions, and important considerations when using machine learning techniques to improve the performance of low-cost air quality monitors.

[*] Corresponding author. Center for Urban Science and Progress, New York University, United States.
E-mail address: ckontokosta@nyu.edu (C.E. Kontokosta).

https://doi.org/10.1016/j.atmosenv.2018.04.019
Received 21 August 2017; Received in revised form 9 April 2018; Accepted 10 April 2018
Available online 13 April 2018

1. Introduction

Air quality is an important quality of life concern with well-established links to serious respiratory illnesses, cardiovascular disease, and increased mortality rates (Pope III and Dockery, 2006). Cities in particular often experience high levels of fine particulate matter (PM2.5), especially in developing countries where industrial expansion has created unprecedented levels of poor air quality (Cheng et al., 2016). In order to monitor and evaluate levels of PM2.5, government agencies often operate air quality monitoring stations that provide ambient PM2.5 concentration measurements. These networks, however, often fail to capture the granular spatiotemporal variations in PM2.5 levels that can occur over short distances (<1 km) (Castell et al., 2017). Urban environments, in particular, contain widely varying mixing ratios with diverse and complex emission sources that require high resolution spatial and temporal monitoring networks to adequately quantify and describe air quality (Mead et al., 2013).

The proliferation of low-cost sensor technologies offers new opportunities to monitor and study air quality in urban environments. A growing body of research has begun to use low-cost aerosol monitors to provide high resolution spatiotemporal measurements by creating dense spatial networks that can inform local and regional emission sources' contribution to total pollution levels, as well as increase the ability to identify pollution hot-spots (Heimann et al., 2015; Jerrett et al., 2005; Shusterman et al., 2016; Manikonda et al., 2016; Moltchanov et al., 2015). Furthermore, these low-cost technologies are often compact, low-powered, and easy to operate, thus offering the ability to establish and facilitate participatory networks (Jovašević-Stojanović et al., 2015; Snyder et al., 2013). High density air quality monitoring networks enable community-based feedback loops that can be used to both protect those individuals susceptible to poor air quality and identify specific causes of particulate matter pollution.

While low-cost devices offer new opportunities for large-scale air quality monitoring, there are several important limitations to be considered. Central to the issue of using low-cost devices is ensuring data quality (Snyder et al., 2013; Kumar et al., 2015). Though federal, state and local monitoring devices operate at significantly higher costs, they also operate under standard procedures for calibration, data collection, and data post-processing methods, which ensure consistency across devices. In contrast, low-cost devices often suffer from a lack of manufacturer information about the specific operation and limitations of the device, as well as employ simplistic sampling techniques that fundamentally inhibit the device's performance ability. Furthermore, low-cost sensors often require individual and frequent calibration, which involves regular access to expensive equipment and expertise, and can be impractical for a large-scale deployments. To address many of these challenges, a number of studies have evaluated multivariate calibration using machine learning techniques (De Vito et al., 2018; Fishbain and Moreno-Centeno, 2016).

In this study, we present the design of a low-cost air quality monitoring platform based on the Shinyei PPD42 aerosol monitor and examine the suitability of the sensor for deployment in a dense spatial network configuration. We assess the sensor's performance during a field calibration campaign from February 7th to March 25th 2017 with a reference instrument in New York City and present a novel calibration approach using a machine learning method that incorporates publicly available meteorological data in order to improve the sensor's performance.

This work is a part of a long-term study, the *Quantified Community*, aimed to understand neighborhood-scale interactions between the environment and man-made infrastructure and their effects on individuals and communities (Kontokosta, 2016). To understand this complex interaction, we aim to leverage low-cost technologies to create a dense sensor network in neighborhoods throughout New York City that provides real-time and granular spatiotemporal environmental data. The air quality monitoring platform described in this work is one aspect of a multi-sensor platform being developed.

2. Materials and methods

2.1. Node design

The *Quantified Community* sensor platform was developed using commodity hardware and designed to capture environmental parameters including fine particulate matter, ambient noise level, air temperature, relative humidity and luminosity. To achieve a high density monitoring network, the selection of sensors and platform hardware required careful consideration in order to find a balance between performance, reliability, accuracy, cost and scalability. Our sensor platform is designed to be deployed in a variety of urban environments, including dense, high-rise neighborhoods with comprehensive digital infrastructure to low density, economically disadvantaged communities with incomplete access to power and wireless network connectivity.

The Shinyei PPD42 was selected to measure PM2.5 because of its low cost, ease of use, and performance capability demonstrated in previous work (Holstius et al., 2014; Gao et al., 2015; Kelly et al., 2017; Austin et al., 2015; Jovašević-Stojanović et al., 2015; Wang et al., 2015). The PPD42 uses a light scattering technique to estimate particle concentration and is capable of measuring particles greater than 1 μm in diameter. Particles pass through a lighting chamber where the combination of a light emitter and photodiode detector measure the amount of light scattered by particles passing through the chamber. A 0.25 W thermal resistor, located at the bottom of the sensing chamber, increases the air temperature inside the chamber relative to the surrounding outside air temperature to create an updraft that draws particles into and through the chamber.

The PPD42 generates two output signals in the form of digital pulses that are referred to by the manufacturer as Low Pulse Occupancy (LPO) and are proportional to particle count concentration. In order to distinguish particle size, output P1 is used to measure particles greater than 1 μm and output P2 is used to measure particles greater than 2.5 μm. Particles with a diameter between 1 μm and 2.5 μm are determined by subtracting P2 from P1. The PPD42 outputs are connected to the interrupt points (INT0 and INT1) of an Atmega microcontroller in order to accurately capture pulses that range from 10 to 90 ms in length. The raw sensor output is converted into LPO readings and sent to a Raspberry Pi microcontroller via USB every 10 s to be stored locally. Though the Raspberry Pi is capable of transmitting the data to a central server for real-time processing, there was no available Wi-Fi connectivity in the study area.

A factory calibrated Bosch SHT31 sensor was used to measure air temperature and relative humidity with an accuracy of ±0.3 °C and ±2% relative humidity. The electronics were contained in a 6″ × 4″ × 2″ gray ABS plastic enclosure with a 5VDC fan attached to the bottom in order to draw air into the enclosure through a 1 1/2″ filtered vent. Based on the manufacturer specifications, we estimate complete air exchange inside the enclosure occurs approximately three times per second.

The PPD42 sensor used in this study cost approximately $15USD. Additional sensors, the microcontroller platform, and enclosure materials added an additional $80 USD resulting in an overall cost of approximately $100 USD, which is orders of magnitude less than reference instruments operated by state and federal agencies.

2.2. Reference instrument

The reference instrument for this study was a Thermo Scientific tapered element oscillating microbalance (TEOM) 1400 that provides continuous PM2.5 mass measurements at hourly intervals. TEOM instruments employ a size selective inlet that accumulates particles on a sampling filter located atop an oscillating element whose resonant frequency changes proportionally to particle mass (Kulkarni et al., 2011; Amaral et al., 2015). The device is owned and operated by the New York State Department of Environmental Conservation (NYS DEC) and costs approximately $30,000. Data from the reference instrument were obtained directly from the Department of Environmental Conservation.[1] It was observed that the data contained 32 observations with negative values due to the processing procedure performed by the NYS DEC; these measurements were subsequently removed from the analysis.

2.3. Study location

The study site was located at an elementary school (PS 104) rooftop on Division Street in Lower Manhattan. The location is a dense urban area with varying infrastructure comprised of approximately 11% commercial buildings, 10% residential buildings, 22% mixed residential and commercial and 2% industrial buildings within 1000 m, based on information from NYC's Primary Land Use Tax Output (PLUTO) database. Table S3 provides a description of the surrounding characteristics. Of important note, the site is located less than 50 m from the Manhattan Bridge with an average of 115,000 vehicles crossing every day (New York State Department of Transportation, 2017). The study area also contains approximately 56 buildings that use oil boiler systems, which are known to be significant sources of particulate matter in New York City (Clougherty et al., 2013; Jain et al., 2014).

The individual nodes were fixed on a custom mounting platform at a height of approximately 1.5 m above the rooftop (approximately 12 m from ground level) and 3 m from the rooftop edge. The design of the mounting platform positioned two devices facing east towards the Manhattan bridge and one device facing west away from the bridge. The devices were located approximately 5 m from the intake of the reference instrument due to logistical reasons.

2.4. PPD42 performance evaluation

An initial evaluation of the PPD42 was conducted to assess the accuracy and precision of the three individual deployed devices. Raw LPO readings were aggregated to an hourly average in order to match data from the reference monitor, and pairwise plots were used to compare individual sensor responses with the reference monitor. To evaluate the linear relationship between individual devices and the reference monitor, an Ordinary Least Squares (OLS) regression was performed on the matched hourly data and the coefficient of determination (R^2) and the root mean squared error (RMSE) values were used to evaluate the strength and accuracy of the relationship. In this study, measurements from the TEOM monitor are used as the dependent variable and measurements from the PPD42 are the independent variable.

A sensitivity analysis was performed using multiple meteorological parameters to determine their potential influence on sensor measurements. The coefficient of determination was used to evaluate the strength of the relationship between meteorological parameters (independent variables) and the PPD42 and TEOM measurements (dependent variables). Temperature and humidity measurements were taken directly from individual sensor platforms using the SHT31 sensor located inside the enclosure directly adjacent to the PPD42. Other meteorological parameters were also assessed including barometric pressure, wind speed, dew point, and precipitation. These measurements were obtained from a nearby weather station located at La Guardia airport. Fig. 1 shows the meteorological conditions during the study period.

In order to determine the device's sensitivity in low concentration environments, the lower limit of detection was calculated as:

$$LOD = 3\sigma_{blk} * \beta_1$$

where σ_{blk} is the standard deviation of the PPD42 measurements obtained when TEOM measurements were below 5.0 μg/m^3, 3.0 μg/m^3 and 1.0 μg/m^3, and β_1 is the slope of the line obtained from the OLS regression analysis. We include multiple calculations of the LOD in order to provide statistically significant results given the small number of samples from the TEOM below 1.0 μg/m^3 (14 samples). This approach was established by Kaiser and Specker (1956) and also used in similar studies (Austin et al., 2015; Wang et al., 2015; Kelly et al., 2017).

2.5. Calibration approaches

Three statistical approaches were evaluated to determine the best-fit calibration model. All three models were based on measurements from the individual sensor platforms, as well as meteorological data that included air temperature, relative humidity, barometric pressure, dew point, and precipitation. As noted in previous work, the PPD42's response is non-linear across the entire range of the device and therefore a quadratic term was also included into the model (Gao et al., 2015; Austin et al., 2015; Wang et al., 2015). A final parameter was added to account for the time of day based on an analysis of diurnal readings from the PPD42 devices, which showed a 1.5 standard deviation difference between the reference instrument during the afternoon hours from 10:00–15:00 (Fig. S1). This difference is likely caused by solar radiation affecting the sensor's optics and the inclusion of a time parameter is intended to capture this phenomenon. R^2 and RMSE were used to compare calibration accuracy.

The first calibration method used a standard multiple linear regression model in the form of:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \varepsilon$$

where y is the reference instrument values, β_0 is the intercept, $x_1 \ldots x_p$ are the predictors including the PPD42 measurements, and ε is the error term. The model was specified using best-subset selection, which iteratively finds the combination of features that result in the greatest reduction in the residual sum of squares for each subset of size k where $k = p - 1 \ldots p$. The single best model from $M_0 \ldots M_k$ was chosen based on Bayesian Information Criterion scores. To detect and account for multicollinearity between variables, the variance inflation factor (VIF) was calculated for all features, and the feature with the highest score was removed. This process was performed recursively until all features' VIF scores were below the threshold of five. The final model included only statistically significant features.

The second calibration technique used a regularization method to address some of the problems with least squares regression. Regularization adds a penalty term (λ) to large model coefficients in order to reduce multicollinearity between features. The Ridge regression model used here applies an $\ell 2$ penalty to the sum of the squared coefficients. Ridge coefficients ($\hat{\beta}^R$) are values that minimize:

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

where λ controls the amount of penalization. The λ parameter was determined through a five-fold cross validation and set to 0.4. In order to evaluate the significance of individual features, we rank each feature based on the absolute value of the coefficient (β_j). The larger the coefficient, the larger the impact on the model and hence the greater significance of the feature.

The final calibration approach used a gradient boosting regression tree (GBRT) model. GBRT is a decision tree-based regression model that implements boosting to improve model performance. Boosting is a statistical technique that sequentially builds many 'weak' models (learners) that are combined into a final consensus model (Schapire,

[1] www.dec.ny.org.

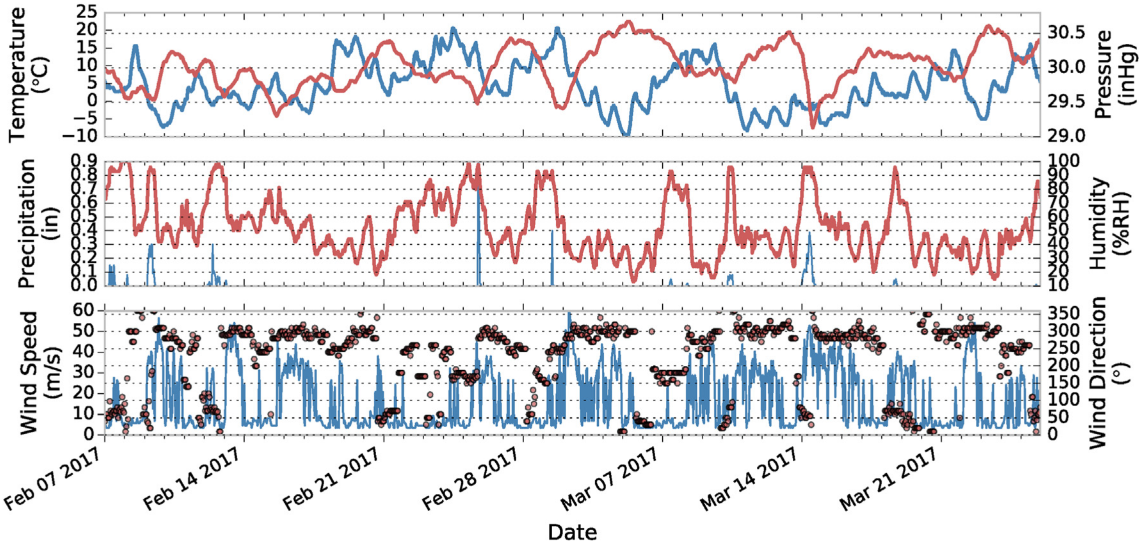


Fig. 1. Meteorological measurements taken from La Guardia airport over the study period. (a) Temperature (blue) and sea level pressure (red), (b) precipitation (blue) and humidity (red line), and (c) wind speed (blue line) and wind direction (red points). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

2003). A ‘weak’ learner is one whose performance is only slightly better than random guessing. The final model is built in an additive forward stagewise manner where at each step a new learner is added that minimizes the negative gradient by least squares. The residuals of the current model are then used as the input for the next tree allowing the model to ’learn' from the errors of the previous models (Friedman et al., 2001).

Parameter tuning is an important element to optimize the GBRT model performance. Tree-specific parameters include the depth of each tree, the minimum number of samples to form a terminal node (leaf), and the maximum number of features included in each tree. Boosting parameters include the number of trees used in the model and the contribution of each tree to the final model (learning rate). Tree depth, the number of trees, and the maximum number of features in each tree control the degree of interaction between features. Since trees are grown sequentially, a large number of shallow trees is preferred in order to fully explore the feature space, at the expense of computation time. The learning rate and the minimum number of samples per leaf are used to control overfitting. A low learning rate is generally preferred, but will require a larger number of trees to maintain performance.

To build the ridge and GBRT models, data were first randomly split into train (80%) and test (20%) sets. The training set was used to evaluate model parameters through an exhaustive grid search with 5-fold cross-validation and the final model was evaluated on the test set. All three models were implemented using the scikit-learn package for Python (Pedregosa et al., 2011).

3. Results and discussion

All three platform nodes collected data continuously throughout the 47-day study period with the exception of four days in which all three devices experienced a power outage. Fig. 2 shows pairwise plots from the co-located PPD42 devices. A total of 1128 hourly observations were recorded from all three devices. Hourly PM2.5 measurements from the TEOM ranged from 1 μg/m^3 to 28.1 μg/m^3 with an average of 7.8 μg/m^3.

Fig. 3 shows a scatter plot of the linear fit model between the TEOM and PPD42 devices. Based on the calculated R^2 values, individual PPD42 devices demonstrate a moderate level of agreement compared to the TEOM with R^2 values of 0.48 and 0.53 for two devices and the third device slightly lower at 0.37. These results are similar to previous work by Holstius et al. (2014) who conducted an eight-day field calibration campaign at a regulatory site in Oakland, California and found that a linear correlation was sufficient to explain 55–60% of the variance (RMSE = 3.4–3.6) in the federal equivalent method instrument at a one hour interval and 72% at a 24 h interval. Kelly et al. (2017) also found moderate correlation (R^2 = 0.59–0.8) between the PPD42 and a commercial grade optical device (TSI DustTrak II Model 8532) during ambient wind tunnel tests, and Gao et al. (2015) found similar correlations (R^2 = 0.53) with 24 h gravimetric measurements during a four-day calibration campaign in Xi'an, China. Gao et al. (2015), however, also observed significantly higher hourly correlations (R^2 = 0.87–0.88) with the DustTrak instrument and suggest the higher correlation is likely due to the increased levels of PM2.5 concentrations observed in Xi'an (range: 77–889μg/m^3) compared to Holstius et al. (2014) (range: 0.3–30μg/m^3) since the PPD42's measurement errors increase at lower concentration levels.

Individual PPD42 devices show high correlation with R^2 values of 0.93–0.96 and a linear response across the concentration range (Fig. S3). This high correlation between PPD42 devices has been largely consistent across studies by Holstius et al. (2014), Gao et al. (2015) and Kelly et al. (2017), who all report high inter-device correlations (R^2> 0.9) with the exception of one experiment by Kelly et al. (2017) reporting a correlation of R^2 = 0.72.

3.1. Ambient conditions

The average temperature during the study period was 4.5 °C (range: −10.0-20.6 °C) with an average humidity of 52% (range: 0–100%). Rapid fluctuations in meteorological conditions were observed throughout the study period. For example, the average temperature during the week of February 9th-17th was 0.8 °C (range: −7.2-8.2 °C) and increased significantly to an average temperature of 10.7 °C (range: 1.7–20.6 °C) the following week. Other extreme weather conditions were also observed including 20 days with high winds (>30 m/s), three separate snow days with a total accumulation of five inches and two days with freezing rain. The observed ranges in temperature, humidity, and precipitation are significantly greater than those of previous field calibration studies.

Table S1 shows the sensitivity test results. Dew point temperature measurements show the highest correlation between both the PPD42 and the TEOM (R^2 = 0.38–0.41 and R^2 = 0.18) compared to other meteorological parameters. Temperature and relative humidity are

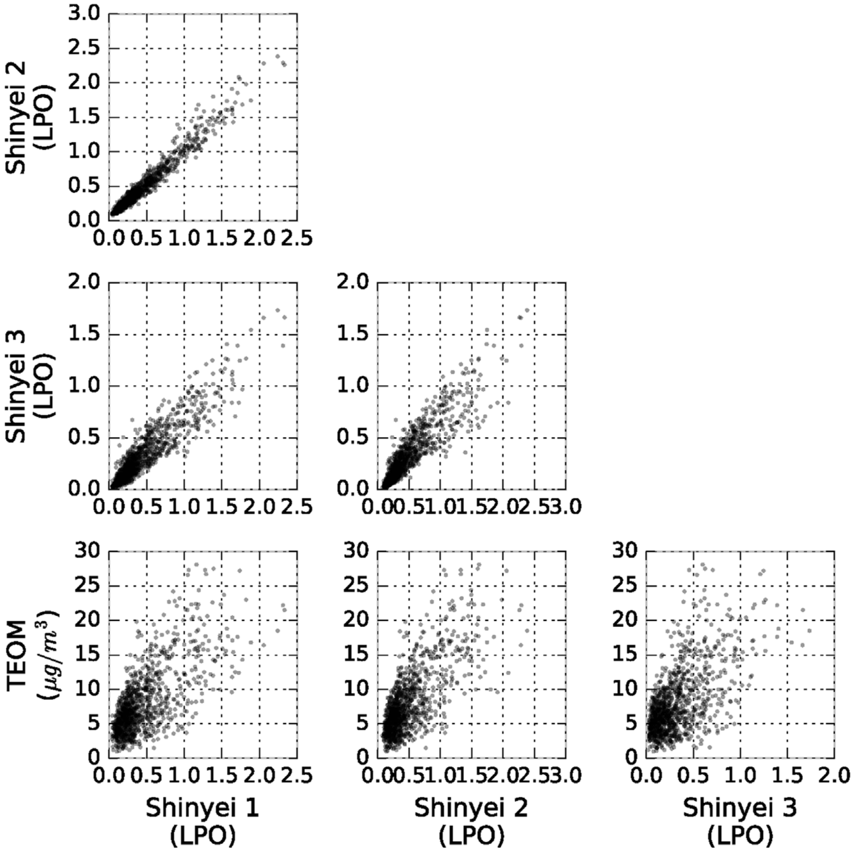


Fig. 2. Pairwise plots between three Shinyei PPD42 devices and a reference TEOM based on hourly data collected from February 7th 2017 to March 25th 2017.

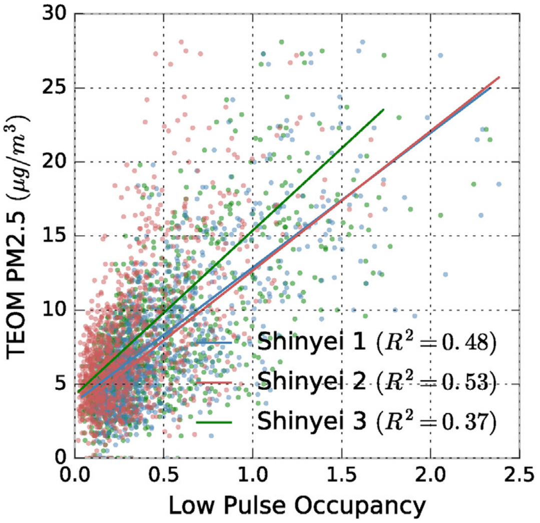


Fig. 3. Linear model fit for hourly data collected from three Shinyei PPD42 sensors and a NYS DEC reference monitor between February 7th 2017 and March 25th, 2017.

both weakly correlated (R^2 = 0.24–0.25 and R^2 = 0.13–0.19) with the PPD42 measurements, and show only minor influence on the TEOM (R^2 = 0.15). Previous work by Holstius et al. (2014) evaluated the affect of temperature, relative humidity and light levels on PPD42 measurements and found only relative humidity had a minor correlation (R^2 = 0.25–0.28). While we observe the effect of relative humidity to be slightly lower and the effect of temperature to be significantly higher than findings by Holstius et al. (2014), it should be noted that the meteorological conditions during the Holstius et al. (2014) study varied significantly from this study with temperatures ranging from 20 to 30 °C and relative humidity ranging between 10 and 60%. Gao et al. (2015) also found that temperature and relative humidity effects were significant, noting the differences in meteorological conditions between their work and findings by Holstius et al. (2014).

Differences between these studies may be explained by the convective technique used to create air flow through the sensing chamber. Since the convective flow generated by the resistor is proportional to the surrounding air temperature, fluctuations in ambient temperature will have a direct effect on the sensor's ability to draw particles through the sensing chamber. As observed in this study, and noted by Gao et al. (2015) and Kelly et al. (2017), cooler ambient temperatures will more significantly affect the PPD42 measurements than higher ambient temperatures. Furthermore, Kelly et al. (2017) also compare the PPD42 with a similar optical aerosol monitor, the Plantower PMS3003, and suggest that the improved performance of the PMS3003 may be due to the use of a fan to control air flow through the sensing chamber.

3.2. *Limit of detection*

Table 1 shows results for the PPD42's lower limit of detection. The average LOD is 4.83μg/m^3 for concentrations below 5.0 μg/m^3 (323

Table 1
Results from calculating the lower limit of detection for the PPD42 during a field calibration campaign with a TEOM reference instrument. Units are in μg/m^3.

Concentration	Sample Size	Shinyei 1	Shinyei 2	Shinyei 3	TEOM
< 1 μg/m^3	14	3.34	2.90	2.30	0.79
< 3 μg/m^3	90	3.35	3.30	4.45	2.75
< 5 μg/m^3	323	4.82	4.65	5.12	3.37

Table 2
Comparison of results from three calibration techniques.

Parameter	OLS				Ridge				GBRT			
	R^2	RMSE	β_0	Slope	R^2	RMSE	β_0	Slope	R^2	RMSE	β_0	Slope
Shinyei 1	0.452	3.28	3.60	0.59	0.466	3.24	3.35	0.62	0.716	2.36	1.84	0.79
Shinyei 2	0.507	3.11	3.28	0.64	0.521	3.07	2.99	0.67	0.762	2.16	1.47	0.83
Shinyei 3	0.360	3.55	4.74	0.44	0.364	3.54	4.31	0.48	0.678	2.52	2.48	0.72

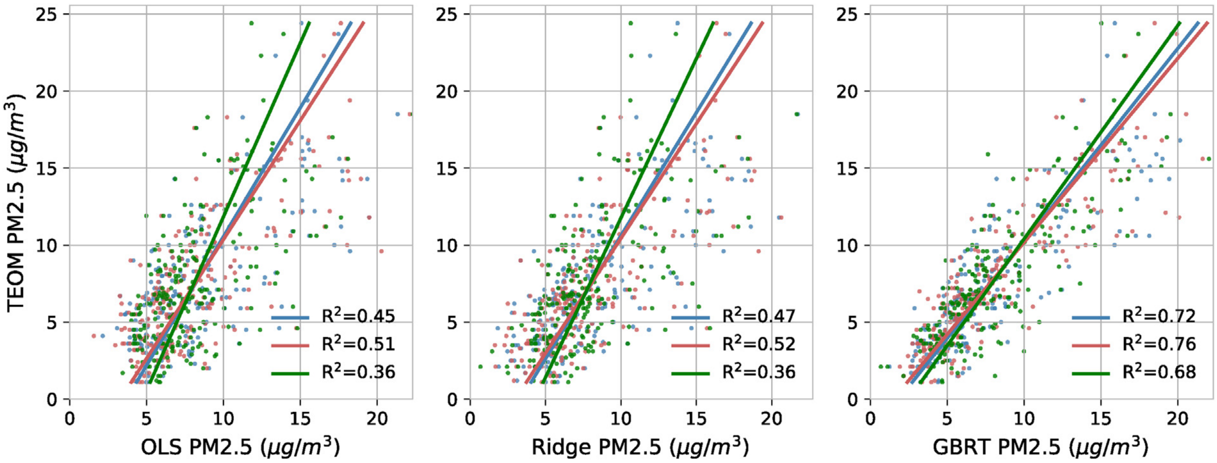


Fig. 4. Scatter plots of three Shinyei PPD42 sensors calibrated with three different techniques. Sensors are calibrated through a multi-linear regression, ridge regression and gradient boosting regression tree model.

samples), 3.6 μg/m^3 for concentrations below 3.0 μg/m^3 (90 samples) and 2.8 μg/m^3 for concentrations below 1.0 μg/m^3 (14 samples). These findings are in the range of laboratory tests performed by Austin et al. (2015) (1.0 μg/m^3) and Wang et al. (2015) (4.59μg/m^3 and 6.44μg/m^3).

3.3. Calibration results

Table 2 and Fig. 4 compare OLS, Ridge, and GBRT results from the hourly test data and show that the GBRT model significantly outperforms both the OLS and Ridge models with an average R^2 of 0.72. While it is expected that the more complex model will outperform other models, there are two observations that should be highlighted. First, the overall magnitude of improvement by the GBRT model is significant, increasing by approximately 20–30% over the Ridge model. Second, the GBRT model also reduces the range of scores between devices from 0.16 points in the Ridge model to 0.08 points in the GBRT model. This ability to reduce device variability is a significant enhancement for relative calibration and large-scale deployments.

Fig. 5 compares OLS, Ridge and GBRT calibrated hourly measurements. Overall, the OLS and Ridge models show similar R^2 values and track well against the TEOM monitor. However, results from the OLS and Ridge models periodically under- and over-estimate TEOM measurements. Significant under-estimates by the PPD42, for example, are observed on February 11th and February 16–19th, in which the TEOM instrument reported higher PM2.5 concentrations during both periods. Over-estimates are often found during the evening hours (e.g. Mar 9–12th) and are likely due to the low PM2.5 concentration levels that fall below the PPD42's lower limit of detection. The GBRT model, however, does not demonstrate the same under- and over-estimates observed in the OLS and Ridge models.

Fig. S2 compares feature importance between the Ridge model and GBRT model. The most significant features in the Ridge model are the PPD42 output, sea level pressure and the squared PPD42 sensor output, while the GBRT model identifies pressure, dew point, the PPD42 output and the squared PPD42 sensor output. These results also show that the Ridge model places greater weight on only a few parameters, while relative feature importance is distributed across features in the GBRT model. This is expected given that the GBRT model is a more robust model capable of learning complex relationships across a large set of input parameters. In this case, the model is able to better establish the relationship between sensor measurements and meteorological conditions to improve the calibration. Table S2 shows the complete OLS model results with computed significance values for each parameter for comparison.

3.4. Main findings

The aim of this study is to examine the viability of a low-cost air quality platform based on the PPD42 aerosol monitor to measure PM2.5 in a dense urban environment. Based on an extensive field calibration campaign, we find the PPD42 performs reasonably well throughout a variety of environmental conditions and can be a suitable device for measuring PM2.5, especially considering the difference in cost from other commercially-available instruments. The high correlation between PPD42 devices is particularly significant for high-density sensor networks that rely on relative measurements to inform the spatial distribution and variability of PM2.5 across a study area. Furthermore, while measurement errors increase at lower PM2.5 concentrations ($<$ 5 μg/m^3), the limit of detection falls below the range of ambient concentration levels expected in many urban environments. For example, New York City's average annual PM2.5 concentration level is 11.55μg/m^3 with a range of 5.17–26.48μg/m^3 (Matte et al., 2013).

An important consideration in evaluating acceptable detection limits is the specific application and use of the recorded particulate matter observations. Larger measurement errors from low-cost devices may still be acceptable to compare ambient PM2.5 levels between communities, identify local hot spots, and provide feedback to local residents. Furthermore, the temporal resolution offered by many low-cost devices, including the PPD42, can be useful in measuring transient

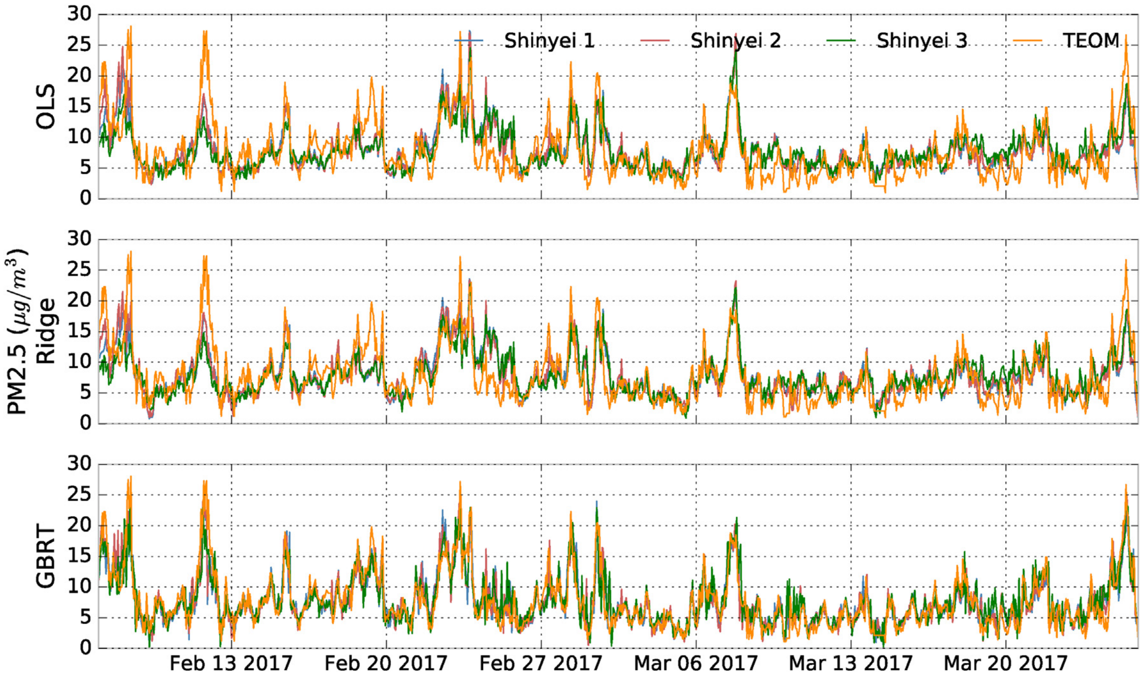


Fig. 5. Comparison of calibration results with a reference instrument using different calibration techniques including multiple linear regression, ridge regression and gradient boosting regression tree models. Hourly PM2.5 measurements were obtained from three Shinyei PPD42 sensors co-located with a TEOM reference instrument from February 7th through March 25th, 2017.

emission sources that may significantly exceed ambient concentration levels over short time periods.

Through comparing various calibration techniques, this study found that a GBRT model that uses publicly available meteorological data can significantly improve the performance of a low-cost aerosol monitor. While this calibration process does not necessarily establish an equivalence between the devices, it does provide a method for converting raw sensor readings into standard units (μg/m^3) and improving the sensor's performance by identifying meteorological conditions that cause measurement error and adjusting the sensor's response accordingly. Furthermore, the implementation of a machine learning model to calibrate low-cost instruments can be a step towards a universal calibration curve and standardized sensor deployments. A properly trained machine learning model could be publicly distributed and implemented in similar hardware deployments by citizen science communities and nonspecialists, which could reduce the need to calibrate devices individually, improve long-term device stability, and standardize data generation and collection methods.

3.5. Limitations

A significant limitation when using the PPD42 is the inability to explain measurement errors and variability between the PPD42 devices. This is largely a result of the optical sensing technique employed. Unlike other sampling techniques, the light scattering approach used by many low-cost aerosol monitors is unable to evaluate the physical properties of particles such as composition, type, mass, or optical characteristics. For example, organic particles tend to absorb moisture from the surrounding environment making them more susceptible to changes in humidity. Similarly, different particle types have different optical properties that can vary depending on the wavelength of light used in the sensor.

This work is also limited by the use and comparison of three sensing units, which limits a full evaluation of inter-device variation. Though our analysis is consistent with previous work showing high correlation (R^2 = 0.93–0.96) between PPD42 devices, a more robust statistical analysis that includes greater than 10 devices has yet to be performed. Similarly, while the calibration campaign does provide sufficient data to assess the sensor's performance in concentration ranges typical for New York City, these ranges may vary significantly in other urban areas around the world. To ensure accurate calibration, especially when using machine learning techniques, the devices should be exposed to the entire range of concentrations expected during deployment in order to include the training data necessary for the model to establish the proper input-response relationship. Furthermore, the study duration also limits an evaluation of long-term stability (>1yr) and time-in-use effects, such as the gradual accumulation of particles inside the sensing chamber, which may effect the sensor's optics.

There are also several important limitations to implementing machine learning algorithms for sensor calibration. One significant challenge is the potential to overfit the model to either the specific environment in which the calibration took place, or to the sample data used for the calibration. The latter is a general concern whenever using machine learning models and can be addressed with various techniques such as cross validation, as implemented in this analysis. Overfitting the calibration environment, however, can occur by incorporating parameters into the calibration model that are either specific to the calibration location, or do not include the full range of conditions that the sensor will be exposed to during deployment. It is essential that individual parameters contain sufficient variance to properly capture potential deployment conditions, while excluding any spatial parameters that could potentially affect the input stimulus (i.e PM2.5). During this study, for example, wind direction was observed to explain 10% of the variance of the TEOM monitor and the inclusion of this parameter in the GBRT model improved results on average by 5%. However, the affect of wind direction on PM2.5 in this specific location may result from variations in the built environment that potentially include PM sources (e.g buildings with specific boiler types), which will likely differ from deployment locations. Including wind direction would therefore train the calibration model based on the specific conditions of the study location instead of identifying the interaction of non-site specific variables that affect the PPD42. Similarly, the inclusion of a time-of-day parameter could led to erroneous calibration errors since diurnal PM2.5 trends may be affected by local emission sources that vary per location.

Furthermore, while a machine learning model can increase overall performance, it is unable to explain measurement error nor does it provide information about particle properties. Feature importance is one method to understand how the model is using features to make predictions and adjust the sensor response, but it does not necessarily describe the impact of certain meteorological parameters, or combinations of parameters, on the sensor's response.

4. Conclusion

This study demonstrates the suitability of a low-cost aerosol monitor to measure intra-urban PM2.5 concentrations. Over a 47-day study period, three PPD42 sensors, integrated with a Raspberry Pi microcontroller and Bosch SHT31 temperature and relative humidity sensor, were deployed on the roof of an approximately 12 m high building proximate to a TEOM instrument installed and operated by the NYS DEC. The devices were exposed to wide variations in ambient temperature, relative humidity, barometric pressure, and precipitation in an environment characterized by a diversity of urban land use types. Potential point sources of pollution included 56 surrounding buildings using oil boilers for heating and the vehicular traffic along the Manhattan Bridge.

We evaluate three machine learning methods to calibrate the deployed sensors, including traditional OLS regression, Ridge regression, and a GBRT decision tree model. Our results indicate that the GBRT model significantly outperforms the OLS and Ridge models. Overall, we find that low-cost aerosol devices can be used to inform community air quality monitoring efforts in heterogeneous urban environments. The GBRT calibration method provides superior performance when combined with meteorological data that can be used to convert raw sensor readings to standard units. Importantly, this machine learning approach can also be used to standardize readings across field-deployed sensors to improve relative performance and support citizen science and participatory sensing campaigns.

Acknowledgements

This research is supported in part by the UK Engineering and Physical Sciences Research Council (EPSRC) Center for Doctoral Training in Urban Science and Progress (EP/L016400/1). We would like to thank the New York State Department of Environmental Conservation for their assistance and expertise in conducting the study. We would also like to thank the reviewers for their insightful comments and feedback.

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.atmosenv.2018.04.019.

References

Amaral, S.S., de Carvalho, J.A., Costa, M.A.M., Pinheiro, C., 2015. An overview of particulate matter measurement instruments. Atmosphere 6 (9), 1327–1345.

Austin, E., Novosselov, I., Seto, E., Yost, M.G., 2015. Laboratory evaluation of the shinyei ppd42ns low-cost particulate matter sensor. PLoS One 10 (9), e0137789.

Castell, N., Dauge, F.R., Schneider, P., Vogt, M., Lerner, U., Fishbain, B., Broday, D., Bartonova, A., 2017. Can commercial low-cost sensor platforms contribute to air quality monitoring and exposure estimates? Environ. Int. 99, 293–302.

Cheng, Z., Luo, L., Wang, S., Wang, Y., Sharma, S., Shimadera, H., Wang, X., Bressi, M., de Miranda, R.M., Jiang, J., et al., 2016. Status and characteristics of ambient pm 2.5 pollution in global megacities. Environ. Int. 89, 212–221.

Clougherty, J.E., Kheirbek, I., Eisl, H.M., Ross, Z., Pezeshki, G., Gorczynski, J.E., Johnson, S., Markowitz, S., Kass, D., Matte, T., 2013. Intra-urban spatial variability in wintertime street-level concentrations of multiple combustion-related air pollutants: the New York city community air survey (nyccas). J. Expo. Sci. Environ. Epidemiol. 23 (3), 232.

De Vito, S., Esposito, E., Salvato, M., Popoola, O., Formisano, F., Jones, R., Di Francia, G., 2018. Calibrating chemical multisensory devices for real world applications: an in-depth comparison of quantitative machine learning approaches. Sensor. Actuator. B Chem. 255, 1191–1210.

Fishbain, B., Moreno-Centeno, E., 2016. Self calibrated wireless distributed environmental sensory networks. Sci. Rep. 6.

Friedman, J., Hastie, T., Tibshirani, R., 2001. The Elements of Statistical Learning, vol. 1 Springer series in statistics New York.

Gao, M., Cao, J., Seto, E., 2015. A distributed network of low-cost continuous reading sensors to measure spatiotemporal variations of pm2. 5 in xi'an, China. Environ. Pollut. 199, 56–65.

Heimann, I., Bright, V., McLeod, M., Mead, M., Popoola, O., Stewart, G., Jones, R., 2015. Source attribution of air pollution by spatial scale separation using high spatial density networks of low cost air quality sensors. Atmos. Environ. 113, 10–19.

Holstius, D.M., Pillarisetti, A., Smith, K., Seto, E., 2014. Field calibrations of a low-cost aerosol sensor at a regulatory monitoring site in California. Atmos. Meas. Tech. 7 (4), 1121–1131.

Jain, R.K., Moura, J.M., Kontokosta, C.E., 2014. Big data + big cities: graph signals of urban air pollution. IEEE Signal Process. Mag. 31 (5), 130–136.

Jerrett, M., Arain, A., Kanaroglou, P., Beckerman, B., Potoglou, D., Sahsuvaroglu, T., Morrison, J., Giovis, C., 2005. A review and evaluation of intraurban air pollution exposure models. J. Expo. Sci. Environ. Epidemiol. 15 (2), 185–204.

Jovašević-Stojanović, M., Bartonova, A., Topalović, D., Lazović, I., Pokrić, B., Ristovski, Z., 2015. On the use of small and cheaper sensors and devices for indicative citizen-based monitoring of respirable particulate matter. Environ. Pollut. 206, 696–704.

Kaiser, H., Specker, H., 1956. Bewertung und vergleich von analysenverfahren. Fresenius' J. Anal. Chem. 149 (1), 46–66.

Kelly, K., Whitaker, J., Petty, A., Widmer, C., Dybwad, A., Sleeth, D., Martin, R., Butterfield, A., 2017. Ambient and laboratory evaluation of a low-cost particulate matter sensor. Environ. Pollut. 221, 491–500.

Kontokosta, C.E., 2016. The quantified community and neighborhood labs: a framework for computational urban science and civic technology innovation. J. Urban Technol. 23 (4), 67–84.

Kulkarni, P., Baron, P.A., Willeke, K., 2011. Aerosol Measurement: Principles, Techniques, and Applications. John Wiley & Sons.

Kumar, P., Morawska, L., Martani, C., Biskos, G., Neophytou, M., Di Sabatino, S., Bell, M., Norford, L., Britter, R., 2015. The rise of low-cost sensing for managing air pollution in cities. Environ. Int. 75, 199–205.

Manikonda, A., Zíková, N., Hopke, P.K., Ferro, A.R., 2016. Laboratory assessment of low-cost pm monitors. J. Aerosol Sci. 102, 29–40.

Matte, T.D., Ross, Z., Kheirbek, I., Eisl, H., Johnson, S., Gorczynski, J.E., Kass, D., Markowitz, S., Pezeshki, G., Clougherty, J.E., 2013. Monitoring intraurban spatial patterns of multiple combustion air pollutants in New York city: design and implementation. J. Expo. Sci. Environ. Epidemiol. 23 (3), 223–231.

Mead, M., Popoola, O., Stewart, G., Landshoff, P., Calleja, M., Hayes, M., Baldovi, J., McLeod, M., Hodgson, T., Dicks, J., et al., 2013. The use of electrochemical sensors for monitoring urban air quality in low-cost, high-density networks. Atmos. Environ. 70, 186–203.

Moltchanov, S., Levy, I., Etzion, Y., Lerner, U., Broday, D.M., Fishbain, B., 2015. On the feasibility of measuring urban air pollution by wireless distributed sensor networks. Sci. Total Environ. 502, 537–547.

New York State Department of Transportation, 2017. Traffic Data Viewer.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.

Pope III, C.A., Dockery, D.W., 2006. Health effects of fine particulate air pollution: lines that connect. J. Air Waste Manag. Assoc. 56 (6), 709–742.

Schapire, R.E., 2003. The boosting approach to machine learning: an overview. In: Nonlinear Estimation and Classification. Springer, pp. 149–171.

Shusterman, A.A., Teige, V.E., Turner, A.J., Newman, C., Kim, J., Cohen, R.C., 2016. The berkeley atmospheric co 2 observation network: initial evaluation. Atmos. Chem. Phys. 16 (21), 13449–13463.

Snyder, E.G., Watkins, T.H., Solomon, P.A., Thoma, E.D., Williams, R.W., Hagler, G.S., Shelow, D., Hindin, D.A., Kilaru, V.J., Preuss, P.W., 2013. The Changing Paradigm of Air Pollution Monitoring.

Wang, Y., Li, J., Jing, H., Zhang, Q., Jiang, J., Biswas, P., 2015. Laboratory evaluation and calibration of three low-cost particle sensors for particulate matter measurement. Aerosol. Sci. Technol. 49 (11), 1063–1077.