Applying machine learning for large scale field calibration of low-cost $PM_{2.5}$ and PM_{10} air pollution sensors

Priscilla Adong [1] Engineer Bainomugisha [1] Deo Okure [1] Richard Sserunjogi [1]

[1]AirQo, Department of Computer Science, College of Computing and Information Sciences Makerere University

Introduction

- Ambient air pollution is a major environmental health risk in cities all over the world with harmful effects on human health and the ecosystem. It causes 4.2 million deaths per year.
- Ambient air quality data collection is done using reference grade monitors,e.g, the Beta Attenuation Monitor (BAM) which measures Particulate Matter (PM).
- They are highly accurate, but remain scarce in many cities in low & middle-income countries.
- Low-cost air quality monitors (LCAQMs) are increasingly being adopted as a complementary approach to fill the air quality data gaps while increasing spatial resolution of air quality data.
- We demonstrate the feasibility of using machine learning (ML) methods for large-scale calibration of AirQo low-cost PM sensors.

The low-cost sensor calibration challenge

LCAQMs are more error prone than reference grade monitors.

- **Their accuracy degrades over time**
- **They can be affected by external factors** such as weather changes
- **They suffer from cross-sensitivities between different ambient pollutants**

Sensor calibration is crucial for LCAQMs to ensure data quality and reliability

- This involves using appropriate statistical methods to correct measurements from low-cost sensors and validating against reference-grade monitors

In this research study, we used **AirQo LCAQMs** and investigated;

- **ML approaches for sensor calibration on a large scale air pollution network in urban environments with relatively high levels of particulate matter concentrations and variations**
- **The issues involved in deploying such ML-based calibration models to a production system**

Materials and Methods

Study Locations

We considered a real world air quality monitoring network with over 120 nodes deployed in cities with in Uganda. The experimental setup for the calibration data included two monitoring sites.

(a) Reference site 1 (b) Reference site 2

Figure 1. Monitoring sites used in this study. Part (a), shows AirQo devices and BAMs installed at Makerere University (`Reference site 1`), part (b), shows AirQo devices and BAM installed at Nakawa (`Reference site 2`)

Data collection and pre-processing

- PM data was collected using a total of 8 AirQo devices & 2 BAMs collocated at reference site 1 between 15th July 2020 & 17th July 2021 & reference site 2 from 30th Sept to 26th Oct 2021
- Met data(temperature & humidity) from the BAMs and from TAHMO stations was used.
- The average data completeness for all devices used in this study was approximately 87.61%.

Algorithm selection and validation

- We evaluated the performance of various ML algorithms for low-cost $PM_{2.5}$ and PM_{10} calibration.
- These included KNN, SVM, Multivariate Linear Regression, Multi-layer Perceptron, Randorm Forest (RF), XGBoost, ridge, lasso and elastic net regression.
- Performance of different algorithms was evaluated using the same training & validation datasets.
- Performance evaluation was done using the RMSE, MAE, R^2 and Pearson's correlation coefficient.

Input variable selection

- We selected the best variable combinations using variables including hourly $PM_{2.5}$ & PM_{10} from the low-cost sensor, atmospheric temperature(AT), RH, features derived from timestamp (month and hour(hr)), features from PM including $errorPM_{2.5}$, $errorPM_{10}$, $PM_{2.5} - PM_{10}$.

Algorithm validation methods

- **Cross unit validation:** We conducted performance evaluation for the proposed models using data from other AirQo devices within the same site.
- **Cross site validation:** We conducted performance evaluation for the proposed models using other AirQo devices collocated with the BAM at another reference site.

Algorithm selection

Best performance was achieved using variable combinations in equations 1 & 2 for PM calibration.

$$TargetPM_{2.5} = RF(PM_{2.5}, AT, RH, PM_{10}, errorPM_{2.5}, errorPM_{10}, PM_{2.5} - PM_{10}, month, hr) \quad (1)$$

$$TargetPM_{10} = Lasso(PM_{10}, AT, RH, PM_{2.5}, hr) \quad (2)$$

RF had the best performance for $PM_{2.5}$ calibration

Lasso regression had the best performance for low-cost PM_{10} calibration

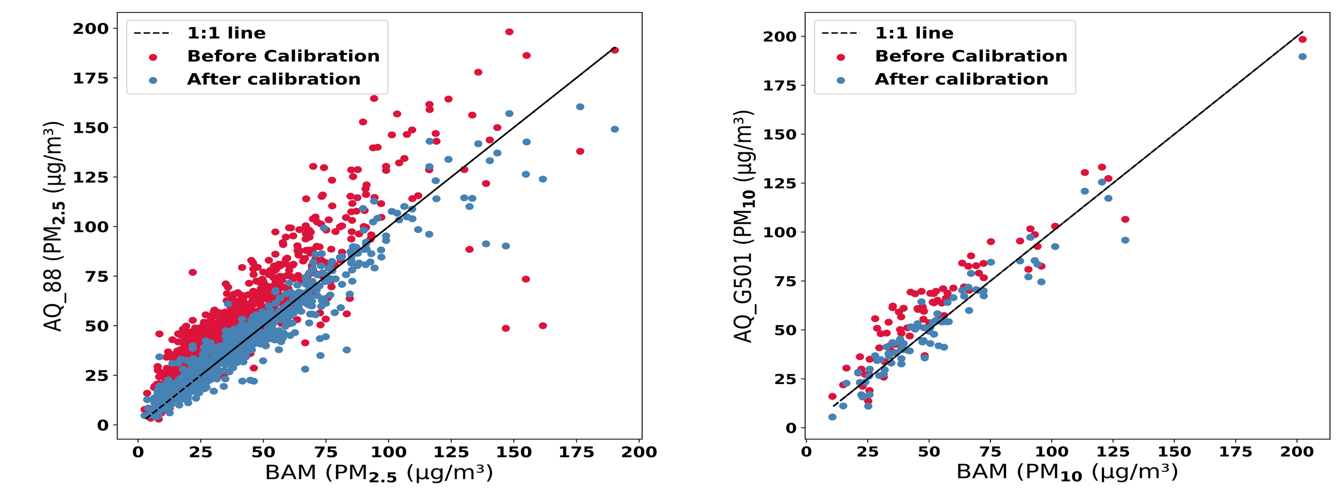
Results

Table 1. Random forest using optimal parameters and various input variable combinations.

Input variables	RMSE ($\mu g/m^3$)	MAE ($\mu g/m^3$)	R^2	Correlation
Factory calibrated (Raw $PM_{2.5}$)	18.6	14.6	0.52	0.9
$PM_{2.5}, AT, RH$	10.4	6.02	0.85	0.92
$PM_{2.5}, AT, RH, PM_{10}$	9.3	5.6	0.88	0.94
$PM_{2.5}, AT, RH, PM_{10}, errorPM_{2.5}$	9.1	5.3	0.88	0.94
$PM_{2.5}, AT, RH, PM_{10}, errorPM_{2.5}, errorPM_{10}$	8.5	5.1	0.90	0.95
$PM_{2.5}, AT, RH, PM_{10}, errorPM_{2.5}, errorPM_{10}, PM_{2.5} - PM_{10}$	7.6	4.8	0.92	0.96
$PM_{2.5}, AT, RH, PM_{10}, errorPM_{2.5}, errorPM_{10}, PM_{2.5} - PM_{10}$, month	7.4	4.7	0.92	0.96
$PM_{2.5}, AT, RH, PM_{10}, errorPM_{2.5}, errorPM_{10}, PM_{2.5} - PM_{10}$, month, hr	7.2	4.6	0.92	0.96
Collocated BAMs (Benchmark)	6.2	4.1	0.92	0.96

Table 2. Lasso regression using optimal parameters and various input variable combinations.

Input Combinations	RMSE ($\mu g/m^3$)	MAE ($\mu g/m^3$)	R^2	Correlation
Factory calibrated (PM_{10})	13.4	11.3	0.72	0.93
PM_{10}, AT, RH	9.0	6.9	0.91	0.96
$PM_{10}, AT, RH, PM_{2.5}$	8.2	6.3	0.93	0.96
$PM_{10}, AT, RH, PM_{2.5}, errorPM_{10}$	8.2	6.3	0.93	0.96
$PM_{10}, AT, RH, PM_{2.5}, errorPM_{10}, errorPM_{2.5}$	8.2	6.3	0.93	0.96
$PM_{10}, AT, RH, PM_{2.5}, errorPM_{10}, errorPM_{2.5}, PM_{2.5} - PM_{10}$	8.2	6.3	0.93	0.96
$PM_{10}, AT, RH, PM_{2.5}, errorPM_{10}, errorPM_{2.5}, PM_{2.5} - PM_{10}$, month	8.2	6.3	0.93	0.96
$PM_{10}, AT, RH, PM_{2.5}, errorPM_{10}, errorPM_{2.5}, PM_{2.5} - PM_{10}$, month, hr	7.9	6.0	0.93	0.97
$PM_{10}, AT, RH, PM_{2.5}, hr$	7.9	6.0	0.93	0.97
Collocated BAMs (Benchmark)	5.1	4.0	0.96	0.98



(a) (b)

Figure 2. Comparison between BAM and low-cost PM from the test set. Part (a) shows the relationship between BAM and lowcost (AQ_88) $PM_{2.5}$ before and after calibration using the proposed RF model. Part (b) shows the relationship between BAM and lowcost (AQ_G501) PM_{10} before and after calibration using the proposed lasso regression model

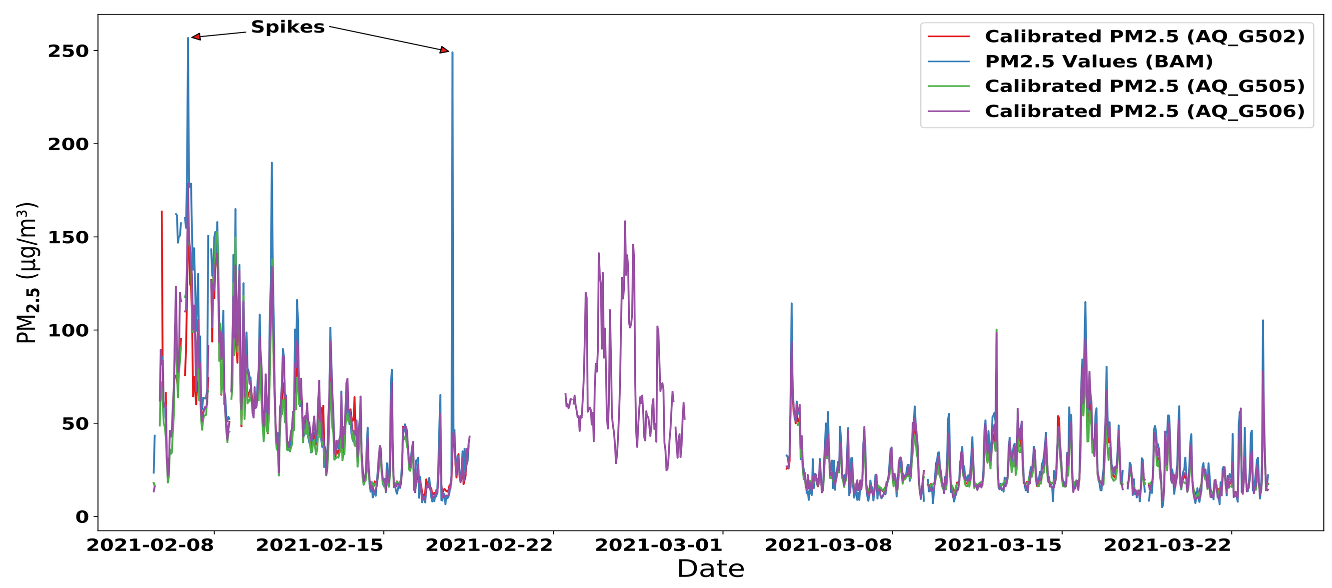


Figure 3. Cross-unit validation results for $PM_{2.5}$ calibration using the RF model. We presents hourly comparison between BAM and calibrated low-cost $PM_{2.5}$ for AirQo devices (AQ_G502), (AQ_G505) and (AQ_G506).

Deployment of calibration models in production

- The models are deployed as part of an urban air quality sensing system that is accessible to users via an open air quality API, an analytics dashboard **https://platform.airqo.net**, and a mobile app.
- The calibration models are encapsulated as a microservice that are exposed as REST APIs.
- Raw measurements from all devices on the network are streamed to a cloud-based IoT platform.
- Raw hourly PM concentrations are fed into the calibration models with corresponding hourly temperature & humidity readings to generate corresponding calibrated PM concentrations.
- **The deployment serves as a demonstration of the use of a Machine Learning system in addressing society challenges, in this case ambient urban air pollution.**

Conclusion and Discussion

- Various ML methods were compared for AirQo device calibration, with RF and lasso regression performing well for $PM_{2.5}$ and PM_{10} calibration respectively.
- RF model tends to under-predict spikes but excluding spikes leads to improved accuracy.
- We achieved reasonable accuracy with cross-unit and cross-site validation hence AirQo monitors do not have to be calibrated individually.
- Periodical retraining of the models is important in order to cater for seasonal and condition-specific dependency of calibration factors

References

[1] Priscilla Adong, Engineer Bainomugisha, Deo Okure, and Richard Sserunjogi. Applying machine learning for large scale field calibration of low-cost pm2. 5 and pm10 air pollution sensors. *Applied AI Letters*, 3(3):e76, 2022.