International Journal of Advances in Soft Computing and Intelligent Systems (IJASCIS)
2023, Vol 02, Issue 01, 07-13
Science Transactions © 2023
Original Paper

HIGH-RESOLUTION MACHINE LEARNING-BASED CALIBRATION OF A LOW-COST PARTICULATE MATTER SENSOR VIA INCORPORATION OF ENVIRONMENTAL PARAMETERS

Adarsh Mishra[1], Gaurav Sarode[1], Shubham Bhange[1],
Roshan Wathore[1],2,
Piyush A. Kokate[1],2
[1]CSIR-National Environmental Engineering Research Institute (NEERI), Nagpur, MH, INDIA - 440020
[2]Academy of Scientific and Innovative Research (AcSIR), Ghaziabad, INDIA-201002
Corresponding Authors: pa_kokate@neeri.res.in; r.wathore@neeri.res.in

ABSTRACT

Low-cost sensors for the purpose of air quality monitoring are gaining widespread attention due to their ease of operation, affordability and their ability to provide high-resolution data in both spatial and temporal scales. However, their performance is sub-par as compared to conventional monitoring methods and is susceptible to environmental/meteorological parameters such as temperature (T) and relative humidity (RH). In this study, the $PM_{2.5}$ measurement from an optical low-cost particulate matter sensor is calibrated using a higher grade optical $PM_{2.5}$ sensor as a secondary reference. Three calibration models (Artificial Neural Networks, Support Vector Machines, and Multilinear Regression) were deployed and their performance was compared; performance on inclusion of environmental parameters T and RH were also compared. It was found that ANN models performed as well as or better than the multilinear model. ANN models in general outperformed the base LR model, indicating with the availability of additional data, further modifications in the model and optimization of hyperparameters, deep learning methods could potentially be used for improving the performance of low-cost environmental sensors.

KEYWORDS

Low-cost particulate matter sensors, machine learning, deep learning, calibration.

1. Introduction

Worldwide, nearly 7 million annual premature deaths are attributed to household and ambient air pollution. India also faces a significant threat from air pollution. In 2019, out of the world's top most polluted 30 cities, India had recorded 21 cities [1]. In India, the Central Pollution Control Board (CPCB) is the principal institution in charge of collecting and reporting data on ambient air quality via the installations of Continuous Ambient Air Quality Monitoring Stations (CAAQMS) across several cities in India. Data from CAAQMS is considered as reference data; however, the high-grade equipment and analyzers used are very expensive (>50,000 USD), including their operation and maintenance. However, CAAQMS in India are available mostly in urban areas, with hardly any installation in rural India, where the majority of the population resides. Due to the high expenses involved in operation and regular maintenance of CAAQMS, dense air quality monitoring is not a practical alternative [2].

A newer generation of low-cost particulate matter sensors (LCPMS) have gained popularity over the past decade for the purpose of air quality monitoring. Since LCPMS are affordable, portable, and require less maintenance. LCPMS can therefore be applied to the monitoring of air quality. They have the capability to provide high spatial and temporal resolutions at a fraction of the cost

of CAAQMS. However, they are not as accurate and it is established that environmental factors such as temperature and humidity affect the performance of LCPMS [2][3].

Studies have shown that performance of low-cost sensors can be improved by the application of machine learning algorithms. More specifically, inclusion of environmental parameters and other derived features have shown potential to reduce the errors associated with the measurements of low-cost sensors when compared to conventional monitoring methods [4][5].

The present work looks into high-resolution data (5 minutes) collected from a low-cost PM sensor to a secondary reference optical PM sensor which has been calibrated against gravimetric PM. We then explore the effect of different calibration models - linear regression (LR), support vector regression (SVR) and artificial neural networks (ANN) as well the effect of environmental parameters - temperature (T) and relative humidity (RH).

2. Materials and Methods

2.1. Study design

The present study aims at deployment of machine learning and deep learning models for the calibration of high resolution (5 minutes) $PM_{2.5}$ data generated from optical sensors. Two optical based PM sensors were used for this study - TSI SidePak AM520i (henceforth referred to as TSI) and Plan tower PMS5003 low-cost particulate matter sensor (henceforth referred to as LCPMS).

The TSI is a small, portable light-scattering laser photometer with the purpose to measure real-time aerosol mass concentration such as smoke, fog and dust. It can be operated both via an internal battery and a power source and has an in-build data-logging capability. It internally converts the raw measurements to real-time aerosol mass concentration. The instrument is on the expensive side (> $3000 USD.) and hence cannot be procured in bulk. It has various options for impactors with capability and provide size fraction cut points ranging from 10 µm to 0.8 µm. For this study, $PM_{2.5}$ was the target pollutant to be calibrated. Depending on the type of aerosol sampled, the TSI has a calibration factor which can be applied via the instrument interface and can be calibrated to specific emissions in a variety of indoor and outdoor settings using gravimetric data. For this study, we have derived our own calibration equation by co-locating the TSI with a gravimetric $PM_{2.5}$ sampler (MiniVol Tactical Air Sampler) for 5 days (8 hours per day). Occasionally, we performed biomass burning via an improved cookstove in the vicinity of the samplers in order to get a range of concentrations to generate a linear calibration equation.

A sensor module was developed which integrated the LCPMS and the DHT22 temperature (T) and relative humidity (RH). Subsequently the LCPMS sensor module was then co-located with the calibrated TSI. Data was collected intermittently from April 21st 2022 to May 13th 2022. The logging frequency of both the instruments was 2 seconds. The data was then collated, mission data points were removed and averaged to 5 minutes. The resulting dataset consisted of 321 data points. This dataset was used for subsequent analysis.

2.2. Development of the LCPMS module:

The low-cost optical sensors were utilized in this comparative analysis. The optical sensor is made up of a laser source and a detector. The dust concentration can be measured using light scattering principle and diameter can be measured using MIE scattering theory. Arduino uno microcontroller was interfaced with the low-cost PM, temperature and humidity sensor to record the concentrations. Figure 1 depicts the schematic diagram of the sensor module developed in this work.

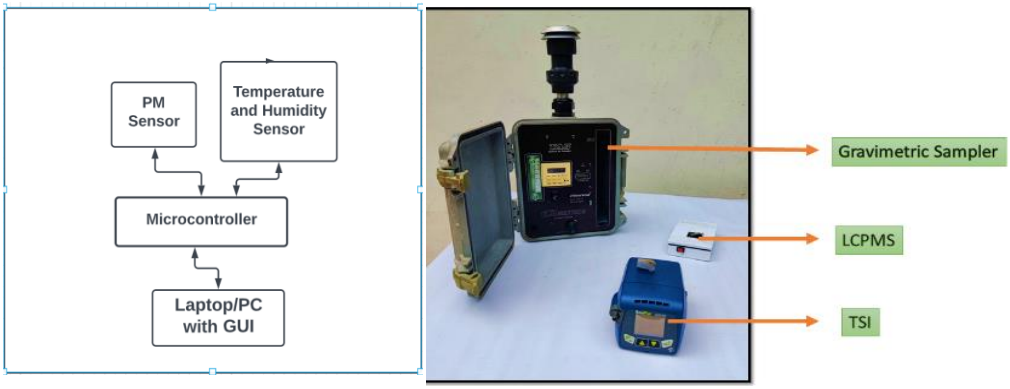

Figure 1. Schematic diagram of the sensor module and the pictures of the instruments used in this study

2.3. Model Development

Linear Regression

We use a linear regression (LR) model as our baseline model. Additionally, we also deploy 3 variants of support vector regression (SVR) models and 2 variants of artificial neural networks (ANN). Table 1 summarizes the model hyperparameters for the SVR models and the architectures for the ANN models. ANN 1 architecture consists of 3 hidden layers in addition to the input and output layers, whereas ANN 2 is a deeper and comparatively more complex model with a higher number of hidden layers and nodes. A brief background on SVR and ANN can be found elsewhere [3], [6], [7]. Subsequently, all the above univariate models were also compared with multivariate models with T and RH as additional inputs. Data analysis and model development was done in Python using the sklearn library for LR and SVR models and keras library for the ANN models.

Table 1. The hyperparameters for the machine learning models deployed in this study.

Model	**Hyperparameters**
SVR poly	kernel = poly, C=100, degree=2
SVR rbf 1	kernel = rbf, C=1.0, epsilon=0.1
SVR rbf 2	kernel = rbf, C=100, epsilon=0.1
ANN 1	Units in hidden layers = 16,4,2,1 Epochs = 500, Patience = 50, learning rate = 0.01, loss = mean_squared_error
ANN 2	Units in hidden layers = 128,64,32,16,8,1 Epochs = 500, Patience = 75, learning rate = 0.005 loss = mean_squared_error

2.4. Metrics used

For this work, the accuracy of the models was determined by the coefficient of correlation (r) and the root mean square error (RMSE). The equations for the metrics are provided below in Equation 1 and Equation 2 respectively:

$$r = \frac{\sum i(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_i (x_i - \overline{x})^2}\sqrt{\sum_i (x_i - \overline{x})^2}} \quad (1)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N}} \quad (2)$$

3. Results and Discussion

3.1 Calibration model for TSI vs gravimetric (on daily data)

Figure 2 shows the calibration equation derived for the TSI against the gravimetric sampler. A strong coefficient of determination ($R^2 > 0.92$) was obtained, indicating a good agreement between the two instruments over a varying concentration range. The calibration equation obtained was applied to the raw TSI values and was used as a secondary reference to calibrate the LCPMS.

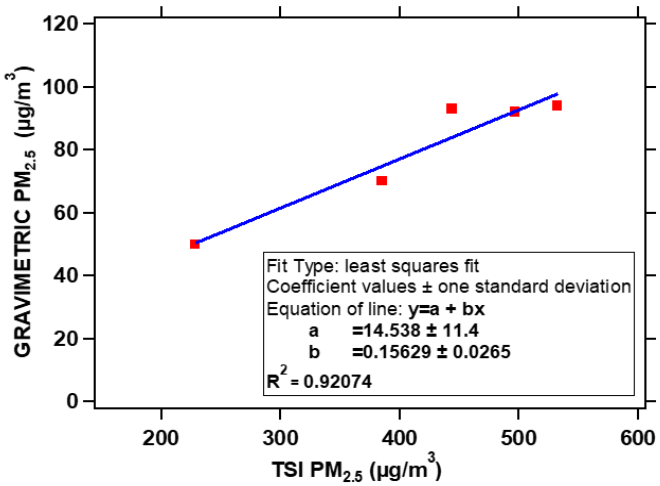

Figure 2. Calibration of the TSI PM2.5 against gravimetric PM2.5

3.2 Calibration of the LCPMS

The results of all the models used in this study are compiled in Table 2. The base model is a multilinear regression with two parameters. For the test data, it is observed that subsequent increase in the number of parameters results in sequential reduction of the RMSE for the multilinear model – this is generally observed for the ML, SVR poly and ANN models. However, for the SVR rbf models, there is hardly any improvement in the model performance from inclusion of the environmental parameters, signifying that the rbf kernel may not be an appropriate option for such calibration approaches.

While the univariate ANN models perform worse than the baseline LR mode, inclusion of environmental parameters leads to better performance than the baseline multivariate LR models. This is likely due to the fact that the model is able to capture the various non-linear responses of the sensor. Among the multivariate ANNs, ANN1 which has a relatively simpler architecture outperforms ANN2, which is a deeper model. However, it is expected that with the inclusion of additional input parameters (including derived parameters) and additional data for training and testing would provide additional insights on the comparative performance of ANNs.

Table 2. Calibration results for the test and train datasets for the various models used in this study.

Model	No. of parameters	TEST		TRAIN	
		r	rmse	r	rmse
LR	1 parameter	0.959	2.922	0.975	5.371
	3 parameters	0.959	2.914	0.976	5.281
SVR poly	1 parameter	0.946	3.407	0.947	21.614
	3 parameters	0.964	2.761	0.955	11.998
SVR rbf 1	1 parameter	0.944	3.791	0.419	12.569
	3 parameters	0.939	4.055	0.455	12.682
SVR rbf 2	1 parameter	0.969	2.611	0.627	11.299
	3 parameters	0.988	1.614	0.481	11.412
ANN 1	1 parameter	0.959	3.292	0.975	6.550
	3 parameters	0.959	3.592	0.975	4.017
ANN 2	1 parameter	0.959	3.134	0.975	6.304
	3 parameters	0.959	2.923	0.975	5.129

4. Conclusions and Future Work

In this study, the $PM_{2.5}$ measurement from an optical LCPMS is calibrated using a higher grade optical $PM_{2.5}$ sensor as a secondary reference. Three models (Artificial Neural Networks, Support Vector Machines, and Multilinear Regression) were deployed and their performance was compared; performance on inclusion of environmental parameters T and RH were also compared. It was found that ANN models performed as well as or better than the multilinear model. ANN models in general outperformed the base LR model, indicating with the availability of additional data, further modifications in the model and optimization of hyperparameters, deep learning methods could potentially be used for improving the performance of low-cost environmental sensors.

This work can be taken forward in the following ways:

• Based on the results obtained, a systematic analysis of the performance of the sensor with respect to environmental parameters would also be explored. This includes error dependence on the environmental parameters – T and RH.

• Also, the accuracy of the sensor to estimate the air quality index (AQI) category would also be explored. Such analysis would also be helpful for community level air quality reporting and increasing public awareness.

Acknowledgements
The authors would like to thank Nitin K Labhsetwar, Chief Scientist and Head, Energy & Resource Management Division, CSIR NEERI, Nagpur for their support. This work was done under the Project No. HCP-42 sponsored by CSIR, New Delhi.

References

[1] IQAir, World Air Quality Report 2020 (2020). file:///C:/Users/HP/Downloads/world-air-quality-report-2020-en%20(1).pdf

[2] Jha, Sonu Kumar, Mohit Kumar, Vipul Arora, Sachchida Nand Tripathi, Vidyanand Motiram Motghare, A. A. Shingare, Karansingh A. Rajput, and Sneha Kamble. "Domain Adaptation-Based Deep Calibration of Low-Cost $PM_{2.5}$ Sensors." IEEE Sensors Journal 21, no. 22 (November 2021): 25941–49. https://doi.org/10.1109/JSEN.2021.3118454.

[3] Kumar, Prashant, Lidia Morawska, Claudio Martani, George Biskos, Marina Neophytou, Silvana Di Sabatino, Margaret Bell, Leslie Norford, and Rex Britter. "The Rise of Low-Cost Sensing for Managing Air Pollution in Cities." Environment International 75 (February 1, 2015): 199–205. https://doi.org/10.1016/j.envint.2014.11.019.

[4] Kumar, Vikas, and Manoranjan Sahu. "Evaluation of Nine Machine Learning Regression Algorithms for Calibration of Low-Cost PM2.5 Sensor." Journal of Aerosol Science 157 (September 1, 2021): 105809. https://doi.org/10.1016/j.jaerosci.2021.105809.

[5] Malings, Carl, Daniel Westervelt, Aliaksei Hauryliuk, Albert A. Presto, Andrew Grieshop, Ashley Bittner, Matthias Beekmann, and R. Subramanian. "Application of Low-Cost Fine Particulate Mass Monitors to Convert Satellite Aerosol Optical Depth Measurements to Surface Concentrations in North America and Africa." Preprint. Aerosols/Remote Sensing/Validation and Intercomparisons, March 3, 2020. https://doi.org/10.5194/amt-2020-67.

[6] Spinelle, Laurent, Michel Gerboles, Gertjan Kok, Stefan Persijn, and Tilman Sauerwald. "Review of Portable and Low-Cost Sensors for the Ambient Air Monitoring of Benzene and Other Volatile Organic Compounds." Sensors 17, no. 7 (July 2017): 1520. https://doi.org/10.3390/s17071520.

[7] Topalović, Dušan B., Miloš D. Davidović, Maja Jovanović, Alena Bartonova, Zoran Ristovski, and Milena Jovašević-Stojanović. "In Search of an Optimal In-Field Calibration Method of Low-Cost Gas Sensors for Ambient Air Pollutants: Comparison of Linear, Multilinear and Artificial Neural Network Approaches." Atmospheric Environment 213 (September 15, 2019): 640–58. https://doi.org/10.1016/j.atmosenv.2019.06.028.

[8] Wathore, Roshan, Samyak Rawlekar, Saima Anjum, Ankit Gupta, Hemant Bherwani, Nitin Labhasetwar, and Rakesh Kumar. "Improving Performance of Deep Learning Predictive Models for COVID-19 by Incorporating Environmental Parameters." Gondwana Research, April 8, 2022. https://doi.org/10.1016/j.gr.2022.03.014.

[9] World Health Organization. WHO Global Air Quality Guidelines: Particulate Matter (PM2.5 and PM10), Ozone, Nitrogen Dioxide, Sulfur Dioxide and Carbon Monoxide. World Health Organization, 2021. https://apps.who.int/iris/handle/10665/345329.

[10] Zimmerman, Naomi, Albert A. Presto, Sriniwasa P. N. Kumar, Jason Gu, Aliaksei Hauryliuk, Ellis S. Robinson, Allen L. Robinson, and R. Subramanian. "A Machine Learning Calibration Model Using Random Forests to Improve Sensor Performance for Lower-Cost Air Quality Monitoring." Atmospheric Measurement Techniques 11, no. 1 (January 15, 2018): 291–313. https://doi.org/10.5194/amt-11-291-2018.

Authors

Er. Adarsh Sushil Mishra received the B.Tech. Degree from Laxminarayan Institute of Technology, RTM Nagpur University. He is a Project Associate – 1 at Energy and Resource Management Division of CSIR-NEERI, Nagpur. Prior to joining CSIR-NEERI, he has worked at JK papers as a Process Engineer. There he was working on an AI/ML project in recovery Island. His expertise includes Data Analysis and related works.

Er. Shubham Sanjayrao Bhange received the B.E. Degree from P.R Pote College of Engineering and Management, SGB Amravati University. He is a Project Associate – 1 at Energy and Resource Management Division of CSIR-NEERI, Nagpur. His research work includes data analysis of research work, AI/ML, Data Structures and Algorithm, and a lot more.

Er. Gaurav Laxmikant Sarode received the B.E. Degree from GHRCE (Autonomous), RTM Nagpur University. He is a Project Associate – 2 at Energy and Resource Management Division of CSIR-NEERI, Nagpur. He has 7 Research Papers related to his work. His research works includes, Arduino-Based Embedded C language, Environment Parameters based sensors, IoT, Energy Management and Data Analysis.

Er. Roshan Wathore is a Scientist in the Energy and Resource Management Division at CSIR-NEERI. He has received his M. Tech Degree from IIT Kanpur and MS degree from North Carolina State University, USA. He has extensive experience in applications of AI and ML in various environmental engineering domains, which includes low-cost sensors, optimizing energy efficiency and policy aiding.

Er. Piyush A. Kokate received the M. Tech Degree in VLSI Technologies from NMU, Maharashtra, India. He is working as a Senior Scientist at Energy and Resource Management Division at CSIR-NEERI, Nagpur. He has more than 15 research papers related to his work. His research area includes applications of Sensors for environment, IoT, UAVs, WSN devices. He has handled more than 09 projects related to environmental engineering and electronics domain. He is a Life Member of Indian Society for Technical Education (ISTE) and Indian Society of Remote Sensing (ISRS), Instrument Society Of India (ISOI), IISc Bangalore.