Optimization and Evaluation of Calibration for Low-cost Air Quality Sensors: Supervised and Unsupervised Machine Learning Models

Petar Zhivkov

Institute of Information and Communication Technologies -
Bulgarian Academy of Sciences (IICT-BAS)
acad. Georgi Bonchev str. Bl. 2, 1113, Sofia, Bulgaria
Email: pzhivkov@iit.bas.bg

Abstract—With the advancement of air pollution management, low-cost sensors are increasingly being used in air quality monitoring, but the data quality of these sensors is still a major source of concern. In this paper, data from five air monitoring stations in Sofia were compared to data from fixed low-cost PM sensors. The values of atmospheric pressure from low-cost sensors and the effects of relative humidity were investigated. A two-step model was created to refine the calibration process for low-cost PM sensors. At first, we calibrated the sensors with five separate supervised machine learning models and then the ANN-final model with anomaly detection completed the results. The ANN-final model improved the R^2 values of the PM10 determined by low-cost sensors from 0.62 to 0.95 as compared to standard instruments. In conclusion, the two-step calibration model proved to be a positive solution to addressing low-cost sensor efficiency issues.

I. INTRODUCTION

AIR POLLUTION is a significant public health problem that has long been a source of anxiety for citizens.An air pollutant is described as any substance that can affect humans, animals, plants, or materials. In the case of humans, an air pollutant may cause or lead to an increase in mortality or serious illness, as well as pose a current or potential health risk [1]. Measurements of air emissions are critical for epidemiology and air quality control, but the scope of ground-based air pollution observations has limitations [2]. PM (Particulate Matter) air pollution is a suspended combination of solid and liquid particles that vary in quantity, size, shape, surface area, chemical composition, solubility, and origin. Total suspended particles (TSPs) have a trimodal size distribution in the ambient air, including coarse particles, fine particles, and ultrafine particles [3]. PM size-selective sampling refers to the collection of particles that are below, above, or within a defined aerodynamic range of sizes, which is commonly chosen to be relevant to inhalation and deposition, causes, or toxicity [4].

A. Air quality monitoring systems

Traditionally, concentrations of air emissions have been monitored by air monitoring stations equipped with standard equipment, allowing for highly reliable monitoring results. However, the high costs of equipment and servicing make meeting the demands of high-resolution surveillance and assessing the extent of personal exposure impossible [5], [6]. Low-cost air quality sensors have been widely used in air monitoring in recent years due to the benefits of low cost, low power usage, quick operation, and rapid response [7].

B. Opportunities and disadvantages of wireless low-cost stations

Wireless sensor networks connect a large number of fixed sensors in multiple places into a single network, enabling long-term, high-resolution surveillance of air contaminants [8]. Its applications are seen often in health-related studies and tracking individual exposure. A sensor tracking network was built on the Hong Kong marathon route in 2015, and it is used to measure the Air Quality Health Index (AQHI) and determine athletes' individual exposure levels [9]. In Rochester, NY, USA, a study used PM sensors to test airborne PM at various locations concurrently and continuously to determine the temporal and spatial variance of PM, as well as the impact of traffic and wood burning on outdoor PM concentrations [10]. By installing sensors on carriers such as vehicles, motorcycles, and drones, the mobile sensor network can provide more compact spatial data than fixed sensors and can achieve stereoscopic tracking of air quality and emission sources [11].

C. Effects from humidity and height

Despite the low-cost sensor's widespread usage, the data's precision has been challenged [12]. Some countries have conducted sensor assessment and calibration tests, and recommendations for the use and evaluation of sensors have been written [13]. However, just a few researchers have tested and calibrated low-cost sensors so far, and their performance under a variety of environmental conditions and time scales is still unknown [14]. Relative humidity (RH) and particle size distribution have been shown to have a significant effect on sensor monitoring results [15]. Other research investigated in foggy conditions at high RH (RH > 80%) the tracked performance of the sensor was higher than that of the normal instrument [16]. Several studies identify that barometric

pressure is important for modeling particulate matter because complicated wind flows may occur, resulting in stagnant/stationary conditions with little circulation. Pollutants accumulate near the ground as a result of these conditions [17]. When barometric pressure was included in the model, the connection between particulate matter and cardiovascular mortality was marginally strengthened [18]. Finding associations between MLH and near-surface pollutant concentrations representative for a city like Berlin (flat terrain) appears to be impractical, particularly when traffic emissions are dominant [19].

II. METHODOLOGY

The efficiency of fixed low-cost sensors was evaluated in this study using five air quality monitoring stations in Sofia (capital city of Bulgaria) which were compared with standard instruments. The effects of AP, RH, and PM size distribution (PM10 ratio) on the performance of PM sensors were analyzed. Taking independent variables (RH, T and AP) and the dependent variable (PM10) as input factors, a two-step model calibration model to adjust fixed sensors was designed. Finally, we evaluated the performance of each model and gave recommendations for the conditions of the model application.

A. Reference instrument

This research used five air quality control stations with traditional measuring methods as a guide. The BETA-attenuation monitor (BAM) was used to calculate PM10. Impactors, cyclones, detection parts, and a dynamic heating system are among the standard instruments. They are situated in Sofia, Bulgaria respectively in the areas of Mladost, Druzhba, Nadezhda, Hipodruma, and Krasno Selo. Only one of these stations is measuring PM 2.5 and due to this PM 2.5 is not used for a reference in this research.

B. PM sensors

The PM sensor used in this research was the NovaFitness SDS011 laser particle sensor. The laser diffraction principle is used to use the sensor. The laser illuminates the trapped ions when capturing dispersed light waves at a certain angle as air flows through the photosensitive region of the sensor. A particle size continuum is created by classifying these pulse signals into various particle size intervals in order to measure the mass concentration of the particles [20].

C. The wireless network

In this research the Wireless Sensor Network (WSN) of Luftdaten was used. It was made up of 300 fixed sensors covering Sofia. Each sensor was installed in a plastic tube that could be mounted on walls, balconies, street light poles, and other structures.
Certain guidelines were developed to aid in obtaining the best representation of PM emissions in the city with the least amount of sensors possible. The WSN used fixed sensors that were mounted in 1 km grids to ensure that the majority of the downtown area was covered with adequate density.
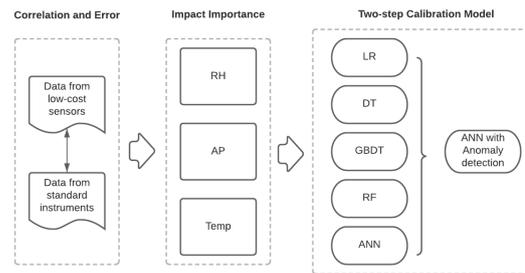

Fig. 1. Calibration Model

D. Calibration model setup and methods

The five sensors were situated right next to the five air monitoring stations. Since the sensor's time resolution differed from that of the regular instrument, the hourly mean was used for calculation and evaluation. A two-step model calibration method was modeled in this analysis, as seen in Fig. 1. Values of PM2.5 were measured only from one reference instrument, therefore, it was decided to exclude this instance from the model. A decision tree (DT) is a decision-making model that employs a tree-like model of decisions and their potential consequences, including the implications of chance events [21]. The Gradient Boosting Decision Tree (GBDT) algorithm is an iterative decision tree algorithm made up of multiple decision trees [22]. To obtain the final answer, the conclusions of all trees are added together. Random Forest (RF) is a tree predictor hybrid in which each tree is based on the values of a random variable sampled independently and with the same distribution for all trees in the forest [23]. In this research a RF with 10 trees is applied. The Artificial Neural Network (ANN) is a mathematical model that simulates neuronal behavior and is automatically modified by back-propagation errors [24]. Anomaly detection is decided It is a method of detecting unusual objects or occurrences in data sets that are out of the ordinary [25].The anomaly detector was used to remove outliers from the training dataset. As this is unsupervised learning, an evaluation with the same ANN setup was made, before and after cleaning the dataset to identify if unsupervised learning is appropriate for this dataset. RF and ANN are stochastic techniques, therefore, several runs have been performed in order to obtain objective results. In addition, statistical test from multiple runs have shown if the difference between the standard ANN and the one with anomaly detection is indeed significant.
The learning process of the model was divided into two stages: learning and testing. The raw data was split into two data sets at random, with 80% for training and 20% for testing. The model was first trained using the training data, and then its output was evaluated by the test data set.

E. Correlational analysis for atmospheric pressure

For evaluating the pressure measurement from the low-cost sensor were used the values from the reference instrument,

the height difference between each sensor and station, and the Barometric formula. To do this 4 low-cost sensors within a perimeter of 500m on different heights were evaluated. The Barometric formula (1) is used to model how the pressure of the air changes with altitude and it is as follows:

$$P = P_b \cdot \left[\frac{T_b + L_b \cdot (h - h_b)}{T_b} \right]^{\frac{-g_0 \cdot M}{R^* \cdot L_b}} \qquad (1)$$

To evaluate the AP results from low-cost sensors a correlation method was used. The comparison statistical methods fall into two categorizations: parametric and nonparametric. Parametric comparisons are based on the premise that the variable is continuous and normally distributed. Nonparametric approaches are used where data is continuous with non-normal distribution or any other form of data other than continuous variables. As our calculation model includes data normality and strict sample size, a parametric method is chosen. Moreover, parametric methods are better ways to measure the difference between the groups relative to their equivalent nonparametric methods.

The parametric Pearson correlation coeficient (2) is used for comparing the two sources of data. It provides a measure of the linear association between the two continuous variables (usually just referred to as correlation coefficient). Correlation coefficients for each (x , y) pair are determined to carry out the evaluation, and the values of x and y are consequently replaced by their ranks. Applying the test findings to a coefficient of correlation ranging from -1 to 1.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \qquad (2)$$

III. RESULTS AND DISCUSSION

A. Comparison between low-cost sensors and standard instruments

For better evaluation of the low-cost sensors together with the coefficient of determination (R squared), the mean absolute error and mean squared error are calculated (see Table I). The mean value of R squared between PM sensors and standard instruments without calibration was 0.62. The LR model showed worst result with a mean R^2 of 0.77. The best correlation for PM10 came from the ANN model. The mean value of the R^2 was 0.94 (PM10), which matched the findings of previous studies [12], [2]. For long-term comparison, the low-cost sensor and the regular instrument were put in the same location, which was a common approach for sensor evaluation in previous studies [26].

B. Relative humidity and air pressure

Relative humidity (RH) and temperature are considered to be the most important impact factors on particle sensor efficiency. High RH has been shown in previous studies to be the catalyst to causing hygroscopic growth of particles and

modifying optical properties, resulting in substantial interference for PM sensors [27]. Moreover, RH turned out to be of the highest importance in the RF and ANN models for the PM10 values.

The low-cost sensors had identical values as compared to normal instruments when RH was below 40%. While PM10 had a poor correlation when RH was above 80%.

Results of the AP from 4 sensors installed on different heights within 500m were compared with the calculations from the barometric formula and the data from the reference instrument. The installation height of the sensors was 3, 6, 8, and 18m while the height of the reference instrument is 2m. Calculations showed a high correlation between the sensors and r the with a mean value of the Pearson correlation coefficient r = 0,92.

C. Results of the calibration model

Table I presents the statistical outcomes of each model's testing, where Mean Absolute Error (MAE), Mean Squared Error (MSE), and the R^2 were determined.

The output of the other five separate models showed that the ANN model performed best. The RF model showed slightly worse results. The R^2 of PM10 increased from 0.62 to 0.9 and 0.94 for RF and ANN respectively. The ANN model performed best out of the 5 models, slightly better than the RF model, therefore, was chosen to be used in comparison with anomaly detection.

D. Improvements of the model through unsupervised learning

The ANN model was used as an autonomous model with sensor data and environmental variables as inputs. With the output values of five independent models as inputs, the ANN-final model conducted an artificial neural network model after filtering the dataset from anomalies . It was set up to find the top 20 anomalous instances within a forest with 256 trees. The highest anomaly score of a tree was 68%. These 20 outliers were removed from the training set and ran an ANN with the cleaned model which was evaluated again. The compared evaluations showed improvement with an R^2 of 0.95. In addition, the MAE and MSE decreased by 5.16% and 14.69% respectively. Therefore, the use of unsupervised learning in this study is considered to be useful. In conclusion, the ANN-final model had the best calibration score, with the largest R squared and the best correlation, indicating that the two-step model was more accurate than a single model in the model calibration of low-cost sensors.

IV. CONCLUSION AND FUTURE RESEARCH

The efficiency of PM sensors was measured by comparing by standard instruments using the wireless sensor network. To calibrate the fixed sensors, a two-step process was developed, and the model's results were evaluated. The following are the major conclusions: The findings of the two-step model were satisfactory. The R^2 of the fixed PM10 sensors increased from 0.62 to 0.95. The ANN model had the strongest impact of the five independent models, followed by the RF model,

TABLE I
RESULTS FROM SUPERVISED LEARNING MODELS.

TYPE OF MODEL	MEAN ABSOLUTE ERROR	MEAN SQUARED ERROR	R SQUARED
LINEAR REGRESSION:	11.19	288.12	0.77
DECISION TREE:	8.89	170.03	0.86
GRADIENT BOOSTING DECISION TREE:	8.68	145.22	0.89
RANDOM FOREST:	7.96	125.57	0.90
ARTIFICIAL NEURAL NETWORK:	6.27	83.90	0.94

while the LR model was ineffective. Anomaly detectors can be an unsupervised alternative to classifiers in an unbalanced dataset and in this research the final result was improved. The atmospheric pressure values of 4 low cost sensors were compared with a standard station by the use of calculations with the Barometric formula. The correlation was strong which means that low-cost sensors may be considered as a good source of modeling air pollution in vertical planning in further researches.

Further studies will beneficial in incorporating gas sensors into the WSN network. In addition, it is useful to analyze automotive emissions with integrated mobile sensors in the vehicles and using the model's improvement from this research.

REFERENCES

[1] M. Kampa and E. Castanas, "Human health effects of air pollution," *Environmental pollution*, vol. 151, no. 2, pp. 362–367, 2008.
[2] X. Qin, L. Hou, J. Gao, and S. Si, "The evaluation and optimization of calibration methods for low-cost particulate matter sensors: Inter-comparison between fixed and mobile methods," *Science of The Total Environment*, vol. 715, p. 136791, 2020.
[3] C. A. Pope III and D. W. Dockery, "Health effects of fine particulate air pollution: lines that connect," *Journal of the air & waste management association*, vol. 56, no. 6, pp. 709–742, 2006.
[4] J. C. Chow, "Measurement methods to determine compliance with ambient air quality standards for suspended particles," *Journal of the Air & Waste Management Association*, vol. 45, no. 5, pp. 320–382, 1995.
[5] P. Mouzourides, P. Kumar, and M. K.-A. Neophytou, "Assessment of long-term measurements of particulate matter and gaseous pollutants in south-east mediterranean," *Atmospheric Environment*, vol. 107, pp. 148–165, 2015.
[6] J. Y. Chin, T. Steinle, T. Wehlus, D. Dregely, T. Weiss, V. I. Belotelov, B. Stritzker, and H. Giessen, "Nonreciprocal plasmonics enables giant enhancement of thin-film faraday rotation," *Nature communications*, vol. 4, no. 1, pp. 1–6, 2013.
[7] Y. Wang, J. Li, H. Jing, Q. Zhang, J. Jiang, and P. Biswas, "Laboratory evaluation and calibration of three low-cost particle sensors for particulate matter measurement," *Aerosol Science and Technology*, vol. 49, no. 11, pp. 1063–1077, 2015.
[8] A. R. Rasyid, N. P. Bhandary, and R. Yatabe, "Performance of frequency ratio and logistic regression model in creating gis based landslides susceptibility map at lompobattang mountain, indonesia," *Geoenvironmental Disasters*, vol. 3, no. 1, pp. 1–16, 2016.
[9] L. Sun, J. Wei, D. Duan, Y. Guo, D. Yang, C. Jia, and X. Mi, "Impact of land-use and land-cover change on urban air quality in representative cities of china," *Journal of Atmospheric and Solar-Terrestrial Physics*, vol. 142, pp. 43–54, 2016.
[10] N. Zikova, M. Masiol, D. C. Chalupa, D. Q. Rich, A. R. Ferro, and P. K. Hopke, "Estimating hourly concentrations of pm2. 5 across a metropolitan area using low-cost particle monitors," *Sensors*, vol. 17, no. 8, p. 1922, 2017.
[11] M. Mead, O. Popoola, G. Stewart, P. Landshoff, M. Calleja, M. Hayes, J. Baldovi, M. McLeod, T. Hodgson, J. Dicks *et al.*, "The use of electrochemical sensors for monitoring urban air quality in low-cost,

high-density networks," *Atmospheric Environment*, vol. 70, pp. 186–203, 2013.
[12] A. C. Rai, P. Kumar, F. Pilla, A. N. Skouloudis, S. Di Sabatino, C. Ratti, A. Yasar, and D. Rickerby, "End-user perspective of low-cost sensors for outdoor air pollution monitoring," *Science of The Total Environment*, vol. 607, pp. 691–705, 2017.
[13] H. Brantley, G. Hagler, E. Kimbrough, R. Williams, S. Mukerjee, and L. Neas, "Mobile air monitoring data-processing strategies and effects on spatial air pollution trends," *Atmospheric measurement techniques*, vol. 7, no. 7, pp. 2169–2183, 2014.
[14] B. Zheng, D. Tong, M. Li, F. Liu, C. Hong, G. Geng, H. Li, X. Li, L. Peng, J. Qi *et al.*, "Trends in china's anthropogenic emissions since 2010 as the consequence of clean air actions," *Atmospheric Chemistry and Physics*, vol. 18, no. 19, pp. 14 095–14 111, 2018.
[15] A. Mukherjee and M. Agrawal, "World air particulate matter: sources, distribution and health effects," *Environmental Chemistry Letters*, vol. 15, no. 2, pp. 283–309, 2017.
[16] R. Jayaratne, X. Liu, P. Thai, M. Dunbabin, and L. Morawska, "The influence of humidity on the performance of a low-cost air particle mass sensor and the effect of atmospheric fog," *Atmospheric Measurement Techniques*, vol. 11, no. 8, pp. 4883–4890, 2018.
[17] B. Murthy, R. Latha, A. Tiwari, A. Rathod, S. Singh, and G. Beig, "Impact of mixing layer height on air quality in winter," *Journal of Atmospheric and Solar-Terrestrial Physics*, vol. 197, p. 105157, 2020.
[18] N. Janssen, P. Fischer, M. Marra, C. Ameling, and F. Cassee, "Short-term effects of pm2. 5, pm10 and pm2. 5–10 on daily mortality in the netherlands," *Science of the Total Environment*, vol. 463, pp. 20–26, 2013.
[19] A. Geiß, M. Wiegner, B. Bonn, K. Schäfer, R. Forkel, E. v. Schneidemesser, C. Münkel, K. L. Chan, and R. Nothard, "Mixing layer height as an indicator for urban air quality?" *Atmospheric Measurement Techniques*, vol. 10, no. 8, pp. 2969–2988, 2017.
[20] K. A. Koehler and T. M. Peters, "New methods for personal exposure monitoring for airborne particles," *Current environmental health reports*, vol. 2, no. 4, pp. 399–411, 2015.
[21] S. R. Safavian and D. Landgrebe, "A survey of decision tree classifier methodology," *IEEE transactions on systems, man, and cybernetics*, vol. 21, no. 3, pp. 660–674, 1991.
[22] S. W. Kwok and C. Carter, "Multiple decision trees," in *Machine Intelligence and Pattern Recognition*. Elsevier, 1990, vol. 9, pp. 327–335.
[23] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.
[24] M. Gevrey, I. Dimopoulos, and S. Lek, "Review and comparison of methods to study the contribution of variables in artificial neural network models," *Ecological modelling*, vol. 160, no. 3, pp. 249–264, 2003.
[25] D. Savage, X. Zhang, X. Yu, P. Chou, and Q. Wang, "Anomaly detection in online social networks," *Social Networks*, vol. 39, pp. 62–70, 2014.
[26] D. H. Hagan, G. Isaacman-VanWertz, J. P. Franklin, L. M. Wallace, B. D. Kocar, C. L. Heald, and J. H. Kroll, "Calibration and assessment of electrochemical air quality sensors by co-location with regulatory-grade instruments," *Atmospheric Measurement Techniques*, vol. 11, no. 1, pp. 315–328, 2018.
[27] X. Liu, R. Jayaratne, P. Thai, T. Kuhn, I. Zing, B. Christensen, R. Lamont, M. Dunbabin, S. Zhu, J. Gao *et al.*, "Low-cost sensors as an alternative for long-term air quality monitoring," *Environmental research*, vol. 185, p. 109438, 2020.