



Regression methods in the calibration of low-cost sensors for ambient particulate matter measurements

Marek Badura¹ · Piotr Batog² · Anetta Drzeniecka-Osiadacz³ · Piotr Modzel³© The Author(s) 2019 [OPEN](#)

Abstract

The article presents comparison of regression methods used to obtain calibration formulas for low-cost optical particulate matter sensors. Data for analysis were taken from 1-year collocation study of PMS7003 sensors (Plantower) with research-grade instrument TEOM 1400a. The $PM_{2.5}$ fraction was considered in this study. The results of measurements showed that PMS7003 was characterized by high reproducibility between units (coefficient of variation was lower than 10%), but the raw sensor outputs significantly overestimated $PM_{2.5}$ concentrations. Data analysis revealed that simple univariate models were sufficient to obtain a good fitting quality to TEOM data; however, the best results were achieved for raw PM_1 outputs ($R^2 \approx 0.81$). The fitting quality was improved when multi-variable equations were examined ($R^2 \approx 0.84$). The addition of temperature and relative humidity in the models was also beneficial ($R^2 \approx 0.87$). Stepwise selection algorithm was used to choose the best subset of variables in the model. The results of that method were compared with “all possible regression” approach, demonstrating the convenience of stepwise regression. Data from Plantower sensor were also used for training of artificial neural network. That algorithm proved to be very effective for fitting data from one sensor ($R^2 \approx 0.9$), but it was susceptible to deviations in the data from the other units. In general, regression analysis proved to be useful for sensor systems for ambient particulate matter measurements.

Keywords Air pollution · $PM_{2.5}$ · Sensor network · Data analysis · Stepwise regression

1 Introduction

In recent years, the progress in the field of electrical engineering has led to the expansion of air pollution sensing devices [1, 2]. Different sensors are currently available on the market, allowing the measurements of various gaseous species and particulate matter (PM) [3].

Generally, sensor devices are characterized by small size and small weight, relatively low power requirements and short response time [2–4]. What is significant, their price is few orders of magnitude lower than the price of

traditional air quality measurement instruments. Those so-called low-cost sensors are changing the paradigm of air pollution monitoring, hitherto based on expensive and complex instruments, operated by governmental, industry or research agencies [2, 5].

The possibilities of using sensors in the measurements of air quality are very wide. Those inexpensive devices might be used for the improvement in the spatial coverage of ambient air pollution data. Therefore, they can supplement the conventional monitoring stations networks [2, 6]. They could also provide data in real time (or near real time)

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s42452-019-0630-1>) contains supplementary material, which is available to authorized users.

✉ Marek Badura, marek.badura@pwr.edu.pl | ¹Department of Air Conditioning, Heating, Gas Engineering and Air Protection, Faculty of Environmental Engineering, Wrocław University of Science and Technology, 50-373 Wrocław, Poland. ²INSYSPOM, 54-427 Wrocław, Poland. ³Department of Climatology and Atmosphere Protection, Institute of Geography and Regional Development, Faculty of Earth Science and Environmental Management, University of Wrocław, 51-621 Wrocław, Poland.

SN Applied Sciences (2019) 1:622 | <https://doi.org/10.1007/s42452-019-0630-1>

Received: 31 January 2019 / Accepted: 16 May 2019 / Published online: 22 May 2019

and increase the temporal resolution of measurements. This feature is useful for “hot spot” detection or indication of elevated pollutant events [7, 8].

Many low-cost sensors are usually easy to use and often adopted by citizen scientist. This trend is particularly important in raising public awareness about the air pollution [9–11]. Compact and lightweight versions of sensors with high-resolution data acquisition could be also used for personal exposure monitoring. Such monitors might be helpful in finding link between short-term pollutant exposures and health effects [11–14].

The application of low-cost sensors is not limited to atmospheric air only and sensor techniques might be used for indoor air quality assessment as well. Characterization of indoor concentrations, identification of emitting sources and management of ventilation rates and energy are just a few examples of the use of sensors in indoor spaces [15–19].

It should be noted that sensor is always only a constituent of a larger whole—a sensor system [4]. Such system might have a form of a stand-alone monitor (stationary, hand-held, portable, mobile or wearable [10, 20–22]) or might be integrated into a node of a widespread network [23–26]. Sensor system may contain one or many pollutant sensors and often includes additional sensors for temperature and/or relative humidity measurements [4]. Besides the sensing devices, the system includes other, configuration-dependent, elements: housing, sampling probe, power source, control board, data acquisition and data analysis module, data transmission module, positioning system [2, 4, 23].

In general, sensors have an analogue output (voltage or current) or digital output (e.g. in the form of mass or volume concentration). However, the sensor response might be largely influenced by cross-sensitivities (in case of gas sensors), particles properties (in case of particulate matter sensors) or environmental factors (in both cases). Therefore, data quality is a critical issue in the usage of low-cost sensors and the calibration or recalibration of sensors before the deployment is necessary in many situations [3, 4]. The most popular way of such data adjustment is based on a field collocation with a reference-grade or research-grade instrument [27–31]. During this “training” period, the relationship between raw sensor data and reference data is established and the data correction algorithm is developed [32]. For this reason, the most important part of sensor system might be the data analysis module, where data processing occurs.

Overall, different approaches are used to create calibration formulas. In some cases, simple linear models are sufficient to adjust the raw data [8, 33]. In other cases, nonlinear equations or multi-parameter methods are necessary to obtain results close to Ref. [25, 30, 34, 35]. More

sophisticated techniques, from the field of machine learning, are also utilized for this purpose [36, 37].

This article presents comparison of different algorithms for the adjustment of data from low-cost optical particulate matter sensors. Data for testing have been collected during 1-year collocation study with research-grade instrument (TEOM 1400a) for $PM_{2.5}$ measurements. On the basis of the previous analyses [27], PMS7003 sensor from Plantower was chosen for this investigation. This sensor has proved to work stable for several months of measurements, showed high linear correlation with comparison instrument and was precise in terms of reproducibility between units [27].

The paper focuses on the linear regression methods (univariate and multiple regressions); however, comparison with nonlinear algorithm (artificial neural network) was made too.

2 Materials and methods

2.1 Measurement site and control instrument

The collocation study took place in Poland at the Meteorological Observatory of Department of Climatology and Atmosphere Protection of University of Wrocław. In the vicinity of the observatory, there are detached houses and allotments and a large municipal park. In this area, the main sources of particulate matter are the individual heating systems in households.

The observatory is equipped with instruments for PM_{10} and $PM_{2.5}$ measurements (TEOMs); however, operational problems with PM_{10} unit have led to the exclusion of this device from analysis. TEOM 1400a analyser is an example of tapered element oscillating microbalance [38]—a research-grade instrument, with the possibility of near real-time monitoring, which proved to be useful for low-cost sensors testing [27, 39, 40]. TEOM with a $PM_{2.5}$ inlet provided 1-min averaged data that were stored in the database.

2.2 Measurement set-up for PM sensor

Special measurement box was designed for the purpose of testing different sensors under the same measurement conditions. The box was made from PVC and was equipped with rainproof lid, air inlets and a fan, forcing the air flow. Power suppliers, microcomputer and USB hubs for connecting the sensors were placed inside this enclosure. The measurement set-up included also data logger with temperature and relative humidity (RH) sensor for the measurements of those parameters in the vicinity of PM sensors. The box was placed near TEOM intake (circa 1.5–1.8 m

below). Construction details of the measurement box may be found in [27].

PMS7003 (Beijing Plantower Co., Ltd, China) is a small and lightweight sensor ($48 \times 37 \times 12$ mm, ~ 30 g), which can be classified as low-cost device (approximate price at the level of 15–20 \$). PMS7003 is a light-scattering optical sensor that composes of a small measurement chamber with light-emitting diode, light detector (photodiode) and a set of focusing lenses. This sensor uses also a microfan to induce the flow of air.

According to the PMS7003 datasheet, the minimum detectable particle diameter is $0.3 \mu\text{m}$. The sensor contains a microprocessor that provides digital signals in two forms:

1. Mass concentration ($\mu\text{g}/\text{m}^3$) of PM_{10} , $\text{PM}_{2.5}$ and $\text{PM}_{1.0}$ fractions with correction factor for “factory environment” (“FE”) and for “atmospheric environment” (“AE”);
2. Number of particles per unit volume (0.1 l of air) for 6 size bins: beyond $0.3 \mu\text{m}$ (bin 1), beyond $0.5 \mu\text{m}$ (bin 2), beyond $1.0 \mu\text{m}$ (bin 3), beyond $2.5 \mu\text{m}$ (bin 4), beyond $5.0 \mu\text{m}$ (bin 5) and beyond $10.0 \mu\text{m}$ (bin 6). The product manual contains information that particles diameters and the number of particles in size bins are estimated on the basis of light-scattering intensities and light-scattering signal distribution, with the use of Mie theory. It can be deduced that the PM_{10} , $\text{PM}_{2.5}$ and $\text{PM}_{1.0}$ concentrations are calculated in the subsequent step. However, the details of the calculations and also the factory calibration procedures and the type of particles used for calibration are not specified in the datasheet.

Three copies of PMS7003 sensor were mounted inside the measurement box and connected via USB hub with microcomputer. Sensor signals were averaged in 1-min intervals and stored in the database for further analysis.

3 Data analysis

3.1 Data preparation and preliminary analysis

Data from TEOM and PM sensors registered from 21/08/2017 to 20/08/2018 were utilized in this study. 1-min averaged TEOM outputs and Plantower signals were used to create a new set of 1-h averaged data. This type of data is usually provided by automated measuring systems [41] from governmental monitoring stations and is very popular in informing the public about the air quality. Averaging was made only for hours with at least 75% completeness of data.

The preliminary analysis covered the evaluation of reproducibility between units of PMS7003 sensor. 1-min

and 1-h averaged $\text{PM}_{2.5}$ outputs with “FE” correction factor were used to calculate the correlations of PMS7003 units and coefficient of variation (CV). Low CV value indicates high reproducibility of sensor units, and CV value below 10% is considered acceptable in the low-cost sensor studies [21, 42]. The $\text{PM}_{2.5}$ output was chosen on the basis of an assumption that it should reflect $\text{PM}_{2.5}$ concentrations in the best way.

The other aspect of preliminary study was the assessment of sensor signals: mass concentrations and number of particles in bins. Additionally, combinations of bins in form of differences between bin 1 and the other bins were taken into account. Sensor outputs were investigated on the basis of Pearson’s correlation coefficient (r).

After the preliminary investigation, the hourly averaged dataset was randomly divided into training set (70% of data) and validation set (30%) that were used to create and evaluate different calibration models. There were 6116 samples in training set and 2621 samples in validation set. Both datasets contained similar range of PM concentrations from TEOM device. All data processing was performed in MATLAB environment.

3.2 Evaluation criteria

The performance of calibration equations was assessed for two datasets: training and validation. Two popular goodness-of-fit indicators were used for that purpose: coefficient of determination (R^2 , dimensionless) and root mean square error (RMSE, expressed in $\mu\text{g}/\text{m}^3$). The R^2 value near 1 reflects very good agreement with control measurements and small R^2 indicates poor fitting quality. In turn, small value of RMSE demonstrates small error of fitting.

3.3 Regression methods for PM sensor calibration

3.3.1 Univariate regression

Generally speaking, regression analysis is used for determining the relationship between two or more variables. Regression model includes the dependent variable (response) and other variables, which are thought to provide information on the behaviour of response-independent variables (also called predictors or explanatory variables) [43]. In this study, the concept of so-called inverse calibration was adopted [44]. TEOM readouts were chosen as dependent variable and PM sensor data as predictor variables.

Firstly, the univariate regression with only one independent variable was tested. On the basis of the previous study [27], the linear models were assumed to be sufficient

to describe the relationship between TEOM and sensor signals. The following linear equations were examined:

$$y = a_0 + a_1x + \varepsilon,$$

where y denotes the dependent variable (TEOM response), x is the independent variable (one of the sensor outputs: one type of mass concentration or number of particles for size bins or bins combination) and ε is a random term. Regression coefficients a_0 (the intercept) and a_1 (the slope) were estimated by ordinary least squares procedure [43].

3.3.2 Multiple regression

Linear additive models with several independent variables were examined in this study as well. The general equation taken into account had form:

$$y = a_0 + a_1x_1 + a_2x_2 + \dots + a_kx_k + \varepsilon,$$

where k is the number of independent variables ($x_1 \dots x_k$) and $a_0 \dots a_k$ are the regression coefficients.

Two types of models were tested:

1. Model that included different forms of sensor outputs (mass concentrations, particles number in size bins, combination of bins);
2. Model with the mentioned sensor outputs and also with temperature and relative humidity.

In the case of PMS7003, small impact of high levels of RH was previously noticed [27], so the second approach was aimed to assess the validity of including environmental factors in calibration equation.

3.3.3 Variable selection for multiple regression models

The previously described models contain a quite number of variables and some of them may be irrelevant and could be eliminated. Generally, the multi-variable models may be fitted to get simpler formulas, easier to interpret and to implement. Also, the removal of redundant variables may simplify the data acquisition and signal processing.

One of the possible strategies for variable selection is stepwise regression [43, 44]. In the stepwise selection process, variables are sequentially added or removed from the model, on the basis of their statistical significance. It should be noted that this algorithm finds variable subsets that are locally optimal—the selection of the globally best subset is not guaranteed.

The algorithm applied in this study started from constant (intercept) term and added and removed predictors in subsequent steps. Only linear additive models were examined, and the F-test was employed for judging the importance of variables. Some stringent criteria to obtain

fitted models were applied: p value for a term to be added to a model was set to 0.005 and p value for removing variables was equal 0.010.

The results of stepwise regression were compared to the results of “all possible regressions” approach [43], where models with all possible subsets of variables were created and tested. The discussion on the choice of the best subset size was based on two information criteria: Akaike information criterion (AIC) and Bayesian information criterion (BIC). Both criteria are used to find the trade-off between accuracy of fit and the number of predictors used in the model [43]. The model with the minimum value of AIC or BIC is the most appropriate in relation to the concerned criterion.

3.4 Neural networks for PM sensor calibration

Neural network (NN) is a computation system consisting of a number of highly interconnected units (neurons), organized in layers. Each neuron converts received information by means of activation function and produces output value, which might be processed by neurons in the next layer. The most popular NN approach is feedforward network with input, hidden and output layers. The NN training process is based on updating the weights of neurons via supervised learning. After the training, NN gains the unique approximation capabilities [45, 46].

Feedforward NN with 10 neurons with sigmoid transfer function in hidden layer and linear output neuron was used in this study. Backpropagation method with Levenberg–Marquardt algorithm was adopted for training. Patterns for learning and testing were taken only from training dataset.

4 Results and discussion

4.1 Measurements results

Figure 1 presents the results of 1-year $PM_{2.5}$ measurements with TEOM control device and PMS7003 sensor. ($PM_{2.5}$ signals with “FE” and “AE” correction factors for one unit were plotted for clarity.) During this period, some power outages and data acquisition problems were noticed and caused data gaps for both types of devices. An error in data transfer script has resulted in loss of bin 6 data (number of particles beyond $10.0 \mu m$), and this type of data was excluded from further analyses.

Nonetheless, all units of Plantower PMS7003 were stable during this campaign and the trends of their outputs were similar to TEOM signals. However, the 1-h TEOM averages were in the range $1\text{--}120 \mu g/m^3$ and the $PM_{2.5}$ values from raw sensor outputs were significantly overestimated

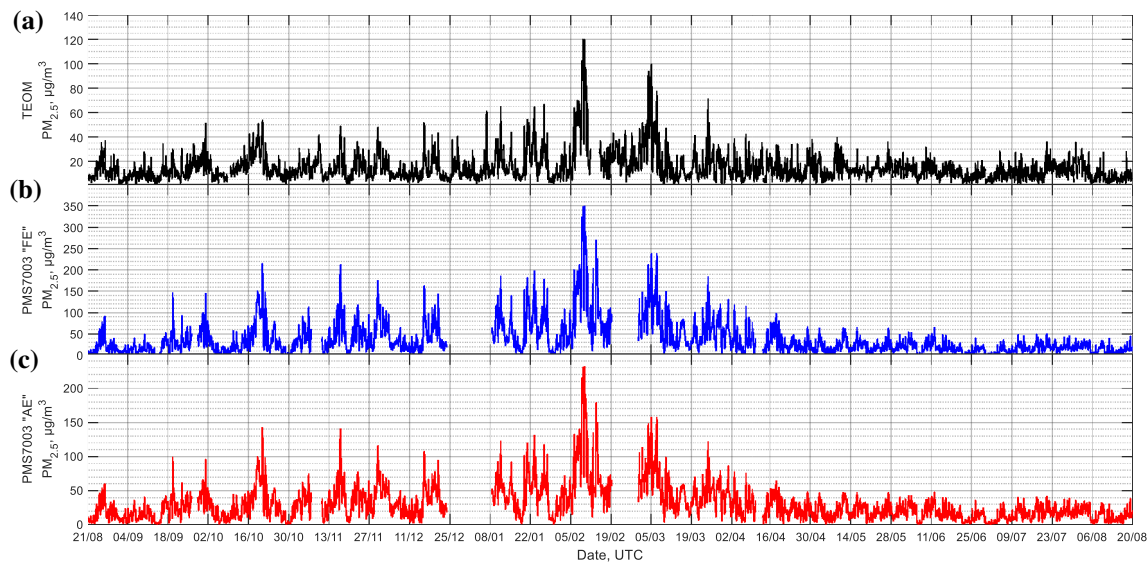


Fig. 1 Results of 1-year $PM_{2.5}$ measuring campaign: (a) TEOM 1400a outputs, (b) PMS7003 outputs with “FE” correction factor, (c) PMS7003 outputs with “AE” correction factor. 1-h averages for selected sensor were plotted for clarity. Please note different scales on y-axes

(about three times in case of “FE” outputs). The “atmospheric” (“AE”) $PM_{2.5}$ output was also not well suited for field measurements—overestimation by a factor of 2.2 was observed. This situation may derive from factory calibration using particles with completely different properties than PM in ambient air. Thus, it was confirmed that this low-cost sensor needs calibration in the final environment of measurements.

Regarding reproducibility, it was high in case of Plan-tower sensor units. The correlation coefficients between all units were higher than 0.990, and the lowest variability was observed for units no. 1 and no. 2—the correlation was at the level of 0.996 for 1-min data and 0.998 for 1-h data. Outputs from unit no. 3 were to some extent distant from other unit signals, but also highly correlated with them ($r \approx 0.99$). The scatter of PMS7003 data is presented in Fig. S1 in Supplementary Material.

The coefficient of variation (CV) computed for all units for 1-min averages was equal to 8.43% and for hourly data was equal to 6.47%. Taking into account the high reproducibility of PMS7003, only one unit of that sensor was used in further analyses (unit no. 2 was chosen arbitrarily).

Table S1 in Supplementary Material presents the correlation coefficients (r) for 1-min raw outputs from PMS7003 sensor (mass concentrations and number of particles in bins) and differences between bin 1 and the other bins. The highest correlations were computed for “FE” mass concentrations—the r value between $PM_{2.5}$ and PM_{10} was equal to 0.995, and for $PM_{2.5}$ and PM_{10} , it was 0.998. This results show very high linear relationships between PM mass outputs. The ratio between

PM_{10} and $PM_{2.5}$ was generally constant and was equal about 1.1, and the ratio between $PM_{2.5}$ and PM_{10} was at the level of 1.5. Such simple relationships might be not adequate for ambient air monitoring, where PM mass ratios depend on the pollutant sources and may change during the year [47, 48].

Very high linear correlations and similar ratios were observed for “AE” mass concentrations as well. Generally, the “FE” and “AE” outputs were highly correlated (e.g. for $PM_{2.5}$ $r = 0.988$), but there existed some transfer functions that changed the relationships of “FE” and “AE” signals. In the case of $PM_{2.5}$ concentration, “FE” and “AE” outputs were the same up to about 25–30 $\mu\text{g}/\text{m}^3$, and above that level, nonlinear relationship was observed up to about 100 $\mu\text{g}/\text{m}^3$ of “FE” output. Above that threshold, the mass concentrations were again highly linearly correlated, but the “FE” values were 1.5 times higher the “AE”. The relationship between the discussed outputs is presented in Fig. S2 in Supplementary Material.

$PM_{2.5}$ output was highly linearly correlated ($r = 0.989$) with bin 3 (particles beyond 1.0 μm). This bin was also the most correlated with PM_{10} data ($r = 0.990$). Similar situation was observed for “AE” signals ($r \approx 0.977$). In case of combinations of bins, the difference between bin 1 and bin 5 (i.e. particles number beyond 5.0 μm subtracted from particles number beyond 0.3 μm) had the highest r value for “FE” $PM_{2.5}$ (0.982). The linear correlations between bins and mass concentrations were in general very high, but it seems that some nonlinear function might be responsible for calculations on mass concentrations and this issue requires further consideration.

4.2 Univariate regression results

Table 1 presents the results of simple regression fittings for TEOM 1400a and PMS7003 outputs. All regression coefficients are provided in Table S2 in Supplementary Material.

The highest values of coefficient of determination and smallest RMSE levels were observed for both datasets for mass concentration outputs with “FE” factors. The best results were observed for PM₁ data ($R^2=0.815$ and $RMSE=5.09 \mu\text{g}/\text{m}^3$ for training set/ $R^2=0.801$ and $RMSE=4.97 \mu\text{g}/\text{m}^3$ for validation data). The PM_{2.5} output, which appears to be dedicated to the measurements of that PM fraction, had a somewhat worse fit.

In case of bins and bins combinations, better results were obtained in most situations for the latter ones. The highest R^2 was at the level of 0.78 in training set for difference between bin 1 and bin 2 (all particles number beyond 0.5 μm subtracted from particles number beyond 0.3 μm). Regarding raw bins, model with bin 1 showed the smallest value of error, suggesting that all

particles detected by the sensor are mainly related to PM_{2.5} mass concentration. However, it should be mentioned that the quality of fitting depends also on the quality of control instrument used for comparison. This aspect is especially significant for TEOM device, which is susceptible to measurement errors under certain conditions [49–51]. The detection possibilities of PMS7003 should be therefore further investigated by means of other reference instrument.

4.3 Multiple regression results

Complex multiple regression models were created with full set of mass concentration outputs (“FE” and “AE” types) and bin differences, which have proved to be more correlated with TEOM outputs than raw bins data. The comparison of results of such regression fittings and regression that included temperature and relative humidity is shown in Table 2.

The tested multiple regression models were substantially better fitted to TEOM data than univariate models. The R^2 for the first type of multi-parameter model was equal 0.853 for training set, and the result for validation set was 0.837. RMSE errors were lower than in the previous case and equal about 4.5 $\mu\text{g}/\text{m}^3$.

The addition of temperature and relative humidity to the equation resulted in further improvement in goodness of fit. The value of R^2 increased by approximately 0.02 (to the level of 0.87), and the RMSE error has decreased by around 0.3 $\mu\text{g}/\text{m}^3$ (to the level of 4.2 $\mu\text{g}/\text{m}^3$). It should be noted that temperature and RH were registered inside the measurement box and may reflect only the environment in the vicinity of sensors, only in conditions of that study. The inclusion of RH to the model may be beneficial, because of some small impact of high humidity levels on performance of PMS7003 [27]. As regards the temperature, that parameter was moderately correlated with RH during this measuring campaign and the impact of temperature for that sensor has not been investigated so far. For this reason, the incorporation of temperature into the calibration equation may be questionable.

Table 1 Results of univariate regression for training set and validation set for different types of predictors in regression models

| Model structure | Training set | | Validation set | |
|------------------------|--------------|-----------------------------------|----------------|-----------------------------------|
| | R^2 | RMSE ($\mu\text{g}/\text{m}^3$) | R^2 | RMSE ($\mu\text{g}/\text{m}^3$) |
| PM ₁ “FE” | 0.815 | 5.09 | 0.801 | 4.97 |
| PM _{2.5} “FE” | 0.812 | 5.13 | 0.799 | 5.00 |
| PM ₁₀ “FE” | 0.811 | 5.14 | 0.797 | 5.02 |
| PM ₁ “AE” | 0.785 | 5.48 | 0.768 | 5.37 |
| PM _{2.5} “AE” | 0.786 | 5.47 | 0.770 | 5.35 |
| PM ₁₀ “AE” | 0.763 | 5.76 | 0.749 | 5.59 |
| Bin 1 | 0.776 | 5.59 | 0.761 | 5.45 |
| Bin 2 | 0.768 | 5.70 | 0.757 | 5.49 |
| Bin 3 | 0.766 | 5.72 | 0.753 | 5.54 |
| Bin 4 | 0.760 | 5.80 | 0.744 | 5.64 |
| Bin 5 | 0.758 | 5.81 | 0.745 | 5.63 |
| Bin 1–bin 2 | 0.779 | 5.56 | 0.762 | 5.44 |
| Bin 1–bin 3 | 0.776 | 5.60 | 0.761 | 5.46 |
| Bin 1–bin 4 | 0.776 | 5.60 | 0.761 | 5.45 |
| Bin 1–bin 5 | 0.776 | 5.60 | 0.761 | 5.45 |

Table 2 Results of multiple regression for training set and validation set for different types of predictors in regression models

| Model structure | Predictors in model | Number of predictors | Training set | | Validation set | |
|---|---------------------|----------------------|--------------|-----------------------------------|----------------|-----------------------------------|
| | | | R^2 | RMSE ($\mu\text{g}/\text{m}^3$) | R^2 | RMSE ($\mu\text{g}/\text{m}^3$) |
| PM ₁ “FE”, PM _{2.5} “FE”, PM ₁₀ “FE”, PM ₁ “AE”, PM _{2.5} “AE”, PM ₁₀ “AE”, bin 1–bin 2, bin 1–bin 3, bin 1–bin 4, bin 1–bin 5 | | 10 | 0.853 | 4.54 | 0.837 | 4.50 |
| PM ₁ “FE”, PM _{2.5} “FE”, PM ₁₀ “FE”, PM ₁ “AE”, PM _{2.5} “AE”, PM ₁₀ “AE”, bin 1–bin 2, bin 1–bin 3, bin 1–bin 4, bin 1–bin 5, temperature, RH | | 12 | 0.874 | 4.20 | 0.861 | 4.17 |

4.4 Stepwise regression results and selection of the best subset of variables

Stepwise regression algorithm was utilized for dataset with 12 variables: all types of mass concentration, bin differences and both additional environmental factors: temperature and RH. The algorithm performed 12 steps, resulting in a model with eight independent variables and an intercept. The following predictors were chosen by this algorithm: PM₁₀ "FE", PM₁ "AE", PM_{2.5} "AE", PM₁₀ "AE", bin 1–bin 3, bin 1–bin 5 and also temperature and RH. The value of R² did not significantly decreased as compared to previously described equation and was at the level of 0.874 for testing set and 0.860 for validation set. The change in RMSE was unnoticeable. Regression coefficients for that model are given in Table S3 in Supplementary Material.

The results of stepwise selection were compared to the selection based on all possible regressions and information criteria: AIC and BIC (Table 3). Generally, all selection methods gave similar results in terms of goodness-of-fit indicators. The choice based on AIC criterion gave model with the lowest error for validation set, but that model consisted of the largest number of predictors (10). BIC criterion pointed to more truncated model with seven variables. It should be noted that BIC value from stepwise algorithm was only slightly higher than that model and R² and RMSE for both equations were practically the same. The other important issue is that all of the presented models did not include the raw PM_{2.5} value with "FE" factor, but include "AE" mass concentrations. Moreover, temperature was also included to those models with relative humidity.

4.5 Neural network results

Neural network was created with inputs in form of all types of mass concentration, bins, bin differences and temperature and RH (17 variables). The training of neural network took 31 epochs, and the results of fitting to TEOM data are presented in Table 4. This algorithm gave the best results in terms of values of R² and RMSE, when compared to other

Table 4 Results of neural network fitting performance for training set and validation set

| Model structure | Training set | | Validation set | |
|-----------------|----------------|---------------------------|----------------|---------------------------|
| | R ² | RMSE (µg/m ³) | R ² | RMSE (µg/m ³) |
| 17 inputs | 0.901 | 3.73 | 0.881 | 3.84 |

methods of fitting. In training set, the R² value exceeded 0.9 and good approximation was observed also for validation set (R² ≈ 0.88), pointing to satisfying generalization capabilities of that structure. RMSE was below 4 µg/m³ in case of both datasets.

4.6 Comparison of regression methods

Figure 2 presents the comparison of fitting possibilities of four developed models: (a) univariate regression model with PM₁ "FE" data, (b) multiple regression model with 12 variables: mass concentrations, bin differences, temperature and RH, (c) multiple regression model from stepwise selection with eight variables and (d) neural network with 17 inputs: mass concentrations, bins, bin differences, temperature and RH. The goodness of fit for presented models was good (R² > 0.8) or very good (R² > 0.9). Neural network gave overall better fitting results for training and validation data, but some deviation from ideal relationship was observed above 100 µg/m³ (Fig. 2d). The linearity of outputs was the highest for multiple regression models (Fig. 2b and c), and in the case of fitting with only one predictor (Fig. 2a), the performance of data adjustment might be improved by means of some nonlinear equation.

In addition, it has been observed that all models were characterized with some larger data scatter for concentration range below ~ 30 µg/m³. The reason for that dispersion may not necessarily derive from the sensor operation, but may arise from the performance of TEOM analyser, as discussed in [27].

Additional evaluation of developed models was made for measurement results from the two other PMS7003

Table 3 Results of choice of the best multiple regression models. The first model was chosen with the stepwise regression algorithm, the second was chosen on the basis of AIC criterion, and the third was chosen on the basis of BIC criterion

| Model structure | Predictors in model | Number of predictors | Training set | | Validation set | | Criteria | |
|--|---------------------|----------------------|----------------|-------------------------|----------------|-------------------------|----------|-----|
| | | | R ² | RMSE, µg/m ³ | R ² | RMSE, µg/m ³ | AIC | BIC |
| PM ₁₀ "FE", PM ₁ "AE", PM _{2.5} "AE", PM ₁₀ "AE", bin 1–bin 3, bin 1–bin 5, temperature, RH | 8 | 0.874 | 4.20 | 0.860 | 4.17 | 30660 | 30719 | |
| PM ₁ "FE", PM ₁₀ "FE", PM ₁ "AE", PM _{2.5} "AE", PM ₁₀ "AE", bin 1–bin 2, bin 1–bin 3, bin 1–bin 5, temperature, RH | 10 | 0.874 | 4.20 | 0.861 | 4.16 | 30656 | 30728 | |
| PM ₁₀ "FE", PM _{2.5} "AE", PM ₁₀ "AE", bin 1–bin 3, bin 1–bin 5, temperature, RH | 7 | 0.874 | 4.20 | 0.860 | 4.17 | 30666 | 30718 | |

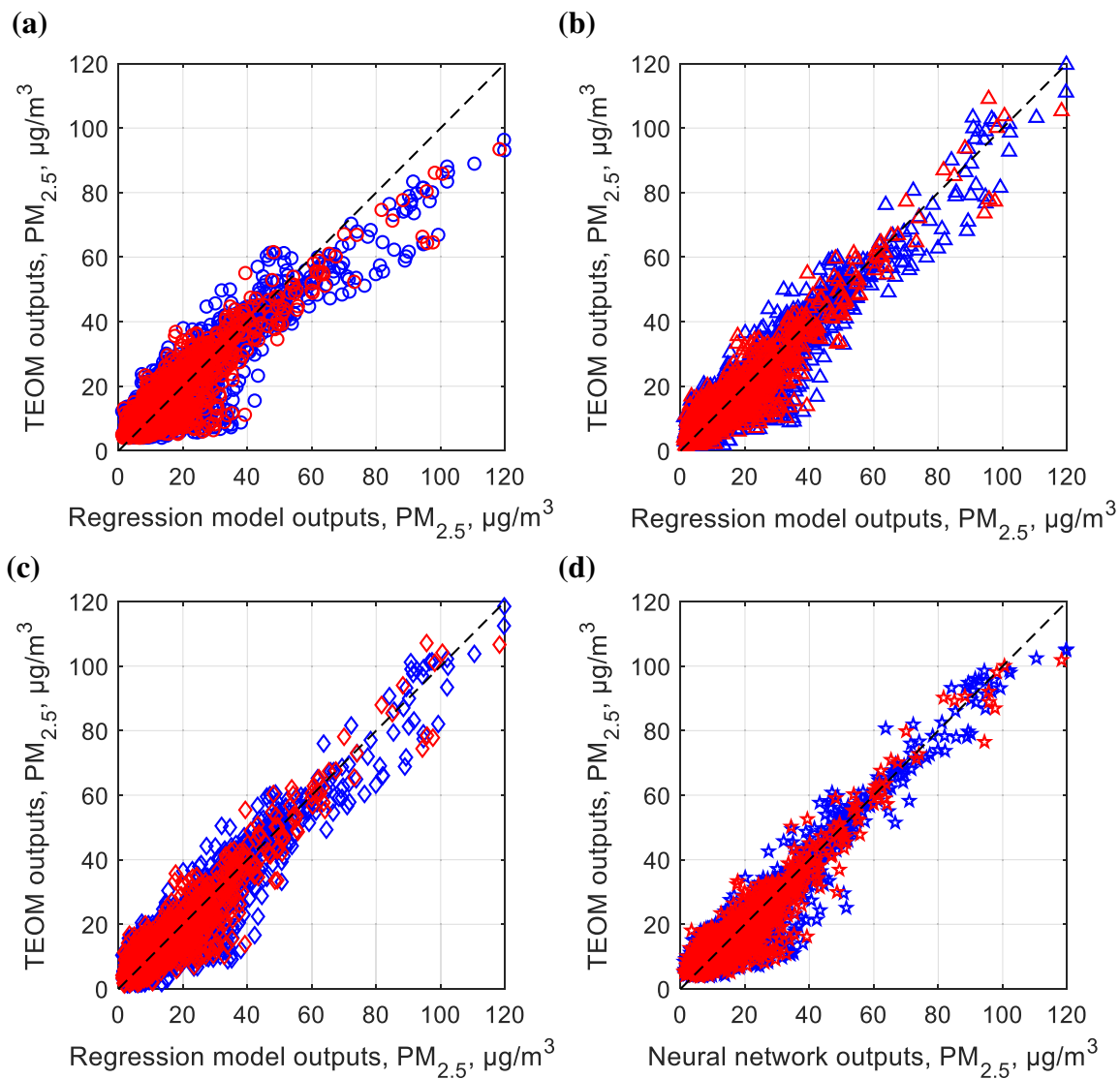


Fig. 2 Comparison of fitting possibilities for: **a** univariate regression model, **b** multiple regression model with 12 variables, **c** multiple regression model from stepwise selection with eight variables,

d neural network with 17 inputs. Red indicators refer to data from training set and blue indicators refer to validation set. The black dashed line indicates the ideal (1:1) relationship

units. That test has been carried out to examine the generalization possibilities of described algorithms. The results of comparison made on full datasets are presented

in Table 5. The simplest model with only one variable was fitted well to signals from units no. 2 and also no. 1 ($R^2 \approx 0.81$, $RMSE \approx 5.1 \mu\text{g}/\text{m}^3$), but was characterized with

Table 5 Results of comparison of data adjustment methods performed on full datasets from three units of PMS7003 sensor

| Model structure | Unit 1 | | Unit 2 | | Unit 3 | |
|---|--------|-----------------------------------|--------|-----------------------------------|--------|-----------------------------------|
| | R^2 | RMSE ($\mu\text{g}/\text{m}^3$) | R^2 | RMSE ($\mu\text{g}/\text{m}^3$) | R^2 | RMSE ($\mu\text{g}/\text{m}^3$) |
| Univariate | 0.807 | 5.11 | 0.811 | 5.06 | 0.769 | 5.59 |
| Multi-variable with 12 variables | 0.858 | 4.39 | 0.870 | 4.19 | 0.834 | 4.73 |
| Multi-variable with eight variables (from stepwise selection) | 0.856 | 4.42 | 0.870 | 4.19 | 0.845 | 4.58 |
| Neural network with 17 inputs | 0.815 | 5.01 | 0.895 | 3.76 | 0.667 | 6.71 |

greater error when signals from unit no. 3 were taken into account ($R^2 \approx 0.77$, $RMSE \approx 5.6 \mu\text{g}/\text{m}^3$). The similar situation was observed for neural network—that structure manifested the smallest error for dataset from sensor no. 2 ($R^2 \approx 0.9$, $RMSE < 4 \mu\text{g}/\text{m}^3$) and good fitting to unit no. 1 data ($R^2 \approx 0.81$, $RMSE \approx 5.0 \mu\text{g}/\text{m}^3$), but the performance on dataset from unit no. 3 was considerably worse ($R^2 \approx 0.67$, $RMSE \approx 6.7 \mu\text{g}/\text{m}^3$). It might be thought that the trained NN structure was overfitted to unit no. 2 signals and functioned still well with comparable data from sensors no. 1, but the dissimilar outputs from unit no. 3 have resulted in higher inaccuracy.

Such behaviour was not noticed for equations from multiple regression—in particular, the model selected by the stepwise regression was robust to slightly different data from sensor no. 3. The R^2 value was in the range 0.85–0.87 and RMSE reached 4.2–4.6 $\mu\text{g}/\text{m}^3$ when all units of PMS7003 were considered.

5 Conclusions

The results of the 1-year collocation study confirmed that low-cost optical sensors may be a useful tool for indicative monitoring of $\text{PM}_{2.5}$ changes in the ambient air. In particular, the sensors like PMS7003 from Plantower could be used in nodes of widely dispersed networks, because of the high reproducibility between units. In such case, calibration equations developed for one unit might be used for others, with a negligible loss of accuracy.

In this paper, different calibration equations were evaluated. The results of univariate regression showed that the raw sensor outputs dedicated to $\text{PM}_{2.5}$ mass concentration do not have to be the best option to establish relationship with the control instrument at all. Regarding the PMS7003, the output for PM_1 was better in terms of higher R^2 value and lower RMSE than $\text{PM}_{2.5}$ output.

The fitting quality might be improved when multiple regression is taken into account. A set of outputs from PMS7003 (mass concentrations, number of particles in bins or bin differences) may be used to construct more complex models, more suited to reference data. The inclusion of additional variables, like temperature and relative humidity, could also be beneficial. Furthermore, the results of conducted comparison demonstrated that stepwise regression might be used to select models that represent the compromise between simplicity and accuracy of fit. What is important, the selected models did not contain the raw $\text{PM}_{2.5}$ output.

Regression models were also contrasted with neural network. That algorithm proved to be very effective in adjustment of signals from sensor selected for training. Nevertheless, it was more susceptible to deviations in data

from other units. That problem was not so acute in case of regression models.

Overall, the study showed that raw signals from low-cost sensors have to be adjusted to obtain outputs matched with reference devices. In such case, regression analysis can support the development of calibration equation for data processing module of the sensor system.

Acknowledgements The work was partially financed by the European Union and co-financed by NFOSiGW within LIFE + Program, Project LIFE-APIS/PL-Air Pollution and biometeorological forecast and Information System (LIFE12 ENV/PL/000056).

Funding This work was co-financed within the specific subsidies granted for the Faculty of Environmental Engineering, Wrocław University of Science and Technology, by the Ministry of Science and Higher Education (Statutory Project Nos. 0402/0130/17, 0401/0055/18).

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

1. Hagler GSW, Solomon PA, Hunt SW (2014) New technology for low-cost, real-time air monitoring. EM January 2014, pp 6–9
2. Snyder EG, Watkins TH, Solomon PA, Thoma ED, Williams RW, Hagler GSW, Shelow D, Hindin DA, Kilaru VJ, Preuss PW (2013) The changing paradigm of air pollution monitoring. Environ Sci Technol 47(20):11369–11377. <https://doi.org/10.1021/es4022602>
3. Rai AC, Kumar P, Pilla F, Skouloudis AN, Di Sabatino S, Ratti C, Yasar A, Rickerby D (2017) End-user perspective of low-cost sensors for outdoor air pollution monitoring. Sci Total Environ 607–608:691–705. <https://doi.org/10.1016/j.scitotenv.2017.06.266>
4. World Meteorological Organization (2018) Low-cost sensors for the measurement of atmospheric composition: overview of topic and future applications. WMO-No. 1215
5. 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. <https://doi.org/10.1016/j.envint.2014.11.019>
6. Budde M, Zhang L, Beigl M (2014) Distributed, low-cost particulate matter sensing: scenarios, challenges, approaches. In: Proceedings of the 1st international conference on atmospheric dust; Castellana Marina, Italy, 1–6 June 2014, pp 230–236
7. Mannshardt E, Benedict K, Jenkins S, Keating M, Mintz D, Stone S, Wayland R (2017) Analysis of short-term ozone and $\text{PM}_{2.5}$ measurements: characteristics and relationships for air sensor

- messaging. *J Air Waste Manag Assoc* 67(4):462–474. <https://doi.org/10.1080/10962247.2016.1251995>
8. Ly BT, Matsumi Y, Nakayama T, Nghiem DT (2018) Characterizing PM_{2.5} in Hanoi with new high temporal resolution sensor. *Aerosol Air Qual Res* 18:2487–2497. <https://doi.org/10.4209/aaqr.2017.10.0435>
 9. Budde M, Köpke M, Beigl M (2016) Design of a light-scattering particle sensor for citizen science air quality monitoring with smartphones: tradeoffs and experiences. In: ProScience 3: conference proceedings: 2nd international conference on atmospheric dust—DUST2016, pp 13–20
 10. Curto A, Donaire-Gonzalez D, Barrera-Gómez J, Marshall JD, Nieuwenhuijsen MJ, Wellenius GA, Tonne C (2018) Performance of low-cost monitors to assess household air pollution. *Environ Res* 163:53–63. <https://doi.org/10.1016/j.scitotenv.2017.11.275>
 11. Jerrett M, Donaire-Gonzalez D, Popoola O, Jones R, Cohen RC, Almanza E, de Nazelle A, Mead I, Carrasco-Turigas G, Cole-Hunter T, Triguero-Mas M, Seto E, Nieuwenhuijsen M (2017) Validating novel air pollution sensors to improve exposure estimates for epidemiological analyses and citizen science. *Environ Res* 158:286–294. <https://doi.org/10.1016/j.envres.2017.04.023>
 12. Morawska L, Thai PK, Liu X, Asumadu-Sakyi A, Ayoko G, Bartonova A, Bedini A, Chai F, Christensen B, Dunbabin M, Gao J, Hagler GSW, Jayaratne R, Kumar P, Lau AKH, Louie PKK, Mazaheri M, Ning Z, Motta N, Mullins B, Rahman MM, Ristovski Z, Shafiei M, Tjondronegoro D, Westerdaal D, Williams R (2018) Applications of low-cost sensing technologies for air quality monitoring and exposure assessment: how far have they gone? *Environ Int* 116:286–299. <https://doi.org/10.1016/j.envint.2018.04.018>
 13. Steinle S, Reis S, Sabel CE, Semple S, Twigg MM, Braban CF, Leeson SR, Heal MR, Harrison D, Lin C, Wu H (2015) Personal exposure monitoring of PM_{2.5} in indoor and outdoor microenvironments. *Sci Total Environ* 508:383–394. <https://doi.org/10.1016/j.scitotenv.2014.12.003>
 14. Mazaheri M, Clifford S, Yeganeh B, Viana M, Rizza V, Flament R, Buonanno G, Morawska L (2018) Investigations into factors affecting personal exposure to particles in urban microenvironments using low-cost sensors. *Environ Int* 120:496–504. <https://doi.org/10.1016/j.envint.2018.08.033>
 15. Kumar P, Skouloudis AN, Bell M, Viana M, Carotta MC, Biskos G, Morawska L (2016) Real-time sensors for indoor air monitoring and challenges ahead in deploying them to urban buildings. *Sci Total Environ* 560–561:150–159. <https://doi.org/10.1016/j.scitotenv.2016.04.032>
 16. Kumar P, Martani C, Morawska L, Norford L, Choudhary R, Bell M, Leach M (2016) Indoor air quality and energy management through real-time sensing in commercial buildings. *Energy Build* 111:145–153. <https://doi.org/10.1016/j.enbuild.2015.11.037>
 17. Tiele A, Esfahani S, Covington J (2018) Design and development of a low-cost, portable monitoring device for indoor environment quality. *J Sens.*, Article ID 5353816. <https://doi.org/10.1155/2018/5353816>
 18. Saad SM, Andrew AM, Shakaff AY, Saad AR, Kamarudin AM, Zakaria A (2015) Classifying sources influencing Indoor Air Quality (IAQ) using Artificial Neural Network (ANN). *Sensors-Basel* 15(5):11665–11684. <https://doi.org/10.3390/s150511665>
 19. Szczurek A, Dolega A, Maciejewska M (2018) Profile of occupant activity impact on indoor air—method of its determination. *Energy Build* 158:1564–1575. <https://doi.org/10.1016/j.enbuild.2017.11.052>
 20. Manikonda A, Zíková N, Hopke PK, Ferro AR (2016) Laboratory assessment of low-cost PM monitors. *J Aerosol Sci* 102:29–40. <https://doi.org/10.1016/j.jaerosci.2016.08.010>
 21. Sousan S, Koehler K, Hallett L, Peters TM (2017) Evaluation of consumer monitors to measure particulate matter. *J Aerosol Sci* 107:123–133. <https://doi.org/10.1016/j.jaerosci.2017.02.013>
 22. McKercher GR, Salmond JA, Vanos JK (2017) Characteristics and applications of small, portable gaseous air pollution monitors. *Environ Pollut* 223:102–110. <https://doi.org/10.1016/j.envpol.2016.12.045>
 23. Castell N, Dauge FR, 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. <https://doi.org/10.1016/j.envint.2016.12.007>
 24. Popoola OAM, Carruthers D, Lad C, Bright VB, Mead MI, Stettler MEJ, Saffell JR, Jones RL (2018) Use of networks of low cost air quality sensors to quantify air quality in urban settings. *Atmos Environ* 194:58–70. <https://doi.org/10.1016/j.atmosenv.2018.09.030>
 25. Gao M, Cao J, Seto E (2015) A distributed network of low-cost continuous reading sensors to measure spatiotemporal variations of PM_{2.5} in Xi'an, China. *Environ Pollut* 199:56–65. <https://doi.org/10.1016/j.envpol.2015.01.013>
 26. Moltchanov S, Levy I, Etzion Y, Lerner U, Fishbain B (2015) On the feasibility of measuring urban air pollution by wireless distributed sensor networks. *Sci Total Environ* 502:537–547. <https://doi.org/10.1016/j.scitotenv.2014.09.059>
 27. Badura M, Batog P, Drzeniecka-Osiadacz A, Modzel P (2018) Evaluation of low-cost sensors for ambient PM_{2.5} monitoring. *J Sens.*, Article ID 5096540. <https://doi.org/10.1155/2018/5096540>
 28. Lin C, Gillespie J, Schuder MD, Duberstein W, Beverland IJ, Heal MR (2015) Evaluation and calibration of Aeroqual series 500 portable gas sensors for accurate measurement of ambient ozone and nitrogen dioxide. *Atmos Environ* 100:111–116. <https://doi.org/10.1016/j.atmosenv.2014.11.002>
 29. Masey N, Gillespie J, Ezani E, Lin C, Wu H, Ferguson NS, Hamilton S, Heal MR, Beverland IJ (2018) Temporal changes in field calibration relationships for Aeroqual S500 O₃ and NO₂ sensor-based monitors. *Sens Actuators B Chem* 273:1800–1806. <https://doi.org/10.1016/j.snb.2018.07.087>
 30. Masiol M, Squizzato S, Chalupa D, Rich DQ, Hopke PK (2018) Evaluation and field calibration of a low-cost ozone monitor at a regulatory urban monitoring station. *Aerosol Air Qual Res* 18:2029–2037. <https://doi.org/10.4209/aaqr.2018.02.0056>
 31. Lin C, Masey N, Wu H, Jackson M, Carruthers DJ, Reis S, Doherty RM, Beverland IJ, Heal MR (2017) Practical field calibration of portable monitors for mobile measurements of multiple air pollutants. *Atmosphere-Basel* 8(12):231. <https://doi.org/10.3390/atmos8120231>
 32. Hagler GSW, Williams R, Papapostolou V, Polidori A (2018) Air quality sensors and data adjustment algorithms: when is it no longer a measurement? *Environ Sci Technol* 52(10):5530–5531. <https://doi.org/10.1021/acs.est.8b01826>
 33. Nakayama T, Matsumi Y, Kawahito K, Watabe Y (2018) Development and evaluation of a palm-sized optical sensor. *Aerosol Sci Technol* 52(2):2–12. <https://doi.org/10.1080/02786826.2017.1375078>
 34. Spinelle L, Gerboles M, Villani MG, Aleixandre M, Bonavitacola F (2017) Field calibration of a cluster of low-cost available sensors for air quality monitoring. Part A: ozone and nitrogen dioxide. *Sens. Actuators B Chem* 215:249–257. <https://doi.org/10.1016/j.snb.2015.03.031>
 35. Kelly KE, 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. <https://doi.org/10.1016/j.envpol.2016.12.039>
 36. Zimmerman N, Presto AA, Kumar SPN, Gu J, Hauryliuk A, Robinson ES, Robinson AL, Subramanian R (2018) A machine learning calibration model using random forests to improve sensor

- performance for lower-cost air quality monitoring. *Atmos Meas Tech* 11:291–313. <https://doi.org/10.5194/amt-11-291-2018>
37. Liu H-Y, Schneider P, Haugen R, Vogt M (2019) Performance assessment of a low-cost PM_{2.5} sensor for a near four-month period in Oslo, Norway. *Atmosphere-Basel* 10(2):41. <https://doi.org/10.3390/atmos10020041>
 38. Patashnick H, Rupprecht EG (1991) Continuous PM-10 measurements using the tapered element oscillating microbalance. *J Air Waste Manag Assoc* 41(8):1079–1083. <https://doi.org/10.1080/10473289.1991.10466903>
 39. Johnson KK, Bergin MH, Russell AG, Hagler GSW (2016) Using low cost sensors to measure ambient particulate matter concentrations and on-road emissions factors. *Atmos Meas Tech*. <https://doi.org/10.5194/amt-2015-331>
 40. Johnson KK, Bergin MH, Russell AG, Hagler GSW (2018) Field test of several low-cost particulate matter sensors in high and low concentration urban environments. *Aerosol Air Qual Res* 18:565–578. <https://doi.org/10.4209/aaqr.2017.10.0418>
 41. European Committee for Standardization (2017) EN 16450:2017, ambient air—automated measuring systems for the measurement of the concentration of particulate matter (PM₁₀; PM_{2.5})
 42. Sousan S, Koehler K, Thomas G, Park JH, Hillman M, Halterman A, Peters TM (2016) Inter-comparison of low-cost sensors for measuring the mass concentration of occupational aerosols. *Aerosol Sci Technol* 50(5):462–473. <https://doi.org/10.1080/02786826.2016.1162901>
 43. Rawlings JO, Pantula SG, Dickey DA (1998) *Applied regression analysis: a research tool*. Springer, New York
 44. Naes T, Isaksson T, Fearn T, Davies T (2002) *A user friendly guide to multivariate calibration and classification*. NIR Publications, Chichester
 45. Zhang GP (2010) Neural networks for data mining. In: Maimon O, Rokach L (eds) *Data mining and knowledge discovery handbook*, 2nd edn. Springer, New York, pp 419–444
 46. Badura M, Szczurek A, Szczówka P (2013) Statistical assessment of quantification methods used in gas sensor system. *Sens Actuators B Chem* 188:815–823. <https://doi.org/10.1016/j.snb.2013.07.105>
 47. Li X, Chen X, Yuan X, Zeng G, León T, Liang J, Chen G, Yuan X (2017) Characteristics of particulate pollution (PM_{2.5} and PM₁₀) and their spacescale-dependent relationships with meteorological elements in China. *Sustainability* 9(12):2330. <https://doi.org/10.3390/su9122330>
 48. Munir S (2017) Analysing temporal trends in the ratios of PM_{2.5}/PM₁₀ in the UK. *Aerosol Air Qual Res* 17:34–48. <https://doi.org/10.4209/aaqr.2016.02.0081>
 49. Wanjura JD, Shaw BW, Parnell CB Jr., Lacey RE, Capareda SC (2008) Comparison of continuous monitor (TEOM) and gravimetric sampler particulate matter concentrations. *Trans ASABE* 51(1):251–257. <https://doi.org/10.13031/2013.24218>
 50. Allen G, Sioutas C, Koutrakis P, Reiss R, Lurmann FW, Roberts PT (1997) Evaluation of the TEOM[®] method for measurement of ambient particulate mass in urban areas. *J Air Waste Manag Assoc* 47(6):682–689. <https://doi.org/10.1080/10473289.1997.10463923>
 51. Tortajada-Genaro LA, Borrás E (2011) Temperature effect of tapered element oscillating microbalance (TEOM) system measuring semi-volatile organic particulate matter. *J Environ Monit* 13(4):1017–1026. <https://doi.org/10.1039/c0em00451k>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.