# LOW-PROCESSING DATA ENRICHMENT AND CALIBRATION FOR $PM_{2.5}$ LOW-COST SENSORS

by

***Danka B. STOJANOVIĆ[a*], Duška N. KLEUT[a], Miloš D. DAVIDOVIĆ[a], Saverio DE VITO[c], Milena V. JOVASEVIĆ-STOJANOVIĆ[a], Alena BARTONOVA[b], and Jean-Marie LEPIOUFLE[b]***

[a]Vinca Institute of Nuclear Sciences, National Institute of the Republic of Serbia, University of Belgrade, Belgrade, Serbia
[b]NILU—Norwegian Institute for Air Research, Kjeller, Norway
[c]ENEA – Agenzia per le Nuove Tecnologie, l' Energia e lo Sviluppo Economico Sostenibile, C.R. Portici, P.le E. Fermi, Portici, Naples, Italy

Original scientific paper
https://doi.org/10.2298/TSCI221109221S

*Particulate matter (PM) in air has been proven to be hazardous to human health. Here we focused on analysis of PM data we obtained from the same campaign which was presented in our previous study. Multivariate linear and random forest models were used for the calibration and analysis. In our linear regression model the inputs were PM, temperature and humidity measured with low-cost sensors, and the target was the reference PM measurements obtained from SEPA in the same timeframe.*

Key words*: $PM_{2.5}$, low-cost sensors, data enrichment, calibration, low-processing*

## Introduction

Pollutant concentrations in ambient air have been measured and regulated for many decades now, to prevent negative effects of air pollution to health and the environment [1]. Evidence on adverse effects of ambient air pollution has been mounting steadily, and public interest in the quality of the air we breathe has been on the rise especially since the outbreak of COVID-19 pandemic [2, 3]. While until recently, monitoring of air quality has been done by professional agencies, today, the availability of portable, low cost microsensor devices and the exponential growth of Internet of Things (IoT) in everyday life has enabled widespread monitoring of air quality also by lay people [4, 5]. This development raises a number of technical and scientific challenges, among them, comparability, accuracy and repeatability of measurements, but also provides a number of opportunities to obtain data from locations that are important for human exposure but not well assessed with traditional methods.

A number of new low-cost devices for monitoring air quality are commercially available. They are equipped with multiple sensors for measuring the concentration of pollu-

*Corresponding author, e-mail: dankas@vin.bg.ac.rs

tant gases (CO, NO, $NO_2$, $SO_2$) as well as PM, air pressure, temperature, humidity. Low-cost sensors measuring PM mostly use light scattering principle. However, the light scattering method is sensitive to external environment and in particular, to meteorological conditions (especially to temperature, humidity). This can lead to accuracy and stability drawbacks [6-9]. For instance, PM low-cost sensors often report PM concentration increasing in presence of water droplets in humid meteorological conditions.

Field calibration is an approach to correct the signal of low-cost sensors for the influences of their interfering environment [10-12]. This technique is based on models such as multi-linear or machine learning regression, using the measurements of the low-cost sensor and variables from external environment as features and monitoring stations as reference [13-17]. To improve the quality of data, the raw signal from the low-cost sensor is post-processed [18]. Removing noise is often the main task in post-processing and data enrichment the second one, missing or incomplete data can be supplemented [19], and existing data can be enhanced either with surrounded pattern [20, 21] or with its decomposition into several components [22]. For data enrichment, to reduce noise and to derive calibration curves, several techniques can be used both at the device level and at a cloud level, including mathematical, statistical or machine learning approaches. The common drawback of data processing using machine learning is the processing requirements for *e.g.*, power consumption or processing time that make them unsuitable for use on individual low-cost platforms. Such in-situ data processing would however alleviate the large-scale system performance by reducing data transfer and connectivity requirements [23].

Could we, by a sequence of low-processing data enrichment and a simple calibration method, reach accuracy as close as a calibration based on machine learning? We introduce the term low-processing as used in [23] as an opposition to heavy computational processing. The main objectives of this paper are:

- to propose low-processing data enrichment for low-cost $PM_{2.5}$ sensor measurement, and
- to test the approach on existing data, comparing a computationally simple and a more complex calibration model.

The results indicate possible future direction for signal low-processing to achieve the required quality of data from low-cost sensor devices monitoring air quality.

## Materials and methods

### *The CITI-SENSE air monitoring campaign in Belgrade, Serbia*

Our data come from campaigns performed in the CITI-SENSE project [24]. The CITI-SENSE explored ways how to increase the involvement of the public in environmental decisions, both directly and through provision of citizen collected data. In the City of Belgrade people participated both in outdoor measurements and in air quality measurements around schools.

In 2015, two CITI-SENSE campaigns were carried out with 25 units of AQ MESH pods equipped with a $PM_{2.5}$ sensor. We co-located these 25 low-cost units to an automatic monitoring station (AMS) at Stari Grad (GPS: 44.82113, 20.45912) which belongs to the State Network run by the Serbian Environmental Protection Agency (SEPA). The area of the study, fig. 1(c), was located about 150 m from a high-traffic arterial road and represented a typical street in the city center. It was frequented by pedestrians and cars, but without public transport, and the local ambient air pollution levels are representative of a residential-traffic area of the Belgrade city center. The AQMESH, fig. 1(d), were located at 20 m from the sampling inlet of the automatic monitoring station.

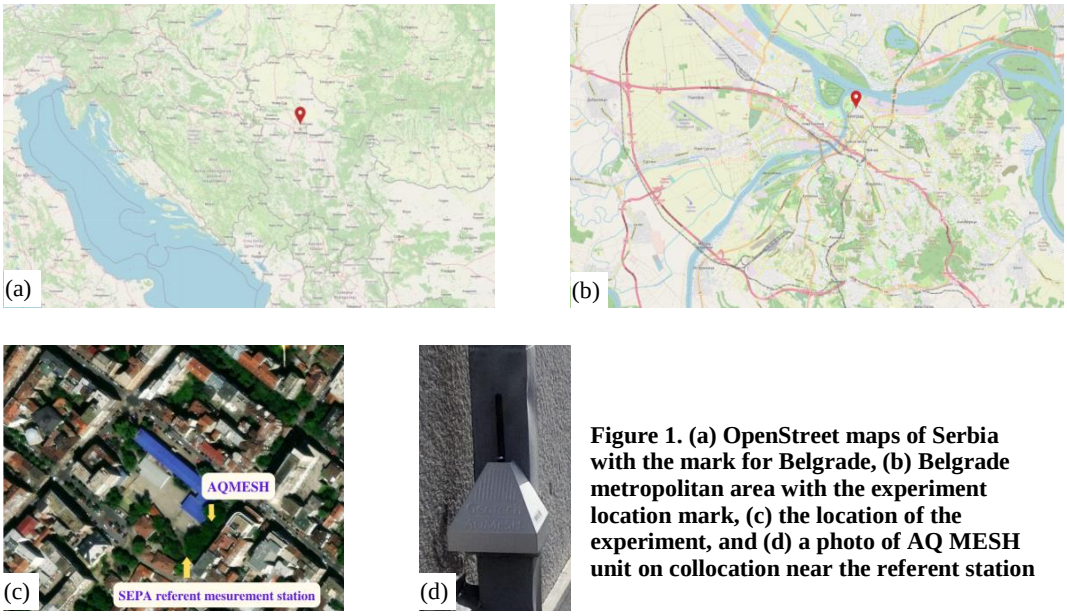


**Figure 1. (a) OpenStreet maps of Serbia with the mark for Belgrade, (b) Belgrade metropolitan area with the experiment location mark, (c) the location of the experiment, and (d) a photo of AQ MESH unit on collocation near the referent station**

### *Data from $PM_{2.5}$ SEPA monitoring station and AQ MESH low-cost sensor devices*

The AMS at Stari Grad performed a continuous measurement of $PM_{2.5}$ with a GRIMM Aerosol Spectrometer EDM 180. This method is an equivalent method for $PM_{2.5\text{-}10}$ and $PM_{2.5}$ [25, 26]. In our study, this station provided the air quality measurements as 1 minute averages. An illustration of its 1 minute $PM_{2.5}$ signal is shown in fig. 2.

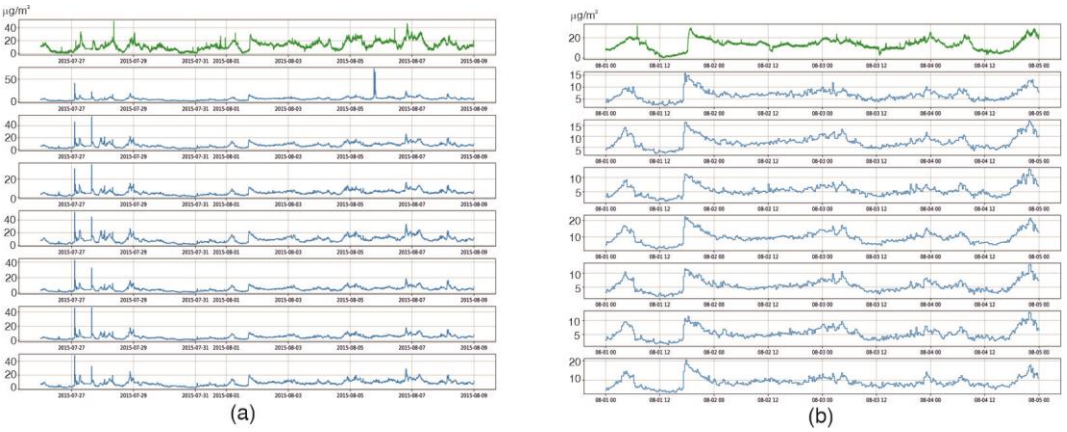


**Figure 2. (a) Timeseries of 1 minute $PM_{2.5}$ from 2015-07-26 00:00 to 2015-08-27 00:00 and (b) timeseries of 1 minute $PM_{2.5}$ for subset between 2015-08-01 00:00 to 2015-08-05 00:00 from the monitoring station (green) and the seven AQMESH low-cost sensors; from top to down: #702150, #716150, #722150, #810150, #839150', #870150, #875150**
(*for color image see journal web site*)

The AQ MESH v. 3.5 units combine new sensor technologies, hardware platform, General Packet Radio Service (GPRS) communications and cloud-based data post processing

in a monitoring unit [27]. The $PM_{2.5}$ value are obtained by the conversion from number to mass concentration, performed by the manufacturer. These PM sensors and their internal pumps were idle most of the time, measuring continuously for one minute every 15 minutes. The AQ MESH measurements of the CITI-SENSE air monitoring campaign in Belgrade, Serbia were presented in our previous paper [17]. For this study, we chose 7 devices that provided the most complete data, device numbers #702150, #716150, #722150, #810150, #839150', #870150, and #875150. An illustration of their 1 minute $PM_{2.5}$ signal is shown in fig. 2(a). A subset for the period 2015-08-01 to 2015-08-05 is presented in fig. 2(b).

### *Data enrichment and model calibration*

We focused our study on using low-processing methods. This section presents a methodology to resample a signal, then to build-up new features describing the concentration of $PM_{2.5}$ from the reference monitoring station, to extract all the relevant information from the signal of the low-cost sensors, and finally to present a simple calibration model.

#### *Supplementing missing values and removal of the highest spikes*

The AQ MESH devices provide a $PM_{2.5}$ value obtained from a measurement of 1 minute duration every 15 minutes. In order to keep data processing to a minimum, we have replaced the missing measurements by their last preceding valid value. The highest spikes from the low-cost $PM_{2.5}$ signal were removed when they were higher than seven times the median of the $PM_{2.5}$ concentration within the CITI-SENSE measurement campaign period. As done previously, these removed values were supplemented by their last preceding valid value.

#### *Encoding periodic time-related features using B-splines*

The 1 minutes $PM_{2.5}$ SEPA monitoring station data exhibit two main pattern, hourly and daily, fig. 3. These patterns are directly connected to the emission strengths of $PM_{2.5}$ and its precursors, and to the diurnal meteorological patterns. Without having any direct quantitative information about $PM_{2.5}$ emission, encoding the periodic pattern of the concentration into features will bring valuable information into calibration models. These features are related to: hours within a day and day of week.

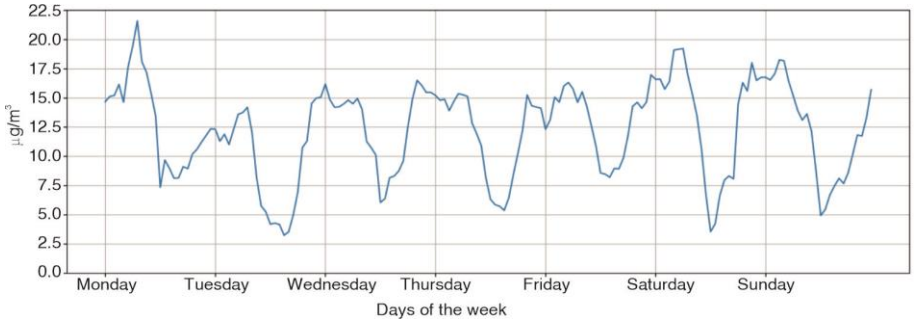


**Figure 3. Hourly and daily patterns of $PM_{2.5}$ concentration at the SEPA monitoring station, Stari Grad**

Hours are naturally described as values from 1 to 24, and days of week as labels from Monday to Sunday, often described as values from 1 to 7 in data science. We paid attention avoiding any jump between the first and the last value of these periodic ranges and avoid-

ing giving more weight to hour 24 than hour 1, or day 7 than day 1. For this purpose, we encoded each of these two periodic ranges into two sets of B-splines [28, 29].

Splines are piecewise polynomials, parametrized by their polynomial degree and the positions of the knots. The periodic encoding function, *f*, of the time-related, *x*, is a spline that reads:

$$f(x) = \sum_{s=1}^{k-1} \beta_s B_s^d(x) \tag{1}$$

where $B_s^d$ is a basis function of polynomial of order $d$, $\beta_s$ – the associated spline coefficients of the $s^{\text{th}}$ spline, and $k$ – the number of knots $\xi$ of the splines.

We used B-splines to encode the periodic ranges. For any $d > 0$, B-spline basis functions of degree $d$ are defined by:

$$B_s^d(x) = \frac{x-\xi_s}{\xi_{s+d}-\xi_s} B_s^{d-1}(x) - \frac{\xi_{s+d+1}-x}{\xi_{s+d+1}-\xi_{s+1}} B_{s+1}^{d-1}(x),\ s = 1 \text{ to } k-1 \tag{2}$$

with the following condition:

$$B_s^0(x) = \begin{cases} 1 & \text{if } \xi_s \le x < \xi_{s+1} \\ 0 & \text{else} \end{cases} \text{ and } B_s^0(x) = 0 \text{ if } \xi_s = \xi_{s+1}$$

In our case, we used B-splines of order 3 and spline coefficients equal to 1. Features of the hours of day are described by a set of 12 splines with 13 knots regularly spread from 0 to 24, and features of the weekdays are described by a set of 3 splines with 4 knots regularly spread from 0 to 7, fig. 4.

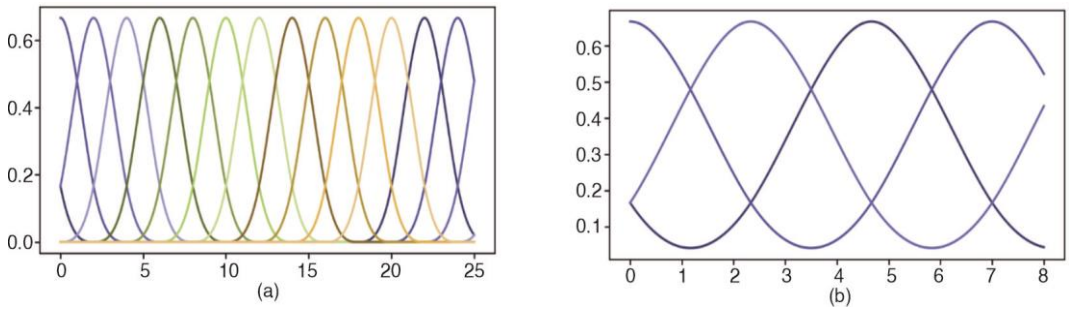


**Figure 4. Encoded periodic features as splines for the hours of day (a) and days of week (b)**

### *Decomposition of the low-cost sensor signal with low-pass Butterworth filter*

The signals from the low-cost sensor were noisy and had unexplained spikes. Nonetheless, they could provide relevant information at different timescales [22]. To increase the amount of information a low-cost sensor signal could give, we decided to extract the signals corresponding to 24 hours, 12 hours, 6 hours, 1 hours, 30 minutes, 15 minutes, and 5 minutes. Extracting such patterns by processing rolling window average would require too much processing. An efficient alternative is the use of low-pass filter for each frequency of interest and recursively subtracting their outcomes to keep the signals of interest. In this study, and because of its low processing requirement, we used the low-pass Butterworth filter [30]. The transfer function of the low-pass Butterworth filter reads:

$$H(\mathrm{s}) = G_0 \prod_{k=1}^{n} \frac{\omega_c}{s - s_k} \tag{3}$$

where $G_0$ is the gain at zero frequency, $s$ – the frequency of the signal, $\omega_c$ – the cutoff frequency, and $n$ – the order of the transfer. In our study, $G_0$ was chosen equal to one and the order $n$ to one as well. The term $s_k$ is called the $k^{\text{th}}$ pole and reads:

$$s_k = \omega_c \mathrm{e}^{\frac{j(2k-1)\pi}{2n}}, \; k = 1, ..., n \tag{4}$$

given the time resolution of signal $f_s$, the choice of the timescale to be cut-off $f_c$ and the Nyquist frequency, the cutoff frequency reads:

$$\omega_c = \frac{f_s}{2 f_c} \tag{5}$$

given the 1 minute time resolution of the low-cost signal, and given the timescale of interest, the timescale cut-off, $f_c$, got the values: 1440, 720, 360, 60, 30, 15 and 5.

Finally, the residual signal at the timescale of interest is retrieved by subtracting to the filtered signals the one at higher timescale. For instance, the residual signal of 720 minutes resulted as a difference between the signal filtered at 720 minutes and the signal filtered at 1440 minutes. An illustration of these seven features for the low-cost station #702150 is presented in fig. 5.

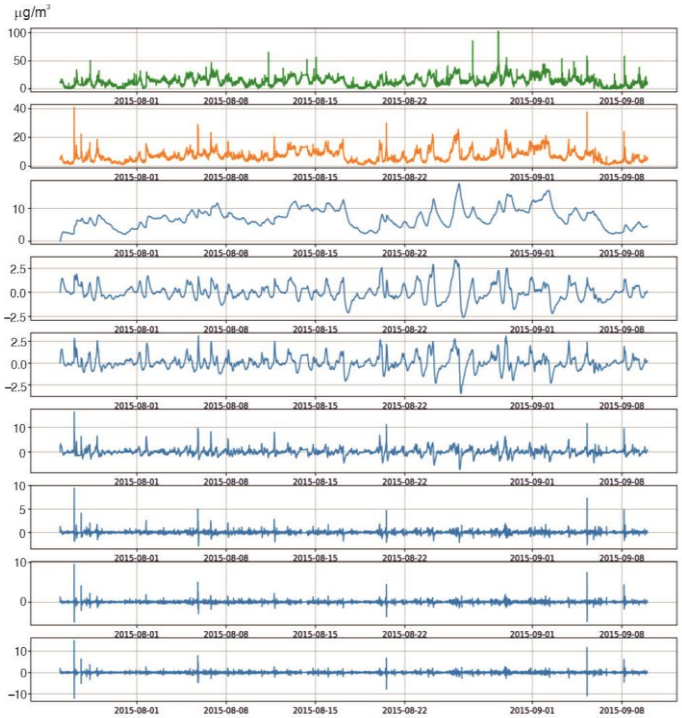


**Figure 5. Filtered signals at a timescale 1 day, and residual signals at 12 hours, 6 hours, 1 hour, 30 minutes, 15 minutes, and 5 minutes for instrument # 702150**

*Calibration with a multivariate linear model*

Calibration based on a multi-linear regression model [31, 32] respects the low-processing condition. The multivariate linear calibration model reads:

$$Y = \beta_0 + \sum_{i=1}^{N} \beta_i X_i \tag{6}$$

where $Y$ is the target, $X_i$ – the features, and $\beta_i$ – the coefficients. In our case the target was the reference monitoring station and the features were the ones described in this section.

### *Experimentation plan*

The dataset between 27/07/2015 and 26/08/2015 was used for the training and the testing of the calibration model; the dataset between 31/08/2015 and 09/09/2015 was used for its validation. The first period was randomly split in two parts with a ratio 60/40 for the training and the testing steps. Several experiments were done in this study to analyse the effects of each data enrichment on the calibration model. In addition to a multivariate linear model, we used Random Forest (RF) [33] to get a comparison. The RF is well known for providing prediction with high accuracy but at a higher processing cost. In total, we performed seven experiments, the overview is presented in tab. 1.

**Table 1. Overview of the seven experimentations**

| Experimentation ID | Data enrichment | Calibration model type |
|---|---|---|
| E1 | S | None |
| E2 | S | ML |
| E3 | S | RF |
| E4 | S + C | ML |
| E5 | S + C | RF |
| E6 | S + C + T | ML |
| E7 | S + C + T | RF |

Note: S – supplementing, C – composition from low-band Butterworth filter, T – Temporal features, ML – Multivariate-linear, and RF – Random Forest

We applied two usual metrics to quantify the accuracy of the output; the root mean square error (RMSE), and the coefficient of determination, $R^2$. The Pearson correlation coefficient is processed to compare our results with previous studies. The experimentations were run on an Intel(R) Core (TM) i7-6600U CPU at 2.60 GHz 2.81 GHz. We measured the time processing for experimentations E2 to E7. We did not implement any CPU parallelisation to ease comparisons in-between experimentations.

## Results and discussion

An illustration of the calibration of E2 to E7 at AQ MESH # 702150 is shown in fig. 6 and the quantile-quantile plot in fig. 7.

Metrics of the seven experiments for the evaluation phase are presented in tab. 2. For both multivariate linear and RF models, increasing the number of features with the composition of signal at different time scale increase most of the time both metrics RMSE and $R^2$. Adding periodic features have not always a positive consequence: for *e.g.*, with node #722150 or # 716150. For experimentation E6 and E7, RF sees its coefficient of determination between 0.57 and 0.78 and multivariate linear model between 0.42 and 0.85. For experimentation E6 and E7, RF sees its RMSE between 4.23 and 5.91 and multivariate linear model between 3.44 and 6.86.

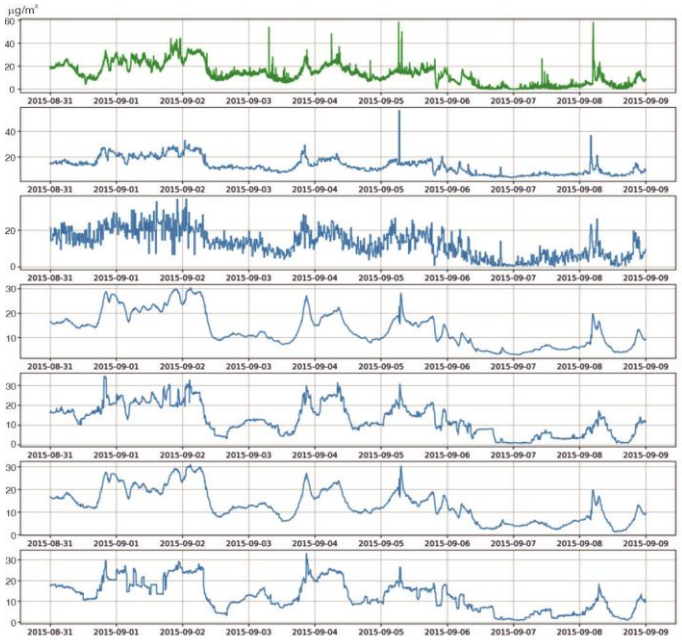

**Figure 6. Timeseries for the period of the evaluation phase for the reference monitoring station (top) and the calibrated AQMESH # 702150 for experiment E2 (second line), E3 (third line), E4 (fourth line), E5 (fifth line), and E6 (sixth line)**

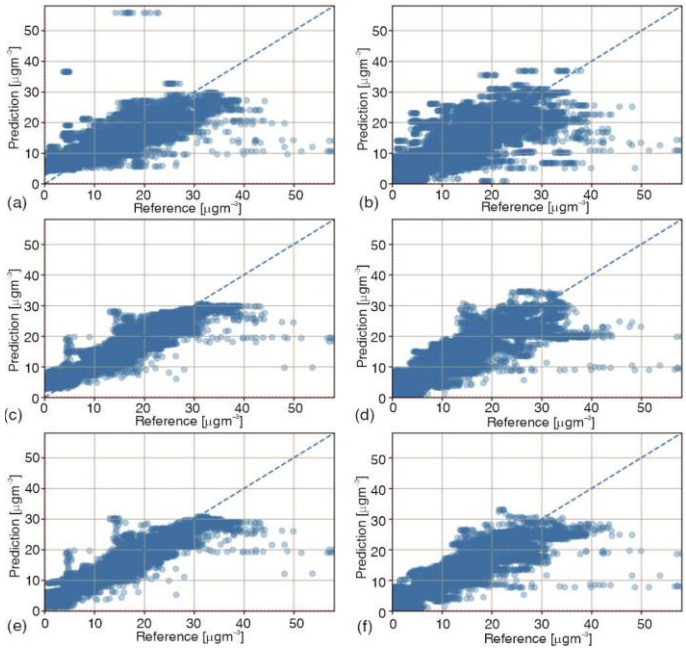


**Figure 7. Quantile-quantile plots of the calibrated node # 702150 against a reference monitoring air quality station; each plot corresponds to one experiment; (a) E2, (b) E3, (c) E4, (d) E5, (e) E6, and (f) E7**

**Table 2. Metrics of the seven experimentations for the seven nodes during the evaluation phase**

| ID | Metrics | E1 | E2 | E3 | E4 | E5 | E6 | E7 |
|---|---|---|---|---|---|---|---|---|
| 702150 | $R^2$ | 0.16 | 0.66 | 0.53 | 0.84 | 0.72 | **0.85** | 0.76 |
| | RMSE | 8.26 | 5.25 | 6.19 | 3.59 | 4.73 | 3.44 | 4.41 |
| | *r* | 0.82 | 0.82 | 0.73 | 0.93 | 0.85 | 0.93 | 0.88 |
| 716150 | $R^2$ | 0.01 | 0.28 | 0.24 | 0.39 | **0.59** | 0.42 | 0.57 |
| | RMSE | 8.97 | 7.64 | 7.84 | 7.02 | 5.76 | 6.86 | 5.91 |
| | *r* | 0.53 | 0.53 | 0.57 | 0.63 | 0.78 | 0.65 | 0.76 |
| 722150 | $R^2$ | -0.19 | 0.39 | 0.42 | 0.53 | 0.60 | 0.52 | **0.63** |
| | RMSE | 9.81 | 7.04 | 6.88 | 6.18 | 5.67 | 6.20 | 5.46 |
| | *r* | 0.66 | 0.66 | 0.68 | 0.76 | 0.77 | 0.73 | 0.8 |
| 810150 | $R^2$ | 0.20 | 0.36 | 0.37 | 0.47 | **0.66** | 0.47 | 0.65 |
| | RMSE | 8.05 | 7.20 | 7.16 | 6.57 | 5.24 | 6.53 | 5.29 |
| | *r* | 0.6 | 0.6 | 0.66 | 0.69 | 0.81 | 0.69 | 0.81 |
| 839150 | $R^2$ | -0.28 | 0.47 | 0.52 | 0.63 | 0.75 | 0.64 | **0.78** |
| | RMSE | 10.19 | 6.57 | 6.25 | 5.48 | 4.49 | 5.36 | 4.23 |
| | *r* | 0.76 | 0.76 | 0.73 | 0.9 | 0.87 | 0.87 | 0.89 |
| 870150 | $R^2$ | -0.41 | 0.42 | 0.46 | 0.56 | **0.70** | 0.58 | 0.64 |
| | RMSE | 10.67 | 6.84 | 6.58 | 5.96 | 4.95 | 5.83 | 5.41 |
| | *r* | 0.78 | 0.78 | 0.72 | 0.91 | 0.88 | 0.88 | 0.84 |
| 875150 | $R^2$ | 0.48 | 0.67 | 0.47 | 0.83 | 0.72 | **0.85** | 0.77 |
| | RMSE | 6.46 | 5.20 | 6.57 | 3.67 | 4.79 | 3.50 | 4.32 |
| | *r* | 0.83 | 0.83 | 0.71 | 0.93 | 0.85 | 0.93 | 0.89 |

Average processing time of the seven experiments for the training phase are presented in tab. 3. Experiments with a multivariate linear based calibration sees its processing time from 0.004 seconds to 0.046 seconds on average, and experiments with a Random Forest based calibration sees its processing time from 5 seconds to 34 seconds on average.

**Table 3. Average processing time of the seven experimentations for the seven nodes during the training phase**

| Experiments | E2 | E3 | E4 | E5 | E6 | E7 |
|---|---|---|---|---|---|---|
| Processing time [s] | 0.004 ±0.007 | 5.000 ±0.415 | 0.015 ±0.00 | 27.845 ±2.302 | 0.046 ±0.022 | 34.587 ±1.269 |

Low-processing data enrichment such as supplementing missing values, encoding periodic time-related features, and making a composition of the initial low-cost signal at different time scale showed convincing results by improving the metrics of calibration with both multi-linear regression and RF regression. Moreover, using data enrichment with a multi-linear regression outperforms RF with no data enrichment.

Without any surprise, for most of the experiments and with identical data enrichment, calibration with RF got better metrics than multi-linear prediction. Nonetheless, it is worth to point out two drawbacks:

– Training a RF regression model is CPU consuming. Comparing experiments E6 and E7 for the training phase, shows that processing of a RF calibration model took 751 more time than a multi-linear calibration model.
– Calibration methods such as RF would only predict values from features it has been trained for and will struggle for phenomena outside of its knowledge [34]. As shown for

AQ MESH sensors #702150 and #875150, experiments with a multi-linear calibration outperform the ones with a RF calibration. Indeed, the period of measurement used for training the calibration model was one month long. This short period did not cover all the complexity a $PM_{2.5}$ concentration phenomena can get all year long.

Previous research [35] studied the basic statistics obtained during the co-location of 24 identical AQ MESH nodes for the period April 13$^{th}$-June 24$^{th}$ 2015, at the reference station of Kirkeveien. The results showed that even for identical sensors and platform, the performance can vary from sensor to sensor. The average $PM_{2.5}$ RMSE for all nodes was 5.57 and a Pearson correlation coefficient, *r*, of 0.51. By comparing these former results with sensor #716150 characterized by a similar Pearson correlation coefficient of 0.53, we show that data enrichment improves this metric to 0.65 with a multi-linear model and to 0.76 with a RF model.

The results of combining data enrichment with calibration demonstrated in this paper provide greater improvement in the metrics, and some solutions to optimize the calibration process considering the CPU processing time and the data complexity. Our future work includes understanding the signal from the low-cost sensor and having a low-cost platform with meteorological sensors. More data from inside the platform might give insight to predict spikes: *e.g*. Voltage at the power supply, intensity at the PM sensor, CPU-processing rate, input-output rate.

## Conclusions

Our objective was to evaluate low-processing data enrichment and calibration of AQ MESH $PM_{2.5}$ low-cost sensors which were co-located with a SEPA reference station. Low-processing data enrichment such as resampling, encoding periodic time-related features and making a composition of the initial low-cost signal at different time scale showed convincing results on calibration both multi-linear and based on a RF regression. Besides in our case where period of measurement is short (one month), combining data enrichment with a multi-linear can outperform RF.

## Acknowledgment

Funding for this work has been provided by the EU H2020 Framework Programme for research and innovation under grant agreement no 952433 (VIDIS); EU 7$^{th}$ Framework Programme for research, technological development and demonstration under grant agreement no 308524 (CITI-SENSE); Ministry of Education, Science and Technological Development of the Republic of Serbia (Grant 451-03-68/2022-14/ 200017). The authors gratefully acknowledge the Serbian Environmental Protection Agency (SEPA) for access to their automatic air-quality monitoring station and data.

## References

[1] ***, WHO global air quality guidelines 2021.
[2] Comunian, S., *et al*., Air Pollution and COVID-19: The Role of Particulate Matter in the Spread and Increase of COVID-19's Morbidity and Mortality, *International Journal Environ. Res. Public Health, 17* (2020.), 4487
[3] Bai, Y,. *et al*., Changes in Stoichiometric Characteristics of Ambient Air Pollutants Pre-To Post-COVID-19 in China, *Environmental Research, 209* (2022), 112806
[4] Yang, C.-T., *et al*., Current Advances and Future Challenges of AIoT Applications in Particulate Matters (PM) Monitoring and Control, *J. Hazard. Mat., 419* (2021), 126442

[5] Jovasevic-Stojanovic, M., *et al*., On the Use of Small and Cheaper Sensors and Devices for Indicative Citizen-Based Monitoring of Respirable Particulate Matter, *Enviromental Pollution, 206-205* (2015), Nov., pp. 696-704
[6] Morawska, L., *et al*., Applications of Low-Cost Sensing Technologies for Air Quality Monitoring and Exposure Assessment: How Far Have they Gone?, *Environment International, 116* (2018), July, pp. 286-299
[7] Cho, H., Baek, Y., Practical Particulate Matter Sensing and Accurate Calibration System Using Low-Cost Commercial Sensor, *Sensors, 21* (2021), 6162
[8] Rai, A. C., *et al*., End-User Perspective of Low-Cost Sensors for Outdoor Air Pollution Monitoring, *Science of the Total Environment, 607-608* (2017), Dec., pp. 607-608
[9] Giordano, M. R., *et al*., From Low-Cost Sensors to High-Quality Data: A Summary of Challenges and Best Practices for Effectively Calibrating Low-Cost Particulate Matter Mass Sensors, *Journal of Aerosol Science, 158* (2021), 105833
[10] Liang, L., Daniels, J., What Influences Low-cost Sensor Data Calibration?- A Systematic Assessment of Algorithms, Duration, and Predictor Selection, *Aerosol and Air Quality Research, 22* (2022), 220076
[11] Liang, L., Calibrating Low-Cost Sensors for Ambient Air Monitoring: Techniques, Trends, and Challenges, *Enviromental Research, 197* (2021), 111163
[12] Wei, P., *et al*., Impact Analysis of Temperature and Humidity Conditions on Electrochemical Sensor Response in Ambient Air Quality Monitoring, *Sensors, 18* (2018), 59
[13] Badura, M., *et al*., Regression Methods in the Calibration of Low-Cost Sensors for Ambient Particulate Matter Measurements, *SN Appl. Sci., 1* (2019), 622
[14] Jiao, W., *et al*., Community Air Sensor Network (CAIRSENSE) Project: Evaluation of Low-Cost Sensor Performance in a Suburban Environment in the Southeastern United States, *Atmos. Meas. Tech., 9* (2016), 11, pp. 5281-5292
[15] Loh, B. G., Choi, G. H., Calibration of Portable Particulate Matter–Monitoring Device Using Web Query and Machine Learning, *Saf. Health Work., 10* (2019), 4, pp. 452-460
[16] Chojer, H., *et al*., Can Data Reliability of Low-Cost Sensor Devices for Indoor Air Particulate Matter Monitoring Be Improved? – An Approach Using Machine Learning, *Atmospheric Environment, 286* (2022), 119251
[17] Topalovic, D., *et al*., 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* (2019), Sept., pp. 640-658
[18] Schneider, P., *et al*., Toward a Unified Terminology of Processing Levels for Low-Cost Air-Quality Sensors(2019) Environ, *Sci. Technol., 53* (2019), 15, pp. 8485-8487
[19] Allen, M., Cervo, D., Multi-Domain Master Data Management: Advanced MDM and Data Governance in Practice, Morgan Kaufmann, Burlington, Mass., USA, 2015
[20] Eamonn, J. K., Pazzani, M. J., An Enhanced Representation of Time Series which Allows Fast and Accurate Classification, Clustering and Relevance Feedback, *KDD-98 Proceedings, 98*(1998), Aug., pp. 239-243
[21] Knapp, E. D., Langill, J., Industrial Network Security: Securing Critical Infrastructure Networks for Smart Grid, SCADA, and other Industrial Control Systems, Syngress, Elsevier, Amsterdam, The Netherlands, 2014
[22] Rhif, M., *et al*.,Wavelet Transform Application for/in Non-Stationary Time-Series Analysis: A Review, *Applied Sciences, 9* (2019), 7, 1345
[23] Wojcikowski, M., *et al*., A Surrogate-Assisted Measurement Correction Method for Accurate and Low-Cost Monitoring of Particulate Matter Pollutants, *Measurement, 200* (2022), 111601
[24] ***, http://co.citi-sense.eu
[25] Gilliam, J., Hall, E., Reference and Equivalent Methods Used to Measure National Ambient Air Quality Standards (NAAQS) Criteria Air Pollutants, vol. I, U.S. Environmental Protection Agency, Washington DC, 2016
[26] Polidori, A., *et al*., *Field Evaluation Aqmesh Monitor (v.4.0)*, South Coast Air Quality Performance Evaluation Center, http://www.aqmd.gov/docs/default-source/aq-spec/field-evaluations/aqmesh-(v-4-0)---field-evaluation.pdf?sfvrsn=10, 2016
[27] ***, AQMESH Technical Specification http://www.AQMESH.com/produt/technical-details/
[28] Eilers, P., Marx, B., Flexible Smoothing with B-Splines and Penalties, *Statist. Sci., 11* (1996), 2, pp. 89-121

[29] Perperoglou, A., *et al*., A Review of Spline Function Procedures in R, *BMC Med Res Methodol, 19* (2019), 46
[30] Butterworth, S., On the Theory of Filter Amplifiers, *Experimental Wireless and the Wireless Engineer, 7* (1930), May, pp. 536-541
[31] Badura, M., *et al*., Regression Methods in the Calibration of Low-Cost Sensors for Ambient Particulate Matter Measurements, *SN Appl. Sci., 1* (2019), 622
[32] Thomas, E.V., Haaland, D. M., Comparison of Multivariate Calibration Methods for Quantitative Spectral Analysis, *Anal. Chem., 62* (1990), 7, pp. 1091-1099
[33] Breiman, L., Random Forests, *Mach. Learn., 45* (2001), Oct., pp. 5-32
[34] Lepioufle, J.-M., *et al*., Error Prediction of Air Quality at Monitoring Stations Using Random Forest in a Total Error Framework, *Sensors, 21*(2021), 2160
[35] Castell, N., *et al*., Can Commercial Low-Cost Sensor Platforms Contribute to Air Quality Monitoring and Exposure Estimates? *Environment International, 99* (2017), Feb., pp. 293-302

Paper submitted: November 11, 2022
Paper revised: December 1, 2022
Paper accepted: December 5, 2022