Improving Data Quality of Low-Cost Light-Scattering PM Sensors: Toward Automatic Air Quality Monitoring in Urban Environments

Gustavo Ramirez-Espinosa, *Graduate Student Member, IEEE*, Pietro Chiavassa, *Graduate Student Member, IEEE*, Edoardo Giusto, *Member, IEEE*, Stefano Quer, *Member, IEEE*, Bartolomeo Montrucchio, *Senior Member, IEEE*, and Maurizio Rebaudengo, *Senior Member, IEEE*

***Abstract*—Low-cost light-scattering particulate matter (PM) sensors are often advocated for the dense monitoring networks. Recent literature has focused on evaluating their performance. Nonetheless, low-cost sensors are also considered unreliable and imprecise. Consequently, exploring techniques for anomaly detection, resilient calibration, and data quality improvement should be discussed more. In this study, we analyse a year-long acquisition campaign by positioning 56 low-cost light-scattering sensors near the inlet of an official PM monitoring station. We use the collected measurements to design and test a data processing pipeline composed of different stages, including fault detection, filtering, outlier removal, and calibration. These can be used in large-scale deployment scenarios where the quantity of sensors' data can be too high to be analysed manually. Our framework also exploits sensor redundancy to improve reliability and accuracy. Our results show that the proposed data processing framework produces more reliable measurements, reduces errors, and increases the correlation with the official reference.**

***Index Terms*—Air monitoring, air quality, light-scattering sensor, particulate matter (PM), sensor calibration.**

I. Introduction

PARTICULATE matter (PM) is a relevant air pollutant with substantial negative impacts on human health. PM comprises solid particles and liquid droplets suspended in the air and can be easily inhaled. PM mass concentrations are usually measured for different particle sizes. Common size classifications are PM_{10}, considering all the particles with a diameter smaller than 10 μm, and $PM_{2.5}$ with a maximum diameter of 2.5 μm.

Conventional approaches to tracking air quality and PM particles are based on the sparse networks of static reference-grade detectors, as specified by the European regulations 50/2008/EC. The high cost of instrumentation has limited the spatial coverage of these networks.

Consequently, there has recently been a significant increase in the development and application of low-cost PM (LCPM) light-scattering sensors. Due to their low price, up to a few hundred dollars [1], and small sizes, they can be used as the building block for creating denser monitoring networks. These sensing nodes can be adopted in city-wide applications [2], [3], on bike sharing fleets [4], in location-aware places [5], inside moving vehicles [6], [7], to monitor the conservation state of historic buildings [8], [9], and to detect fire in forests [10] using novel networking solution for sparsely populated areas [11]. Furthermore, mobile monitoring enables participatory sensing approaches.

However, the accuracy and precision of the LCPM sensors are limited due to the miniaturization of the light-scattering technology. This characteristic is reflected in a reduced percentage of detected particles and smaller size ranges. In addition, the measurement of PM concentrations is affected by multiple approximations and assumptions, such as the refractive index (RI) and particle density, which are unknown a priori. LCPM sensors are also negatively influenced by environmental factors, such as high humidity levels, which cause hygroscopic growth of PM particles. Consequently, the evaluation of their performance is one of the main priorities in current research [12], [13], [14], [15], [16], [17]. However, most of the studies limit their scope to the evaluation of the sensors without proposing techniques to detect anomalies, perform resilient calibration, and improve the quality of the measured data. Therefore, this work presents a novel data processing pipeline to improve the accuracy of $PM_{2.5}$ measurement of LCPM, eliminating the requirement for substantial human intervention.

Manuscript received 17 April 2024; accepted 10 May 2024. Date of publication 27 May 2024; date of current version 23 August 2024. This work was supported in part by the Spoke 9 of the ICSC National Research Centre for High Performance Computing, Big Data and Quantum Computing. *(Corresponding author: Stefano Quer.)*

Gustavo Ramirez-Espinosa is with the Department of Electronis Engineering, Pontificia Universidad Javeriana, Bogotá 110231, Colombia (e-mail: ramirez.g@javeriana.edu.co).

Pietro Chiavassa, Stefano Quer, Bartolomeo Montrucchio, and Maurizio Rebaudengo are with the Department of Control and Computer Engineering, Politecnico di Torino, 10129 Turin, Italy (e-mail: pietro.chiavassa@polito.it; stefano.quer@polito.it; bartolomeo.montrucchio@polito.it; maurizio.rebaudengo@polito.it).

Edoardo Giusto is with the Department of Electrical Engineering and Information Technology, University of Naples Federico II, 80138 Naples, Italy (e-mail: egiusto@ieee.org).

Digital Object Identifier 10.1109/JIOT.2024.3405623

The experiment setup includes 14 monitoring stations containing four LCPM sensors. We positioned these stations near the inlet of an official beta attenuation monitor; we used this official station as a reference. The experiment lasted over a year, with hourly $PM_{2.5}$ measurements performed by the reference instrument and our LCPM sensors collecting estimates every second.

We used the data collected to evaluate the different steps of the data processing pipeline. In the first step, we tuned a simple algorithm to detect the most recurring permanent sensor failures. Then, we tested different filters to mitigate the effects of point anomalies in the sensor readings. After that, we performed an enhanced sensor calibration method using a multivariate linear regression model, considering both PM and relative humidity as independent variables. In this method, we excluded outliers before training the calibration models by modeling the sensor data and the official reference using a multivariate Gaussian model (GM). Finally, we computed the median of each station's sensors to increase each monitoring node's reliability.

Results show that the proposed data processing framework improves the performance of the LCPM with respect to the official reference on multiple metrics, i.e., RMSE, MAE, R^2, and correlation. This improvement holds the potential for implementing appropriate security strategies that ensure data immutability and privacy following the energy constraints [18].

This article is organized as follows. Section II illustrates the previous related work. Section III reports some background information on the low-cost light-scattering technology and the area under consideration. Section IV describes our experimental setting and analyses the collected data. Section V describes the different steps in the proposed data processing framework. In Section VI each step of the pipeline is separately evaluated, whereas in Section VII the global framework performance is discussed. Finally, Section VIII draws the study's main findings.

II. Related Work

Key experts in the area mainly perform anomaly detection in the sensor networks. A comprehensive review of anomaly detection techniques applied to the IoT data is provided by Cook et al. [19]. This article encompasses a broad spectrum of strategies and summarizes the prevailing challenges in the domain. Similarly, Chen et al. [20] introduced an anomaly detection framework engineered for large real-world sensor networks. They initially identify spatiotemporal anomalies and regional emission sources, then proceed to rank sensing devices and subsequently discern malfunctioning devices. The authors show an outstanding capability of their framework to detect outliers and infer anomalous events.

Various studies have contributed to advancing our understanding and methodologies in sensor calibration. Brattich et al. [16] characterized the performances and reproducibility of different types of inexpensive sensors and compare them to reference instruments. Moreover, the authors assess the variability of the different sensors and perform a comparative analysis of the various optical particle counters (OPCs) under different meteorological conditions. Hasenfratz et al. [21] introduced an unique approach to leverage mobile sensor platforms on public transportation in Zurich, Switzerland. In their work, the authors collected ultrafine particle measurements over two years. This endeavor led to the creation of pollution maps and a reduction in spatial errors. Liu et al. [22] conducted calibration on several LCPM sensors, demonstrating the importance of a steady particle mass concentration during the calibration process. Further, Maag et al. [23] provided a comprehensive review of state-of-the-art low-cost air pollution sensors, identifying primary error sources, exploring suitable calibration models, and analysing network recalibration strategies. On a related note, Rumburg et al. [24] delved into regulatory statistics to determine the magnitude of the error when sensors do not perform daily samplings.

The calibration presented in [23] is particularly relevant when researchers adopt the LCPM techniques to monitor air quality. Budde et al. [25] juxtaposed the performance of a high-accuracy measure device with a cheap off-the-shelf sensor combined with a mobile phone. They show the potentiality of inexpensive devices through accurate calibration and a processing procedure adopting multisensor data fusion. Furthermore, Montrucchio et al. [26] presented an outdoor calibration model based on the multivariable linear regressions and evaluated its performance in different urban scenarios. Similarly, Concas et al. [27] outlined how rapidly low-cost sensor technologies are expanding and emphasize the role of machine learning (ML) techniques in sensor calibration. Their work also sheds light on open research challenges and future directions. In a focused case study, Crilley et al. [12] appraised the Alphasense OPC-N2, a low-cost OPC, for monitoring ambient airborne particles in typical urban background sites in the U.K. Their study investigates interunit precision, variation in measured particle mass concentration, and comparison with standard commercial OPCs, thus offering valuable insights into the performance and reliability of low-cost sensors. Several studies investigated the performance of the LCPM sensors using the Honeywell PM sensors. Giusto et al. [28] analysed a particular type of sensor and its coherence under the design of an IoT device. Subsequently, related to energy usage and efficient data acquisition, authors, such as Chiavassa et al. [29] and Espinosa et al. [30] explored various techniques on the same type of device, examining their impact on accuracy and data reduction. Additionally, Mao et al. [31] have introduced a computational offload policy optimization tailored explicitly for IoT.

However, most cited studies only perform sensor calibration and simple data processing to evaluate sensor performance and their dependence on external factors. On the contrary, in our work, we try to tackle the problems of an actual deployment scenario, where manual inspection of all the sensor data is impossible. We present a data processing pipeline that helps detect device faults, filter noise, point anomalies, and calibrate the sensor without extensive human supervision. All of these aspects are often not discussed together in the current literature. In addition, for fault and anomaly detection, our work explicitly targets the LCPM sensors, exploiting their

characteristics and the properties of the measured quantity. At the same time, traditional approaches discuss these issues in general terms.

III. Background

This section provides background information on PM monitoring technologies, focusing on LCPM. The section also discusses PM sources and the official monitoring network in the considered region.

A. Official Monitoring

In Italy, the legislative decree 155/2010, an actuation of the European Directive 2008/50/EC, regulates air pollution. The decree defines the minimum size and structure of the Italian territory's monitoring network, indicating the placement of stations. In addition, it specifies the reference methods for sampling and measurement of air pollutants.

Official measurement techniques for PM adopt well-known physical principles, such as gravimetry and β-attenuation. The European standard EN12341:2014 regulates the former, while the Commission only approves the latter if a valid demonstration of equivalence is provided.

Gravimetric instruments use a filter to capture PM dispersed in the air sample, which is drawn in via a vacuum pump. The filter is periodically replaced and weighted to determine the PM mass concentrations. The air inlet, inertial impactors, and filter perform the size selection. On the contrary, β-attenuation devices evaluate the mass of the PM deposit by measuring the attenuation of the radiation of a small radioactive source when shined on the filter.

Other high-precision monitoring instruments adopt different approaches, such as tapered element oscillating microbalance (TEOM) and high-precision light-scattering devices. Nicklin and Darabkhani [32] reported an overview of the PM monitoring technologies.

B. Low-Cost Light-Scattering PM Sensors

LCPM sensors have been introduced in the market in the past few years. These devices are cheaper, lighter, and more compact than the high-precision instruments, making them suitable for IoT applications. They can be used as standalone sensors, integrated into handheld apparatus, or inserted in more complex IoT solutions.

LCPM sensors draw air inside the device via a small fan. A laser beam is shined on the air sample, and a photodiode, positioned at a specific angle on the opposite side, measures the intensity of the scattered light. Two different sensor technologies are available: Nephelometers and OPCs. Manufacturers often do not disclose the technology adopted by the sensors.

Nephelometers correlate the intensity of the scattered light of the whole air sample to the PM mass concentration according to a predefined calibration curve. On the contrary, OPCs can detect single particles and measure their diameter. Particles are classified according to their diameter in different size bins, whose number and size intervals depend on the specific implementation. The total PM mass can be computed by assuming spherical particles and computing the particle density and the size distribution inside each interval. Density and distribution are normalized for the air sample volume to obtain the concentration of the PM mass.

Fig. 1. Metropolitan city of Turin and ARPA monitoring stations (in red).

These measurement procedures introduce multiple assumptions and approximations. The optical properties of particles are strictly related to their RI, which is unknown a priori and depends on the aerosol under analysis. OPCs use Mie Theory, which models light scattered by a perfect sphere, to measure particle diameters. However, the theory assumes spherical particles and a known RI; this conjecture may be invalid for real-world scenarios. In addition, converting from the PM volume to mass requires knowledge of PM density.

Consequently, calibration curves of both the nephelometers and OPCs strictly depend on the particulate type adopted during the calibration procedure. Factory calibration is often performed using the artificial PM composed of polystyrene spherical latex particles (PSLs) of a known diameter and RI. Alternatively, researchers can use more realistic but less comprehensive PM compositions, such as those from cigarette smoke.

Low-cost devices may not detect the whole number of particles in the sampling volume. Thus, they rely on statistics and extrapolation to compute their actual number. Accuracy becomes worse for increased particle sizes since their number decreases dramatically. For this reason, significant concentrations of larger particles, such as PM_{10} and PM_4, are often estimated from PM_1 and $PM_{2.5}$. A minimum size threshold also limits particle detection.

High levels of relative humidity significantly affect the LCPM sensors. Hygroscopic growth increases the size of the particles, leading to an overestimation of the PM mass and influencing their optical properties. Full-size particle counters solve this problem by heating the air before performing the measurement.

C. Area Under Consideration

Turin, shown in Fig. 1, is situated in the northwest of Italy and is the capital city of the Piedmont Region. This area is surrounded by the Alps on the West and North sides and by a big hill on the East side, which favors air stagnation.

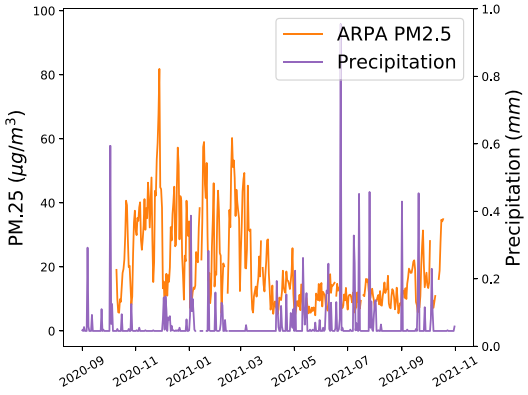

Fig. 2. Daily $PM_{2.5}$ and precipitations measured by the Torino–Rubino station.

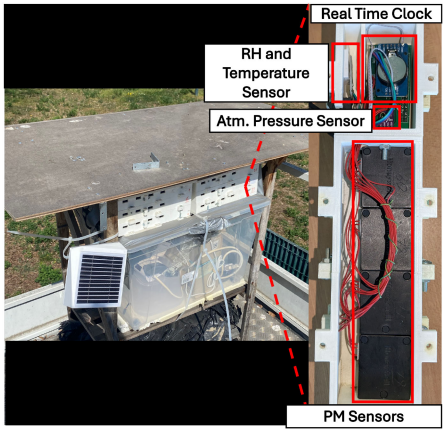

Fig. 3. Our air monitoring station hardware and experiment deployment over the ARPA reference station.

Air quality monitoring is performed by the governmental agency ARPA[1] Piemonte, using a network of fixed stations. Each station's position is classified according to two parameters: 1) the level of urbanization and 2) the sources of pollution. The level of urbanization includes *urban, suburban, and rural areas*. Regarding the sources of pollution, the station can be a *traffic* station or a *background* station. The European Directive 2008/50/EC regulates ARPA's monitoring stations and instruments, ensuring the highest measurement quality and reliability.

According to some analysis of ARPA [33], [34], [35], the primary sources of PM_{10} in Turin are vehicular traffic and domestic heating. Regarding domestic heating, PM is generated by the incineration of wood and pellet fuel. This fuel is mainly used outside the city, where district heating is less common. An essential part of PM from vehicular traffic is caused by NO_x, which acts as a precursor for its formation. Direct exhaust emission, tire wear, and particulate resuspension from traffic also contribute to the high PM levels in the city.

The official measurements of $PM_{2.5}$ at the Torino–Rubino station, a suburban background station, at the time of the experimental campaign, can be seen in Fig. 2). The trend shows an increase in background $PM_{2.5}$ concentrations in winter, starting mid-October. This phenomenon is worsened by thermal inversion, which causes air stagnation and is mainly present in winter. $PM_{2.5}$ levels decrease until they reach a minimum during the summer period from late winter to early spring. Most drops in $PM_{2.5}$ during winter coincide with fast winds and heavy precipitations. The latter's effect is also shown in Fig. 2.

IV. Experiment Campaign

A. Experiment Setup

The research presented in this article utilizes the air monitoring stations developed by Montrucchio et al. [26]. These stations use Raspberry as the main computing component, running on a Linux operating system for ARM. The system collects data from six different sensors. Four Honeywell HPMA115S0-XXX sensor measure the concentration of $PM_{2.5}$ and PM_{10} particles using light scattering. We consider only $PM_{2.5}$ in this work, because the sensor does not measure PM_{10} directly but estimates it from $PM_{2.5}$. The DHT22 sensor measures temperature and relative humidity, while the BME280 sensor captures the atmospheric pressure (see Fig. 3). The air monitoring stations perform measurements of $PM_{2.5}$ every second, temperature and relative humidity every 2 s, and atmospheric pressure every 5 s.

To conduct the study, we placed fourteen monitoring stations on the rooftop of the Rubino ARPA station, as shown in Fig. 3. These stationary sensors remained at the ARPA reference station location from October 10, 2020, to November 1, 2021. They were positioned 1.5 $\sim$ 2 m away from the air inlets of the reference grade devices on the premises.

We considered measurements recorded from the β-attenuation monitoring device taken every hour for the reference instrument. These measurements are recorded and validated by ARPA. The reference instrument employed by ARPA does not provide reliable measurements below 4 μ g/m^3. For this reason, we exclude from our analysis reference values equal to or less than 4 μ g/m^3.

The data collected during the measurement campaign data is published on Zenodo in open access mode [36]. We provide our data collection as a long-term database for validation and to facilitate comparisons with our approach.

B. Sensor Failures

After the measurement campaign, we manually inspected the 56 LCPM sensors in the 14 air monitoring boards to determine their functionality. We identified the sensors that worked adequately (without significant issues of long-term erroneous data, noisy measurements, or values fixed in the lower range of the sensor scale) and those that stopped working due to random and uncorrelated measures.

The inspection showed that 13 PM sensors (23.21%) functioned adequately throughout the experiment. In contrast, 10 sensors (17.86%) failed at the beginning of the experiment, and 33 sensors (58.93%) started to fail during the experiment. We did not notice any electronic issues during data logging or transmission to the central processor, and we attributed malfunctions to errors in the optical part of the sensor. The

[1]Agenzia Regionale per la Protezione Ambientale.

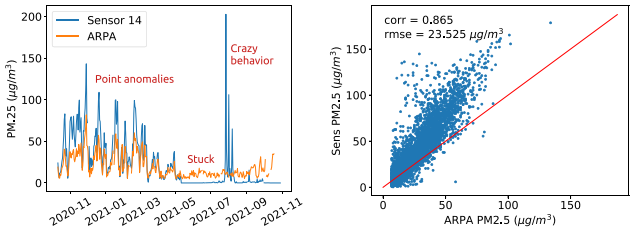

Fig. 4. Faults of sensors 14 compared to beta reference on day averages (left). Scatter plot of sensor 14 with faults removed and beta reference on hour average showing point anomalies (right).

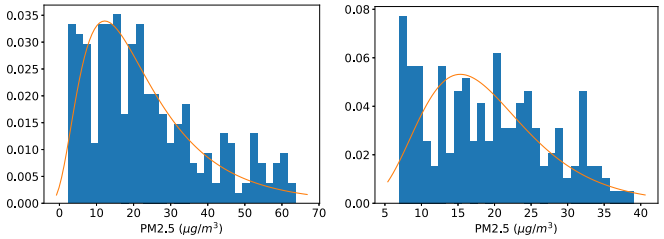

Fig. 5. Distribution of hour averages of $PM_{2.5}$ during the calibration period: sensor 14 (left) and reference beta instrument (the right-hand side).

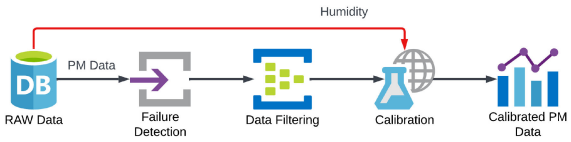

Fig. 6. Framework overview: improving the accuracy and reliability of our pipeline data flow.

statistic confirms the need for more reliability of the LCPM sensors when exposed to long periods of outdoor operation.

Fig. 4 (the left-hand side) shows an overview of the faults encountered. In most cases, the pattern of failure starts with measurements fixed at the lowest values (0 or 1) and remains stable for a variable time. The sensors sometimes show non-deterministic behavior, producing high readings uncorrelated with the reference. Once we removed extensive failures, the sensors were still affected occasionally by point anomalies, as shown in Fig. 4 (right-hand side).

C. Calibration Issues

Previous work [26] showed that the multivariate linear regression, using $PM_{2.5}$ and relative humidity as independent variables, can efficiently calibrate this type of sensor and platform. Once we removed extensive failures, sensors were still affected occasionally by point anomalies, as shown in Fig. 4 (the right-hand side). Introducing humidity correction to the model can help solve this problem, but a more robust system is required, which will be discussed in this work.

Another essential aspect to take into account is seasonal variability. During summer, when $PM_{2.5}$ concentrations are lower, the precision of the sensors is comparable to the measured value, resulting in low correlation. For this reason, calibration during this period produces poor results. This work fixed the sensor calibration period in the first three weeks of the experiment, from October 10 to October 31, 2020.

D. Data Distribution

According to [21] and [24], PM measurements are often log normally distributed. We also observed this property in the data collected during the experiment. This property is exploited by the data processing pipeline presented in this work to improve the data quality and calibration performance. Fig. 5 shows the data distribution of a sensor (left) and the reference instrument (right) for the calibration period.

V. Proposed Framework

Fault detection and sensor calibration are critical tasks that can affect the reliability and performance of PM measurements [37]. As described in the previous section, the failures significantly impact the calibration process. We propose the novel framework presented in Fig. 6 to address these issues. This framework leverages multiple algorithms to detect the most common sensor failures, eliminate outliers, and calibrate the sensors. The overall goal of the pipelined system is to provide more accurate measurements, even in the presence of outliers and sensor failures, without extensive human supervision. Indeed, in large-scale deployment scenarios, the quantity of sensors' data can be too high to be analysed manually.

This work does not adopt complex ML and artificial intelligence (AI) methodologies, such as deep learning neural networks. The proposed approach is based on the simple ML algorithms which do not require extensive training. These algorithms tend to be more general and explainable, which helps avoid unpredictable and undesirable behaviors. This feature is important because air quality measurements are critical since they define urban policies and development plans. In addition, due to the low computational cost, the proposed models could be trained and deployed directly on the sensors without requiring dedicated hardware or specific power necessities. AI methodologies can be more powerful but require complex training phases, which may not be feasible in our scenario.

The proposed framework operates as follows. The raw PM data is processed using a fault detection algorithm to remove the failing sensors, followed by a filtering step to eliminate the data outliers and reduce the sensor noise. The preprocessed data, in conjunction with the ground truth PM data and humidity measures, is used with a newly proposed calibration algorithm that identifies and removes outliers to improve the precision and accuracy of the PM sensors. Finally, we computed the median $PM_{2.5}$ values from the sensors within each air pollution monitoring station.

To provide a more comprehensive exposition of the methodology, we describe all the data processing stages in the following sections, highlighting their contributions to generating precise and reliable PM measurements.

A. Failure Detection

During this step, we analysed the raw readings obtained from the sensors to identify the sensor failures. The primary

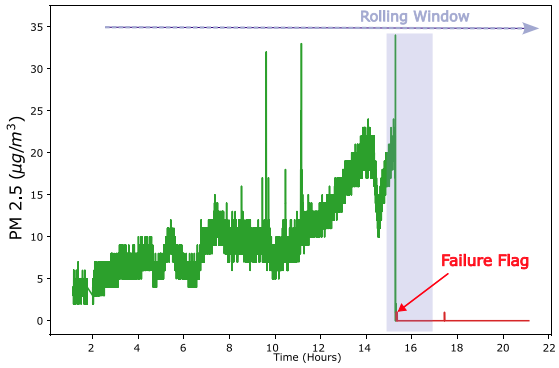

Fig. 7. Fault detection process: plot of the PM density ($\mu g/m^3$) as a function of the time. The algorithm uses a sliding window approach through the data to identify periods with abnormal values, subsequently marking the failure point.

cause of these sensors' malfunction is faults in the embedded photosensor, which remains obscure due to the sensor's uninterrupted data production, without any warning signal.

Considering the difficulty posed by determining errors with nondeterministic or random patterns, especially in the absence of failure signs, detecting the fault becomes challenging. To address this issue, we designed the proposed algorithm to detect the most prevalent pattern: stationary in the lowest values over a specific time (refer to Section IV-B).

As depicted in Fig. 7, the algorithm implements a rolling window technique to identify when the values remained at their lowest (0 or 1 μ g/m^3). If the sensor's data remained at these values for the entire window length, the sensor was flagged as damaged. Any data obtained after the flag was disregarded for subsequent steps. This procedure is applied to the raw sensor data without any aggregation.

We tested multiple time windows to optimize this process, ranging from 1 to 12 h. The aim was to determine the ideal window that would offer the highest accuracy in detecting actual events while minimizing false positives.

B. Data Filtering

Once the data has been analysed for failures, the next step in our pipeline requires filtering out point anomalies. In this phase, the goal is to remove all the readings affected by the noise or external factors. Those events produce high-frequency changes, often impulsive, increasing errors and leading to the data outliers. They can also lead to rapid and sustained changes in the PM values for several minutes. Consequently, our filter must be designed to effectively reduce the number of outliers and minimize the errors with respect to the ground truth values. We performed several tests to identify the optimal filter for removing the noise and maintaining accuracy while selecting filter parameters that reduce the data error and increase the correlation of the measured values. Filters were applied to the per-second data produced by the sensors, while the evaluation was performed on hour aggregations to match the sampling time of the ARPA instrumentation.

During these tests, several factors are taken into consideration. First, some errors may be random, resulting in high-frequency noise, whereas others may be impulse noise. We proposed two filters to reduce the noise in the measurement. A low-pass filter and a nonlinear filter (a median filter) to detect and mitigate the effect of a particle stuck inside the sensor. To ensure that the signal remained unchanged in the bandpass and to avoid a stringent cut-off band that would alter the signal shape, we selected an eighth-order Butterworth IIR filter implemented as a low-pass filter. We chose this filter to achieve a slower roll off, which helps to preserve the signal's shape. The cut-off frequencies and kernel sizes for both the filters were selected according to the values described by Espinosa et al. [38].

Through an initial manual examination, it has been observed that specific erroneous measurements occur within the time intervals of a few minutes. This phenomenon is attributed to the nondeterministic nature, where an element may remain or get stuck within the sensor for a random period. Consequently, determining a general parameter to correct these measurements becomes challenging because a specific frequency or kernel cannot be selected. Moreover, the filters cannot remove these measurements using the previously established predefined setting values.

Despite the effectiveness of these filters, we acknowledged that their determination of an outlier could have been more effective. To address this limitation, we employed a Z-score filter to remove the signal outliers.

The Z-score filter is a statistical filter used to remove outliers in a data set based on the standard deviation and mean of the set. However, the application of the Z-score requires data with a normal distribution. As discussed in Section IV-B, the PM data follows a log-normal distribution. Hence, to implement this filter, the data is first transformed to a normal distribution, the filter is applied, and then inverse transformed to preserve its physical interpretation. The normal-to-log-normal transformation is performed by computing the natural logarithm of the data.

The Z-score is calculated using the formula

$$z = \frac{x - \mu}{\sigma}$$

where μ represents the mean value and σ the standard deviation of the data set. The resulting Z-score z represents a scalar number of how many standard deviations a data point is away from the mean. Positive values indicate points above and negative values points below the mean.

Based on a threshold value, every point with an absolute value of the Z-score above the threshold is considered an outlier and is either removed or replaced. However, defining the static mean and standard deviation values is inappropriate for the PM data, which may exhibit significant variations over time. Therefore, a time-dependent windowing of the database is proposed to prevent false positives in the Z-score filter. A one-week length window is chosen based on the repeatability of the weekly events. In contrast, other window lengths, such as one day, were tested but exhibited inferior performance due to potential variations in dynamics, especially on weekends. For this study, the overlap between the windows is not considered. Specifically, z is computed based on the period containing the evaluated data point.

This approach can accurately remove outliers and preserve the data integrity. Thus, to select and apply the most appropriate filter, we had to consider the unique characteristics of the noise and the outlier patterns.

C. Calibration Model

The final step in the deployment process is sensor calibration. This work presents and evaluates a "push-button" calibration phase that does not require any human intervention. Indeed, in large-scale deployment scenarios, the quantity of sensors' data can be too high to be analysed manually. Multivariate linear regression is an efficient method to calibrate this type of sensor and platform. In this regression, we selected as independent variables [26] $PM_{2.5}$ and relative humidity.

The measurements collected during the first three weeks of the experiment, from October 10th to October 31st, 2020, are used for training the model. Then, the trained model calibrates the sensor data acquired during the remaining part of the data acquisition campaign to simulate a real deployment scenario. The performance of the calibrated sensors is evaluated against the reference instrument. Since, the finest data granularity available from the reference station are hour measurements, we performed the calibration process by hourly aggregating the sensors' readings.

We observed that the model over-compensated the corrections by selecting excessively small $PM_{2.5}$ coefficients, this behavior could be attributed to the presence of outliers. In addition, removing remaining outliers can provide a more general calibration model without being influenced by the occasional sensor faults occurring during the calibration period.

For this reason, an automatic outlier-detection method based on a multivariate GM was introduced. The final objective of this phase is to select the data points to remove from the training set of the calibration model. The points were only excluded for the training procedure but not from the original data set.

The proposed method fits a multivariate Gaussian distribution to the 2-D data points composed of the $PM_{2.5}$ measurements of the sensors and the corresponding values from the reference station, the latter being available when training the model. This association requires the measurement to have the same time granularity, so hour averages were considered.

The next step is to set a threshold probability. A cumulative probability function for the 2-D Gaussian distribution can be defined as a function of the Mahalanobis distance from the sample mean

$$\text{dist}(x_i) = \sqrt{(x_i - x_{\text{mean}})V^{-1}(x_i - x_{\text{mean}})^T}.$$

Intuitively, the farther the points are from the sample mean, the lower the probability of being measured. Given a data point, if the probability of measuring values at a distance greater or equal to the one of the data point is lower than the threshold, the data point should be removed. This probability can be computed following [39]:

$$p(x|\text{dist}(x) \geq \text{dist}(x_i)) = e^{-(\text{dist}(x_i)^2)/2}.$$

Instead of removing a fixed percentage of the less probable data, a probability threshold is used to better adapt our model to the changes in behavior between the sensors.

However, according to what is discussed in Section IV-D, measurements coming from both the sensors and the reference station are better represented by a log-normal distribution rather than a Gaussian. For this reason, before applying the outlier detection model, the hourly readings of the reference station for every single sensor are fitted with a log-normal distribution using the `scypi` software package (`scipy.stats.lognorm.fit`). The fitting allows us to estimate the shift parameter of the log-normal distribution and to apply a transformation to the data so that it follows a normal distribution:

$$\text{normal} = \ln(\text{lognormal} - \text{shift}).$$

Once the data is transformed, the multivariate GM is applied to remove outliers. Finally, the remaining data is transformed back and used to train the calibration model. The double transformation does not alter the measurements in any way; it just improves the filtering capability of the GM. A normality test can be performed on the transformed data to ensure correct outlier removal.

For the analysis presented in this work, we set the threshold probability to 5%. In addition, we even removed the sensors, which showed a correlation with the reference lower than 0.65 during the calibration period. The reason is that a sensor is not functioning correctly during the calibration period should be identified and discarded before deployment.

D. Experiment Methodology

We classified sensors into three categories: 1) those that functioned adequately; 2) those that failed from the beginning; and 3) those that failed during the measurement period. This classification was then used to evaluate the failure detection performance of the framework for both the processes.

After classifying the sensors, we compared each board's readings and the ARPA's ground truth reference. However, it should be noted that ARPA provides new measures hourly, whereas our sensor network produces new measures every second. We aggregated the per-second data to obtain equivalent hourly averages to achieve comparability. The resulting values are then compared with the ARPA's reference values to assess the precision of our system. This comparison is conducted for both the raw input data and the framework output data, allowing us to evaluate the level of precision enhancement provided by the proposed framework.

To assess the contribution of each process within the framework, we evaluated each step under different parameters to identify the optimal set. Subsequently, we conducted a collective evaluation to determine the level of data accuracy improvement in comparison to the reference values.

A confusion matrix is analysed for the fault detection process to identify the window type with the highest precision in detecting the sensor faults. In the anomaly detection and calibration processes, we assessed the performance through the analysis of error metrics (RMSE and MAE), correlation

TABLE I
SUMMARY OF FRAMEWORK'S PERFORMANCE

Metric	Framework Step			
	Failure Detection	Outlier Detection	Calibration Model	Global Performance
RMSE		✓	✓	✓
MAE		✓	✓	✓
r^2		✓	✓	✓
R^2		✓	✓	✓
Confusion Matrix	✓			

TABLE II
CONFUSION MATRICES: RESULT SUMMARY

		True Positives	False Positive	False Negatives	True Negatives
Window Size (hours)	1	21	23	7	5
	2	28	13	7	8
	3	31	7	9	9
	4	32	5	9	10
	8	32	3	10	11
	12	32	0	11	13

(r^2), and coefficient of determination (R^2). Table I provides a summary of the performance metrics that the framework evaluates.

VI. FRAMEWORK PERFORMANCE

This section evaluates each stage of our pipeline framework to determine its contribution to the final performance. We follow the description flow introduced in Section V.

A. Failure Detection

The noisy behavior exhibited by the PM sensors poses a challenge when defining a time window that accurately detects sensor failures. On the one hand, a window with an overly short period might generate many false detections, especially during the summer when PM levels are typically relatively low. On the other hand, very long windows can omit erroneous behavior of the sensors (as explained in Section IV-B), leading to a loss of sensitivity and generating false negatives. Based on empirical observations, we experimented with six different time windows, including 1, 2, 4, 8, and 12 h. To ascertain false positives and negatives in the detection process, we define an erroneous detection as the one in which the algorithm identifies a failure after two weeks with respect to the failure time determined by manual inspection.

Table II presents the fault detection results for the 56 sensors and each time window. Table III shows the metrics used to evaluate the algorithm's performance. These results emphasize the inherent tradeoff between accuracy and sensitivity. Indeed, short windows may be prone to false positives when the PM level is low, whereas long time windows reduce the probability of detection. Notably, the windows lasting 8 and 12 h demonstrate superior performance in the fault detection task. Moreover, adopting a time window longer than 12 h leads to an increase in false negative detections. Table III also shows that the 12-h window, albeit less sensitive, yields fewer errors in fault detection, as indicated by the higher F1-score value. The faults not detected within this window pertain to sensors that exhibited highly noisy data before stabilizing at their lowest values or sensors that consistently measured noisy values throughout the measurement period. Detecting such faults is challenging and necessitates additional analysis to accurately determine the underlying cause.

TABLE III
CONFUSION MATRICES: METRIC COMPARISONS

		Accuracy	Precision	Recall	F1-Score
Window Size (hours)	1	0,464	0,477	0,750	0,583
	2	0,643	0,683	0,800	0,737
	3	0,714	0,816	0,438	0,570
	4	0,750	0,865	0,780	0,821
	8	0,768	0,914	0,762	0,831
	12	0,804	1,000	0,744	0,853

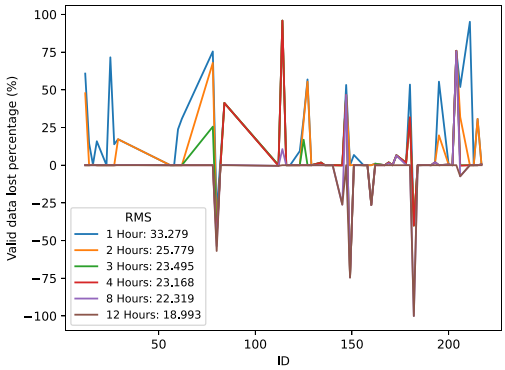

Fig. 8. Valid data loss for each sensor ID in different time windows. Negative values indicate erroneous data that was not removed after a failure. A root mean square (RMS) value close to zero represents the ideal scenario.

Fig. 8 deepens our analysis by plotting, for each sensor, the percentage of invalid readings with respect to the total number of values, as detected by the filter. Positive values on the y-axis denote the percentage of valid information removed due to false positives, leading to the exclusion of genuine PM data. On the contrary, values under zero represent false negatives, which results in the incorporation of erroneous information into the framework.

An optimal detection scenario is indicated by a value equal to zero. We also compute the root mean square of the valid data lost percentage for all the sensors to assess the relative effectiveness of the different time windows. Notably, the 12-h window demonstrates the closest approximation to the ideal scenario.

B. Data Filtering

Following the failure detection phase, we checked various filters to detect and eliminate the abnormal values that impair the measurements. Filters were applied to the per-second data produced by the sensors. After applying the proposed filters, we analysed the reduction in error and its impact on the correlation index compared to the reference values. The

TABLE IV
OUTLIER DETECTION: FILTERS PERFORMANCE COMPARISON

	RMSE $\mu g/m^3$	MAE $\mu g/m^3$	r^2	% Points Corrected	% Points Worsened
Raw Data	18,642	13,226	0,810	-	-
Low Pass Filter	18,724	13,268	0,808	1,160	0,553
Median Filter	18,558	13,162	0,812	0,592	0,178
Z-Score	18,042	13,176	0,822	2,161	2,151
Median + Z-Score	18,443	13,169	0,815	1,873	2,320

evaluation is computed on hour aggregations to match the sampling time of the ARPA instrumentation. Additionally, we compare the percentage of summarized hourly data points that show a significant change in value, showing either a decreasing (*% Points Corrected*) or increasing error (*% Points Worsened*). Table IV summarizes the evaluation of these parameters for each filter. Each value represents the mean of the metrics applied to each sensor over the entire experimental period.

Based on these values, the *Z*-score filter demonstrates better performance in terms of the error reduction, an improvement in the Pearson's coefficient (r^2) value, and a more significant impact on the percentage of corrected summarized data points. The median filter demonstrates the second most favorable performance, and to further evaluate this, we assess the sequential application of the most effective filters (Median + *Z*-score). However, this final strategy exhibits worse performance than applying one filter individually. Given these findings, we selected the *Z*-score filter for our pipeline framework to detect and remove the abnormal values before applying the calibration process. The peculiarity of this filter is that it does not alter the measurements per second but discards the ones it identifies as not compliant. Nonetheless, even with a *Z*-score threshold set to two, the removed data was minimal, and all the hour aggregations could still be computed.

C. Calibration Model

To conduct an in-depth evaluation of each calibration model, the sensors that exhibited failures but were not detected (false negatives) during the fault detection phase were intentionally excluded from the calibration performance analysis. We decided to streamline the metrics and provide a more accurate depiction of the overall performance, eliminating the interference these failed sensors might have introduced.

Figs. 9 and 10 display the distribution of two error metrics, i.e., the RMSE and R^2 (the coefficient of determination). Each violin represents the distribution of the metric by the sensors, the black dots represent each sensor metric average during the experiment and the white line shows the mean value of all dots. This framework aims to reduce the sensor variance (broader and shorter violins are preferred) and obtain the mean values close to the ideal values (0 for RMSE and 1 for r^2). We can observe that the calibration models incorporating a filtering process for outlier detection exhibited superior performance

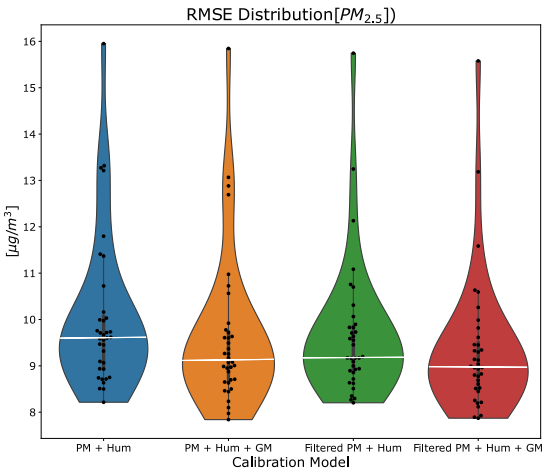

Fig. 9. RMSE distribution for each calibration model. Black dots indicate the RMSE of each sensor. The white lines represent the average error of each sensor.

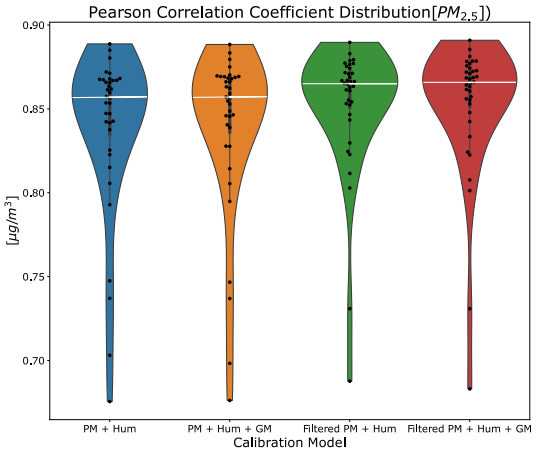

Fig. 10. Pearson's correlation index (r^2) distribution for each calibration model. Black dots indicate the correlation of each sensor. The white lines represent the average r^2 of each sensor.

metrics compared to the reference model (PM + Humidity) [26]. This result highlights the efficacy of outlier detection in enhancing the model calibration.

Additionally, we explicitly designed the GM for outlier detection to minimize the impact of the abnormal or highly variable values (see graphics for "PM + Hum + GM" and "Filtered PM + Hum + GM"). The implementation of this algorithm, in tandem with the other methods, led to a calibration model that was more robust and resilient to the errors.

Table V presents the performance metrics for each calibration model, providing a global view of their performance. This table shows the median values of each calibration process results applied to the selected sensors throughout the experimental period. The implemented filtering process substantially improved the correlation between the variables, whereas the GM algorithm successfully decreased the error value. This also shows the efficacy of the data transformations.

Ultimately, the filtered model PM + Hum + GM demonstrated the highest efficacy in calibration, with notable

TABLE V
COMPARISON OF THE PERFORMANCE METRICS FOR EACH CALIBRATION MODEL

	RMSE $\mu g/m^3$	MAE $\mu g/m^3$	r^2	R^2
PM + Hum	9,608	7,149	0,857	0,670
PM + Hum + GM	9,125	6,744	0,857	0,692
Filtered PM + Hum	9,185	6,960	0,865	0,677
Filtered PM + Hum + GM	8,976	6,681	0,866	0,700

TABLE VI
GLOBAL FRAMEWORK PERFORMANCE METRICS VERSUS REFERENCE MODEL

	RMSE $\mu g/m^3$	MAE $\mu g/m^3$	r^2	R^2
PM + Hum	9,006	6,757	0,858	0,654
Filtered PM + Hum + GM	8,462	6,241	0,868	0,725

improvements in the error reduction (RMSE and MAE by 6.6%) and an increase in the correlation index (R^2) by 4.5%. These improvements highlight the value of using outlier detection and the GM in optimizing the calibration methods.

VII. GLOBAL PERFORMANCE

In our comprehensive evaluation of the proposed framework, each component of the process chain has been considered in unison, encompassing the four PM sensors integrated within each of the 14 distinct monitoring stations. We computed the hourly average for each station for each PM sensor. We obtained the median values from these averages, i.e., an aggregate measure of the performance of all the PM sensors operating at each station within the given period. This approach captures the most reliable and representative PM measures per monitoring station, thus increasing the robustness of the analysis. Subsequently, these resulting measures are utilized as the foundation upon which performance metrics are calculated and compared against the reference calibration model. Table VI summarizes each station's metrics' average, providing a holistic overview of our system's performance.

Our findings show the efficacy of our system in controlling the impact of possible undetected failures. Thanks to the inbuilt redundancy of the sensors within each monitoring station, in most cases, the system demonstrates either equivalent or superior performance when compared with the values presented in Table V. This behavior is a clear indicator of our framework's robustness and its ability to enhance the system's precision despite the challenges posed by the high failure rate of our sensors observed throughout the year-long deployment period.

Analysing the distributions portrayed by the box plots in Figs. 11 and 12, we observe that the dispersion in our measures remains remarkably stable. This result highlights the high coherence and consistency of our measurement process, further bolstering the reliability of our framework.

However, notice that our framework has its limitations. Notably, the presence of more than two faulty sensors within a single station can pose challenges to the system's reliability.

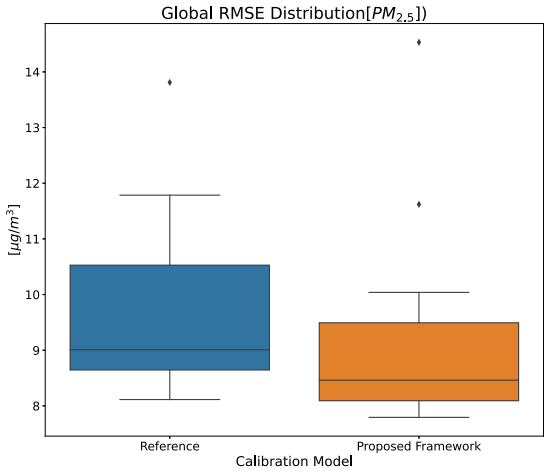

Fig. 11. RMSE distribution for all monitoring stations: comparison between the proposed framework and the reference calibration model.

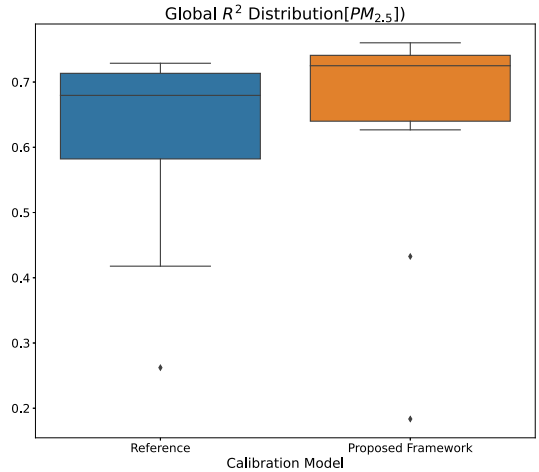

Fig. 12. Coefficient of determination (R^2) distribution for all monitoring stations: comparison between the proposed framework and the reference calibration model.

While relatively rare, this occurrence can result in outlier points in terms of error and the R^2 as demonstrated in Fig. 12). These anomalous points, although infrequent, underscore the importance of the sensor redundancy and the robust error-handling mechanisms in ensuring the overall performance and reliability of the system.

VIII. CONCLUSION

This work proposes a data processing pipeline for improving the data quality of the LCPM sensors. The overall framework leverages several algorithms to detect the most common failure in the sensors, remove outliers, and perform sensor calibration without the need for extensive human supervision.

We conducted a year-long experiment by positioning 56 LCPM sensors, divided into 14 low-cost monitoring stations, near the inlet of an official monitoring device. The collected data is used to design and test the proposed data processing pipeline. The first step uses a simple failure detection algorithm to identify the most common pattern sensor failure, i.e., the sensors remaining stuck to low values. In the second phase, we tested different filters to remove high-frequency and impulse

noise, with the Z-score filter being the most effective. Due to the data's log-normality, we computed the measurements' natural logarithm before applying the filter. In the following step, sensor calibration is carried out via a multivariate liner regression model, considering both PM and relative humidity as independent variables, on the first three weeks of the experiment. To remove outliers that affect the linear regression, we applied a multivariate GM to the measurements and their reference. Like before, in this case, it was necessary to transform the data to follow a normal distribution. Finally, the redundancy of the four sensors installed in each low-cost monitoring station is exploited by computing their median.

Results show the proposed system provides accurate measurements, even when the sensors encounter anomalies due to partial or complete failure. Filtering and failure detection processes are critical to ensure that the sensor readings are reliable and reflect pollution conditions. Additionally, our pipeline minimizes measurement variations, improving consistency compared to the other calibration systems. Moreover, the measurement median of each station surpasses the error levels specified by the manufacturer. Our achievements offer an optimistic perspective for using these sensors to measure PM in smart city settings. The capability to detect faults also facilitates timely maintenance or replacements, ensuring the sensors' durability for prolonged measurement periods.

Future work should further test the generality and effectiveness of the model and the selected parameters by calibrating different periods and analysing separately different environmental conditions. Sensors should also undergo testing in various scenarios, including high-traffic areas where PM levels change more frequently, and their performance should be evaluated in a completely unsupervised deployment. Finally, more advanced and complex ML models, such as deep learning neural networks, should also be tested for anomaly detection and calibration.

ACKNOWLEDGMENT

The authors thank ARPA Piemonte for putting at disposal official public hourly pollution data, used as a reference, and for letting them to install and test boards inside the "Rubino" station. Data provided in this article [36] can not in any way be considered as official pollution data unlike the ones provided by ARPA. ARPA Piemonte can not be ascribed for any mistake contained in this article, as well as for any error in the experimental values.

REFERENCES

[1] M. Gerboles, L. Spinelle, and A. Borowiak, "Measuring air pollution with low-cost sensors," Eur. Comm., Brussels, Belgium, Rep. JRC107461, 2017.

[2] M. Fadda, M. Anedda, R. Girau, G. Pau, and D. D. Giusto, "A social Internet of Things smart city solution for traffic and pollution monitoring in Cagliari," *IEEE Internet Things J.*, vol. 10, no. 3, pp. 2373–2390, Feb. 2023.

[3] M. Anedda, M. Fadda, R. Girau, G. Pau, and D. Giusto, "A social smart city for public and private mobility: A real case study," *Comput. Netw.*, vol. 220, Jan. 2023, Art. no. 109464.

[4] D. Aguiari et al., "Canarin II: Designing a smart e-bike ecosystem," in *Proc. 15th IEEE Annual Consum. Commun. Netwo. Conf. (CCNC)*, 2018, pp. 1–6.

[5] R. Tse, D. Aguiari, L. Monti, G. Pau, C. Prandi, and P. Salomoni, "On assessing the accuracy of air pollution models exploiting a strategic sensors deployment," in *Proc. 4th EAI Int. Conf. Smart Objects Technol. Soc. Good*, 2018, pp. 55–58. [Online]. Available: http://dl.acm.org/citation.cfm?doid=3284869.3284880

[6] T. Baldi, G. Delnevo, R. Girau, and S. Mirri, "On the prediction of air quality within vehicles using outdoor air pollution: Sensors and machine learning algorithms," in *Proc. ACM SIGCOMM Workshop Netw. Sens. Syst. Sustain. Soc.*, 2022, pp. 14–19.

[7] L. Russi, P. Guidorzi, B. Pulvirenti, D. Aguiari, G. Pau, and G. Semprini, "Air quality and comfort characterisation within an electric vehicle cabin in heating and cooling operations," *Sensors*, vol. 22, no. 2, p. 543, 2022.

[8] R. Tse, M. Im, S.-K. Tang, L. Menezes, A. Dias, and G. Pau, "Self-adaptive sensing IoT platform for conserving historic buildings and collections in museums," in *Proc. 5th Int. Conf. Internet Things, Big Data Secur.*, 2020, pp. 392–398. [Online]. Available: http://www.scitepress.org/DigitalLibrary/Link.aspx?doi=10.5220/0009470203920398

[9] R. Tse, D. Aguiari, K.-S. Chou, S.-K. Tang, D. Giusto, and G. Pau, "Monitoring cultural heritage buildings via low-cost edge computing/sensing platforms: The Biblioteca Joanina de Coimbra case study," in *Proc. 4th EAI Int. Conf. Smart Objects Technol. Soc. Good*, 2018, pp. 148–152.

[10] G. Pettorru, M. Fadda, R. Girau, M. Anedda, and D. Giusto, "An IoT-based electronic sniffing for forest fire detection," in *Proc. IEEE Int. Conf. Consum. Electron. (ICCE)*, 2023, pp. 1–5.

[11] Z. M. Fadlullah and N. Kato, "On smart IoT remote sensing over integrated terrestrial-aerial-space networks: An asynchronous federated learning approach," *IEEE Netw.*, vol. 35, no. 5, pp. 129–135, Sep./Oct. 2021.

[12] L. R. Crilley et al., "Evaluation of a low-cost optical particle counter (Alphasense OPC-N2) for ambient air monitoring," *Atmos. Meas. Techn.*, vol. 11, no. 2, pp. 709–720, Feb. 2018. [Online]. Available: https://amt.copernicus.org/articles/11/709/2018/

[13] F. M. J. Bulot et al., "Long-term field comparison of multiple low-cost particulate matter sensors in an outdoor urban environment," *Sci. Rep.*, vol. 9, no. 1, p. 7497, May 2019. [Online]. Available: https://www.nature.com/articles/s41598-019-43716-3

[14] 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 Sci. Technol.*, vol. 49, no. 11, pp. 1063–1077, Nov. 2015. [Online]. Available: https://doi.org/10.1080/02786826.2015.1100710

[15] D. H. Hagan and J. H. Kroll, "Assessing the accuracy of low-cost optical particle sensors using a physics-based approach," *Atmos. Meas. Techn.*, vol. 13, no. 11, pp. 6343–6355, Nov. 2020. [Online]. Available: https://amt.copernicus.org/articles/13/6343/2020/

[16] E. Brattich et al., "How to get the best from low-cost particulate matter sensors: Guidelines and practical recommendations," *Sensors*, vol. 20, no. 11, p. 3073, Jan. 2020. [Online]. Available: https://www.mdpi.com/1424-8220/20/11/3073

[17] M. R. Giordano 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," *J. Aerosol Sci.*, vol. 158, Nov. 2021, Art. no. 105833. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0021850221005644

[18] B. Mao, Y. Kawamoto, and N. Kato, "AI-based joint optimization of QoS and security for 6G energy harvesting Internet of Things," *IEEE Internet Things J.*, vol. 7, no. 8, pp. 7032–7042, Aug. 2020.

[19] A. A. Cook, G. Mısırlı, and Z. Fan, "Anomaly detection for IoT time-series data: A survey," *IEEE Internet Things J.*, vol. 7, no. 7, pp. 6481–6494, Jul. 2020.

[20] L.-J. Chen, Y.-H. Ho, H.-H. Hsieh, S.-T. Huang, H.-C. Lee, and S. Mahajan, "ADF: An anomaly detection framework for large-scale PM2.5 sensing systems," *IEEE Internet Things J.*, vol. 5, no. 2, pp. 559–570, Apr. 2018.

[21] D. Hasenfratz et al., "Deriving high-resolution urban air pollution maps using mobile sensor nodes," *Pervasive Mobile Comput.*, vol. 16, pp. 268–285, Jan. 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1574119214001928

[22] D. Liu, Q. Zhang, J. Jiang, and D.-R. Chen, "Performance calibration of low-cost and portable particular matter (PM) sensors," *J. Aerosol Sci.*, vol. 112, pp. 1–10, Oct. 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0021850217300848

[23] B. Maag, Z. Zhou, and L. Thiele, "A survey on sensor calibration in air pollution monitoring deployments," *IEEE Internet Things J.*, vol. 5, no. 6, pp. 4857–4870, Dec. 2018.

[24] B. Rumburg, R. Alldredge, and C. Claiborn, "Statistical distributions of particulate matter and the error associated with sampling frequency," *Atmos. Environ.*, vol. 35, pp. 2907–2920, Jun. 2001.
[25] M. Budde, R. El Masri, T. Riedel, and M. Beigl, "Enabling low-cost particulate matter measurement for participatory sensing scenarios," in *Proc. 12th Int. Conf. Mobile Ubiquitous Multimedia*, 2013, pp. 1–10. [Online]. Available: https://doi.org/10.1145/2541831.2541859
[26] B. Montrucchio, E. Giusto, M. G. Vakili, S. Quer, R. Ferrero, and C. Fornaro, "A densely-deployed, high sampling rate, open-source air pollution monitoring WSN," *IEEE Trans. Veh. Technol.*, vol. 69, no. 12, pp. 15786–15799, Dec. 2020.
[27] F. Concas et al., "Low-cost outdoor air quality monitoring and sensor calibration: A survey and critical analysis," *ACM Trans. Sens. Netw.*, vol. 17, no. 2, pp. 1–44, May 2021. [Online]. Available: https://doi.org/10.1145/3446005
[28] E. Giusto, R. Ferrero, F. Gandino, B. Montrucchio, M. Rebaudengo, and M. Zhang, "Particulate matter monitoring in mixed indoor/outdoor industrial applications: A case study," in *Proc. IEEE 23rd Int. Conf. Emerg. Technol. Fact. Autom. (ETFA)*, vol. 1, 2018, pp. 838–844.
[29] P. Chiavassa, F. Gandino, and E. Giusto, "An investigation on duty-cycle for particulate matter monitoring with light-scattering sensors," in *Proc. 6th Int. Conf. Smart Sustain. Technol. (SpliTech)*, 2021, pp. 1–6.
[30] G. R. Espinosa, B. Montrucchio, E. Giusto, and M. Rebaudengo, "Low-cost PM sensor behaviour based on duty-cycle analysis," in *Proc. 26th IEEE Int. Conf. Emerg. Technol. Fact. Autom. (ETFA)*, 2021, pp. 1–8.
[31] B. Mao, F. Tang, Y. Kawamoto, and N. Kato, "Optimizing computation offloading in satellite-UAV-served 6G IoT: A deep learning approach," *IEEE Netw.*, vol. 35, no. 4, pp. 102–108, Jul./Aug. 2021.
[32] D. Nicklin and H. G. Darabkhani, "Techniques to measure particulate matter emissions from stationary sources: A critical technology review using multi criteria decision analysis (MCDA)," *J. Environ. Manag.*, vol. 296, Oct. 2021, Art. no. 113167.
[33] "Inquinamento da particolato PM10: Le sorgenti." 2021. [Online]. Available: https://www.arpa.piemonte.it/news/inquinamento-da-particolato-pm10-le-fonti
[34] "Inquinamento da particolato PM10: Il riscaldamento domestico." 2019. [Online]. Available: http://www.arpa.piemonte.it/news/inquinamento-da-particolato-pm10-il-riscaldamento-domestico
[35] "Inquinamento da particolato PM10: Il trasporto su strada." 2019. [Online]. Available: https://www.arpa.piemonte.it/news/inquinamento-da-particolato-pm10-il-trasporto-su-strada
[36] P. Chiavassa et al., Sep. 2023. "Dataset for 'improving data quality of low-cost light-scattering PM sensors: Towards automatic air quality monitoring in urban environments'," Dataset, Zenodo. [Online]. Available: https://doi.org/10.5281/zenodo.8329133
[37] K. Kelly et al., "Ambient and laboratory evaluation of a low-cost particulate matter sensor," *Environ. Pollut.*, vol. 221, pp. 491–500, Feb. 2017. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S026974911632718X
[38] G. R. Espinosa, B. Montrucchio, F. Gandino, and M. Rebaudengo, "Frequency analysis of particulate matter in urban environments under low-cost sensors," in *Proc. Int. Conf. Comput. Commun. Artif. Intell. (CCAI)*, 2021, pp. 97–105. [Online]. Available: https://ieeexplore.ieee.org/document/9447517/
[39] "N-dimensional cumulative function, and other useful facts about Gaussians and normal densities." 2009. [Online]. Available: https://upload.wikimedia.org/wikipedia/commons/a/a2/Cumulative_function_n_dimensional_Gaussians_12.2013.pdf

Gustavo Ramirez-Espinosa (Graduate Student Member, IEEE) received the M.S. degree in electronic engineering from Pontificia Universidad Javeriana, Bogota, Colombia, in 2013, and the Ph.D. degree in computer engineering and control from the Politecnico di Torino, Turin, Italy, in 2023.

He is currently an Assistant Professor with the Department of Electronics, Pontificia Universidad Javeriana. His research interests include IoT, embedded systems, machine learning, and computer networks.

Pietro Chiavassa (Graduate Student Member, IEEE) received the M.S. degree in computer engineering from Politecnico di Torino, Turin, Italy, in 2020, where he is currently pursuing the Ph.D. degree with the Dipartimento di Automatica e Informatica.

His research interests include IoT, security and privacy, and quantum computing.

Edoardo Giusto (Member, IEEE) received the M.S. degree in computer engineering and the Ph.D. degree in computer and systems engineering from the Politecnico di Torino, Turin, Italy, in 2017 and 2021, respectively.

He is an Assistant Professor with the Department of Electrical Engineering and Information Technology, University of Naples Federico II, Naples, Italy. He was a Visiting Postdoctoral Fellow with the Superconducting Quantum Materials and Systems Center, Fermilab, Batavia, IL, USA. This postdoctoral position was part of the Next Generation Internet Transatlantic Fellowship Program, funded by the European Commission under Horizon Europe. His research interests revolve around quantum computing, encompassing applications of QC, problem mapping, and reliability and fault tolerance of QC devices, as well as the integration of QC in high-performance computing infrastructures.

Dr. Giusto actively contributes to the field as a Technical Committee Member for the IEEE QCE—Quantum Week conference and the IEEE CTSoc Quantum in Consumer Technology He is a member of ACM.

Stefano Quer (Member, IEEE) received the M.S. degree in electronic engineering and the Ph.D. degree in computer engineering from the Politecnico di Torino, Turin, Italy, in 1991 and 1996, respectively.

He has been with the "Advanced Technology Group," Synopsys Inc., Mountain View, SA, USA; the "Alpha Development Group," Compaq Computer Corporation, Shrewsbury, MA, USA; and a Compaq Computer Corporation Consultant. He is a Professor with the Department of Control and Computer Engineering, Politecnico di Torino, Turin, Italy. His main research interests include CAD tools for VLSI, formal methods, concurrent algorithms, and optimization techniques.

Bartolomeo Montrucchio (Senior Member, IEEE) received the M.S. degree in electronic engineering and the Ph.D. degree in computer engineering from the Politecnico di Torino, Turin, Italy, in 1998 and 2002, respectively.

He is currently a Full Professor of Computer Engineering, Department of Control and Computer Engineering, Politecnico di Torino. His research interests include image analysis and synthesis techniques, scientific visualization, sensor networks, RFIDs, and quantum computing.

Maurizio Rebaudengo (Senior Member, IEEE) received the M.S. degree in electronics and the Ph.D. degree in computer engineering from the Politecnico di Torino, Turin, Italy, in 1991 and 1995, respectively.

He is currently a Full Professor with the Department of Control and Computer Engineering, Politecnico di Torino.

His research interests include ubiquitous computing, IoT platforms, and dependability analysis of computer-based systems.

Open Access funding provided by "Politecnico di Torino" within the CRUI CARE Agreement