A Densely-Deployed, High Sampling Rate, Open-Source Air Pollution Monitoring WSN

# A Densely-Deployed, High Sampling Rate, Open-Source Air Pollution Monitoring WSN

Bartolomeo Montrucchio, *Member, IEEE,* Edoardo Giusto, *Student Member, IEEE,* Mohammad Ghazi Vakili, *Student Member, IEEE,* Stefano Quer, Renato Ferrero, *Senior Member, IEEE,* and Claudio Fornaro

***Abstract*—Air quality, especially particulate matter, has recently attracted a lot of attention from governments, industry, and academia, motivating the use of denser air quality monitoring networks based on low-cost sensing strategies. However, low-cost sensors are frequently sensitive to aging, environmental conditions, and pollutant cross-sensitivities. These issues have been only partially addressed, limiting their usage.**

**In this study, we develop a low-cost particulate matter monitoring system based on special-purpose acquisition boards, deployed for monitoring air quality on both stationary and mobile sensor platforms. We explore the influence of all model variables, the quality of different calibration strategies, the accuracy across different concentration ranges, and the usefulness of redundant sensors placed in each station. The collected sensor data amounts to about 50GB of data, gathered in six months during the winter season. Tests of statically immovable stations include an analysis of accuracy and sensors' reliability made by comparing our results with more accurate and expensive standard $\beta$-radiation sensors. Tests on mobile stations have been designed to analyze the reactivity of our system to unexpected and abrupt events. These experiments embrace traffic analysis, pollution investigation using different means of transport and pollution analysis during peculiar events.**

**With respect to other approaches, our methodology has been proved to be extremely easy to calibrate, to offer a very high sample rate (one sample per second), and to be based on an open-source software architecture. Database and software are available as open source in [1].**

***Index Terms*—WSN, Pollution Measurement, Open-source.**

## I. Introduction

Conventional approaches to track the air quality are based on very sparse networks of static reference-grade detectors. The spatial coverage of these networks has been limited by the high cost of instrumentation. From the one side, micro-balance Particulate Matter (PM) monitoring stations are very accurate, but they are large and cost on the order of 50K-100K dollars. On the other one, portable light-scattering based PM detectors have varying accuracy and costs between 300-2K dollars [2]. Moreover, air pollutant concentrations often exhibit significant spatial variability depending on local sources and features of the built environment, which may not be well captured by the existing sparse monitoring networks. As a consequence, there has recently been a significant increase in developing and applying low-cost sensor-based technology which could enable much denser air quality networks at a comparable cost with the existing ones (see, for example, Giusto et al. [3]). These sensing nodes can be adopted not only in city-wide applications, but they can also be used in more strategic, location-aware deployments [4] or even to monitor the conservation state of historic buildings [5]. Furthermore, mobile monitoring enables participatory sensing approaches, which in turn are well-suited to address many of the above aspects, as they intrinsically involve empowering citizens by providing individuals with low-cost measurement devices. Unfortunately, mobile sensing devices have also several drawbacks. Among these disadvantages, we recall their necessity to be battery-supplied and their limited processing capability. These limitations present challenges that can be overcome only adopting persistent engineering solutions [6], [7].

Following the previous considerations, in this work we design, build, and verify a low-cost, open-source air-quality system which is based on special-purpose acquisition boards, it is deployed on stationary and mobile platforms, and it is devised for participatory sensing strategies. We follow article 18.5 of Italian Decree 155/2010 on the dissemination of air quality data, which absorbs EU directive 2008/50/CE. Therefore we declare, in the acknowledgment section, that our data cannot be considered as official.

First of all, we describe the design of our sensor platforms. Each station contains 4 particulate matter ($PM_{10}$ and $PM_{2.5}$) sensors plus single sensors for temperature, humidity, and pressure. Each platform is powered by a Raspberry Pi Zero Wireless board. This configuration enables our application to work with a quite high sampling rate (i.e., one sample per second) which is able to accurately follow sudden atmospheric phenomena. We deployed about 100 sensors, both on stationary and mobile platforms, for a period of 5 months, from October 2018 to February 2019. We focused upon the city center of Turin (located in the north-west region of Italy) inside and outside the limited traffic area to reflect the variation in traffic density, driving speed, and street configuration. For the sake of completeness, we also included in our experiments a recreational area (i.e., a park) with very low traffic density at its border. The monitoring hours cover the entire day, with specific experiments running from 11.00 a.m. to 05.00 p.m. The collected database amounts to about 50GB of data, corresponding to about $700{\cdot}10^6$ data tuples.

Secondly, we describe the entire software architecture. Our application manipulates a heterogeneous set of input data coming from our sensor stations and the reference platforms. The public air quality station uses expensive instruments deliv-



Authors are with Politecnico di Torino, Department of Control and Computer Engineering, Turin, 10129, Italy

ering very accurate measures. Unfortunately, these samples are available only from very sparse locations and they are gathered only once per hour. We also collect public weather data to study the relationship between the air quality as assessed by our sensor stations and other weather information. To store multiple kinds of data sources in a uniform format, we store them in a my-SQL database.

Thirdly, once the database has been collected, we concentrate on data calibration and data validation. As far as calibration is concerned, accurate and precise calibration models are particularly critical to the success of dense sensor networks deployed in urban areas of developed countries. In these situations, pollutant concentrations are often at the low end of the spectrum of global pollutant concentrations, and poor signal-to-noise ratio and cross-sensitivity may hamper the ability of the network to deliver reliable results. Keeping in mind this consideration, we perform our calibration phase with great care, using both the Multivariate Linear Regression Model and the Random Forest machine learning algorithm [8], [9]. The latter one, to the best of our knowledge, has been rarely applied to low-cost air quality monitor calibrations. Moreover, we compare the results gathered with different calibration phases and we validate them over different periods to assess the quality of the result in time.

Finally, we perform some mobile tests. These include testing a semaphore capability of managing traffic and the connection with pollution, the pollution level to which dwellers are exposed during their daily mobility (using different means of transport), and the effect of sudden events on the air quality (such as wind and the New Year's eve fireworks). Even if the results gathered in these scenarios often seem to confirm common conjectures on pollution levels, in some specific case they represent unexpected situation deserving further analysis.

### A. Contributions

To sum up, the main contributions of our work are the following:

- A special-purpose designed air-quality station, hosting several low-cost sensors, with simple hardware and software architectures, and a quite high sample rate (one sample per second).
- A careful deployment of our stations in the city center, for a period of over 5 months. This provides valuable training and testing data for our models, enabling a long-term evaluation of the entire system.
- Results on several calibration strategies with related validation data over different periods of time.
- Several experiments to investigate the air-quality on mobile and dynamic related events. These experiments are enabled by our high sample rate, well suited to analyze fast transient phenomena such as the pollution variation at the traffic light, or the dynamic pollution variation during New Year's fireworks.
- A completely open-source architecture. The entire hardware architecture, software implementation, and the entire data-set collected during 5 months are made available [1].

### B. Roadmap

The rest of this paper is structured as follows. Section II describes the related work about air quality monitoring. Section III presents an overview of our system hardware and software architecture. Section IV describes our acquisition methodology and our measurement sites. Section V and Section VI focus on the calibration and validation strategy adopted for our sensor stations. Sections VII and VIII illustrate data gathered by stationary and mobile station boards, respectively. Section IX reports some final considerations and discussion on the lessons that can be learned from our analysis. Finally, Section X concludes the paper with some summarizing remarks and it gives some directions for future works.

## II. Related Works

Urban air pollution has attracted great attention in recent years as it has been shown to be of a significant risk to city dwellers. At present, air pollution concentrations are mainly collected by environmental or government authorities using networks of fixed monitoring stations. Fixed stations obtain flawless air quality data, as they can provide very accurate measurements at the deployment locations. However, these stations usually require significant investments and human resources to be built and maintained. Thus, several alternatives have been proposed over the years.

Randall [10] demonstrates that coarse-grained information about air quality of the Earth's surface can be obtained by remote sensing using satellites. Although a large-scale area can be easily covered by only one satellite, the accuracy of this strategy highly depends on factors like weather conditions and land-use characteristics. Following Kawamoto et al. [11], a satellite-routed sensor system can increase accuracy, as data can be accumulated by a large number of sensor terminals, then gathered by the satellite, and finally transferred to the ground station. The problem of data collisions may be solved by adopting a "divide and conquer" approach to collect data on demand. The method achieves efficient data collection from numerous sensor terminals and it minimizes all operational delays in the system. Nevertheless, the cost of any solution exploiting satellites remains extremely high.

In order to find an alternative, cheaper than the previous strategies, to the problem of air pollution, many recent works concentrate on deploying low-cost sensors. A large number of publications have reported the use of stationary or mobile laboratories with low-cost sensors to collect air quality data for specific purposes. For example, it has been shown that distributed or mobile personal measurement devices equipped with cheap commercial off-the-shelf dust sensors can reach meaningful accuracy at a cost one to two orders of magnitude lower than the one of current hand-held solutions [12]. The same study also shows that participatory sensing, where co-located measurements are shared across different devices, can help reaching a high measurement accuracy. Moreover, the participatory sensing paradigm includes also subjective perceptions, such as posts by citizens on Online Social Network (OSN) platforms, which can enrich the mere sampling of the data [13]. Overall, two main research topics can be identified: Managing a distributed network for local sensing and developing low-cost sensors of air pollutants. These topics are deeply

investigated in recent scientific literature, as discussed in the following paragraphs.

A network of vehicles carrying sensors for a flexible air quality monitoring is called a *vehicular sensor network* (VSN). For example, a VSN may consist of a set of cars equipped with gas sensors, wireless connection, and GPS receivers. Gathering big data efficiently in such densely distributed sensor networks is challenging. Adjusting the data sampling rates of cars can balance monitoring accuracy and communication cost as it prevents the transmission of similar data collected from close positions, thus alleviating network congestion [14]. Moreover, the wireless transmission should be optimized to avoid excessive energy consumption. The network is usually divided into sub-networks because of limited wireless communication range. In this case, a proper clustering algorithm increases energy efficiency of data gathering by managing the mobile sink routing in the sub-networks [15]. Evidence shows that there are other successful implementations of VSNs aside from cars. For example, bicycles can carry sensors for air pollution monitoring [16], [17]. It is shown that even a limited set of mobile measurements makes it possible to map locations with systematically higher or lower ultra-fine particles and $PM_{10}$ concentrations in urban environments. Unfortunately, the use of semi-professional equipment to monitor $PM_{10}$ levels makes this experiment unsuitable for low-cost urban sensing scenarios. A different implementation exploits smart devices integrating sensors to build an architecture for people-centric environmental sensing platforms [18]. Smart objects and virtual node technology establish closed loops of interactions among people and physical devices. By aggregating on-demand user data from smart devices, it is possible to measure the space-time distribution of particulate matter. A case study of particulate matter exposure in New York City illustrates the potential application of such a system.

There are several issues in developing low-cost sensors for air pollution management in cities, among which, reliability, sensitivity, selectivity (different gases can contribute to the response of the sensors), stability, longevity of operation before replacement [19]. Low sensitivity and poor signal quality can be addressed with sliding window and a low pass filter [20]. This approach is adopted in a real case study, where a wireless network of low-cost particle sensors is deployed in a woodworking shop. Data quality can be improved by identifying outliers in raw measurement data and inferring anomalous events. This task can be achieved by means of an anomaly detection framework composed of four modules [21]: Time-sliced anomaly detector (detecting spatial, temporal, and spatio-temporal anomalies in real-time sensor measurement data stream), a real-time emission detector (detecting potential regional emission sources), a device ranker (providing a ranking for each sensing device), and a malfunction detection (identifying malfunctioning devices).

Finally, particular attention must be paid to calibration, as this is a necessary step to obtain accurate measures. Zimmerman et al. [9] proposes a multi-pollutant sensor package, which measures CO, $NO_2$, $O_3$, and $CO_2$, on which they compare three different calibration methods: Laboratory univariate linear regression, empirical multiple linear regression, and machine-learning-based calibration models using random forests. The evaluation reveals that only the sensors calibrated with random-forest approach meet the US EPA Air Sensors Guidebook [22] recommendations of minimum data quality for personal exposure measurement. A similar study by Bigi et al. [8] investigates the medium-term performance of a set of NO and $NO_2$ electrochemical sensors using three different calibration approaches: Multivariate linear regression, support vector regression, and random forest. The behavior of the sensing devices over time and after a relocation was studied. It was noted that the performance of many algorithms strongly depends on the comparability of calibration and on the deployment area. The suitability of the devices for mapping intra-urban pollution gradients of NO and $NO_2$ was also studied. The devices could not reliably map small intra-urban gradients, thus they are not suitable for cleaner urban areas. Nevertheless, they can quantitatively resolve intra-urban concentration gradients on a hourly basis in higher polluted cities. Time and cost of the calibration of low-cost sensors can be reduced by firstly selecting sensors with similar responses. Then, a single on-site calibration for one sensor could be used for all sensors as the computed percent differences in the field are similar to laboratory results [23].

Although mobile sensing can be successfully applied to measure concentration of gases, such as ozone [24], and ultra-fine particles [25], it is more frequently adopted in monitoring air pollutants like $PM_{2.5}$. For example, AirCloud is a cloud-based monitoring system for $PM_{2.5}$ concentration using affordable sensors [26]. At the front-end, two types of Internet-connected particulate matter monitors are adopted, i.e., AQM and miniAQM, with a mechanical structure optimized for inlet air-flow. On the cloud-side, a novel air quality analytic engine calibrates the sensed data to improve accuracy. Overall, the project enables the adoption of low-cost sensors based on light-scattering for public air quality monitoring. In Mosaic, another low-cost urban $PM_{2.5}$ monitoring system based on mobile sensing, monitoring nodes are first built with a novel constructive airflow-disturbance design based on a carefully tuned airflow structure and a GPS-assisted filtering method [27]. Then, the buses used for system deployment are selected by a novel algorithm that achieves both high coverage and low computation overhead.

## III. System Overview

This section includes a description of our architecture from several points of view, going from the hardware and software architecture, to the communication protocols.

### A. Hardware Architecture

We target the following key characteristics for our system: (1) rapid and easy prototyping capabilities, (2) flexibility in connection scenarios, and (3) cheapness but also robustness of components. As each board has to include a limited number of modules, to facilitate our prototype development, we select the Raspberry Pi (RPi) [28] single-board computer as the monitoring board. Due to our constraints in terms of cost, size, and power consumption, we chose the Zero Wireless [29] version based on the ARM® 11 microprocessor. All sensors will be plugged to it as shown in Fig. 1.

The basic operating principle of the system is the following. The data gathered from the sensors are stored in the MicroSD card of the RPi board. At certain time intervals the RPi tries to connect to a Wi-Fi network and, if such a connection is established, it uploads the newly acquired data to a remote server. The creation of the Wi-Fi network is achieved using a mobile phone set to operate a personal hot-spot, while on the remote server the database storing all the performed measurements resides.

### B. Software Architecture

Wi-Fi connectivity was one of the requirements for the system, but at the same time, the system itself should not produce unnecessary electromagnetic noise, possibly impacting the operating ability of the host's appliances. To reduce the time in which the Wi-Fi connection was active, the Linux OS was set to activate the specific interface at predefined time instants in order to connect to the portable hot-spot. Once connected to the network, the system performed the following tasks: (1) synchronization of the system and Real-Time Clock with a remote Network Time Protocol (NTP) server [30], (2) synchronization of the local samples directory with the remote directory residing on the server. The latter task is performed using the UNIX `rsync` utility, which has to be installed on both machines.

To gather data from the sensors, a `C` program has been implemented, which runs continuously with a separate process reading from each physical sensor plugged to the board and writing on the MicroSD card. It has to be noted that for what concerns the PM sensors, since the UART communication had to take place using GPIOs, a `Pigpiod` daemon [31] has been exploited, to create digital serial ports over the Pi's pins.

The directories on the remote server are a simple copy of the MicroSD cards mounted on the boards. Data in these directories have been inserted in a MySQL database with the structure depicted in Fig. 2.

### C. Mechanical Design and Hardware Components

In order to easily stack more than one device together, a 3D printed modular case has been designed. Several enclosing frames can be tied together using nuts and bolts, with the use of a single cap on top. Fig. 3 shows the 3D board design, together with the final sensor and board configurations.

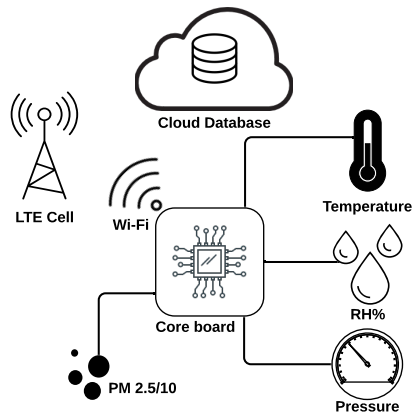



Fig. 1: Architecture of the proposed system. The data coming from the sensors are first stored in Raspberry Pi, and then transferred to a remote server over the Wi-Fi network.

Each platform is equipped with 4 PM sensors (a good trade-off between size and redundancy), 1 Temperature (T) and Relative Humidity (HT) sensor and 1 Pressure (P) sensor. As our target is to capture significant data sampling for the particulate matter we adopt the following sensors:

- The Honeywell® HPMA115S0-XXX [32] as PM sensor. As one of our targets is to evaluate these sensors' suitability for air pollution monitoring applications, we insert 4 instances of this sensor in every single platform. This sort of redundancy allows us to detect strange phenomena and to avoid several kinds of malfunctions, making more stable the overall system.
- The DHT22 [33] as temperature and relative humidity sensor. This is very widespread in prototyping applications, with several open-source implementations of its library, publicly available on the internet [34].
- The Bosch® BME280 [35] as a pressure sensor. This is a cheap but precise barometric pressure and temperature sensor which comes pre-soldered on a small PCB for easy prototyping.

The system also includes a Real Time Clock (RTC) module for the operating system to retrieve the correct time after a sudden power loss, i.e., the DS3231 module. The DS3231 communicates via *I2C* interface and has native support from the Linux kernel.

As a last comment, a Printed Circuit Board (PCB) was designed to facilitate connections and soldering of the various sensors and other components.

## IV. Data Acquisition and Measurement Sites

Our data acquisition campaign was carried out during autumn and winter, from October 2018 to February 2019, in the city of Turin in the north-west region of Italy. We deployed our stations to include a wide range of environments, such as residential boroughs, commercial areas, and parks. Each location corresponds to a particular GPS coordinate, POI (Point Of Interest), and sensor readings. Fig. 4 shows a map of the city of Turin. It represents the deployment positions of the stationary stations of some of our sensors and the paths followed by a few other sensor platforms during the analysis of dynamic and mobile events. Blue pins represent stationary sensors outside and inside the limited traffic zone. The orange pin represents the traffic light position in which we performed some dynamic analysis. The green path shows the roadway followed by our sensors on a bus, in a car, on a bicycle, and on foot. The red pin indicates the position of the ARPA reference station.

Fig. 5 focuses on the $PM_{2.5}$ pollutant levels registered by all our stationary sensors, when placed inside the ARPA station, over a period of 5 months. Fig. 5a plots the readings of all sensors before calibration but with all plots vertically shifted to start from the same initial position. Plots consider our original data sample rate, i.e., one reading per second. This sample rate is definitely high, allowing us to analyze pollution levels both statically ad dynamically, in terms of

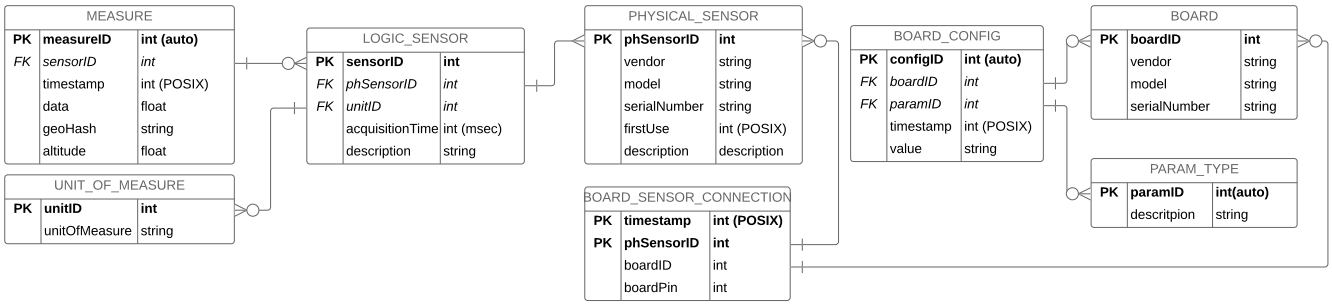
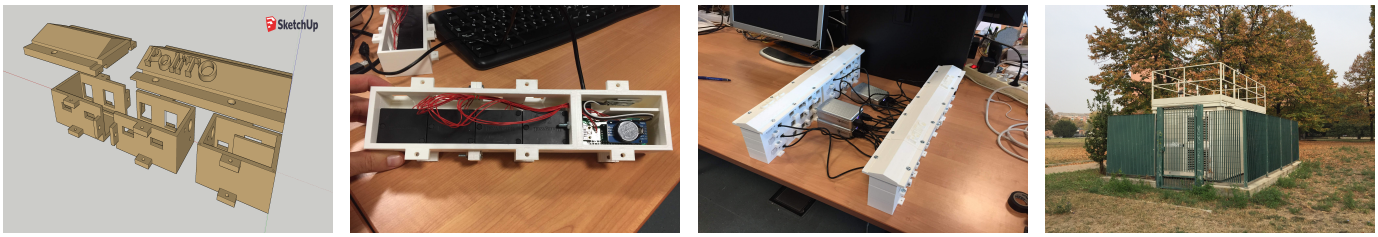



Fig. 2: ER-diagram of the database, in Crow's Foot notation.




(a) (b) (c) (d)

Fig. 3: Board design, sensors, and final measuring station. The stackable modular 3D printed case (a), a single sensor board with 4 sensors (b), the set of boards used during the calibration phase (c), and ARPA Rubino monitoring station (d).

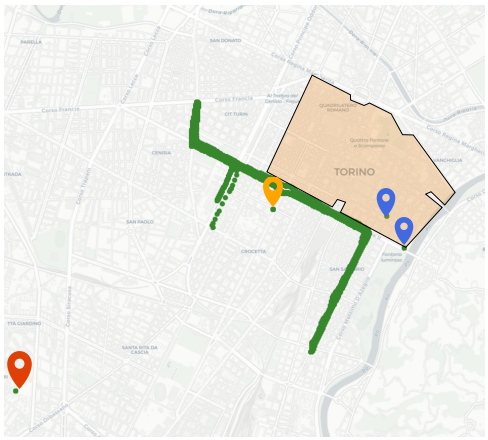



Fig. 4: The pins represent the following objects: The ARPA reference station (red), the stationary sensors outside and inside the limited traffic zone (blue), the traffic light (orange), The green path shows the roadway followed by our sensors during the dynamic analysis.

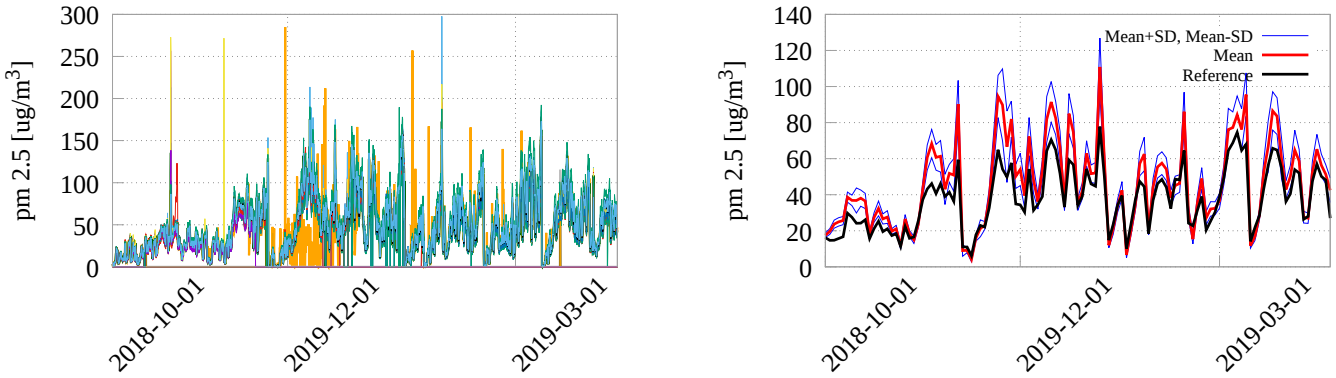
air pollution spatio-temporal variation. Fig. 5b represents the reference sensor readings with the mean values of all our sensors and the mean standard deviation, i.e., the mean values incremented and decremented by the standard deviation. To make the graph more readable than the one of Fig. 5a, we report daily averaged values. Fig. 6 reports the same data of Fig. 5b, i.e., the daily mean of the reference readings and of our sensor readings, within a scatter X-Y plot. The coefficient of determination (as computed by the `SciKitLearn` Python library [36] `r2_score`) is equal to 0.8267.

## V. The Calibration Phase

As low-cost sensors are prone to cross-sensitivities with other ambient variables [9], one of the main primary requirements in ambient measurement is their calibration. In general, temperature and relative humidity follow linear patterns, and linear regression has been the main technique used for low-cost sensors. Multivariate linear regression (MLR) analysis has been used to investigate several aspects of the air pollution over the years. For example, Chaloulakou et al. [37] used the regression models to investigate the complex relationships between the meteorological and time period parameters as factors controlling the PM levels. However, even if a sensor is calibrated, non-linearities sometimes appear due to the impurity and aging of low-cost sensing techniques. In these cases, accurate and precise calibration models are particularly critical, and there has been increasing interest in more sophisticated algorithms for low-cost sensor calibration [8]. Moreover, as reported by several researchers [38]–[40] artificial neural networks may give more accurate results than the multivariate linear regression model, mostly for $PM_{10}$ forecasting, even though the difference is often not remarkable. As a consequence, we experimented with three different calibration methods, namely Multivariate Linear Regression (MLR), Random Forest (RF), and the Support Vector Regression (SVR) model. As the last two methods delivered close results, we just concentrate on MLR and RF in the sequel.

### A. Calibration Strategies

*1) The Multivariate Linear Regression (MLR) Model:* In Multivariate Linear Regression Models, regression analysis is used to predict the value of one or more responses from a




(a) (b)

Fig. 5: Collected data for all statically deployed sensors (in the ARPA station) measuring $PM_{2.5}$ (48 overall) for a period of 5 months (from October 2018 to February 2019). All sensors are uncalibrated but their initial offset is modified to make graphs coincide in the origin. Fig. 5a reports the time series for all sensors. Fig. 5b plots the reference, the mean and the mean standard deviation (mean $\pm$ SD) for all our sensors.

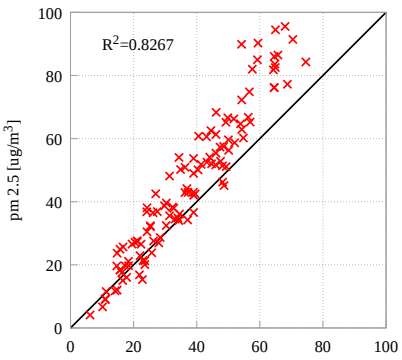



Fig. 6: Scatter plot comparing the reference data with our sensor mean represented in Fig. 5b.

set of predictors. Let $(x_1, x_2, \ldots, x_n)$ be a set of predictors (dependent variables) believed to be related to a response (independent) variable $Y$. The linear regression model for the j-th sample unit has the form

$$Y_j \quad = \quad \beta_0 + \beta_1 \cdot x_{j1} + \beta_2 \cdot x_{j2} + \ldots + \beta_r \cdot x_{jr} + \epsilon_j$$

where $\epsilon_j$ is a random error and the $\beta_i$ are unknown (and fixed) regression coefficients. The value $\beta_0$ is the intercept. With $n$ independent observations, we can write one model for each sample unit or we can organize everything into vectors and matrices as:

$$Y \quad = \quad X \cdot \beta + \epsilon.$$

The training data are used to calculate the model coefficients, and the model performance is evaluated on withheld testing data. Separate MLR models are usually developed for each sensor and each measure.

*2) The Random Forest (RF) Model:* A Random Forest Model is a machine learning algorithm for solving regression or classification problems. It works by constructing an ensemble of decision trees using a training data set. The mean value from that ensemble of decision trees is then used to predict the value for new input data.

To develop a random forest model, we must specify the maximum number of trees that make up the forest, and each tree is constructed using a bootstrapped random sample from the training data set. The origin node of the decision tree is split into sub-nodes by considering a random subset of the possible explanatory variables. The training algorithm splits the tree based on which of the explanatory variables in each random subset is the strongest predictor of the response. This process of node splitting is repeated until a terminal node is reached.

The user can specify the number of random explanatory variables considered at each node, the maximum number of sub-nodes or the minimum number of data points in the node as the indication to terminate the tree.

### B. Our Calibration Process

As $PM_{2.5}$ is heavily influenced by meteorology factors, we exploit the dependencies between the sensor error and meteorology factors. More specifically, the Honeywell® HPMA115S0-XXX Particulate Matter sensor has a relatively high precision ($\pm 15 \mu g/_{m^3}$ from 0 to 100 $\mu g/_{m^3}$), considered the extremely low price and the technology used.

Nonetheless, it is required to calibrate the collected data to remove possible offsets and linearity errors.

For that reason, during the calibration period, all sensors have been placed near the stationary ARPA station in the city of Turin, which exploits the $\beta$-radiation technology to provide high precision measures. ARPA provides hourly average data which have been used as a reference for all data collected from the sensor boards. Please note that hourly-average data, obtained with $\beta$-radiation approach, are fully consistent with gravimetric sensor measurements.

As far as the boards are concerned, we first apply different filters to remove outliers and possible undesired data. Then, we compute the window average with variable width to smooth the samples. After that, we group the values collected on a per-second basis to have hourly measures directly comparable with the ARPA values. Most samples fall in the $\pm 15 \mu g/_{m^3}$ range, which is reasonable considering the sensitivity of the sensor. Finally, we perform calibration on the hourly average samples. As previously introduced in this section, we consider Multivariate Linear Regression and Random Forest. Nevertheless, we apply three types of MLR, i.e., using only the temperature, only the humidity, and both temperature and humidity. In all cases, the calibration using RF considers using both temperature and humidity.

## VI. The Validation Phase

### A. Validation Strategies

The way to quantify the accuracy of a fitting model is by minimizing some error function that measures the misfit between the output and the response function for any given value of the data set. In the following, we will use several metrics defined as in the `SciKitLearn` Python library [36]. We will indicate with $y_i$ the true value of the $i$-th sample, $\hat{y}_i$ the corresponding predicted value, and $\bar{y}$ as the mean of the true samples:

- The coefficient of determination is the proportion of the variance in the dependent variable that is predictable from the independent variable:

$$\mathrm{R}^2(y, \hat{y}) \quad = \quad \frac{\sum_{i=0}^{n_{\text{samples}}-1} (\hat{y}_i - \bar{y})^2}{\sum_{i=0}^{n_{\text{samples}}-1} (y_i - \bar{y})^2} \tag{1}$$

- The mean squared error (MSE) measures the average of the squares of the errors. It is the second moment (about the origin) of the error and thus incorporates the variance of the calibration curve:

$$\mathrm{MSE}(y, \hat{y}) \quad = \quad \frac{1}{n_{\text{samples}}} \cdot \sum_{i=0}^{n_{\text{samples}}-1} (y_i - \hat{y}_i)^2 , \tag{2}$$

- The Mean Bias Error (MBE) is usually adopted to capture the average bias in a prediction.

$$\mathrm{MBE}(y, \hat{y}) \quad = \quad \frac{1}{n_{\text{samples}}} \cdot \sum_{i=0}^{n_{\text{samples}}-1} (y_i - \hat{y}_i) . \tag{3}$$

- The root mean squared error (RMSE) allows comparing different sizes of data sets, since it is measured on the same scale as the target value. It is obtained as the square root of the MSE, i.e.,

$$\mathrm{RMSE}(y, \hat{y}) \quad = \quad \sqrt{\mathrm{MSE}(y, \hat{y})}. \tag{4}$$

- The CRMSE is the Root Mean Square Error (RMSE) corrected for bias, i.e., it is defined as:

$$CRMSE \quad = \quad RMSE \cdot sign(\sigma_{model} - \sigma_{reference}) \tag{5}$$

where $\sigma$ is the standard deviation of the measure.

- The correlation coefficient (Pearson product-moment correlation coefficient) is defined as the covariance of the variables divided by the product of their standard deviations.

$$\rho_{y,\hat{y}} \quad = \quad \frac{cov(y,\hat{y})}{\sigma_y \cdot \sigma_{\hat{y}}} \tag{6}$$

### B. Our Validation Process

To test the performance of the two different calibration models, we first calibrate our sensors using the data collected in the first 2 weeks of October 2018. Then, we validate these calibration methods using samples collected in the last 2 weeks of the same month. In this period, we compare the concentrations obtained after the calibration with the measured reference concentrations.

For the sake of simplicity, Fig. 7 shows our results for one single sensor (sensor 34), randomly selected, using the MLR model. For this model, the three plots present the data gathered using as dependent variable only the temperature, only the humidity, and both the variables as free variables. For all graphics, the calibrated plot is far more stable than the original one, but there is no clear winner among the three strategies.

To deepen our analysis, Fig. 8 compares the MLR model with the RF one (again using sensor 34). As for Fig. 7 calibration is performed during the first 2 weeks of October and validation during the last 2 weeks. In this case, we use both the temperature and the humidity as free variables. The charts report time series (Fig. 8a and 8c) and scatter plots (Fig. 8b and 8d). Somehow unexpectedly (please see Zimmerman et al. [9]) our results show no advantage for the RF model with respect to the MLR one. On the contrary, the MLR model seems to outperform the RF model.

To better evaluate our results and to better assess the overall model performance, we performed calibration and validation tests for longer periods. A secondary target of this analysis is to find the best trade-off between the calibration effort and the error obtained. We consider calibration periods varying from the 2 weeks used so far up to 12 weeks, starting in October and ending in November 2018. In all the cases, the validation period has been selected in December 2018.

Table I reports the RMSE (i.e., the Root Mean Square Error, computed as defined by Equation 4) and the data correlation (computed following Equation 6) for a representative sensor with different calibration periods (2, 6, and 12 weeks, respectively).

### C. Final Considerations

MLR has been used during calibration because it is the strategy historically adopted for this phase and it is the one that delivers the best results in our environment. Anyway, we also experimented with the RF and the Support Vector Regression (SVR) models. These latter methodologies are supposed to perform better with non-linear measures [8], [9] and they

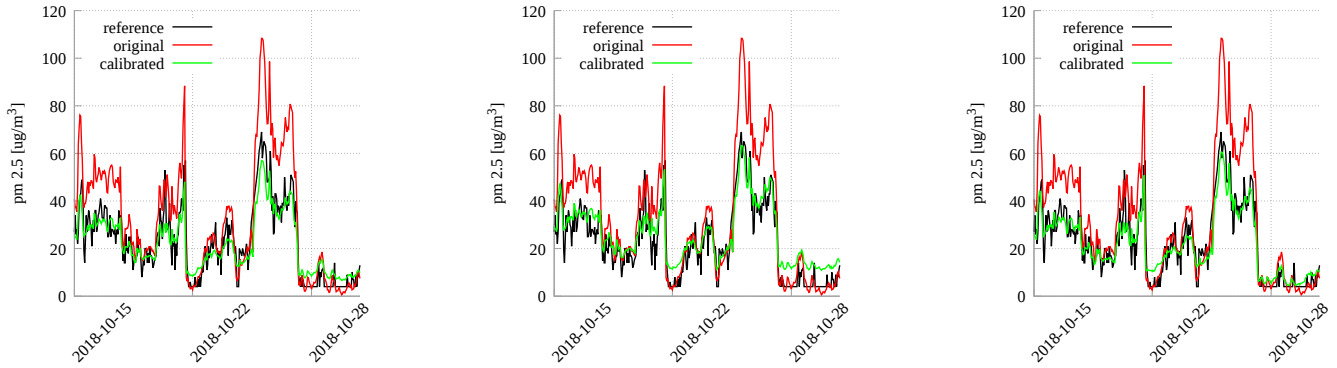



Fig. 7: Time series comparing the reference, the raw, and the calibrated data using sensor 34, randomly selected. Calibration is performed using MLR. The different plots show MLR with different dependent variables, namely temperature (a), humidity (b), and temperature plus humidity (c).

| | | 2 weeks 323 samples | | | 6 weeks 840 samples | | | 12 weeks 1851 samples | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | T | H | T+H | T | H | T+H | T | H | T+H |
| LR | RMSE | 19.03 | 14.17 | 14.42 | 18.96 | 12.58 | 14.70 | 13.49 | 10.04 | 11.66 |
| | Correlation | 0.89 | 0.88 | 0.89 | 0.89 | 0.88 | 0.89 | 0.88 | 0.88 | 0.89 |
| RF | RMSE | 25.41 | 21.77 | 23.50 | 25.28 | 18.75 | 19.37 | 13.85 | 12.63 | 11.33 |
| | Correlation | 0.82 | 0.80 | 0.78 | 0.77 | 0.82 | 0.80 | 0.85 | 0.88 | 0.89 |

TABLE I: Comparing different calibration techniques over different calibration periods (2, 4, and 6 weeks, respectively), adopting the RMSE (Root Mean Square Error) metric. We consider all stationary sensors.

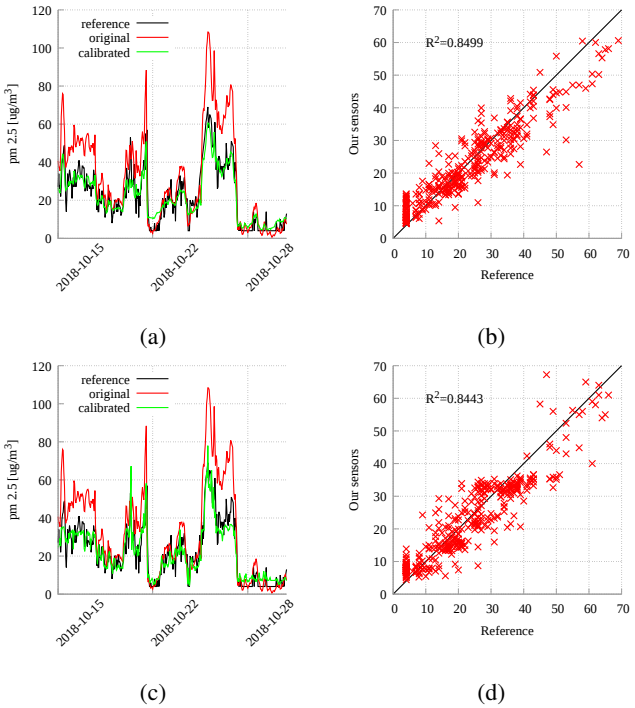



Fig. 8: Comparing different calibration techniques on our sensor network using sensor 34, randomly selected. The graphics compare linear regressions, with temperature and humidity as the dependent variables (a and b), versus random forest (c and d). The figure reports time series (a and c) and scatter plots (b and d).

delivered very similar results. Actually, the results we gathered using RF look slightly disappointing in our environment. In our opinion, this may be due to three different reasons. First of all, an important difference between MLR and RF is that MLR allows to extrapolate outside the range of its input dataset, while the estimates provided by RF can only be within the bounds of the calibration space. This behavior can motivate our results as in our framework short calibration periods have been followed by long working sessions with wider measure variations. This consideration is strengthen by the data of Table I, which show how the computed errors (RMSE and correlation) seem to decrease the longer the calibration period. Secondly, we focused on particulate matter, whereas Bigi et al. and Zimmerman et al. [8], [9], who reported favorable results for non-linear models, concentrated on CO, $NO_2$, $O_3$, and $CO_2$. We can then conclude that MLR works at its best for our low-cost sensors measuring particulate matter. Thirdly, many works estimating the concentration of $PM_{10}$ obtained a much lower accuracy than the one gathered by us. For example, Alam et al. [40] claimed a $R^2$ coefficient equal to 66% in the best case. On the contrary, we obtained values close to 90% for the same metric. Thus, with this level of accuracy non-linearity effects may have a much lower impact and RF may not be able to improve the results obtained with MLR.

Overall, another remarkable result is that we obtained better performance of the regression process using only the relative humidity instead of both the humidity and the temperature. This makes it possible to simplify the hardware (and software) structure of our platform, by removing the temperature and the pressure sensors, as they do not provide any significant

advantage when performing calibration. It is also important to note that measuring temperature is not easy in our case, because the sensors present cross-heating, mainly due to the heat given by the particulate matter sensors, which use more electrical power. Better results in measuring temperature could be obtained by putting a separate sensor in a different box, but this would negatively impact on the size and portability of the global system. Atmospheric pressure is simpler to be measured, but it is less significant than temperature. Relative humidity is by far the most important parameter for calibrating PM sensors. Furthermore, relatively short calibration periods (2 to 4 weeks) perform quite well compared to considerably longer periods (10 to 12 weeks), thus reducing the necessity of performing long calibration sessions.

## VII. Stationary Platforms

Once the previously described calibration and validation phases have been performed, we analyze air quality and intra-urban pollutant gradients over long periods of time.

Following Fig. 5, Fig. 9 and Fig. 10 focus on the $PM_{2.5}$ air pollutant levels registered by all our stationary sensors, placed in the ARPA station, over a period of 5 months. All sensors have been calibrated using MLR with humidity as the dependent variable. The graphics report: The time-series plot for all sensors with one reading per second (Fig. 9a), the mean and the mean standard deviation using daily averaged values (Fig. 9b), and the scatter X-Y plot (Fig. 10) comparing the reference values with our mean ones. Compared with Fig. 5, Fig. 9 shows much lower peaks, a much lower deviation, and more regular behavior. In fact, the coefficient of determination based on the data plotted in Fig. 9b is 0.9177, higher than the one computed for Fig. 5b.

Fig. 11 compares our readings with the reference one for a period of about 2 months. Given a randomly selected platform, we first calibrated its 4 sensors (namely sensors 123, 125, 127, and 129). We adopt the MLR calibration model, with relative humidity as the dependent variable, over 2 weeks. Then, we analyze the behavior of the 4 sensors over 12 weeks. Fig. 11a reports the time-series, and Fig. 11b the corresponding scatter plot, for sensor 123. As it can be easily noticed, readings and fluctuations are very similar to the reference ones for the entire period. The other sensors are not reported for the sake of brevity, but they show a very similar behavior. This can be verified on the target diagram [41] of Fig. 11c, which considers all 4 sensors. In a target diagram, the x axis indicates the CRMSE (computed as in Equation 5) and the y axis the MBE (please, refer to Equation 3), both normalized by the standard deviation of the reference $\sigma_{reference}$. As a consequence, the vector distance between any given point and the origin is the RMSE normalized by the standard deviation of the reference measurements. Again, all sensors deliver points within the unit circle both before and after calibrations, with the second set of points closer to the origin than the first one.

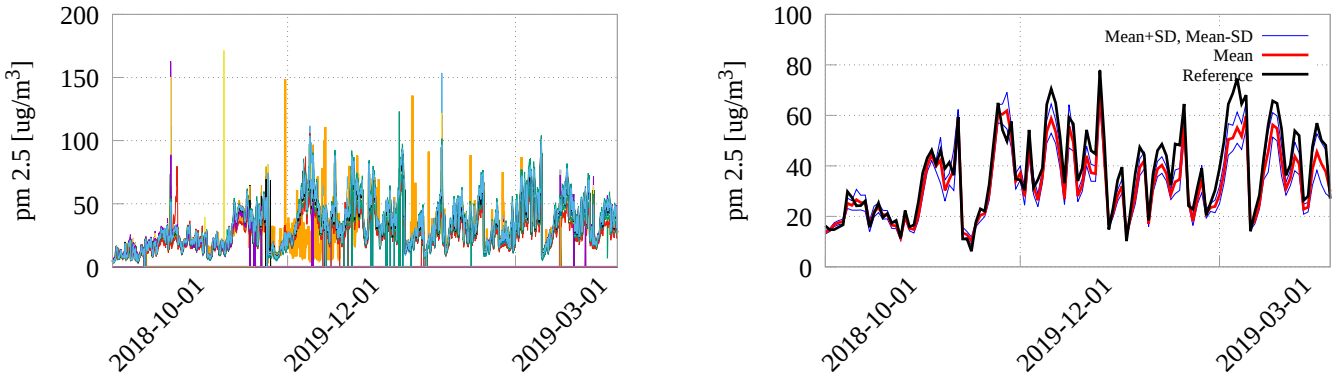
To analyze intra-urban pollutant gradients, Fig. 12 shows a test performed with two platforms (each one including 3 sensors) placed at the border and inside the Limited Traffic Area (LTA) in the city of Turin. This test has been repeated for two days, on December the 3rd, and on December the 4th, 2018, from 4:00 p.m. to 4:30 p.m. (local time). These dates have been selected due to the fact that they were at the beginning of an emergency traffic control, which has been in progress until December 8th, due to unhealthy pollution levels in the previous days. The time is representative of the rush hour, with significant traffic increments with respect to other hours. At first glance, it can be noted that albeit all sensors of each station delivered very similar readings, the measurements carried out inside the LTA are significantly higher than those collected outside. A few considerations may address this issue. The location outside LTA is indeed a very crowded crossing, but it is located in an open area, in the immediate proximity of the Po river, the longest river in Italy. These factors may greatly impact the concentration of PM in the air since the area is very wide and the river could induce current flow just by its own motion. On the contrary, the location inside the LTA is represented by the crossing of two narrow streets. This situation, combined with the height of the surrounding buildings (up to 7-story), could greatly reduce air circulation, resulting in concentration values that are higher with respect to the outside location. It has to be pointed out that this test does not undermine the effectiveness of LTA regulations in reducing pollution concentrations in the urban center. Nevertheless, it gives some insights and hints on possible studies which could be carried out in order to increase the level of understanding of the phenomena happening in city areas.

## VIII. Mobile Platforms

In parallel to the systematic data acquisition and calibration campaign, we carried-out targeted experiments to investigate the robustness and accuracy of the system in dynamic applications. As our architecture is able to collect one sample per second, we are able to analyze sharp events in terms of air pollution spatio-temporal variation. This study also allows us to investigate to what extent a limited set of mobile measurements permit us to draw conclusions on global urban air pollution.

*1) Traffic Light Case Study:* Fig. 13 analyzes the pollution levels registered in close proximity of a traffic light within the city center at rush hour. It reports the level of particulate matter registered for a period of 15 minutes (from about 08:00 to about 8:15 a.m.), including several traffic-light cycles. The time-series shows a wave behaviors essentially highlighting 3 different traffic conditions at the traffic light. The higher levels, around 102 $\mu g/_{m^3}$ for $PM_{10}$ and 77 for $PM_{2.5}$, were recorded with a long queue during red traffic lights. Intermediate levels, around 97 $\mu g/_{m^3}$ for $PM_{10}$ and 72 for $PM_{2.5}$, were recorded with average traffic, i.e., a few cars passing by. The lower levels, with a concentration of around 93 $\mu g/_{m^3}$ for $PM_{10}$ and 69 for $PM_{2.5}$ , were recorded with no traffic. Overall, data shows high variability in time, as it can be expected due to the vehicles' stop-and-go at the traffic light [42], and our platforms demonstrate a very good reactivity to transient events.

*2) Citizens' Mobility Case Study:* Following [43], Fig. 14 focuses on pollution levels recorder with portable platforms using different means of transportation along the same route. The path followed, i.e., a 4 Km path crossing the city center of Turin, is represented by the green route in Fig. 4. The different concentrations are plotted with distinct colors depending on




(a) (b)

Fig. 9: Collected data for all statically deployed sensors (in the ARPA station) measuring $PM_{2.5}$ (48 overall) for a period of 5 months (from October 2018to February 2019). All sensors are calibrated during the first two weeks of October, using MLR with humidity as the dependent variable. Fig. 9a reports the time series for all sensors. Fig. 9b plots the reference, the mean and the mean standard deviation (mean $\pm$ SD) for all our sensors.

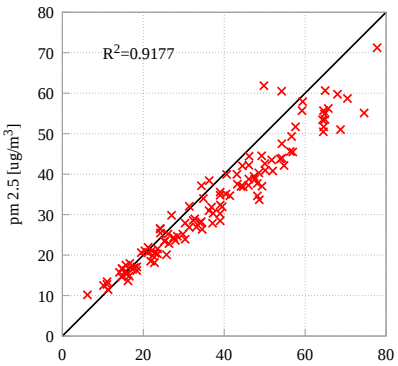



Fig. 10: Scatter plot comparing the reference data with our sensor mean represented in Fig. 9b.

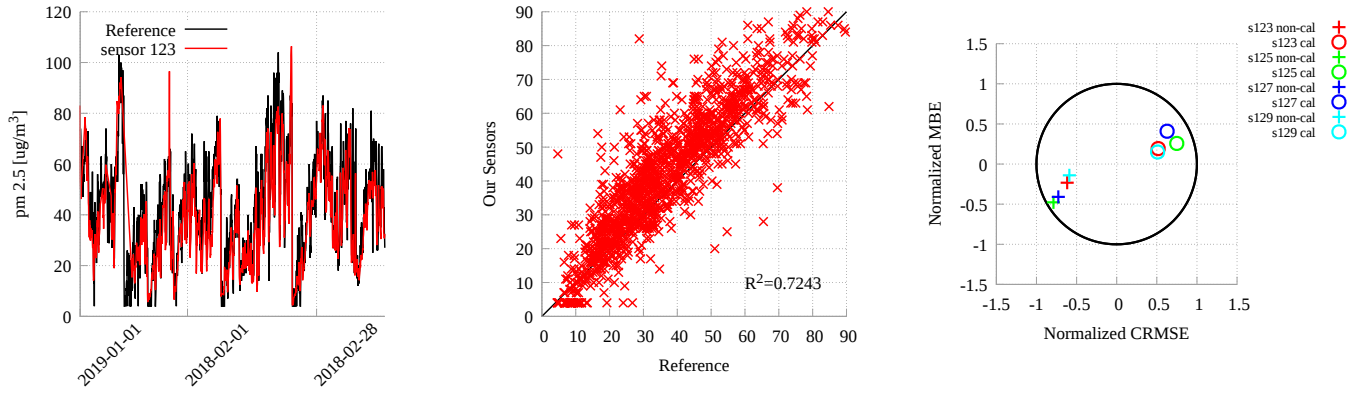
the transportation means. The two transport means which registered the highest pollution concentrations are the tram and the bus. Both of them carry dozens of passengers and run in the street, alongside the usual fossil fuel-propelled vehicles. Analyzing the bike plot, it is possible to guess that the fluctuations of the pollution level mainly depend on the varying number of cars and trucks the user had to follow on his path. Anyhow, if we compare the bike plot with the one obtained on foot, the bike user is exposed to a lower pollution level. The data gathered underground are almost constant, thanks to the air re-circulation and filtering devices used in metro stations and metro trains.

The plot of the car shows concentration levels in the cabin gradually decreasing, due to the car's air re-circulation active. From the driver's point of view, the car looks like the best mean of transport among the ones analyzed. To further analyze this issue, Fig. 15 shows the effect of opening the car window during the previous experiment. It is possible to observe an abrupt increase of pollutant in the middle of the graph.

Finally, Fig. 16 shows the pollution registered on a bike trip, running the bike at different speeds. The target was to verify whether the bicycle speed has some influence on the pollution level to which the rider is exposed. The bike was running at about: 8 Km/h from 15:50 to 15.54, 13 Km/h from 15:54 to 15.58, 23 Km/h from 15:58 to 16.02, and at 27 Km/h from 16:02 to 16.06. The graphic shows that there is no much difference in the pollution level encountered while running the bike at different speeds.

*3) Wind analysis:* Fig. 17 compares pollution before, during, and after the föhn wind arrived in the city of Turin, on October 28th, 2018. This brought an increase in the temperature, but also naturally helped cleaning the air in the city. We follow the event with 2 stations, generating the graphics in Fig. 17a and 17b respectively. The two readings are very coherent, considering that Fig. 17a reports one sensor readings per second, whereas Fig. 17b considers hourly averaged values.

*4) New Year's eve fireworks:* Fig. 18 shows the pollution level during the New Year Firecrackers in Turin. The quick rise of PM concentration is reported about 10 minutes after midnight, with a peak of about 4 times higher than the average value before the event. Moreover, the level remains twice as higher than before the event for more than 1 hour.




(a) (b) (c)

Fig. 11: Time-series (a) and scatter plots (b) for the data coming from sensor 123 over 2 months. Calibration was performed with MLR, with humidity as the dependent variable, for a period of 2 weeks. The target diagram of figure (c) show the 4 sensors (123, 125, 127, and 129) of the board, considering both uncalibrated and calibrated measures.

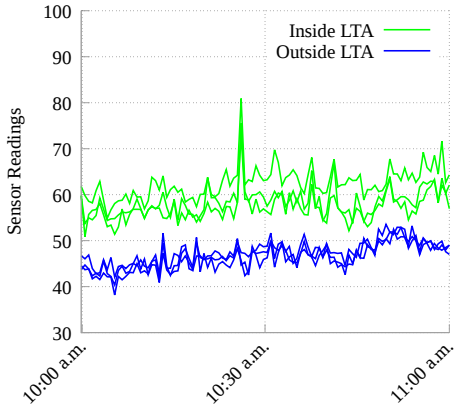


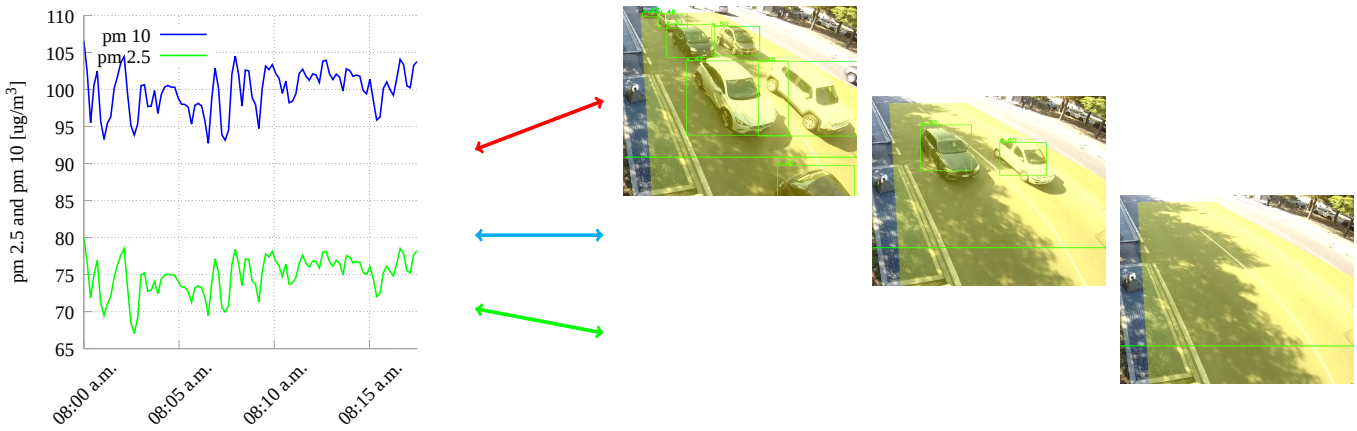

Fig. 12: A time-series comparison between two different sites, one inside and the other one outside the Limited Traffic Area.

## IX. Data Analysis and Lessons Learned

Since our system has been designed to appropriately scale to difficult environments (like factories), it is important to underline the practical difficulties we encountered to design and to build it. Developing the hardware and the software system has required a considerable team effort in terms of man-hours. From the hardware point of view, using more than one PM sensor for each station (in our case 4) has been very useful in order to better compare the behavior of each low cost sensor. The high number of sensors stacked in each board is for sure a point of strength of this study when compared with previous ones, in particular for its use over very long time periods. From the software point of view, the database grows very fast and it requires a powerful computer to be managed efficiently. In order to have a fast response to SQL queries, we used a Dell computer with 20 cores, 384 GB of internal memory, and some TB of hard-disk. Smaller computers would have potentially implied very slow query evaluation. Many different discussions can be done using the data and the system. The main focus here is on the system itself and on some feasibility studies in order to demonstrate how well and accurately it works, and how reliable it is. We plan to use the system for specific purposes, like environment studies of the air quality in interior settings, in addition to all uses already introduced here. For that reason, all data (i.e., the entire database), software and hardware designs are available on IEEE *DataPort* in Open Access mode [1] in order to make easier further evaluations and facilitate comparisons with other approaches.

As far as the calibration phase is concerned, we can make the following observations. Even if RF has the ability to build a non-linear regression model and has been proved to be superior to MLR in handling measures on CO, $NO_2$, $O_3$, and $CO_2$, in our analysis MLR performs better. This somehow motivates its wide use to calibrate low-cost sensors. A difference between MLR and RF is that MLR allows to extrapolate outside the range of its input data-set, while the estimates provided by RF can only be within the bounds of the calibration space. This behavior, due to the intrinsic nature of RF, based on tree manipulations, can partially motivate our data, as short calibration periods have been followed by long working sessions with wider measure ranges. Another important aspect of our calibration phase is that MLR proved to be more accurate when using only the relative humidity as dependent variable instead that the humidity and the temperature together or only the temperature. This was somehow unexpected, and it may lead to a possible simplification of the hardware and software platforms which may avoid collecting and storing data coming from the temperature and the pressure sensors. As previously stated, simplifying the hardware platform may be very useful and measuring the temperature is difficult for the cross-heating effect between sensors due to the electrical power used by the PM sensors.

Another important aspect we focused on was to understand whether the tested sensor units are appropriate to capture particulate matter concentration with high resolution. From that point of view, our results are comfortable, as our platforms may significantly improve our ability to resolve spatial and




(a) (b) (c) (d)

Fig. 13: Time-series comparing different traffic conditions at a traffic light. Three different conditions (high traffic (b), average (c) and low traffic (d) levels) repeat themselves continuously. The test has been performed the 8th of January 2020.

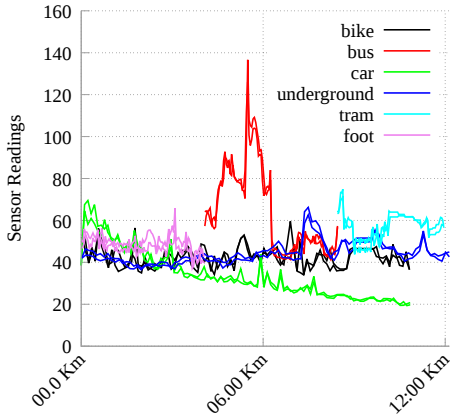



Fig. 14: A time-series comparison among different means of transport along the same path (reported in Fig 4).

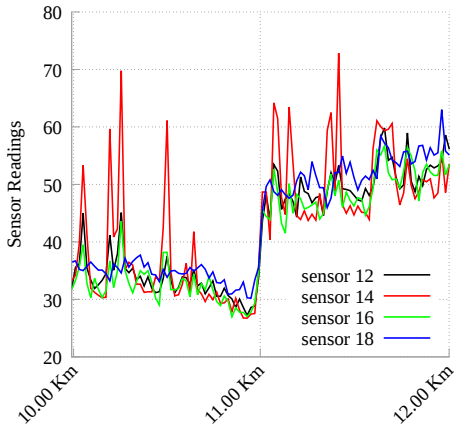



Fig. 15: Effect of opening a window (halfway on the $x$-axis) on a car traveling along the green path of Fig. 4.

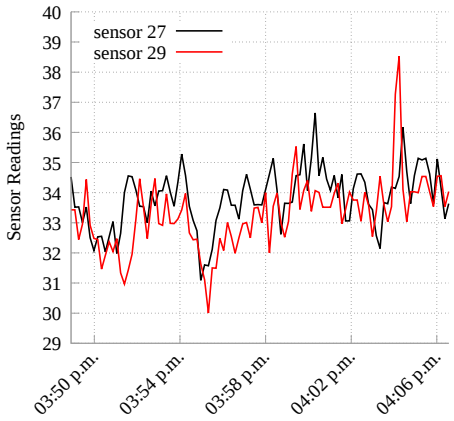
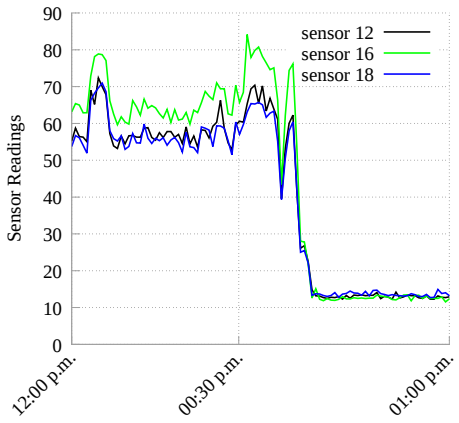


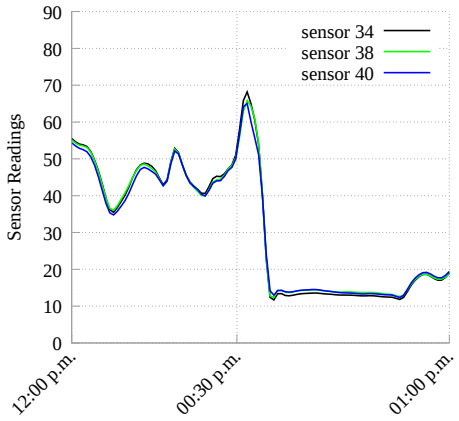

Fig. 16: Registered pollution on a bike running at different speeds (from about 8 Km/h to about 27 Km/h).

time heterogeneity in air pollutant concentrations. We have to consider that numerous factors are involved in the variation of the atmospheric pollution, and often the relationship between the intensity of emission produced by the polluting source and the resulting pollution is not immediate. For that reason, our sample rate (one sample per second) may even be considered higher than necessary to capture even dynamic phenomena. Among the future works, we would like to mention the necessity to evaluate the best possible trade-off between the sample-rate, the dimension of the data-base (that can become really large), and the ability to follow high gradients.

As our system is based on boards including 4 sensors, another feature of our platform is the combination of information coming from a multitude of sources. There are several issues that arise when fusing information from multiple sources, but the most fundamental one is to combine information in a coherent and synergistic manner to obtain a robust, accurate and reliable description of the quantities of interest. One of the feature we are working on at the moment is to model




(a) Platform A (sensors 12, 16, and 18).




(b) Platform B (sensors 34, 38, 40).

Fig. 17: Time-series representing the pollution variation when the wind comes on two stations each one equipped with three $PM_{2.5}$ sensors.

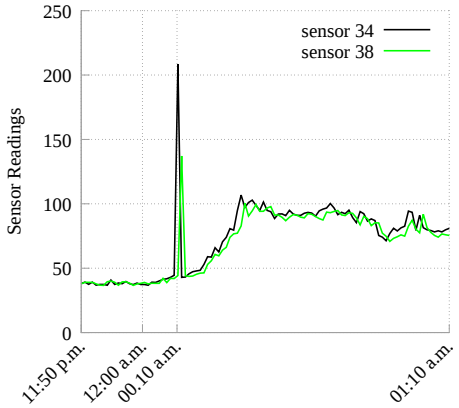



Fig. 18: Pollution level during the New Year Firecrackers (data measured near the ARPA station).

the inherent uncertainties in the sensor measurements due to sensor aging, and to spurious data readings. Even if there is little literature regarding recent advanced fusion techniques and emerging applications, we are developing a fusion strategy based on Bayesian methods that can identify the inconsistency in sensor data so that spurious data can be eliminated from the sensor fusion process.

## X. Conclusion

We approach the challenging problem of accurate and affordable $PM_{2.5}$ monitoring by orchestrating several low-cost sensor units deployed within an urban scenario. More specifically, we first carefully design and build our sensor stations with a high level of sensor redundancy (4 particulate matter sensors are embedded in every station). Then, we placed them on the field and we field-calibrated them. Finally, we deployed them to measure the air quality on stationary monitoring sites and during specific and sudden events.

From a broad perspective, our findings investigate whether the adopted sensors are fit for the intended purpose and the intended environment. Given their low-cost it is possible to deploy a large number of monitoring stations throughout a city providing a spatial dense coverage. Sensor redundancy reduces errors and measurement problems. They proved to be extremely easy to calibrate, as short calibration time (less than 2 weeks) and simple calibration methods (MLR) are sufficient. As they offer a very high sample rate (one sample per second) they are able to follow sudden changes in the environment, providing a temporal dense coverage. One of our goals is that the developed system can be used by researchers and practitioners in order to estimate the level of air pollution and investigate the general behavior of particulate matter.

As future work, we hope to use our sensor stations to further improve our model, and to solve a number of environmental problems, such as identifying pollution sources and air-quality prediction. Moreover, we would like to better analyze the issue of sensor aging and de-calibration and better use the redundancy present in each monitoring station to make each station more reliable and resilient. Furthermore, we would like to improve the experiments in the Participatory Sensing direction, enabling university students to install a monitoring station at their home.

## Acknowledgment

We thank Mauro Guerrera for his effort in the implementation and testing phase. We also thank Prof. Maurizio Rebaudengo for his useful comments and suggestions on the topic.

Authors thank ARPA Piemonte for putting at disposal official public hourly pollution data, used as a reference, and for letting us install and test boards inside the “Rubino” station. Data provided in this paper (see [1]) 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 paper, as well as for any error in the experimental values.

## REFERENCES


[1] B. Montrucchio, E. Giusto, M. Ghazi Vakili, S. Quer, R. Ferrero, and C. Fornaro, "A densely-deployed, high sampling rate, open-source air pollution monitoring wsn," 2020. [Online]. Available: http://dx.doi.org/10.21227/m4pb-g538

[2] "European Union Tech Report: Measuring air pollution with low-cost sensors. Thoughts on the quality of data measured by sensors," Tech. Rep. [Online]. Available: https://ec.europa.eu/environment/air/pdf/Brochure%20lower-cost%20sensors.pdf

[3] 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 *IEEE International Conference on Emerging Technologies and Factory Automation, ETFA*, vol. 2018-September. Institute of Electrical and Electronics Engineers Inc., oct 2018, pp. 838–844.

[4] 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 *ACM International Conference Proceeding Series*. New York, New York, USA: Association for Computing Machinery, nov 2018, pp. 55–58. [Online]. Available: http://dl.acm.org/citation.cfm?doid=3284869.3284880

[5] 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 *Proceedings of the 5th International Conference on Internet of Things, Big Data and Security*. SCITEPRESS - Science and Technology Publications, may 2020, pp. 392–398. [Online]. Available: http://www.scitepress.org/DigitalLibrary/Link.aspx?doi=10.5220/0009470203920398

[6] D. J. Pagliari, M. Poncino, and E. Macii, "Energy-efficient digital processing via approximate computing," in *Smart Systems Integration and Simulation*, 2016.

[7] Y. Chen, D. Jahier Pagliari, E. Macii, and M. Poncino, "Battery-aware design exploration of scheduling policies for multi-sensor devices," in *Proceedings of the ACM Great Lakes Symposium on VLSI, GLSVLSI*, 2018.

[8] A. Bigi, M. Mueller, S. K. Grange, G. Ghermandi, and C. Hueglin, "Performance of NO, $NO_2$ low cost sensors and three calibration approaches within a real world application," *Atmospheric Measurement Techniques*, vol. 11, no. 6, pp. 3717–3735, 2018.

[9] N. Zimmerman, A. A. Presto, S. P. N. Kumar, J. Gu, A. Hauryliuk, E. S. Robinson, A. L. Robinson, and R. Subramanian, "A machine learning calibration model using random forests to improve sensor performance for lower-cost air quality monitoring," *Atmospheric Measurement Techniques*, vol. 11, no. 1, pp. 291–313, 2018.

[10] R. V. Martin, "Satellite remote sensing of surface air quality," *Atmospheric environment*, vol. 42, no. 34, pp. 7823–7843, 2008.

[11] Y. Kawamoto, H. Nishiyama, Z. M. Fadlullah, and N. Kato, "Effective data collection via satellite-routed sensor system (SRSS) to realize global-scaled Internet of Things," *IEEE Sensors Journal*, vol. 13, no. 10, pp. 3645–3654, 2013.

[12] M. Budde, R. El Masri, T. Riedel, and M. Beigl, "Enabling low-cost particulate matter measurement for participatory sensing scenarios," in *12th international conference on mobile and ubiquitous multimedia*, 2013, pp. 1–10.

[13] M. Sammarco, R. Tse, G. Pau, and G. Marfia, "Using geosocial search for urban air pollution monitoring," *Pervasive and Mobile Computing*, vol. 35, pp. 15–31, feb 2017.

[14] Y.-C. Wang and G.-W. Chen, "Efficient data gathering and estimation for metropolitan air quality monitoring by using vehicular sensor networks," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 8, pp. 7234–7248, 2017.

[15] D. Takaishi, H. Nishiyama, N. Kato, and R. Miura, "Toward energy efficient big data gathering in densely distributed sensor networks," *IEEE transactions on emerging topics in computing*, vol. 2, no. 3, pp. 388–397, 2014.

[16] J. Peters, J. Theunis, M. Van Poppel, and P. Berghmans, "Monitoring PM10 and ultrafine particles in urban environments using mobile measurements," *Aerosol and Air Quality Research*, vol. 13, p. 509–522, Apr. 2013.

[17] D. Aguiari, G. Delnevo, L. Monti, V. Ghini, S. Mirri, P. Salomoni, G. Pau, M. Im, R. Tse, M. Ekpanyapong, and R. Battistini, "Canarin II: Designing a smart e-bike eco-system," in *CCNC 2018 - 2018 15th IEEE Annual Consumer Communications and Networking Conference*, vol. 2018-Janua. Institute of Electrical and Electronics Engineers Inc., mar 2018, pp. 1–6.

[18] L. Yang, W. Li, M. Ghandehari, and G. Fortino, "People-centric cognitive internet of things for the quantitative analysis of environmental exposure," *IEEE Internet of Things Journal*, vol. 5, no. 4, pp. 2353–2366, 2018.

[19] P. Kumar, L. Morawska, C. Martani, G. Biskos, M. Neophytou, S. Di Sabatino, M. Bell, L. Norford, and R. Britter, "The rise of low-cost sensing for managing air pollution in cities," *Environment International*, vol. 75, pp. 199–205, Feb. 2015.

[20] J. Li, H. Li, Y. Ma, Y. Wang, A. A. Abokifa, C. Lu, and P. Biswas, "Spatiotemporal distribution of indoor particulate matter concentration with a low-cost sensor network," *Building and Environment*, 2018.

[21] 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 of Things Journal*, vol. 5, no. 2, pp. 559–570, 2017.

[22] "US EPA Air SensorsGuidebook." [Online]. Available: https://www.epa.gov/air-sensor-toolbox/how-use-air-sensors-air-sensor-guidebook

[23] S. Sousan, A. Gray, C. Zuidema, L. Stebounova, G. Thomas, K. Koehler, and T. Peters, "Sensor selection to improve estimates of particulate matter concentration from a low-cost network," *Sensors*, vol. 18, no. 9, Sep. 2018.

[24] D. Hasenfratz, O. Saukh, S. Sturzenegger, and L. Thiele, "Participatory air pollution monitoring using smartphones," *Mobile Sensing*, vol. 1, pp. 1–5, 2012.

[25] D. Hasenfratz, O. Saukh, C. Walser, C. Hueglin, M. Fierz, and L. Thiele, "Pushing the spatio-temporal resolution limit of urban air pollution maps," in *2014 IEEE International Conference on Pervasive Computing and Communications, PerCom 2014*, 2014.

[26] Y. Cheng, X. Li, Z. Li, S. Jiang, Y. Li, J. Jia, and X. Jiang, "AirCloud: a cloud-based air-quality monitoring system for everyone," *12th ACM Conference on Embedded Network Sensor Systems*, pp. 251–265, 2014.

[27] Y. Gao, W. Dong, K. Guo, X. Liu, Y. Chen, X. Liu, J. Bu, and C. Chen, "Mosaic: A low-cost mobile sensing system for urban air quality monitoring," in *35th Annual IEEE International Conference on Computer Communications (IEEE INFOCOM)*. IEEE, 2016, pp. 1–9.

[28] "Raspberry Pi." [Online]. Available: https://www.raspberrypi.org/

[29] "Raspberry Pi Zero W." [Online]. Available: https://www.raspberrypi.org/products/raspberry-pi-zero-w/

[30] "pool.ntp.org." [Online]. Available: https://www.pool.ntp.org/en/

[31] "pigpio library." [Online]. Available: http://abyz.me.uk/rpi/pigpio/index.html

[32] "HPMA115S0 Particulate Matter Sensors - Honeywell." [Online]. Available: https://sensing.honeywell.com/hpma115s0-particulate-matter-sensors

[33] T. Liu, "DHT22 humidity and temperature module/sensor," Tech. Rep. [Online]. Available: https://www.sparkfun.com/datasheets/Sensors/Temperature/DHT22.pdf

[34] "DHT22 library for Raspberry Pi." [Online]. Available: https://github.com/technion/loldht22

[35] "BME280." [Online]. Available: https://www.bosch-sensortec.com/bst/products/allproducts/bme280

[36] "API Reference — scikit-learn 0.21.3 documentation." [Online]. Available: https://scikit-learn.org/stable/modules/classes.html#module-sklearn.metrics

[37] A. Demokritou, P. Kasspmenos, P. Koutrakis, and N. Speyrellis, "Measurements of $PM_{10}$ and $PM_{2.5}$ particle concentrations in Athens, Greece," *Atmospheric Environment*, vol. 37, pp. 649—660, 2003.

[38] M. Caselli, L. Trizio, G. De Gennaro, and P. Ielpo, "A simple feed-forward neural network for the pm 10 forecasting: Comparison with a radial basis function network and a multivariate linear regression model," *Water Air and Soil Pollution - WATER AIR SOIL POLLUT*, vol. 201, pp. 365–377, 07 2009.

[39] K. Polat and S. Durduran, "Usage of output-dependent data scaling in modeling and prediction of air pollution daily concentration values ($PM_10$) in the city of Konya," *Neural Computing and Applications - NCA*, vol. 21, 11 2011.

[40] M. S. Alam and A. McNabola, "Exploring the modeling of spatiotemporal variations in ambient air pollution within the land use reression framework: Estimation of $PM_{10}$ concentrations on a daily basis," *Journal of the Air & Waste Management Association*, vol. 65, no. 5, pp. 628–640, 2015, pMID: 25947321. [Online]. Available: https://doi.org/10.1080/10962247.2015.1006377

[41] J. Jolliff, J. Kindle, I. Shulman, B. Penta, M. Friedrichs, R. Helber, and R. Arnone, "Summary diagrams for coupled hydrodynamic-ecosystem model skill assessment," *Journal of Marine Systems*, vol. 76, pp. 64–82, 02 2009.

[42] C. Li and S. Shimamoto, "An open traffic light control model for reducing vehicles' CO2 emissions based on etc vehicles," *IEEE transactions on vehicular technology*, vol. 61, no. 1, pp. 97–110, 2011.

[43] J. Gulliver and D. J. Briggs, "Personal exposure to particulate air pollution in transport microenvironments," *Atmospheric Environment*, vol. 38, no. 1, pp. 1–8, 2004. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1352231003007982

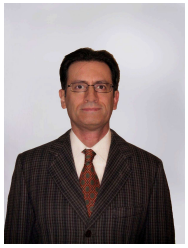
**Bartolomeo Montrucchio** (Member, IEEE) received the M.S. degree in electronic engineering and the Ph.D. degree in computer engineering from the Politecnico di Torino, Turin, Italy, in 1998 and 2002, respectively. He is currently an Associate Professor of computer engineering with the Dipartimento di Automatica e Informatica, Politecnico di Torino. His current research interests include image analysis and synthesis techniques, scientific visualization, sensor networks, RFIDs, and quantum computing.

**Edoardo Giusto** (Student Member IEEE) obtained the B.S. degree in 2015 and M.S. degree in 2017 from Politecnico di Torino. He is currently a Ph.D. Student at the Department of Control and Computer Engineering at Politecnico di Torino. His research interests include Wireless Sensor Networks, Internet of Things, Smart Societies and Quantum Computing.

**Mohammad Ghazi Vakili** received a B.Sc in Telecommunication Engineering and an M.Sc. degree in Mechatronic Engineering from Politecnico di Torino University, Italy, in 2017. He is currently a Ph.D. student at the Department of Control and Computer Engineering, Politecnico di Torino, Italy. His research interests concern Industry 4.0 (future of factories) focus on Automation, Industrial Networks, and Quantum Computing in Industry 4.0.

**Stefano Quer** received an M.S. in Electronic Engineering from Politecnico di Torino and a Ph.D. degree in Computer Engineering, from the Ministry of University and Scientific and Technological Research in Rome. He has been a Visiting Faculty in the Department of Electronic Engineering and Computer Science of the University of California in Berkeley, an intern with the "Advanced Technology Group" at Synopsys Inc., Mountain View, California, and a member of the "Alpha Development Group" at Compaq Computer Corporation, Shrewsbury, Massachusetts. He has been a Compaq Computer Corporation consultant. He is currently a professor with the Department of Control and Computer Engineering at Politecnico di Torino, Torino, Italy. His main research interests include systems and tools for CAD for VLSI, sequential and concurrent algorithms, and optimization techniques able to achieve acceptably solutions with limited resources.

**Renato Ferrero** (M'13, SM'20) received the M.S. degree in Computer Engineering in 2004 and the Ph.D. degree in Computer and System Engineering in 2012, both from Politecnico di Torino, Italy. He is currently an Assistant Professor at the Dipartimento di Automatica e Informatica of Politecnico di Torino. His research interests include ubiquitous computing, wireless sensor networks and RFID systems.



**Claudio Fornaro** received the M.Sc. degree in electronic engineering and the Ph.D. degree in computer engineering from the Politecnico di Torino, Torino, Italy, in 1994 and 1998, respectively. He is currently a Researcher with UniNettuno University, Rome, Italy. He has coauthored several articles in various areas, such as embedded system (for astrophysics instruments), mathematics (special functions), education, computer security, and computer graphics.