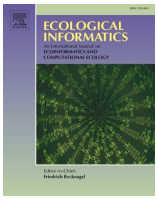
Ecological Informatics 81 (2024) 102568

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf



HypeAIR: A novel framework for real-time low-cost sensor calibration for air quality monitoring in smart cities

Chiara Bachechi, Federica Rollo *, Laura Po

"Enzo Ferrari" Engineering Department, University of Modena and Reggio Emilia, Via P. Vivarelli, 10, 41125 Modena, Italy

ARTICLE INFO

Keywords:
Real-time
Sensor calibration
Air quality monitoring
Smart cities
Air pollution monitoring
Low cost sensors
Time series
Framework
Air quality
LSTM
Random forest

ABSTRACT

While less reliable than authorized air quality stations, low-cost sensors help monitor air quality in areas overlooked by traditional devices. A calibration process in the same environment as the sensor is crucial to enhance their accuracy. Furthermore, low-cost sensors deteriorate over time, necessitating repeated calibration for sustained performance. HypeAIR is a novel open-source framework for the management of sensor calibration in real-time. It incorporates two calibration methodologies: a combination of machine learning models (Voting Regressor and Support Vector Regression) and the Long Short-Term Memory deep learning model. To evaluate the framework, three extensive experiments were conducted over a 2-year period in the city of Modena, Italy, to monitor NO, NO_2, and O_3 gases. Both calibration methodologies outperform the manufacturer calibration and our baseline (i.e., a variation of the Random Forest algorithm) and maintain efficiency over time. The availability of the source code facilitates customization for monitoring additional pollutants, while shared air quality datasets ensure reproducibility.

1. Introduction

As reported by WHO (2021), around the world, only 1 in 10 people breathe healthy air, while an estimated 4.2 million individuals die each year due to exposure to ambient air pollution. Monitoring air quality (AQ) is the first step to raising awareness among citizens and promoting actions from authorities to reduce pollutant emissions. Conventional environmental monitoring, through expensive stations has limited coverage and low spatial resolution, therefore it is insufficient to quantify hyper-local AQ conditions. Advances in low-cost sensors have enabled large-scale AQ monitoring with increased spatiotemporal resolution; however, low-cost air quality (LCAQ) sensors rarely achieve the accuracy required by regulatory standards as they are sensitive to environmental conditions, to the presence of other pollutants, and they suffer degradation of the quality of measurements over time.

A calibration model for air quality is a mathematical or statistical framework designed to adjust and refine the raw data produced by LCAQ sensors. This calibration process involves placing the sensor in close proximity to a reference station or instrument for a co-location period, as highlighted by Maag et al. (2016)). The co-location period needs to be as short as possible to enable quick deployment of sensors in various locations, yet long enough to ensure the reliability of the calibrated data.

Periodic re-calibration is essential, involving placing the sensor near the reference station or instrument for a new co-location period. The additional data collected during these calibration processes are integrated into the training of the calibration model, preserving the quality of measurements. Given that the relation between the raw measurements of an LCAQ sensor and pollutant concentration at the legal station can vary significantly for each sensor, individual sensors require unique calibration models. This complexity underscores the challenging nature of sensor calibration, a critical process that commands considerable attention and effort from the scientific community. When implementing calibration models, several research inquiries arise:

R1 Can a tool be developed that is not merely an ad hoc calibration solution for a specific sensor and pollutant but is adaptable to a variety of sensors?

R2 Can LCAQ sensor measurements achieve the same reliability as legal station measurements?

R3 What is the best solution to ensure that calibration performance is maintained over time in a real environment?

R4 What factors exert the most significant influence on the performance of calibration models?

R5 Is it possible to define essential guidelines to be followed by scientists embarking on the calibration of low-cost sensors for the first

* Corresponding author.
E-mail addresses: chiara.bachechi@unimore.it (C. Bachechi), federica.rollo@unimore.it (F. Rollo), laura.po@unimore.it (L. Po).

https://doi.org/10.1016/j.ecoinf.2024.102568
Received 7 July 2023; Received in revised form 15 February 2024; Accepted 17 March 2024
Available online 27 March 2024

time?

In this paper, we present HypeAIR, a novel framework for LCAQ sensor calibration based on co-location to simplify and automatically manage the calibration process; the framework is open-source, easily adaptable to different sensors and pollutants, and enables real-time hyper-local AQ monitoring. We propose two methodologies for calibration embedded in HypeAIR: one based on a combination of machine learning algorithms that consider the time dependency of observations, and one based on the deep learning algorithm Long Short-Term Memory (LSTM) that was not previously employed for LCAQ calibration. We define an approach to evaluate the performance of the methodologies taking into account the life cycle of sensors and the context (e.g., temperature, humidity) in which the observations are collected. Then, we compare the two methodologies for real-time sensor calibration to discuss their performances in different environmental conditions. Three experiments are conducted, each for a precise purpose:

- **Exp.1**: to measure the robustness of the methodologies even in the presence of few training data,
- **Exp.2**: to investigate the performance of the generated models on a large dataset where observations are collected under heterogeneous weather conditions,
- **Exp.3**: to check the reliability of calibration models several months after the last training period.

We deeply discuss the results of the experiments considering several factors: the influence of seasons, the presence of out-of-range concentration values, the occurrence of big errors (i.e., calibrated data that deviate significantly from the correct value), and the presence of anomalies in the legal station's measurements. The outline of the paper is the following: Section 2 presents the state of the art, then in Section 3 background and motivations are reported. Section 4 describes the HypeAIR framework, the data structure and the calibration algorithms. Section 5 delineates the real case scenario, the sensors' network, and the configuration of the local models; Section 6 outlines the three experiments applied to the real case and reports their results. Section 7 discusses the results of the experiments and the behavior of each sensor, highlighting the main factors that influence the performance of the calibration methodologies. In the end, Section 8 sketches conclusion and future work.

2. Related work

In the last decade, several projects have demonstrated the validity and effectiveness of the widespread use of LCAQ sensors. EuNetAir[1] was a COST Action focused on developing new sensing technologies for affordable AQ control at low cost, and defining innovative approaches to AQ modelling. The UrPolSens project (Boubrima et al. (2017)) implemented a low-cost, energy-efficient AQ monitoring platform employing NO_2 sensors. iSCAPE (Improving the Smart Control of Air Pollution in Europe)[2] was a European research initiative that focused on integrating and advancing AQ control and carbon emissions in cities. It aimed to provide scientific guidance to end-users for deploying LCAQ sensors to monitor AQ and people's exposure with reasonable data quality.

Estimating high-quality pollutant concentrations for LCAQ sensors is becoming an urgent need in smart cities. Given that chemical sensors require calibration algorithms to estimate gas concentrations, various approaches have been proposed and tested on a range of proprietary devices and datasets in the literature. As described by Concas et al. (2021), several issues arise when attempting to calibrate a sensor. Firstly, a model needs to operate effectively in different seasons, geographic locations, and in the presence of a different combination of pollutants. Secondly, calibration models should avoid from making strong assumptions about the distribution of input values since air quality data are rarely independent and identically distributed. Moreover, they generally exhibit significant temporal correlations. Lastly, given the relatively short lifespan of a sensor (typically 2 years at most), the training period cannot be lengthy. Nevertheless, training data should be sufficient to ensure that the model can learn how the cross-sensitivities between different pollutants and the impact of weather conditions affect the concentration of each pollutant.

Numerous studies have investigated and compared calibration algorithms for LCAQ sensors (De Vito et al. (2018); Motlagh et al. (2020); De Vito et al. (2021); Mead et al. (2013)). Vito et al. (2017) conducted a comprehensive review and assessment of five machine learning approaches and their dynamic implementations, revealing that SVR techniques demonstrated the best performance across most scenarios. In a similar vein, Zaytar and Amrani (2020) provided a review of contemporary machine learning algorithms employed for various AQ monitoring tasks, emphasizing the widespread use of ensemble learners (e.g., random forest, boosted trees, extreme gradient boosting) for shallow learning, fully-connected Neural Networks, and Convolutional Neural Networks for deep learning. Zimmerman et al. (2018) proposed training Random Forest models through three to five days of co-location every two months, for a total training period of four weeks, yielding encouraging results.

However, findings reported by Sinha et al. (2019) suggested that tree-based ensemble regressors among machine learning algorithms struggle to predict values outside the training range, particularly for seasonal pollutants. Spinelle et al. (2015) applied several field calibration methods to measure nitrogen dioxide (NO_2) and ozone (O_3) in rural areas, with artificial neural networks proving effective in addressing cross-sensitivity with other pollutants. The calibration of sensors for tropospheric ozone (O_3), prevalent in European summers, poses a challenge due to the short calibration period, as noted by Ferrer-Cid et al. (2019). Furthermore, they compared calibration methods on a large and limited training dataset, highlighting their effects on long-range predictions.

In the context of multivariate time series data obtained from sensors, LSTM has been recently tested for AQ prediction (Fang et al. (2023); Chang et al. (2020); Seng et al. (2021)).

3. Background and motivation

This research was prompted by the TRAFAIR[3] project ("Understanding traffic flow to improve air quality"), a European initiative aimed at monitoring AQ in six EU cities using LCAQ sensors. The project also involved making AQ predictions through simulation models (Po et al. (2019a, 2019b)). A significant challenge for the project's sustainability was maintaining sensor calibration months after the co-location period. The project underscored the strong need for a framework capable of managing different types of LCAQ sensors and automating the creation of calibration models for each sensor and pollutant.

Multiple LCAQ sensors can be integrated in a single low-cost device for monitoring different pollutants. This device is typically housed in a box, with sensors placed inside the box (see Fig. 1.a) or connected externally to the box through specific sockets (see Fig. 1.b). Each sensor, also known as a cell, is dedicated to measuring a specific pollutant with a focus on NO, NO_2, and O_3 in this paper. These sensors detect gases through oxidation-reduction reactions using an electrolyte substance between electrodes. The working electrode, in contact with both the electrolyte and a porous membrane connected to ambient air, measures the electrical current produced by the reaction, converting it into tension. Electrochemical cells provide two raw measures in millivolt (mV): one through the working (we) electrode and the other through the

[1] http://www.eunetair.it/
[2] https://www.iscapeproject.eu/
[3] https://www.trafair.eu/

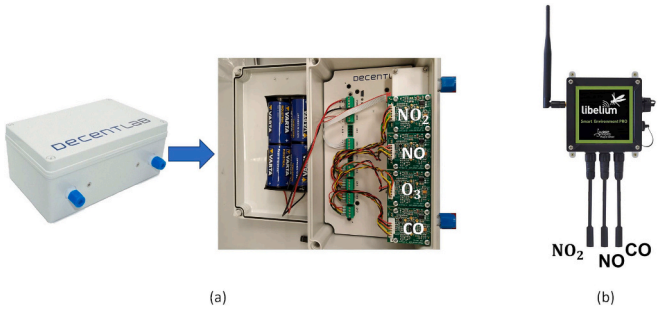


Fig. 1. Two exemplar LCAQ devices: on the left, the sensors are placed inside the device and on the right, they are connected externally.

auxiliary (*aux*) electrode. However, this technology has drawbacks, including potential interference from other gases in the atmosphere and the electrolyte drying out in low humidity and high-temperature conditions, leading to sensor cell breakage, as described by Concas et al. (2021).

Typically, the sensor manufacturer provides a formula along with specific parameters for each cell to estimate pollutant concentration from raw measurements. Alternatively, in some cases, calibrated measurements may be directly provided by the manufacturer. These values are commonly derived using a multivariate regression model, and in certain cases, they can be employed to infer air quality information for a broader area based on a limited number of air quality sensors, Hofman et al. (2022) details this approach. Nevertheless, these concentrations may lack reliability, as they do not stem from a model trained under the same environmental conditions as the sensor's deployment. Demonstrating the significance of developing calibration models that account for relevant environmental factors, Miech et al. (2021) highlighted the potential unreliability of concentrations derived from models not trained in the sensor's specific conditions. We recommend testing the performance of the calibration provided by the manufacturer by comparing concentrations with measurements from legal stations in the field before assuming their reliability. The assessment of the manufacturer calibration in our specific use case is discussed in Section 5.1.

The TRAFAIR project employs a calibration approach that involves a co-location period during which the LCAQ sensor is placed near a legal station. This period gathers concentrations of pollutants from the legal station and raw measurements (in millivolts) from the LCAQ sensor, creating two aligned datasets: the raw dataset, comprising aggregated raw measurements, and the reference dataset, coupled with concentration values from the legal station. These datasets are then used to train a calibration model for each specific pollutant of each sensor. Once generated, this calibration model allows the device to estimate pollutant concentrations in real time from raw measurements anywhere in the city.

To maintain measurement quality, sensors are periodically moved back close to the legal station for a new co-location period to collect updated training data for calibration model updates. The calibration process, guided by project and environmental expert requirements, must adhere to the following criteria:

(a) the co-location period should be as short as possible to reduce the start-up time, enabling quick deployment of sensors across different city points. Simultaneously, it must be long enough to ensure the reliability of the calibrated data generated by the calibration model;
(b) the model's performance should remain roughly unchanged over a long time, minimizing the need for frequent co-location periods;
(c) each sensor must undergo individual calibration, considering that the relation between raw measurements and the pollutant concentrations can vary significantly for each sensor cell.

4. HypeAIR framework

HypeAIR is a suite of calibration methodologies that can be applied to any type of LCAQ sensor, allowing for different approaches: calibration by co-location, calibration in the laboratory, or blind calibration (Maag et al. (2018)).

The main functionalities of HypeAIR are depicted in Fig. 2. After ingesting raw data from the LCAQ sensor and the reference pollutant concentrations, the framework creates a model for the specific device and pollutant (Fig. 2.a). When LCAQ sensors are moved again close to legal stations for a new calibration period, the framework collects other reference data and refines the calibration model. The calibration model is applied to raw data to estimate pollutant concentrations (Fig. 2.b). Finally, the performances of the calibration model can be evaluated comparing them with target concentrations (Fig. 2.c).

Two distinct versions of the open source framework are available: the file-based HypeAIR framework on Code Ocean[4] and the database HypeAIR framework on GitHub.[5] The file-based version accepts "csv" files for raw or reference data as input and produces files as output. This version stands out for its versatility, easily adapting to various sensor types, gases, and calibration procedures (e.g., in a laboratory); however, it does not track the position and status of the sensors.

The database-oriented variant also includes a robust and efficient data management infrastructure designed for handling managing large amounts of sensor data streams. It includes several features: collecting input data directly from the database, saving all the information related to the generated calibration models, and storing the outputs in specific tables. Unlike the file-based version, this variant systematically monitors the position and status of the sensors. Additionally, data from sensors, legal stations, algorithm parameters, and model configurations are stored in the database, as detailed in Section 4.1, enabling straightforward comparison and performance evaluation.

Various machine learning algorithms can be employed to develop an effective calibration model. We conducted tests on different solutions, and the two most successful methodologies have been integrated into the framework. However, the framework offers an interface for defining additional methodologies, ensuring its adaptability to diverse calibration solutions. The first methodology included involves a combination of tree-based machine learning algorithms with a Voting Regressor (VR) associated with Support Vector Regression (SVR) generating a model named VR + SVR (explained in detail in Section 4.3). The second methodology employs the LSTM deep learning algorithm (further information in Section 4.4). To assess the performance of the calibration models, we tested them during a co-location period not included in the dataset used for model training. The metrics employed for evaluation are detailed in Section 4.2.

4.1. Database structure

The structure of the database for storing information related to the air quality sensors and the legal stations is shown in Fig. 3. The data model was implemented following the standard "ISO 19156:2011 Geographic information - Observations and Measurements" (ISO (2011)) and the TAQE model defined for Traffic and Air Quality Applications in Smart Cities by Martínez et al. (2022). The database is a PostgreSQL database with PostGIS extension to handle geospatial data and Timescale extension to make SQL scalable for time-series data (i.e., sensor measurements).

At each moment, every device is described by a status and is located at a specific point in the city. The status changes when a device is moved from one location to another to collect measurements in the new loca-

[4] https://codeocean.com/capsule/0864495/tree
[5] https://github.com/ChiaraBachechi/AQCalibrationFramework

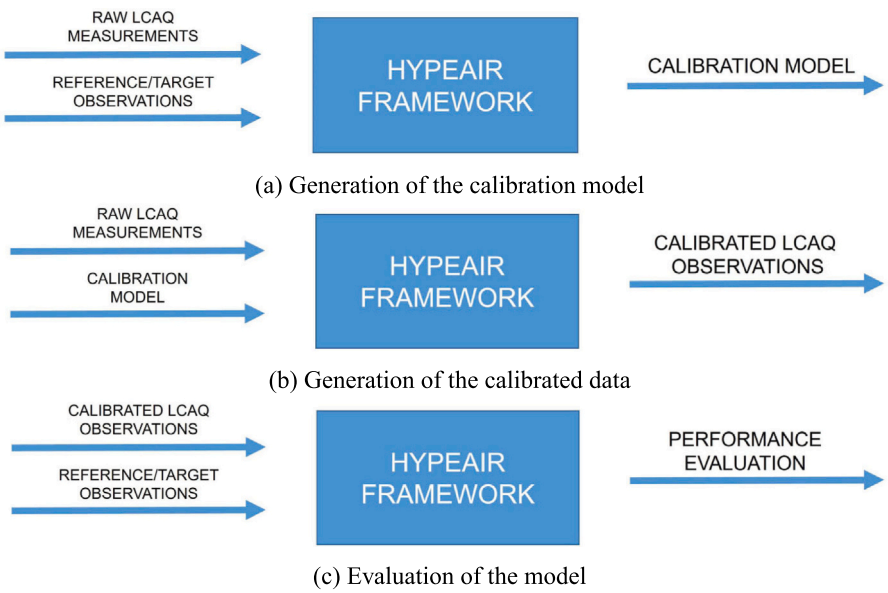


Fig. 2. HypeAIR framework functionalities.

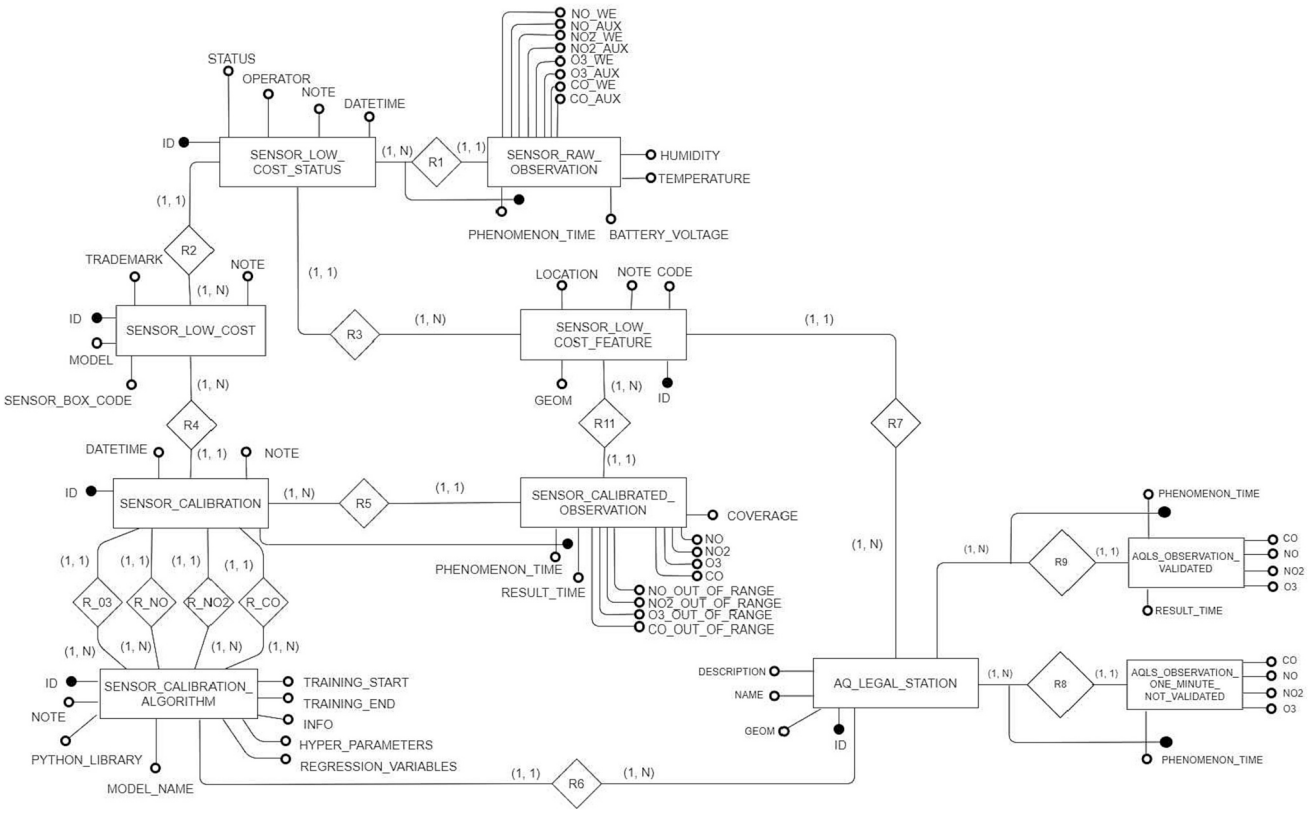


Fig. 3. E/R model of the database to store information related to the air quality sensors, their measurements and calibration and the observations from legal stations.

tion, but also when some problems/failures occur. The device is in "running" status when it collects measurements in a location far from a legal station; when it is moved near a legal station the status changes to "calibration"; "broken" status means that some malfunctions occurred causing unreliable measurements, while "offline" indicates that the device is switched off. Moreover, sensor data, legal station data, calibration algorithm parameters, and model configurations are stored in the database to facilitate comparison and performance evaluation. In table "sensor_low_cost" the identifier of each device is stored along with the model name and trademark; while "sensor_low_cost_feature" collects the name and the GPS coordinates of all possible locations (features) with also the identifier of the legal station if the location is near that legal station. Table "sensor_low_cost_status" gives information about where each device is located at each moment, which is its status, the timestamp of the location change, and the name of the operator who moved the device. The measurements are stored in table "sensor_raw_observation". Each record includes 19 measurements: air temperature, humidity, battery voltage, and 8 raw measurements (2 measurements per 4 gases:

NO, NO_2, CO, and O_x). Each record is associated with a timestamp (phenomenon_time). Each raw measurement can be calibrated by multiple calibration models. Thus, calibrated data are annotated with the date of the measurement, the sensor that has provided it, and the algorithm that was used. This structure allows applying multiple calibration models and comparing their results. The configuration, the hyperparameters, and the training period of each algorithm are stored in table "sensor_calibration_algorithm". The calibration algorithm is associated with a specific gas by the table "sensor_calibration". The name and position of the legal stations are stored in "aq_legal_station"; while "aqls_observation_validated" and "aqls_observation_not_validated" collect the measurements of each gas provided by the legal station. The distinction between not validated measurements and validated measurements is present because the agency that manages the legal stations can decide to share firstly not validated data that then undergo a process of data repairing and validation.

4.2. Assessment metrics

Four metrics are taken into consideration when evaluating the performances of the algorithms: Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Relative Error (MRE), and accuracy.

Given the calibrated value $\widehat{y}_i$ and the corresponding ground truth value y_i, RMSE is calculated by the formula:

$$RMSE = \sqrt{\frac{1}{n}\Sigma_{i=1}^{n}(\widehat{y}_i - y_i)^2}$$

where n is the total number of observations. Since the errors are squared before they are averaged, the RMSE gives a relatively high weight to large errors. On the other hand, MAE measures the average magnitude of the errors weighing all the errors in the same way and without considering the error direction, following the formula:

$$MAE = \frac{1}{n}\Sigma_{i=1}^{n}|\widehat{y}_i - y_i|$$

RMSE and MAE are measured in $\mu g/m^3$ and can range from 0 to ∞. Penalizing large errors has proved to be an effective way to improve the performance of calibration models, as reported by Chai and Draxler (2014). For this reason, RMSE was selected as the main regression metric when evaluating the performances of the methodologies. Moreover, we decided to evaluate MRE which is obtained as the mean of the relative errors:

$$MRE = \frac{1}{n}\Sigma_{i=1}^{n}\frac{|\widehat{y}_i - y_i|}{y_i}$$

MRE can be high when the absolute value of the ground truth is low, penalizing errors in small values. As suggested by Concas et al. (2021), since RMSE, MRE, and MAE are not directly related it can be helpful to consider all of them to measure the performances of the algorithms from different perspectives.

Moreover, it is important to take into consideration that indicative AQ measures are communicated to citizens through color scale maps instead of providing concentrations of pollutants. Several color scales with different thresholds are available, we consider the one provided by the European Environmental Agency (EEA) (Table 1). Real and predicted values are then associated with the corresponding color in the color scale. For each color, a class is created, and we measure the ability of our algorithms to correctly predict the right class/color using accuracy. Accuracy is widely used in classification problems, and it is the ratio between the number of correct predictions and the total number of input samples. For multi-label classification, if $\widehat{y}_i$ is the predicted value of the i^{th} sample and y_i is the corresponding true value, accuracy is defined as:

$$accuracy(y, \widehat{y}) = \frac{1}{n_{samples}}\sum_{i=0}^{n_{samples}-1}1(\widehat{y}_i = y_i)$$

where $1(x)$ is the indicator function. The value of accuracy can give an idea about the significance of the errors in the predicted values.

Table 1
EEA thresholds.

	C1	C2	C3	C4	C5
NO, NO_2 $(\mu g/m^3)$	0–50	50–100	100–200	200–400	> 400
O_3 $(\mu g/m^3)$	0–80	80–100	100–120	120–140	> 140

For LCAQ sensors, the performances are important to establish the application field of the instruments. The two air sensors guidebooks (Williams et al. (2014); Clements et al., 2022), published by the United States Environmental Protection Agency, propose some guidelines to evaluate sensor performances and define their correct use. They also suggest considering precision and bias. Precision measures the agreement among repeated measurements under identical circumstances or substantially similar conditions, and represents the random component of the error. It can be estimated as the ratio between the standard deviation of the raw observations and their average. Precision is calculated for each concentration observed by the reference station (rounding to the nearest unit). Thus, all the concentrations observed by the reference station are rounded to the nearest unit to obtain the expected values $(\widehat{y})$; then, for each concentration $\widehat{y}_c$, the precision is evaluated as follows:

$$Precision_{\widehat{y}_c} = \sigma_{Y_c}/\overline{Y}_c$$

where σ_{Y_c} is the standard deviation of the Y values that the sensor measured for the corresponding concentration $\widehat{y}_c$, and the $\overline{Y}_c$ is their average. The precision of each class is then evaluated, averaging the $Precision_c$ for all the concentrations that belong to the class in the period of observation.

Bias is a systematic error higher or lower than the true value. We evaluate the bias for each concentration $\widehat{y}_c$ rounded to the nearest unit as:

$$Bias_{\widehat{y}_c} = \left(\frac{\overline{Y}_c}{\widehat{y}_c}\right) - 1$$

where $\overline{Y}_c$ is the average of the calibrated values obtained by the low-cost sensors when the legal station measured the reference concentration $\widehat{y}_c$. Then, the total bias is evaluated as the average of the values of $Bias_{\widehat{y}_c}$ for all the concentrations observed by the legal station during the co-location period.

Both precision and bias should be under 0.15 for NO and NO_2 and under 0.7 for O_3 to employ the sensors for regulatory monitoring, based on the requirements expressed by Williams et al. (2014). However, if the error is lower than 0.20 they can be employed as supplemental monitoring together with legal stations.

As recognized by the United States Environmental Protection Agency (Clements et al. (2022)), air sensors serve a crucial role in non-regulatory supplemental and informational monitoring applications, including daily trends, gradient studies, participatory science, education, hotspot detection. These sensors also contribute to long-term changes through epidemiological studies and model verification, emphasizing their diverse applications in understanding and addressing air quality concerns.

4.3. VR + SVR calibration methodology

To define the best-performing algorithms to combine in a Voting Regressor (VR), we tested the performances of some well-known machine learning algorithms: Random Forest (RF) (Breiman (2001)), Extra-Trees (ET) (Geurts et al. (2006)), Gradient Boosting (GB) (Friedman (2002)), Lasso (Tibshirani (1996)), Elastic Net (Zou and Hastie (2005)),

and Logistic regression (Peng et al. (2002)). The VR takes into consideration the three best-performing machine learning models. Each model is trained taking in input the whole observation at the current time interval t, i.e., the values of air temperature and humidity and the raw measurements of the two channels of all the pollutants. In addition, the raw observations in the previous time intervals are added to the input. Indeed, AQ observations exhibit a temporal nature as time series but the above-mentioned machine learning algorithms do not take into account the temporal evolution of the data. Their predictions can be improved by considering information regarding previous time intervals as supplementary features of the algorithm. For instance, to calculate the concentrations of NO at time t, the algorithms can take as input all the features in the current time interval t (i.e., the values of air temperature and humidity, the raw measurements NO_WE_t and NO_AUX_t, $NO_2\_WE_t$ and $NO_2\_AUX_t$, $O_3\_WE_t$ and $O_3\_AUX_t$) plus the raw measurements of the two electrodes of NO in the previous N time intervals (denoted as NO_WE_{t-1}, NO_AUX_{t-1}, NO_WE_{t-2}, NO_AUX_{t-2}, etc.). Features importance (FI) is a built-in property in each employed model and can help to define the number of previous measurements to consider and which features to take into account (a discussion about the feature importance on our specific dataset is reported in Section 5.1).

The VR is trained through a two steps process. Firstly, the 10-fold cross-validation is applied to each regressor to estimate its RMSE. The weight w_i associated with each regressor depends on its RMSE ($RMSE_i$) and is calculated following the formula:

$$w_i = \frac{RMSE_i}{RMSE_{tot}} \quad i = 0, 1, 2$$

where $RMSE_{tot}$ is the sum of the RMSE values of the three models. Then, the entire dataset is used to train the three models separately. Finally, the VR model is generated considering the values predicted by each model separately and giving them the weight associated with the model itself in the first step to evaluate the final result, calculated by the formula:

$$y = \frac{1}{3}\sum_{i=0}^{2} w_i^* y_i$$

where y_i is the value predicted by each model.

Ensemble regressors are not able to predict values where extrapolation is needed, as explained by Hengl et al. (2018); Sinha et al. (2019); Meyer and Pebesma (2021). At best, they can predict an average of training values seen before, because the regressor assumes that the prediction will fall close to the maximum or minimum values in the training set. For this reason, we decided to test the performances of support vector machines. They construct a hyperplane in a high-dimensional feature space and can be used for classification and regression problems. In the latter case, they take the name of Support Vector Regression (SVR), as described by Basak et al. (2007). SVR attempts to minimize the generalized error, rather than minimizing the observed training error. This error bound combines the training error and a regularization term, that controls the complexity of the space. The model only depends on a subset of the training data, since the cost function ignores any data that are close (within a threshold ε) to the model prediction. To overcome this limitation, we exploited the Radial Basis Function kernel that deals with extra range observations. For parameters tuning of SVR, we applied the heuristic presented by Qiujun Huang et al. (2012). Input data were pre-processed using a Standard Scaler that transforms features by removing the mean and scaling to unit variance.

In our final model, we combine the VR with the SVR algorithm to manage out-of-range observations. The SVR algorithm is employed when there is at least one feature out of the range of the training set, while the VR is used in the other cases. An exemplar demonstration of the advantage of using the combination of VR and SVR is reported in Fig. 4, where six out-of-range predictions performed by the VR and the SVR model are compared. As can be observed, SVR generally predicts a value more similar to the actual concentration. However, we conducted several tests showing that even if SVR shows better performances for out-of-range observations, VR has generally a lower RMSE for all the other observations.

To manage missing values in time series, for each sensor two versions of the VR + SVR model are generated: one considering previous observations and another that does not include previous observations as additional features, that is used when previous observations are not available.

4.4. LSTM calibration algorithm

The Long Short-Term Memory (LSTM) architecture, introduced by Hochreiter and Schmidhuber (1997), represents a variant of Recurrent Neural Networks (RNNs) particularly well-suited for handling temporal data sequences. Its distinguishing capability lies in its ability to assimilate extensive historical sequences to forecast future values. These characteristics are achieved through the introduction of a memory cell and gating mechanisms into the RNN architecture. This helps to solve the vanishing gradient problem commonly encountered during the back-propagation process used to update the weights of all the neurons of the network. The mathematical mechanism used for this update is multiplicative, leading to the multiplication of gradients computed in deeper layers through earlier network weights. If the multiplication factor is small, it results in a vanishing gradient, causing the gradient term to approach zero rapidly. This occurrence impedes the network's ability to learn long-term dependencies effectively.

The input data of the LSTM algorithm include a temporal sequence of observations, i.e., the raw measurements of all the pollutants, air temperature and humidity of consecutive time intervals. The length of the sequence, i.e., the number of consecutive observations, can be fixed based on the variability of the data in the use case.

As shown in Fig. 5, the output of an LSTM cell is based on the cell status that keeps a summary of what happened in the past (*Previous Cell State*), the output of the previous element of the sequence (*Previous Hidden State*), and the features of the current observation (*Input Data*). LSTM uses gates (f, i, o) to regulate the amount of past information to keep or discard. The *forget gate layer* is a sigmoid layer that selects the information of the current status that needs to be discarded considering the hidden state and the current observation. The next step decides which new information needs to be stored in the cell state. A tanh *layer* generates a vector that decides how much the cell status components will be updated based on the hidden status and the current observation. If necessary, the tanh values (that range between −1 and 1) allow for reducing the impact of a component in the cell status. Then, this vector is multiplied by the new vector obtained using the *input gate layer*, a sigmoid layer. This gate identifies which components of the input (the current observation) are significant for the cell status. The obtained vector is then added to the cell status updating the long-term memory of the network (*New Cell State*). At this point, the updated status is available. To generate the output we need to apply a filter: the *output gate layer*. This gate is a sigmoid layer that generates a filter vector based on the hidden state and the current observation. The filter is used to select from the updated cell status only the necessary information to generate a prevision. The updated status is forced to assume values between −1 and 1 through a tanh layer and finally multiplied by the filter vector to generate the final output of the model, i.e., the prevision that will be the next hidden state (*New Hidden Layer*).

When a big amount of data have been collected in co-location with the legal station, the LSTM model can be a good method to perform calibration. The generated model has the ability to forget a part of the past, adapting to changes and drifts in the electrochemical cell. Moreover, the capacity to keep a long memory enables to preserve the patterns still present in the time series of measurements (e.g., season and

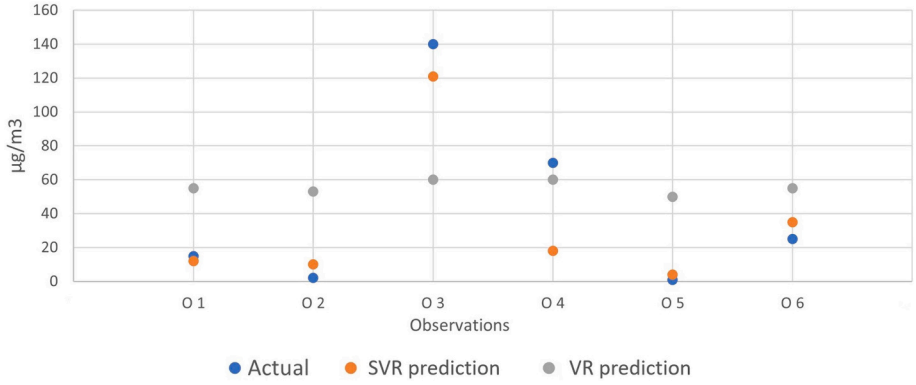


Fig. 4. Comparison of extra range predictions of NO obtained by VR and SVR.

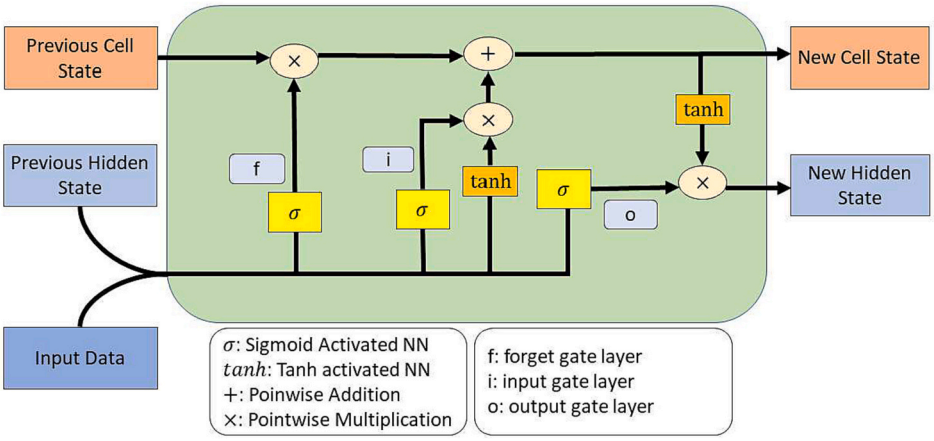


Fig. 5. Structure of the LSTM cell.

weather influence).

Several configurations of the LSTM architecture were tested and compared in Casarotti (2021) showing that a three hidden layer configuration of LSTM does not radically improve the performances, and with a small training dataset might be affected by overfitting.

Since the co-location period needs to be as short as possible and deep learning models usually need many input data, the implemented solution of LSTM is a trade-off between reaching good performances and minimizing the dimension of the training data.

For this reason, the LSTM model implemented has only one hidden layer of y neurons, where y is determined based on the number of training observations:

$$y = \frac{N_s}{\alpha^*(N_i + N_o)}$$

where N_s is the number of samples in the training set, N_i is the number of input neurons, N_o is the number of output neurons, and α is usually a value between 5 and 10.

The LSTM model is trained using the mean squared error (MSE) as loss function and the ADAM optimization (Kingma and Ba (2015)). To avoid overfitting, a dropout layer is added. The dropout layer, described by Özgür and Nar (2020), is a regularization method that randomly excludes some inputs from activation and weight updates while training a network. Inputs not set to zero are scaled up by $1/(1 - rate)$ such that the sum over all inputs remains the same. The training set is pre-processed using a MinMaxScaler that transforms each feature by scaling it to a [0–1] range. The dimension of the temporal window must be selected dynamically considering the performance obtained with different values on the available dataset.

5. Real case scenario

In the specific case of Modena, an Italian city spanning 183 km^2, there are two legal AQ stations denoted by red dots in Fig. 6.

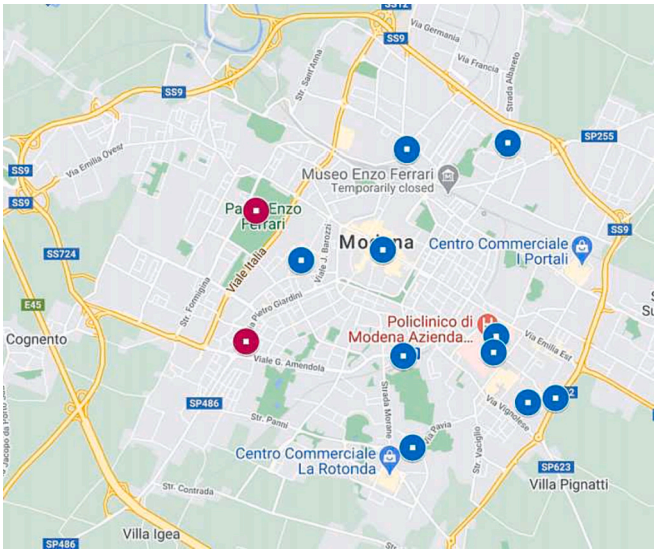


Fig. 6. Points of interest for AQ monitoring (blue dots) and positions of the AQM stations (red dots). Map data: Google, 2020. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

One station, situated within a park, serves as the background station, measuring ambient pollution levels. The other station, located in close proximity to a busy road, functions as the traffic station, specifically monitoring the impact of vehicular traffic on pollution. Both stations measure NO and NO_2 levels every minute, while O_3 is exclusively observed by the background station. Within the urban area of Modena, 12 LCAQ sensors are actively monitoring 10 specific locations, as depicted by the blue dots in Fig. 6. The location of these sensors has been defined by environmental experts. The process of selecting these locations entails a strategic choice of diverse and sparsely distributed points, aiming to comprehensively cover the urban area. This placement is designed to capture a wide range of pollutant concentrations, contributing to a more thorough and nuanced understanding of air quality within the monitored areas.

From August 2019 to April 2021, we collected 3.3 million records of measurements (1.8 GB). For about 20 months, the 12 devices were relocated 108 times and 250 status changes (see Section 4.1) were recorded.

The sensors were positioned near legal stations to collect data employed in the calibration process (calibration status) and then deployed to different points in the city to provide hyper-local AQ monitoring (running status). This process of relocating the sensors near legal stations is iteratively carried out to collect additional data and enhance the calibration. As displayed in Fig. 7, the life cycle is different for each sensor. For example, sensor 4006 was located close to the background station from August to September 2019. Subsequently, it is employed to measure air quality at one designated locations in the city in October 2019. Following this, it undergoes another co-location period, till the beginning of 2020, but this time in close proximity to the traffic station.

The sensor cells, employed in our use case, are produced by Alphasense[6] and assembled in a Decentlab Aircube. Fig. 1.a shows the exterior (on the left) and interior view of the sensor (on the right). Each device provides two raw measures in millivolt (*mV*) for each pollutant through the working (*we*) electrode and the auxiliary (*aux*) electrode in addition to the air temperature (°C), the humidity (%) and the battery voltage (Rollo and Po (2021)). Inside each AQ box, there are sensors for air temperature and humidity monitoring, and 4 AQ electrochemical sensors for NO, NO_2, CO, and O_3. The datasheet of the Decentlab sensors (Alphasense (2015)) is tested on a humidity range between 15% and 85%; thus, observations collected in weather conditions outside that range are less reliable. The Modena climate, with the classic continental characteristics of the Central Eastern Po Valley, sees particularly hot and often muggy summers, and humid climate that often leads to the generation of fog, particularly persistent during the anticyclonic periods, in winters. Therefore, between November and March, the humidity is often higher than 85%. The percentage of these observations vary according to the test season of the experiment: it is less than the 15% of the total observations in **Exp.2** conducted in summer, and, instead, between 29% and 48% for **Exp.1**, and between 24% and 57% for **Exp.3**, that are both conducted in winter. Indeed, the sensor cells have a noise around 18.75 $\mu g/m^3$ for NO and 28.65 $\mu g/m^3$ for NO_2 (Alphasense (2015)); although, the mean concentrations observed by the background station are generally lower during summer. Since the percentage of observations collected in unreliable conditions or underneath the cell noise limit is so high, we cannot remove all of them from the input. For this reason, we decided to test our models in these critical conditions to verify if our calibration methodology can anyway help to provide reliable measurements.

[6] https://www.alphasense.com/

5.1. Calibration algorithms

The pollutant concentrations provided by the manufacturing company of the low-cost sensors are obtained by the formula:

$$Gas\_concentration = \frac{vGas - vAux - \alpha + \beta}{sensitivity}$$

where α, β, and *sensitivity* are parameters specific for each cell, provided by the manufacturer, and *vGas* and *vAux* are the voltages of the working and the auxiliary electrodes, respectively. Considering all the co-location periods of each low-cost device in the whole TRAFAIR dataset the manufacturer calibration presents an average RMSE of 37 for NO, 53 for NO_2, and 109 for O_3, while the accuracy is 0.23 for NO, 0.36 for NO_2, and 0.44 for O_3. The bad performance of the provided calibration demonstrates the need for new calibration models trained in the same environmental conditions in which the sensors will be used.

For the configuration of VR + SVR model, we conducted several experiments using different regression models. Among them, RF, GB, and ET showed the best performances in terms of RMSE and accuracy. Since good results were achieved using standard parameters, no tuning was required. Therefore, we combine these three models in the VR. Since the influence of the previous observations on the actual values can depend on the environment and the atmosphere composition, we defined the number of previous features to take into account through the calculation of the feature importance. Table 2 displays the FI assigned by RF to each feature for the prediction of the pollutant indicated in each column. It can be observed that, for NO, the most relevant channel is NO_AUX, while its past values have decreasing importance: $FI(no\_aux) > FI(no\_aux_{t-1}) > FI(no\_aux_{t-2})$. The values of FI highlight how NO and NO_2 can benefit from a prediction that takes into account previous values, proving the validity of our decision to adapt machine learning algorithms to handle time series. Similar considerations can be extended to ET and GB regressors. Since the importance assigned to past features decreases quickly, only the previous two observations were taken into account. The feature with the highest value of importance for O_3 is the temperature. The values of O_3 in the previous two observations seem to be not significantly correlated to the current value of the pollutant.

For the configuration of LSTM, after evaluating 6, 12, 24, and 48 previous observations in Casarotti (2021), we decided to fix the temporal window to 12 previous observations (i.e., 2 h of observations) for our use case. Given the selected dimension of the temporal window (x) and the number of features (y), the dataset is reshaped such that each observation consists of x rows and y columns. LSTM was also compared with different RNN architectures such as Multilayer Perceptron (Marius et al. (2009)) and Gate Recurrent Unit (Cho et al. (2014)). LSTM proved to have better results than Multilayer Perceptron, reducing RMSE by 5.70% for NO and by 15.23% for NO_2, as shown in Casarotti (2021).

Before applying our calibration methodologies, both legal station data and low-cost sensor data have been aggregated to obtain one value every 10 min. Therefore, the calibration process provides one concentration value for each device and pollutant every 10 min.

5.2. Interpolation maps and data visualization

With the HypeAIR framework, we obtain real-time pollutant concentrations in each location where sensors are placed. However, to perform hyper-local monitoring, we need to generate real-time maps showing the air quality in the whole urban area. Interpolation methodologies are, therefore, applied for estimating the spatial distribution of pollutants to convert spatial discrete information into continuous data. We have tried different methodologies (e.g., Nearest Neighborhood, Ordinary Kriging, and thin plate spline), and in the end, Inverse Distance Weighted shows the best results in predicting the correct values in not observed positions (Li and Heap (2014)). The interpolation maps,

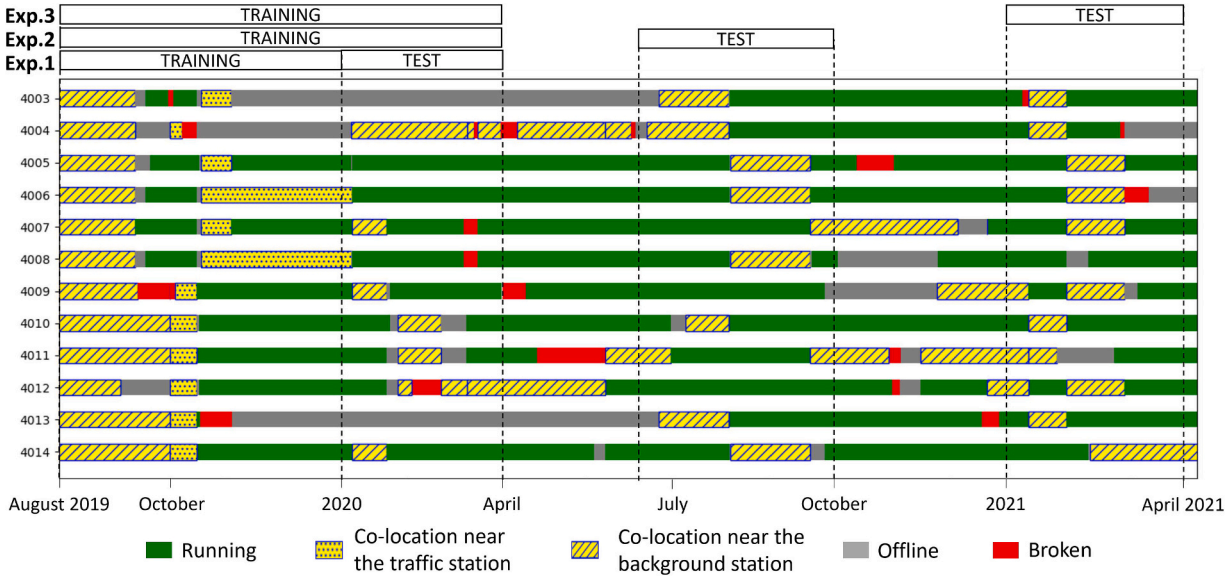


Fig. 7. Experimental periods and sensor status timeline.

Table 2
Example of feature importance in Random Forest calibration for each pollutant.

Feature	NO	NO_2	O_3
NO_WE	**0.0170**	0.0113	0.0100
NO_AUX	**0.6371**	0.0100	0.0138
NO_2_WE	0.0240	**0.4838**	0.0440
NO_2_AUX	0.0031	0.0748	0.0125
O_3_WE	0.0049	0.0221	0.0130
O_3_AUX	0.0025	0.0096	0.0333
humidity	0.0187	0.0191	0.0563
temperature	0.0444	0.0501	**0.7460**
pollutant_WE$_{t-1}$	0.0037	**0.1414**	0.0073
pollutant_AUX$_{t-1}$	**0.1507**	0.0601	0.0058
pollutant_WE$_{t-2}$	0.0029	0.009	0.010
pollutant_AUX$_{t-2}$	0.0110	0.0135	0.0072

the time series of calibrated observations, and statistics about historical values can be investigated using the Trafair Air Quality Dashboard, described in Bachechi et al. (2020, 2022).

6. Experiments and results

We conducted three main groups of experiments on the real case scenario described in Section 5. The periods during which training and testing are performed are co-location periods and they are reported at the top of Fig. 7. Due to the very different life cycles, the training and test periods of the individual sensors vary a lot.

The first experiment (**Exp.1**) evaluates the performances of the calibration methodologies in the period that immediately follows the training period, therefore all the considered training periods are from August to December 2019 and the test periods are from January to March 2020.

In the second experiment (**Exp.2**), we focus on the impact of the duration of the training period on the performance of the two methodologies to answer the **R2** research question; therefore, we considered a longer training period w.r.t. **Exp.1** (from August 2019 till March 2020) while the test period goes from the 15th of June 2020 till the 30th of September 2020.

The purpose of the third and last experiment (**Exp.3**) was to evaluate the performance degradation of calibration methodologies over time, answering the **R3** research question. In **Exp.3**, the same calibration models of **Exp.2** are tested 9 months later, i.e., from the 1st of January 2021 to the 31st of March 2021.

All raw and calibrated data, generated in the experiments, are published as open data and therefore comparable with other calibration methodologies: the raw data are available on the regional,[7] national, and European open data portals; the hourly data of the legal stations are available on the Regional Environment Agency (ARPAE) data portal,[8] while, the calibrated observations are published on the project website.[9]

Table 3 shows the values of RMSE, MAE, MRE and accuracy derived from the averaged performances of sensors within each experiment. Each column reports the performances of the two HypeAIR calibration methodologies and our baseline, i.e., the methodology employed in the TRAFAIR project before the implementation of the HypeAIR framework and described by Baruah et al. (2023). The baseline is the RF algorithm, a standard methodology presented in the literature, where a linear extrapolation function was added to manage out-of-range values. The baseline has a training period similar to HypeAIR training period in **Exp.1** and **Exp.2** and longer in **Exp.3** (until October 2020). As can be seen, in most cases both VR + SVR and LSTM outperform the baseline by reducing RMSE. In **Exp.3**, despite the VR + SVR and LSTM models being trained 9 months before the test period, when compared to RF trained up to 4 months before, the results indicate that in most cases VR + SVR and LSTM achieved superior performance. Moreover, MAE and MRE are reduced by at least one of the two HypeAIR methodologies in most cases. The accuracy is always improved.

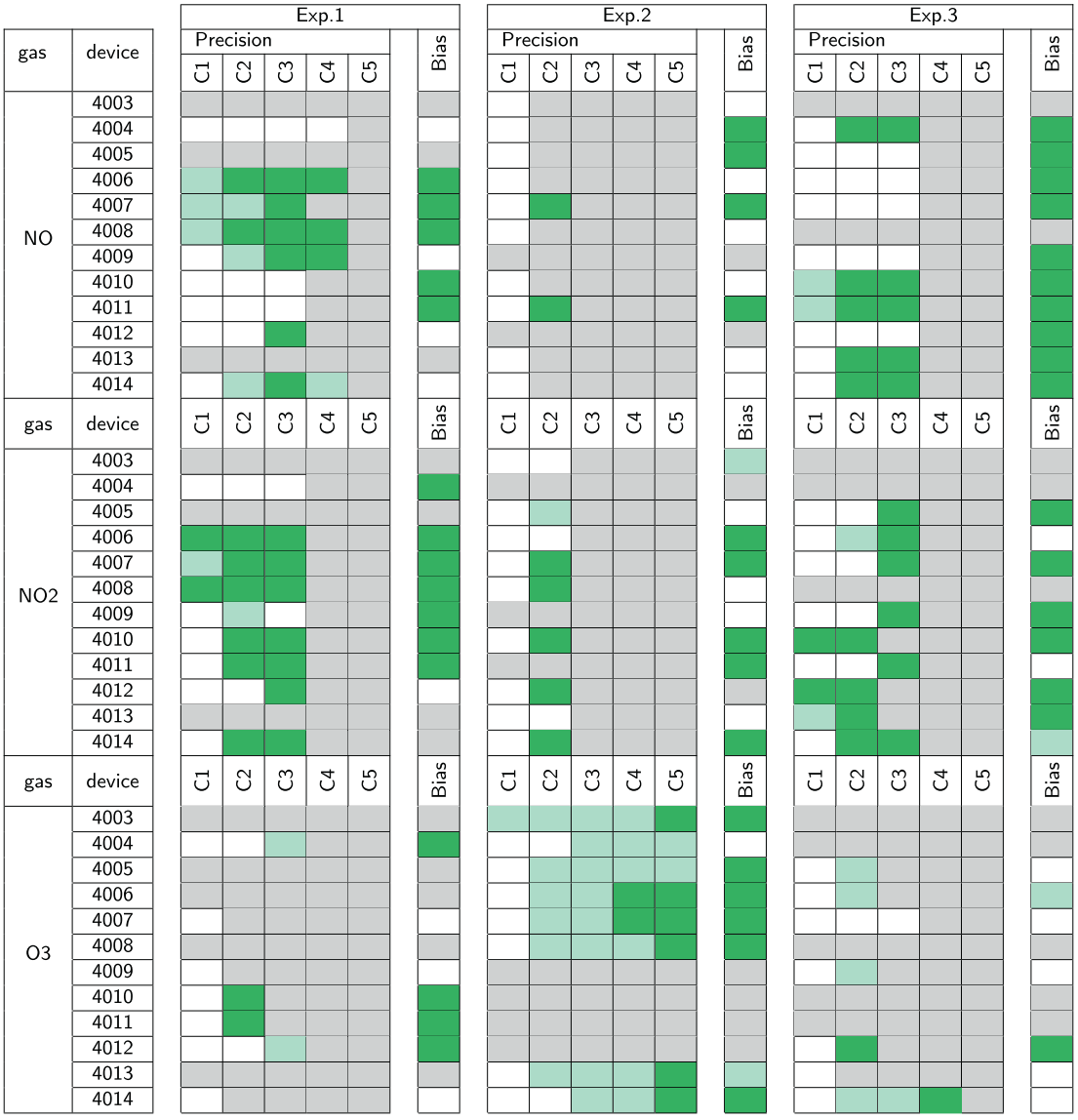
Since LCAQ sensors are primarily used to gather more information about air quality, it is crucial to understand the reliability of their observations, especially as pollutant concentrations increase. To this end, we have prepared Table 4, which reports, for the three experiments, whether the bias and precision values fall within acceptable limits for regulatory monitoring or for supplemental monitoring (as indicated in Section 4.2). The bias and precision levels of sensors across the five EEA classes are recorded. Dark green indicates high reliability, suitable for regulatory monitoring. Light green signifies good reliability for supplemental monitoring, while white indicates insufficient reliability for supplemental monitoring (typically adequate for informative or

[7] https://dati.emilia-romagna.it/
[8] https://dati.arpae.it/dataset/qualita-dell-aria-rete-di-monitoraggio
[9] https://trafair.eu/datasets/modena_airquality_calibration

Table 3
Evaluation of experiments comparing the 10-min predicted concentrations with the non-validated data of legal stations.

		RMSE			MAE			MRE			ACCURACY		
	Gas	VR + SVR	LSTM	baseline	VR + SVR	LSTM	baseline	VR + SVR	LSTM	baseline	VR + SVR	LSTM	baseline
Exp.1	NO	19.07	15.30	36.64	12.04	9.22	22.08	1.16	1.02	0.90	0.87	0.90	0.78
	NO_2	13.52	12.70	19.38	10.46	10.04	12.61	0.45	0.47	0.55	0.76	0.81	0.72
	O_3	22.91	21.19	19.66	17.98	14.84	10.83	3.09	2.25	1.25	0.96	0.95	0.89
Exp.2	NO	3.76	3.09	7.12	2.15	1.62	4.19	1.16	0.68	0.50	1.00	1.00	0.98
	NO_2	9.51	10.25	9.57	6.85	7.55	6.73	0.53	0.54	0.54	0.97	0.98	0.96
	O_3	20.66	22.19	26.86	16.08	17.75	22.49	3.05	1.81	2.38	0.91	0.89	0.56
Exp.3	NO	13.93	13.39	10.67	7.35	7.45	7.14	1.22	0.91	2.07	0.93	0.93	0.93
	NO_2	14.64	13.45	16.35	11.49	10.55	12.63	0.53	0.45	0.50	0.81	0.82	0.74
	O_3	19.99	20.37	21.31	15.62	12.27	17.64	1.54	0.71	2.08	0.91	0.91	0.91

Table 4
Sensor precision with respect to each EEA class and bias: light green background means the value meets the requirements for supplemental monitoring, dark green background for regulatory monitoring.



educational purposes). Gray cells denote that no value is available in the experiment.

The following subsections deepen and detail the results of each experiment for individual sensors in order to compare sensor behaviors and understand in which context each methodology provides the best results. Tables 5, 6 and 7 report the performances obtained in the three experiments by each device with superior performances highlighted in green and inferior ones in red. In particular, we colored in red the values

Table 5
Exp.1 results.

				RMSE		MAE		MRE		ACCURACY	
gas	device	train size	% out of range	VR+ SVR	LSTM	VR+ SVR	LSTM	VR+ SVR	LSTM	VR+ SVR	LSTM
NO	4004	3836	4	26.15	17.41	10.1	10.28	1.08	2.18	0.9	0.92
	4006	12705	0	7.78	7.85	5.21	5.27	0.15	0.12	0.94	0.94
	4007	5689	9	18.01	17.04	12.61	10.76	0.58	0.27	0.86	0.85
	4008	12707	0	7.68	7.28	5.2	4.99	0.15	0.12	0.93	0.93
	4009	5184	26	28.84	18.47	19.04	12.78	0.58	0.41	0.77	0.84
	4010	7667	8	18.14	16.67	10.55	8.23	2.01	1.83	0.9	0.93
	4011	7666	28	17.62	16.1	9.82	8.6	1.59	1.76	0.91	0.92
	4012	3840	46	10.3	8.28	7.84	5.33	3.33	2.03	0.97	0.98
	4014	7668	3	37.09	28.57	27.98	16.7	0.98	0.47	0.63	0.8
	Mean			**19.07**	**15.30**	**12.04**	**9.22**	**1.16**	**1.02**	**0.87**	**0.90**
NO_2	4004	3836	24	20.24	15.88	15.39	12.28	1.05	0.79	0.77	0.84
	4006	12705	0	5.54	5.52	4.15	4.09	0.1	0.09	0.85	0.86
	4007	5689	3	13.49	12.98	11	10.94	0.38	0.32	0.79	0.73
	4008	12707	0	4.3	4.43	3.1	3.37	0.07	0.07	0.89	0.9
	4009	5184	2	18.02	14.8	13.54	12.27	0.36	0.38	0.8	0.71
	4010	7667	3	12.22	13.02	9.18	9.59	0.27	0.3	0.83	0.83
	4011	7666	0	14.82	12.58	11.41	9.5	0.3	0.36	0.79	0.87
	4012	3840	3	16.38	21.47	13.19	16.87	1.14	1.51	0.89	0.82
	4014	7668	1	16.68	13.65	13.16	11.44	0.37	0.38	0.76	0.74
	Mean			**13.52**	**12.70**	**10.46**	**10.04**	**0.45**	**0.47**	**0.76**	**0.81**
O_3	4004	3374	1	31.93	34.09	26.26	26.26	3.3	2.67	0.95	0.94
	4007	3376	0	32.88	32.15	24.83	23.2	6.45	5.59	0.99	0.96
	4009	3590	0	18.82	17.51	14.86	10.4	3.8	2.65	1.00	0.99
	4010	6067	0	10.52	7.24	7.32	5.32	0.87	0.44	0.96	0.98
	4011	6066	0	13.49	10.09	9.82	7.2	1.49	0.58	0.94	0.94
	4012	2240	2	38.16	36.54	30.96	24.72	2.87	2.71	0.88	0.87
	4014	6068	1	14.54	10.68	11.81	6.78	2.87	1.10	1.00	1.00
	Mean			**22.91**	**21.19**	**17.98**	**14.84**	**3.09**	**2.25**	**0.96**	**0.95**

of RMSE and MAE higher than 20 and in green the values lower than 8; while the values of MRE are in green if lower than 0.4 and in red if higher than 1. In the end, we used green for accuracy higher than 0.9 and red for values lower than 0.75. In addition, the number of observations in the training set is reported in the column "train size", while the column "% out of range" shows the number of observations in the test set with the value of at least one of the two channels out of the range between the minimum and the maximum values of the training set.

6.1. *Exp.1*

For this experiment, the training period comprises summer days near the background station and autumn days near the traffic station (which lacks a reference value for O_3) for all sensors. The test periods primarily align with the background station during winter, except for sensors 4006 and 4008, which are in co-location with the traffic station (for these sensors, O_3 evaluation is not possible as the reference station does not measure this gas). Testing the algorithm in winter when the training mainly occurs in summer poses a particular challenge for O_3 calibration. This is due to the gas being highly influenced by radiation, resulting in high concentrations during summer and low concentrations during winter. In contrast, NO and NO_2 concentrations are higher in winter.

Focusing on NO and NO_2, sensors 4006 and 4008 exhibit the most promising performances. Both sensors undergo an extensive training period near both the background station and the traffic station, followed by an immediate test period. The results indicate a very low MRE and over 80% of observations with a relative error below 0.20. For these sensors, LSTM and VR + SVR are both performing well, the difference is negligible.

Considering O_3, the sensors are trained during summer. Consequently, during the test period, the predicted values tend to overestimate O_3 concentrations, leading to a decline in performance compared to NO and NO_2.The top-performing sensors, 4010 and 4011, share the same calibration and testing periods, along with a larger training dataset. Furthermore, both sensors exhibit very similar distributions in O_3 concentrations observed during both the training and test periods. Conversely, sensors 4004, 4007, and 4012 exhibit the poorest performances. Notably, these sensors have smaller training datasets. Additionally, sensors 4004 and 4012 were flagged by environmental experts during the test period for cell malfunctions.

The bias values for **Exp.1**, as presented in Table 4, adhere to regulatory monitoring standards for the majority of sensors. Additionally, the precision for the highest EEA classes of NO and NO_2 (indicative of high concentrations) is notably good and aligns with regulatory monitoring criteria for several sensors. This holds true not only for sensors like 4006 and 4008, which exhibit good performances in terms of RAE and MAE, but also for several others.

6.2. *Exp.2*

In **Exp.2** the training period is longer than in **Exp.1** and covers summer, autumn and winter, while the test period is mainly in summer.

The performance of NO and NO_2 predictions is significantly enhanced for all sensors w.r.t. **Exp.1**. It can be inferred that this is the result of conducting tests during the same season (i.e., the same environmental conditions) of a portion of the training data. Also, for this

Table 6
Results of Exp.2.

				RMSE		MAE		MRE		ACCURACY	
gas	device	train size	% out of range	VR+ SVR	LSTM	VR+ SVR	LSTM	VR+ SVR	LSTM	VR+ SVR	LSTM
NO	4003	5690	0	5.56	3.81	3.07	1.77	1.78	0.88	0.99	0.99
	4004	14987	0	4.1	4.39	2.41	2.05	1.57	0.99	0.99	0.99
	4005	5689	0	2.72	2.1	1.81	1.02	1.01	0.38	1.00	1.00
	4006	13887	0	3.6	3.17	2.13	1.27	1.02	0.39	1.00	1.00
	4007	8368	0	3.15	4.08	1.93	2.3	0.67	0.48	1.00	1.00
	4008	13889	0	2.78	2.19	1.68	1.24	0.83	0.48	1.00	1.00
	4010	11013	0	4.7	3.17	2.53	1.99	1.55	1.09	0.99	0.99
	4011	11010	0	3.78	3.15	2.18	1.48	1.01	0.44	1.00	0.99
	4013	7666	0	4.61	2.99	1.91	1.66	1.14	0.93	0.99	0.99
	4014	10342	0	2.55	1.82	1.88	1.41	1.01	0.74	1.00	1.00
	Mean			**3.76**	**3.09**	**2.15**	**1.62**	**1.16**	**0.68**	**1.00**	**1.00**
NO_2	4003	5690	0	11.79	12.78	7.81	8.28	0.58	0.67	0.98	0.98
	4005	5689	7	12.85	13	10.73	10.6	0.89	0.67	0.96	1.00
	4006	13887	0	11.15	11.33	7.1	7.25	0.48	0.51	0.96	1.00
	4007	8368	0	8.36	9.62	6.08	8.02	0.45	0.52	0.96	0.96
	4008	13889	0	7.91	9.64	5.97	7.54	0.43	0.48	0.97	0.97
	4010	11013	0	7.53	5.99	5.8	4.36	0.52	0.36	0.98	0.98
	4011	11010	0	7.3	9.07	5.43	6.99	0.45	0.58	0.99	0.99
	4013	7666	0	11.73	11.58	7.64	7.36	0.51	0.51	0.98	0.98
	4014	10342	0	6.99	9.23	5.13	7.51	0.46	0.53	0.97	0.97
	Mean			**9.51**	**10.25**	**6.85**	**7.55**	**0.53**	**0.54**	**0.97**	**0.98**
O_3	4003	3377	5	20.39	20.27	16.05	15.5	0.24	0.26	0.58	0.58
	4004	14525	2	29.12	35.91	24.19	32.38	0.29	0.39	0.46	0.46
	4005	3376	0	20.49	30.22	14.86	21.43	7.6	6.91	0.69	0.69
	4006	3376	0	16.85	15.89	13.44	13.19	2.91	0.18	0.66	0.66
	4007	6055	0	14.94	13.45	10.72	10.26	0.27	0.25	0.81	0.81
	4008	3377	0	21.46	19.54	16.04	15.58	6.61	0.48	0.61	0.61
	4013	6066	0	22.91	22.35	18.13	18.33	0.23	0.23	0.5	0.49
	4014	8742	0	21.34	20.05	15.21	15.3	6.27	5.76	0.69	0.69
	Mean			**20.94**	**22.21**	**16.08**	**17.75**	**3.05**	**1.81**	**0.63**	**0.62**

reason, there are just 14 "out-of-range" values in this experiment. For NO, the high values of MRE (higher than 1) depend on the very low concentration of these pollutants during summer. Indeed, the mean concentration observed by the legal station during summer is 3.95 $\mu g/m^3$ for NO and lower than 25 $\mu g/m^3$ for NO_2. The LCAQ sensors have noise around 18.75 $\mu g/m^3$ for NO and 28.65 $\mu g/m^3$ for NO_2, as reported in Alphasense (2015). This explains the high precision errors of NO and NO_2, which does not allow reaching the requirements for supplemental monitoring in most cases, as reported in Table 4. In most cases, VR + SVR and LSTM have comparable performances. If we consider NO, we can notice that VR + SVR always obtained higher values of MRE compared to LSTM.

The prediction of O_3 has worse performance w.r.t. NO and NO_2. However, also for O_3 the performances in terms of MRE are better in this experiment than **Exp.1**. On the other hand, the values of accuracy are very low. The concentration values of O_3 are higher (80.45 $\mu g/m^3$ on average) in the test periods of this experiment compared with the ones of **Exp.1** (25.28 $\mu g/m^3$ on average). Fig. 8 shows the distribution of O_3 concentrations of the exemplar sensor 4007 during the training and test period in the two experiments. In **Exp.2** the distribution of the values observed during the test period and the training are more similar.

Focusing on some particular device, we can report that sensors 4003 and 4013 were not able to correctly measure humidity and were set to broken status during the autumn of 2019 (as can be seen in Fig. 7); thus, the data used for calibration were collected before the maintenance and the sensors were tested after. This can explain the bad performance of these sensors. The NO_2 sensor of the device 4004 broke after the calibration and was changed on the 17th of June 2020, before the test period. Since each cell must be calibrated separately, this experiment demonstrates that using a model calibrated on a different cell on the data collected by a new cell generates erroneous predictions.

In contrast to **Exp.1** and **Exp.3**, which underwent a testing period in winter and spring where low concentrations of O_3 were measured, in **Exp.2**, the values of O_3 span a broad range encompassing all five EEA classes, as presented in Table 4. It is noteworthy that the precision meets supplementary or regulatory monitoring standards, particularly for the four highest classes, despite all sensors registered bad performance in terms of RAE, MAE, MRE, and accuracy.

6.3. Exp.3

In **Exp.3**, we tested the same calibration models of **Exp.2** > 9 months later, during the winter of 2021.

For both NO and NO_2, LSTM shows better predictions for sensors 4005, 4010, 4011, and 4013; while VR + SVR has better performances for sensors 4004, 4007, 4009, 4012, and 4014. Analyzing the distribution of the pollutant concentrations registered by the legal station during the co-location periods, we observed that the difference between the mean values in the training and the test period is lower for sensors 4007, 4009, 4012 and 4014. Thus, VR + SVR has better performances when

Table 7
Results of Exp.3.

				RMSE		MAE		MRE		ACCURACY	
gas	device	train size	% out of range	VR+ SVR	LSTM	VR+ SVR	LSTM	VR+ SVR	LSTM	VR+ SVR	LSTM
NO	4004	14987	0	5.84	7.69	4.59	5.92	0.26	0.3	0.96	0.94
	4005	5689	38	24.03	19.2	17.54	10.14	4.22	1.42	0.88	0.92
	4006	13887	1	20.53	21.9	11.46	11.34	1.93	1.47	0.9	0.9
	4007	8368	4	20.55	23.9	11.24	11.76	1.44	1.73	0.9	0.9
	4009	7857	5	17.39	18.09	8.9	8.67	1.24	1.01	0.92	0.92
	4010	11013	0	5.92	4.58	3.24	3.59	0.17	0.19	0.98	0.97
	4011	11010	0	7.15	4.72	5.52	3.08	0.65	0.17	0.96	0.96
	4012	9518	0	17.87	19.12	8.34	9.41	0.92	0.97	0.92	0.92
	4013	7666	12	14.25	7.69	10.03	5.96	0.48	0.51	0.85	0.94
	4014	10342	1	5.74	6.5	3.43	4.62	0.86	1.28	0.98	0.97
	Mean			**13.93**	**13.34**	**7.35**	**7.45**	**1.22**	**0.91**	**0.93**	**0.93**
NO_2	4005	5689	0	19.23	16.9	15.21	12.72	0.73	0.53	0.76	0.83
	4006	13887	0	20.12	16.64	16.41	13.1	0.97	0.67	0.79	0.83
	4007	8368	0	16.24	17.59	12.4	13.22	0.62	0.62	0.85	0.83
	4009	7857	0	14.26	17.32	10.53	13.44	0.5	0.61	0.87	0.81
	4010	11013	0	5.92	5.87	4.67	4.81	0.15	0.13	0.92	0.86
	4011	11010	0	16.79	13.52	15.21	12.61	0.52	0.39	0.51	0.63
	4012	9518	1	13.01	13.74	8.42	9.47	0.34	0.35	0.88	0.86
	4013	7666	0	14.49	7.58	11.48	5.86	0.29	0.15	0.83	0.85
	4014	10342	0	11.72	11.9	9.1	9.75	0.62	0.63	0.92	0.91
	Mean			**14.64**	**13.45**	**11.49**	**10.55**	**0.53**	**0.45**	**0.81**	**0.82**
O_3	4005	3376	26	37.82	37.88	28.78	17.21	3.42	1.32	0.77	0.85
	4006	3376	0	16.96	18.41	15.02	20.27	1.2	1.48	0.94	0.92
	4007	6055	0	16.94	16.98	11.65	10.75	0.39	0.34	0.93	0.92
	4009	6263	0	14.74	14.26	9.75	9.48	0.4	0.32	0.95	0.95
	4012	7918	0	8.67	8.74	6.62	6.81	0.56	0.41	0.96	0.94
	4013	6066	0	24.32	24.36	21.18	5.94	4.37	0.7	1.00	1.00
	4014	8742	0	20.47	21.94	16.34	15.45	0.42	0.4	0.84	0.82
	Mean			**19.99**	**20.37**	**15.62**	**12.27**	**1.54**	**0.71**	**0.91**	**0.91**

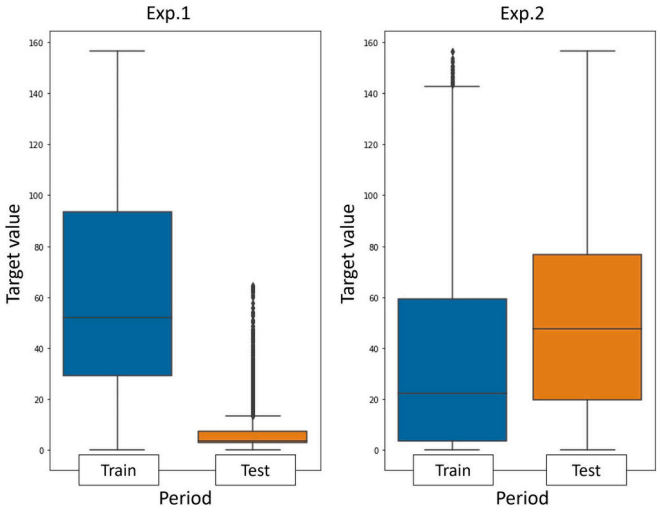


Fig. 8. Comparison of distribution ranges of O_3 concentrations in the train (blue) and test (orange) datasets for **Exp.1** (on the left) and **Exp.2** (on the right). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the distribution of values in the training is more similar to the one of the test period. However, when LSTM is performing better than VR + SVR, the gain in performance is more significant.

In terms of bias, the results are good for both NO and NO_2. The precision of sensors 4005, 4007, 4009, and 4012 is very low for both NO and NO_2. Sensor 4006 shows a high precision error only for NO. The other sensors have very good performances, even reaching the regulatory monitoring required performances. This is an important result, considering that the mean value of the NO concentrations during the test period (17.22 $\mu g/m^3$) was beyond the noise limit of the cell (18.75 $\mu g/m^3$). The mean concentration of NO is low because the test periods are during the co-location near the background station which is located in a green area far away from the most relevant source of pollution, such as road traffic.

For O_3 predictions, even if the mean value of RMSE is slightly lower for VR + SVR, according to the other metrics the LSTM algorithm performs better for the majority of sensors, in particular sensors 4005 and 4013. A unique exception is represented by sensor 4006, in this case, due to the small dimension of the training data for O_3, VR + SVR outperformed the LSTM model. The performances of O_3 prediction are ameliorated for both models compared with previous experiments. In Fig. 9, a comparison between the predicted curves for sensor 4012 in **Exp.1** (March 2020) and **Exp.3** (March 2021) is displayed; we can observe that in **Exp.1** the models were not even able to predict the correct trend. In **Exp.3**, the model has a very similar trend to the ground truth curve. These improvements in the results are a consequence of the

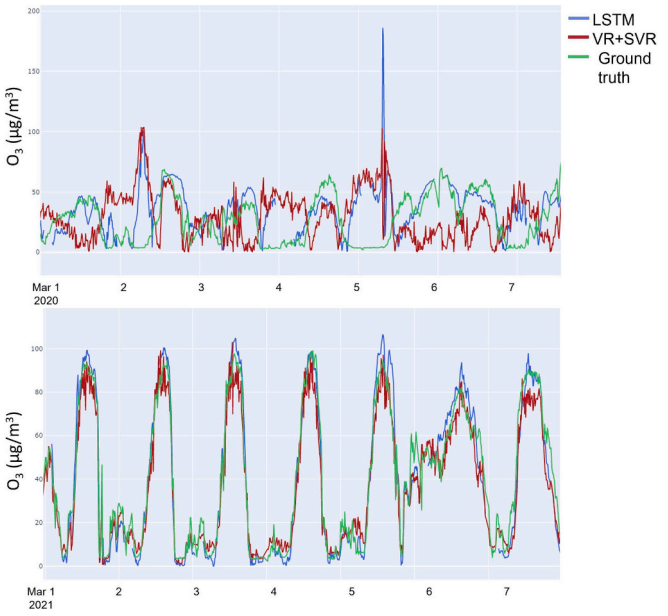


Fig. 9. Calibrated data of Exp.1 (at the top) and Exp.3 (at the bottom) for device 4012. The results refer to the same month (March) of two consecutive years (2020 and 2021).

inclusion of a winter period in the data used to train the model.

Exp.3 underwent a test period similar to **Exp.1**. In terms of bias and precision, similar performances can be observed as reported in Table 4.

7. Discussion

This section is devoted to answering to **R4** research question, i.e. exploring what are the factors that most influence the performance of the calibration models. Firstly, we discuss the results of the three experiments in terms of performances (i.e., assessment metrics) w.r.t. the whole test datasets (Section 7.1) and the percentage of "out-of-range" values (Section 7.2). Then, we study the dependency of the performances on the weather conditions (Section 7.3) and try to identify the cause of "big errors" (Section 7.4). Finally, we compare the concentrations obtained by our algorithms to the validated observations from the legal stations (Section 7.5).

7.1. Performance

Comparing the results of VR + SVR and LSTM, we can observe that, even if in some cases VR + SVR is performing better than LSTM, the increase in performance is usually less significant. Instead, in most cases where LSTM shows better results it significantly increases accuracy and reduces RMSE, MAE and MRE. Nevertheless, in cases where the size of the training dataset is restricted, the performance of LSTM models may prove to be inadequate. In these cases, the VR + SVR can provide better performance than the LSTM. From our experiments, we observed that a training period with <4000 observations generates a model unable to achieve satisfactory performance levels.

Moreover, the LSTM model was not able to predict the values of concentration in the 13% of test observations due to the absence of a long enough sequence of previous observations without missing values in the input features. VR + SVR instead is always able to provide a prediction, satisfying the requirement of data completeness reported in Williams et al. (2014). Therefore, VR + SVR is a good alternative to LSTM when there are frequent holes in the time series of measurements and when there are not enough past observations to successfully employ LSTM.

With VR + SVR we demonstrated how to overcome the problem of tree-based ensemble regressors unable to predict the values whose features are outside the training range. However, the VR + SVR model is affected by a more evident degradation of performance through time and may need more frequent re-calibrations.

7.2. Out-of-range values

The presence of "out-of-range" inputs during the test period can influence the performances of the models. We evaluate separately the performances for "in-range" values and "out-of-range" values.

Due to the limited dimension of the training dataset and the fact that the test period is in a different season w.r.t. the training, **Exp.1** is the one with the highest concentration of "out-of-range" observations. When using the VR + SVR, the MAE on "out-of-range" observations is increased by 115% on average, the MRE by 27%, and the RMSE by 80%; although, the value of accuracy is decreased by 18%. The degradation in performance associated with "out-of-range" values is mitigated through the utilization of LSTM. The MAE is increased by 94%, the MRE by 62%, and the RMSE by 69%, while accuracy undergoes a reduction of 13%.

When the "out-of-range" values are only a few percentages of all the observed values for a certain sensor, we can assume they are anomalous; however, when the percentage of "out-of-range" values is significant (as shown in Fig. 10) it means that we are testing on a period where the distribution of input features is different from the one on which the model is trained; therefore, the model is no more reliable under these new conditions. In **Exp.1**, sensors 4009, 4011, and 4012 have a very high percentage of "out-of-range" values for the NO pollutant (26%, 28%, and 46% respectively). For these sensors, the values of MAE and RMSE are lower and the accuracy is higher in "out-of-range" values than in "in-range" values, and, in all cases, LSTM performs better than VR + SVR. Indeed, the iterative nature of the LSTM models allows learning a new distribution on the fly; even if VR + SVR can learn the relation between a concentration and the features of the current and the two previous observations during training, once the model is trained this relation cannot be updated. As a consequence, the LSTM performs better on "out-of-range" values in 79% of cases.

In **Exp.3**, the percentage of "out-of-range" values notably reduced, suggesting that, in most cases, these deviations can be regarded as anomalies. Evaluating the LSTM model on the "out-of-range" values shows a 32% increase in MAE, a 22% rise in RMSE, more than doubling the MRE by 105%, and a 3% decrease in accuracy. Conversely, when utilizing VR + SVR, there is a 22% increase in MAE, only a 5% rise in RMSE, a 92% spike in MRE, and a decrease in accuracy of <1%. In this case, the VR + SVR models outperform the LSTM model.

7.3. Dependency on weather conditions

Weather conditions (temperature and humidity) strongly influence the pollutant concentrations. Nevertheless, in some cases, due to the lack of appropriate data, it could be necessary to apply the calibration models in a different season w.r.t. the training period. Considering **Exp.1**, sensors 4004, 4007, 4009, and 4014 have a very high percentage of test observations whose weather conditions have not been observed during the training period (65%, 84%, 85%, and 82% respectively). In these cases, the VR + SVR had the worst performances regardless of the pollutants. The LSTM learns the relation between weather conditions and the pollutant concentration dynamically; thus, if the observations have been performed under different weather conditions, the LSTM will have better performances than VR + SVR.

Other important observations need to be made regarding humidity. In **Exp.1**, several observations have been collected with humidity higher than 85%, which is the maximum limit for reliable measurements indicated in Alphasense (2015). We evaluated the assessment metrics of NO prediction considering only reliable observations and observed that the MAE and RMSE values for devices 4007, 4009, and 4014 were halved. Inconsistent weather conditions primarily affect the

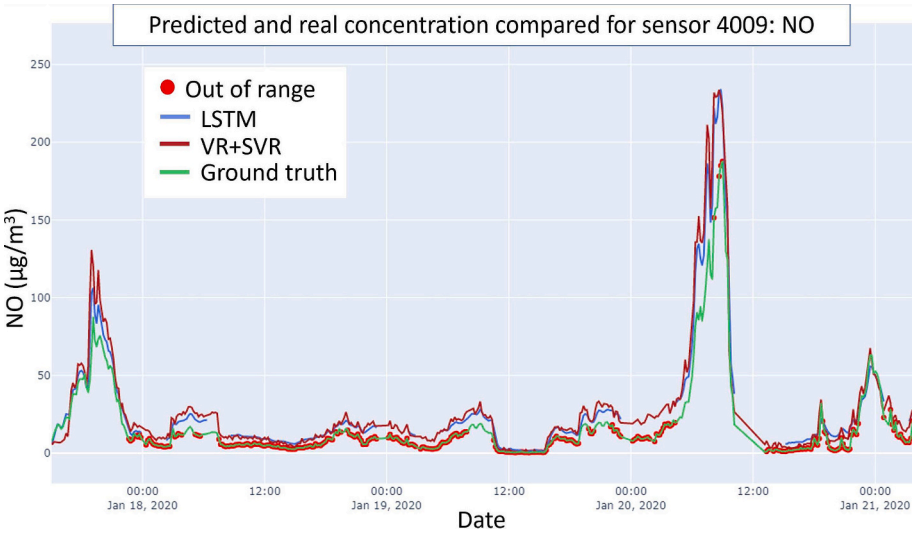


Fig. 10. Comparison of predicted and real concentrations of NO pollutant of the device 4009 in Exp.1.

performance of NO. The VR + SVR model appears to be particularly sensitive to unreliable humidity values, as it demonstrates a noticeable performance improvement when excluding these data points.

Similarly, as reported in Alphasense (2015), the sensors have been tested by the manufacturer only for temperature values below 40 °C. While, in our case, in **Exp.2**, the percentage of observations with the value of temperature higher than 40 °C is between 1% and 3%. The sensors with the highest percentage are 4004 and 4005. When evaluating the performance obtained by removing these observations, the errors are significantly reduced. In the case of the NO sensor of the device 4005, the MAE is reduced by >89% by both models. On average, the NO_2's MAE is reduced by 21% for LSTM and by 36% for the VR + SVR. In the case of O_3, the two models behave very differently: LSTM's MAE is worsened by 34%, and VR + SVR's is increased by 19%. As a consequence, it seems to be quite important to remove the observations that have a temperature above the reliable condition of 40 °C, as the measurements performed by the sensors appear to be erroneous.

In **Exp.3** sensor 4005 has the worst performance for O_3. A possible reason could be that more than half of the observations in the test dataset exhibit temperatures and humidity values exceeding the maximum or falling below the minimum values observed during the model's training phase. Weather conditions hold particular significance for O_3 since the ground-level ozone is generated primarily by photochemical reactions caused by solar radiations. These reactions occur in the presence of precursor pollutants such as volatile organic compounds (VOCs) or nitrogen oxides (NOx).

7.4. Big errors

We evaluated the absolute error for each observation comparing the predicted value and the reference concentration value. Then, we identified "big errors" as those whose absolute value is higher or equal to the MAE plus 5 times the standard deviation of the absolute errors. Our analysis revealed that a majority of these big errors occur concurrently across multiple sensors of different devices. Additionally, some of these errors are consistent across both methodologies.

In the case of **Exp.2**, on the 27th of July 2020, the VR + SVR and LSTM models had very high prediction errors between 2 and 3 p.m.. Looking at the weather conditions, we noticed that the temperature surpassed the reliable threshold of 40 °C, while the humidity was relatively low (around 40%). Low humidity at high temperatures can dry out the dielectric of the sensor, significantly affecting sensor performance (Concas et al. (2021)).

The big errors for NO_2 are very similar for VR + SVR and LSTM; mostly during early morning hours and evening. The majority of the observations corresponding to these big errors had a very high humidity (above 90%) and temperature values significantly lower compared with daily values.

Focusing on NO prediction, the VR + SVR shows big errors that do not occur in the LSTM predictions. These big errors occur when the difference between the observed NO working channel value and the mean of the training observations is high.

7.5. Validated values

The reference values of the legal stations collected every minute may contain anomalies due to malfunctions. The environmental agency in charge of their maintenance applies a data validation process to the one-minute data and aggregates them to obtain hourly validated data. During this process, anomalies are identified and removed. The validation is not performed in real-time, i.e., at the same time the observation is obtained, but it requires some time. Therefore, validated data may be available after some days or months. The number of anomalies in the legal station data is quite low due to the high precision of the instrument, however, in some cases, there is an evident difference between the validated value and the hourly average of the corresponding non-validated values. In the experiments of Section 6, we compared the concentrations predicted by our models with the non-validated values. Now, we aim to compare them w.r.t. the validated data. In our experiments, the period in which the difference between validated and non-validated values is more evident is the second half of January 2020 (Fig. 11) for both NO and NO_2. This period corresponds to the test period of **Exp.1** for all sensors excluding 4010, 4011 and 4012. Since validated values are more reliable, we decided to compare the hourly average concentration predicted by the models with the validated concentrations. The experiment evaluation obtained considering the hourly validated data of the legal stations is reported in Table 8 and can be compared with the values in Table 3.

The errors in **Exp.1** are reduced in all cases except for MRE in NO_2 prediction and the accuracy always increased. We can assume that the high errors of the majority of sensors in NO and NO_2 prediction are caused by the presence of incorrect values in the non-validated data used as ground truth. In **Exp.2**, compared with **Exp.1**, the errors are considerably reduced for all pollutants except for O_3. Observing the performances of each sensor one at a time, we notice that sensors 4003, 4004, and 4013 had very bad performances (their MAE was higher than 55 $\mu g/m^3$ for both models), the other sensors instead have better performances (their average MAE is 14.57 for LSTM and 13.38 $\mu g/m^3$ for

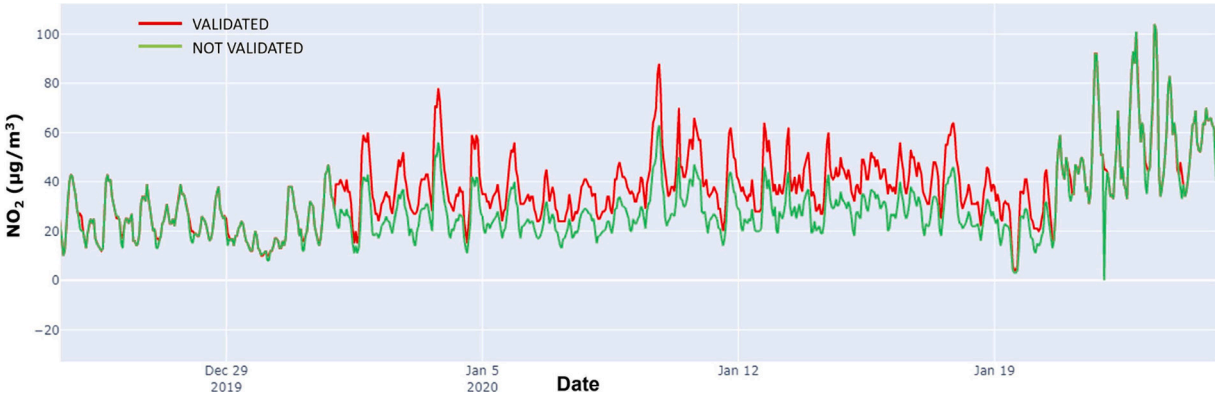


Fig. 11. Comparison between validated values (red) and not validated values (green) of NO_2 concentrations on January 2020. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 8
Evaluation of experiments comparing the hourly averaged predicted concentrations with the hourly validated data of legal stations.

		RMSE		MAE		MRE		ACCURACY	
	Gas	VR + SVR	LSTM	VR + SVR	LSTM	VR + SVR	LSTM	VR + SVR	LSTM
Exp.1	NO	15.48	10.87	9.24	6.65	1.02	0.95	0.90	0.92
	NO_2	13.07	11.01	10.34	8.50	0.58	0.52	0.83	0.84
	O_3	22.10	20.65	17.53	14.51	3.04	2.22	0.97	0.96
Exp.2	NO	3.55	3.02	2.14	1.62	1.29	0.71	1.00	0.92
	NO_2	8.95	9.84	6.49	7.41	0.49	0.52	0.98	0.95
	O_3	40.68	43.67	34.07	36.79	24.88	26.40	0.47	0.46
Exp.3	NO	13.18	12.59	8.12	7.12	1.47	1.02	0.92	0.92
	NO_2	14.69	13.39	11.42	10.45	1.06	0.80	0.81	0.82
	O_3	21.13	19.03	15.89	12.31	2.64	1.59	0.89	0.89

VR + SVR). As already discussed in Section 6.2, 4004 is a special case employed to demonstrate that calibration needs to be repeated when a cell is changed, and 4003 and 4013 both had a broken humidity sensor; thus, bad performances were expected. In the end, the hourly results of **Exp.3** suggest that LSTM is able to reach better performances than VR + SVR for all pollutants.

8. Conclusion

This paper introduced HypeAIR, a new open-source framework designed for real-time low-cost sensor calibration. This framework demonstrates its capability to handle the entire calibration process for any low-cost air quality sensor (LCAQ) and pollutant. Currently integrating two effective methodologies, VR + SVR and LSTM, HypeAIR is designed to be flexible and customizable, allowing for the incorporation of new methodologies.

Through extensive testing with 12 LCAQ sensors, for NO, NO_2, and O_3 monitoring over 21 months, this study reveals that the two methodologies, VR + SVR and LSTM, consistently outperform both the original manufacturer calibration technique and the baseline approach (i.e., a variation of the Random Forest algorithm). Importantly, these methodologies maintain their efficiency over time.

Furthermore, the paper successfully addresses all the research questions initially defined in the Introduction, as reported in the summary boxes below.

Looking ahead, the exploration of innovative sensor calibration methodologies, including the application of neural network Casari et al. (2023) and transfer learning techniques, holds promise. We aim to investigate and leverage cross-dependencies among LCAQ sensors located in different areas within the city.

R1 *Can a tool be developed that is not merely an* ad hoc *calibration solution for a specific sensor and pollutant but is adaptable to a variety of sensors?*

Yes, HypeAir has been demonstrated to be a versatile framework and can effectively manage the calibration process, making it adaptable to a variety of sensors and pollutants.

R2 *Can LCAQ sensor measurements achieve the same reliability as legal station measurements?*

Yes, with adequate training data and consideration of different seasons, LCAQ sensor measurements can achieve the required reliability for supplemental or even regulatory monitoring.

R3 *What is the best solution to ensure that calibration performance is maintained over time in a real environment?*

To ensure the consistency of calibration performance over time, it is crucial to prevent drifts or substitutions of the sensor cell, as such changes would require the generation of a new model. Specifically, the LSTM methodology demonstrates greater resistance to drifts, as it can dynamically adapt and learn the relationship between evolving environmental conditions and pollutant concentrations.

R4 *What factors exert the most significant influence on the performance of calibration models?*

The primary factors influencing the performance of calibration models are changes in weather conditions (seasonality) and the presence of out-of-range values in measurements. Performances are generally better when pollutant concentrations are high, particularly in winter for NO and NO_2, and in summer for O_3.

R5 *Is it possible to define essential guidelines to be followed by scientists embarking on the calibration of low-cost sensors for the first time?*

The essential guidelines for scientists embarking on the calibration of LCAQ sensors, as outlined in this study, can be summarized in the following key steps:

- **Study AQ observations and gas range values**: Begin by examining air quality observations gathered at reference stations. Understand the range values for the gases intended for monitoring and their seasonality.

- **Consider temperature and humidity variability**: Investigate the variability of temperature and humidity in your city, as these factors significantly impact the performance of LCAQ sensors.
- **Collect adequate training data**: Ensure a prolonged co-location period, that couples raw measurements from sensors with observations from legal stations, with a substantial number of measurements. Aim for a minimum of 10,000 observations for robust reliability suitable for regulatory monitoring, or at least 5000 for supplementary monitoring.
- **Diversify training periods and stations**: Conduct training in different periods and at various stations, if available. Prefer seasons characterized by higher pollutant levels and stations with elevated concentrations.
- **Implement data cleaning process**: Apply a data cleaning process to identify and rectify anomalies. Address anomalies in both legal station measurements and LCAQ raw observations to enhance the overall calibration results.

CRediT authorship contribution statement

Chiara Bachechi: Writing – review & editing, Writing – original draft, Validation, Software, Methodology, Data curation. **Federica Rollo:** Writing – review & editing, Writing – original draft, Methodology, Data curation. **Laura Po:** Writing – review & editing, Writing – original draft, Supervision, Project administration, Methodology, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data have been made openly available

Acknowledgements

The research reported in this paper was partially supported by the TRAFAIR project 2017-EU-IA-0167, co-financed by the Connecting Europe Facility of the European Union. The authors would like to thank the LARMA group of the University of Modena and Reggio Emilia for the deployment and maintenance of the sensor network during the project. We also would like to thank Luca Casarotti who developed the first implementation of VR + SVR and LSTM and carried out some initial tests during his Master's Degree internship.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ecoinf.2024.102568.

References

Alphasense, 2015. Technical Specification Doc. Ref. COB4/APR. Alphasense Ltd. https://cdn.decentlab.com/download/datasheets/Alphasense-Gas-Sensor-for-DL-AC-datasheet.pdf. Accessed: 2022-08-08.

Bachechi, C., Desimoni, F., Po, L., Casas, D.M., 2020. Visual analytics for spatio-temporal air quality data. In: 2020 24th International Conference Information Visualisation (IV), pp. 460–466. https://doi.org/10.1109/IV51561.2020.00080.

Bachechi, C., Po, L., Desimoni, F., 2022. Real-time visual analytics for air quality. In: Kovalerchuk, B., Nazemi, K., Andonie, R., Datia, N., Banissi, E. (Eds.), Integrating Artificial Intelligence and Visualization for Visual Knowledge Discovery. Springer International Publishing, Cham, pp. 485–515. https://doi.org/10.1007/978-3-030-93119-3_19.

Baruah, A., Zivan, O., Bigi, A., Ghermandi, G., 2023. Evaluation of low-cost gas sensors to quantify intra-urban variability of atmospheric pollutants. Environ. Sci. Atmosph. https://doi.org/10.1039/d2ea00165a.

Basak, D., Pal, S., Patranabis, D., 2007. Support vector regression. In: Neural Information Processing – Letters and Reviews, 11.

Boubrima, A., Bechkit, W., Rivano, H., 2017. Optimal WSN deployment models for air pollution monitoring. IEEE Trans. Wirel. Commun. 16, 2723–2735. https://doi.org/10.1109/TWC.2017.2658601.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324.

Casari, M., Po, L., Zini, L., 2023. Airmlp: a multilayer perceptron neural network for temporal correction of pm2.5 values in Turin. Sensors 23. URL. https://www.mdpi.com/1424-8220/23/23/9446. https://doi.org/10.3390/s23239446.

Casarotti, L., 2021. Toward Urban Air Quality Monitoring: Machine Learning and Deep Learning Compared. Technical Report. University of Modena and Reggio Emilia.

Chai, T., Draxler, R., 2014. Root mean square error (rmse) or mean absolute error (mae)?– arguments against avoiding rmse in the literature. Geosci. Model Dev. 7, 1247–1250. https://doi.org/10.5194/gmd-7-1247-2014.

Chang, Y.S., Chiao, H.T., Abimannan, S., Huang, Y.P., Tsai, Y.T., Lin, K.M., 2020. An lstm-based aggregated model for air pollution forecasting. Atmospheric. Pollut. Res. 11, 1451–1463. URL. https://www.sciencedirect.com/science/article/pii/S1309104220301215. https://doi.org/10.1016/j.apr.2020.05.015.

Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.. Learning Phrase Representations Using RNN Encoder–Decoder for Statistical Machine Translation. URL. https://www.aclweb.org/anthology/D14-1179.

Clements, A., Duvall, R., Greene, D., Dye, T., 2022. The Enhanced Air Sensor Guidebook. US Environmental Protection Agency. https://cfpub.epa.gov/si/si_public_record_report.cfm?Lab=CEMM&dirEntryId=356426. Accessed: 2023-12-08.

Concas, F., Mineraud, J., Lagerspetz, E., Varjonen, S., Liu, X., Puolamäki, K., Nurmi, P., Tarkoma, S., 2021. Low-cost outdoor air quality monitoring and sensor calibration: a survey and critical analysis. ACM Trans. Sen. Netw. 17 https://doi.org/10.1145/3446005.

De Vito, S., Esposito, E., Salvato, M., Popoola, O., Formisano, F., Jones, R., Di Francia, G., 2018. Calibrating chemical multisensory devices for real world applications: an in-depth comparison of quantitative machine learning approaches. Sensors Actuators B Chem. 255, 1191–1210. URL. https://www.sciencedirect.com/science/article/pii/S0925400517313692. https://doi.org/10.1016/j.snb.2017.07.155.

De Vito, S., Esposito, E., Massera, E., Formisano, F., Fattoruso, G., Ferlito, S., Del Giudice, A., D'Elia, G., Salvato, M., Polichetti, T., D'Auria, P., Ionescu, A.M., Di Francia, G., 2021. Crowdsensing iot architecture for pervasive air quality and exposome monitoring: design, development, calibration, and long-term validation. Sensors 21. URL. https://www.mdpi.com/1424-8220/21/15/5219. https://doi.org/10.3390/s21155219.

Fang, W., Zhu, R., Lin, J.C.W., 2023. An air quality prediction model based on improved vanilla lstm with multichannel input and multiroute output. Expert Syst. Appl. 211, 118422. URL: https://www.sciencedirect.com/science/article/pii/S0957417422015263 https://doi.org/10.1016/j.eswa.2022.118422.

Ferrer-Cid, P., Barceló-Ordinas, J.M., García-Vidal, J., Ripoll, A., Viana, M., 2019. A comparative study of calibration methods for low-cost ozone sensors in iot platforms. IEEE Internet Things J. 6, 9563–9571. https://doi.org/10.1109/JIOT.2019.2929594.

Friedman, J., 2002. Stochastic gradient boosting. Comp. Stat. Data Analys. 38, 367–378. https://doi.org/10.1016/S0167-9473(01)00065-2.

Geurts, P., Ernst, D., Wehenkel, L., 2006. Extremely randomized trees. Mach. Learn. 63, 3–42. https://doi.org/10.1007/s10994-006-6226-1.

Hengl, T., Nussbaum, M., Wright, M.N., Heuvelink, G.B.M., Gräler, B., 2018. Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. PEERJ 6, e5518.

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput. 9, 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735.

Hofman, J., Do, T.H., Qin, X., Bonet, E.R., Philips, W., Deligiannis, N., Manna, V.P.L., 2022. Spatiotemporal air quality inference of low-cost sensor data: evidence from multiple sensor testbeds. Environ. Model Softw. 149, 105306. https://doi.org/10.1016/j.envsoft.2022.105306.

Huang, Qiujun, Mao, Jingli, Liu, Yong, 2012. An improved grid search algorithm of svr parameters optimization. In: 2012 IEEE 14th International Conference on Communication Technology, pp. 1022–1026. https://doi.org/10.1109/ICCT.2012.6511415.

ISO, 2011. ISO 19156:2011 - Geographic Information – Observations and Measurements. https://doi.org/10.13140/2.1.1142.3042.

Kingma, D.P., Ba, J., 2015. Adam: A method for stochastic optimization. In: Bengio, Y., LeCun, Y. (Eds.), 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015. Conference Track Proceedings. URL. http://arxiv.org/abs/1412.6980.

Li, J., Heap, A.D., 2014. Spatial interpolation methods applied in the environmental sciences: a review. Environ. Model Softw. 53, 173–189. URL. http://www.sciencedirect.com/science/article/pii/S1364815213003113. https://doi.org/10.1016/j.envsoft.2013.12.008.

Maag, B., Saukh, O., Hasenfratz, D., Thiele, L., 2016. Pre-deployment testing, augmentation and calibration of cross-sensitive sensors. In: Proceedings of the 2016 International Conference on Embedded Wireless Systems and Networks, Junction Publishing, USA, pp. 169–180.

Maag, B., Zhou, Z., Thiele, L., 2018. A survey on sensor calibration in air pollution monitoring deployments. IEEE Internet Things J. 5, 4857–4870. https://doi.org/10.1109/JIOT.2018.2853660.

Marius, P., Balas, V., Perescu-Popescu, L., Mastorakis, N., 2009. Multilayer perceptron and neural networks. WSEAS Trans. Circuits Syst. 8.

Martínez, D., Po, L., Lado, R.T., Viqueira, J.R.R., 2022. TAQE: a data modeling framework for traffic and air quality applications in smart cities. In: Braun, T., Cristea, D., Jäschke, R. (Eds.), Graph-Based Representation and Reasoning - 27th International Conference on Conceptual Structures, ICCS 2022, Münster, Germany, September 12-15, 2022, Proceedings. Springer, pp. 25–40. https://doi.org/10.1007/978-3-031-16663-1_3.

Mead, M., Popoola, O., Stewart, G., Landshoff, P., Calleja, M., Hayes, M., Baldovi, J., McLeod, M., Hodgson, T., Dicks, J., Lewis, A., Cohen, J., Baron, R., Saffell, J., Jones, R., 2013. The use of electrochemical sensors for monitoring urban air quality in low-cost, high-density networks. Atmos. Environ. 70, 186–203. URL. https://www.sciencedirect.com/science/article/pii/S1352231012011284. https://doi.org/10.1016/j.atmosenv.2012.11.060.

Meyer, H., Pebesma, E., 2021. Predicting into unknown space? Estimating the area of applicability of spatial prediction models. Methods Ecol. Evol. 12 https://doi.org/10.1111/2041-210X.13650.

Miech, J.A., Stanton, L., Gao, M., Micalizzi, P., Uebelherr, J., Herckes, P., Fraser, M.P., 2021. Calibration of low-cost no2 sensors through environmental factor correction. Toxics 9. URL. https://www.mdpi.com/2305-6304/9/11/281. https://doi.org/10.3390/toxics9110281.

Motlagh, N.H., Lagerspetz, E., Nurmi, P., Li, X., Varjonen, S., Mineraud, J., Siekkinen, M., Rebeiro-Hargrave, A., Hussein, T., Petaja, T., Kulmala, M., Tarkoma, S., 2020. Toward massive scale air quality monitoring. IEEE Commun. Mag. 58, 54–59. https://doi.org/10.1109/MCOM.001.1900515.

Özgür, A., Nar, F., 2020. Effect of dropout layer on classical regression problems. In: 2020 28th Signal Processing and Communications Applications Conference (SIU), pp. 1–4. https://doi.org/10.1109/SIU49456.2020.9302054.

Peng, J., Lee, K., Ingersoll, G., 2002. An introduction to logistic regression analysis and reporting. J. Educ. Res. 96, 3–14. https://doi.org/10.1080/00220670209598786.

Po, L., Rollo, F., Bachechi, C., Corni, A., 2019a. From sensors data to urban traffic flow analysis. In: 2019 IEEE International Smart Cities Conference, ISC2 2019, Casablanca, Morocco, October 14-17, 2019. IEEE, pp. 478–485. https://doi.org/10.1109/ISC246665.2019.9071639.

Po, L., Rollo, F., Viqueira, J.R.R., Lado, R.T., Bigi, A., López, J.C., Paolucci, M., Nesi, P., 2019b. TRAFAIR: understanding traffic flow to improve air quality. In: 2019 IEEE International Smart Cities Conference, ISC2 2019, Casablanca, Morocco, October 14-17, 2019. IEEE, pp. 36–43. https://doi.org/10.1109/ISC246665.2019.9071661.

Rollo, F., Po, L., 2021. Senseboard: Sensor monitoring for air quality experts. In: Costa, C., Pitoura, E. (Eds.), Proceedings of the Workshops of the EDBT/ICDT 2021 Joint Conference, Nicosia, Cyprus, March 23, 2021, CEUR-WS.org. URL. http://ceur-ws.org/Vol-2841/BigVis_3.pdf.

Seng, D., Zhang, Q., Zhang, X., Chen, G., Chen, X., 2021. Spatiotemporal prediction of air quality based on lstm neural network. Alex. Eng. J. 60, 2021–2032. URL. https://www.sciencedirect.com/science/article/pii/S1110016820306438. https://doi.org/10.1016/j.aej.2020.12.009.

Sinha, P., Gaughan, A.E., Stevens, F.R., Nieves, J.J., Sorichetta, A., Tatem, A.J., 2019. Assessing the spatial sensitivity of a random forest model: application in gridded population modeling. Comput. Environ. Urban. Syst. 75, 132–145. URL. https://www.sciencedirect.com/science/article/pii/S0198971518302862. https://doi.org/10.1016/j.compenvurbsys.2019.01.006.

Spinelle, L., Gerboles, M., Villani, M.G., Aleixandre, M., Bonavitacola, F., 2015. Field calibration of a cluster of low-cost available sensors for air quality monitoring. Part A: ozone and nitrogen dioxide. Sensors Actuators B Chem. 215, 249–257. URL. https://www.sciencedirect.com/science/article/pii/S092540051500355X. https://doi.org/10.1016/j.snb.2015.03.031.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B Methodol. 58, 267–288. http://www.jstor.org/stable/2346178.

Vito, S.D., Esposito, E., Salvato, M., Popoola, O., Formisano, F., Jones, R., Francia, G.D., 2017. Calibrating chemical multisensory devices for real world applications: an in-depth comparison of quantitative machine learning approaches. CoRR abs/1708.09175. URL. http://arxiv.org/abs/1708.09175. arXiv:1708.09175.

WHO, 2021. World Health Statistics 2021: Monitoring Health for the SDGs. World Health Organization, Geneva (Licence:CC BY-NC-SA 3.0 IGO).

Williams, R., Kilaru, V., Snyder, E., Kaufman, A., Dye, T., Rutter, A., Russell, A., Hafner, H., 2014. Air Sensor Guidebook. US Environmental Protection Agency. https://cfpub.epa.gov/si/si_public_record_report.cfm?Lab=NERL&dirEntryId=277996&simpleSearch=1&searchAll=air+sensor+guidebook. Accessed: 2023-12-08.

Zaytar, M.A., Amrani, C.E., 2020. Machine learning methods for air quality monitoring. In: NISS 2020: The 3rd International Conference on Networking, Information Systems & Security, Marrakech, Morocco, March 31 - April 2, 2020, ACM. https://doi.org/10.1145/3386723.3387835, 16:1–16:5.

Zimmerman, N., Presto, A.A., Kumar, S.P.N., Gu, J., Hauryliuk, A., Robinson, E.S., Robinson, A.L., Subramanian, R., 2018. A machine learning calibration model using random forests to improve sensor performance for lower-cost air quality monitoring. Atmosph. Meas. Tech. 11, 291–313. URL: https://amt.copernicus.org/articles/11/291/2018/ https://doi.org/10.5194/amt-11-291-2018.

Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. J. R. Stat. Soc. Ser. B 67, 301–320.