Low-Cost Outdoor Air Quality Monitoring and In-Field Sensor Calibration

Francesco Concas
Department of Computer Science
University of Helsinki
P.O. 68 (Pietari Kalmin Katu 5)
francesco.concas@helsinki.fi

Julien Mineraud
Department of Computer Science
University of Helsinki
P.O. 68 (Pietari Kalmin Katu 5)
julien.mineraud@helsinki.fi

Eemil Lagerspetz
Department of Computer Science
University of Helsinki
P.O. 68 (Pietari Kalmin Katu 5)
eemil.lagerspetz@cs.helsinki.fi

Samu Varjonen
Department of Computer Science
University of Helsinki
P.O. 68 (Pietari Kalmin Katu 5)
samu.varjonen@helsinki.fi

Kai Puolamäki
Department of Computer Science
University of Helsinki
P.O. 68 (Pietari Kalmin Katu 5)
kai.puolamaki@helsinki.fi

Petteri Nurmi
Department of Computer Science
University of Helsinki
P.O. 68 (Pietari Kalmin Katu 5)
ptnurmi@cs.helsinki.fi

Sasu Tarkoma
Department of Computer Science
University of Helsinki
P.O. 68 (Pietari Kalmin Katu 5)
sasu.tarkoma@helsinki.fi

February 6, 2020

ABSTRACT

The significance of air pollution and problems associated with it is fueling deployments of air quality monitoring stations worldwide. The most common approach for air quality monitoring is to rely on environmental monitoring stations, which unfortunately are very expensive both to acquire and to maintain. Hence, environmental monitoring stations typically are deployed sparsely, resulting in limited spatial resolution for measurements. Recently, low-cost air quality sensors have emerged as an alternative that can improve granularity of monitoring. The use of low-cost air quality sensors, however, presents several challenges: they suffer from cross-sensitivities between different ambient pollutants; they can be affected by external factors such as traffic, weather changes, and human behavior; and their accuracy degrades over time. The accuracy of low-cost sensors can be improved through periodic *re-calibration* with particularly machine learning based calibration having shown great promise due to its capability to calibrate sensors in-field. In this article, we survey the rapidly growing research landscape of low-cost sensor technologies for air quality monitoring, and their calibration using machine learning techniques. We also identify open research challenges and present directions for future research.

Keywords air quality sensors · calibration · low-cost · machine learning · review · survey

1 Introduction

Air pollution is one of the most significant environmental challenges of our time. According to the World Health Organization (WHO), in 2016 air pollution was linked to over 4.2 million deaths per year (11.6% of all deaths) with mortality in low and middle-income countries particularly heavily affected by air pollution [1]. Besides having a direct effect on mortality, air pollution is strongly associated with a broad spectrum of acute and chronic diseases,

including cardiovascular diseases [2, 3, 4], lung diseases [5, 6, 7], several types of cancer [3, 8, 9], and even conditions affecting cognitive capabilities and the central nervous system [10, 11, 12]. Air pollution is also a significant economic burden worldwide with estimates suggesting that 2–5% of overall GDP is spent on treating diseases linked with air pollution [13]. The severity of air pollution is exacerbated by ever-increasing urbanization, with estimates suggesting that 96% of the world's population lives in areas where air pollution exceeds safe limits [14].

Understanding the characteristics of pollutants in urban environments is essential for counteracting problems linked with poor air quality and for assessing the effectiveness of initiatives designed for tackling it. This need for detailed air quality information is driving deployments of air quality monitoring technology worldwide, particularly in metropolitan regions.[1] [2] [3] Traditionally, air pollutant concentrations are monitored using professional air quality monitoring stations that meet strict accuracy criteria [15, 16, 17]. Such stations are highly accurate, but also very expensive, with the cost of a single station often reaching hundreds of thousands or even million dollars [18]. Operating such stations is also costly, requiring periodic maintenance from specially trained engineers. Due to the high deployment and operating costs, professional stations are deployed sparsely with most metropolitan regions only having a single measurement station. While in line with official recommendations, such density is not sufficient, as even a single city block can witness significant variations in pollutant concentrations. For example, congested traffic corridors, such as intersections or bus stops, tend to have significantly higher pollution concentrations than areas around them [19, 20]. To accurately assess the health and environmental risks of pollutants it is also necessary to understand the chemical composition of pollutants, which varies depending on the season, and characteristics of industry and traffic in the region [21, 22, 23]. For these reasons, accurate monitoring of air pollution inside metropolitan regions would require deploying hundreds or even thousands of air quality monitoring stations. In contrast, WHO recommends deploying one air quality monitoring station per square kilometre[4] whereas the EU clean air directive suggests one station per million inhabitants [24].

Low-cost air quality sensors, costing less than $10 000, have recently emerged as a way to reach higher density deployments and achieve higher spatial resolution in air quality monitoring [25, 26, 27]. Low-cost sensors are typically small in size, making it possible to deploy them densely as part of the urban infrastructure. For example, low-cost sensors have been deployed onto light poles and public transport vehicles [28, 29]. The main drawback of low-cost air quality sensors, however, is that their accuracy tends to be poor compared to professional monitoring stations [30, 31, 32, 33]. Indeed, measurements provided by low-cost sensors can vary significantly and have poor correspondence with professional-grade monitoring stations [34], with their performance best suited for specialized tasks where exact measurements are not required, such as detecting pollution hotspots [18].

Accuracy can be improved through periodic re-calibration, with a single calibration cycle improving accuracy for up to a fortnight, before drift [35] and other errors [36] start to decrease their accuracy. Periodic calibration alone, however, is not sufficient, since sensors are vulnerable to cross-sensitivities between different pollutants [31] and variations in atmospheric conditions, with temperature, humidity, and wind direction being examples of factors that influence the performance of sensors [32]. Additionally, the calibration process is highly time-consuming and laborious, making it unfeasible for large-scale deployments [37]. *Machine-learning-based calibration* has recently emerged as a potential solution to improve the generality of calibration techniques and to reduce work effort in the calibration process. The general idea in these approaches is to co-locate low-cost sensors in proximity of a professional station that is used as a reference and to train a model that can estimate the current error of the low-cost sensor from weather and other information sources. While several solutions for machine-learning-based calibration have been proposed[38, 39, 40], currently the overall research landscape around machine-learning-based calibration is poorly understood. Indeed, currently, there is limited information about which methods are best suited for tacking the specific research challenges posed by low cost air quality monitoring stations, how to best evaluate calibration techniques for low-cost sensors, and what are the other major research challenges in the area.

In this article, we contribute by presenting a survey and a gap analysis of the current research landscape on *low-cost outdoor air quality monitoring stations* and their *in-field calibration using machine learning techniques*. We focus specifically on low-cost technology aimed at *improving the resolution* of monitoring and *increasing the density* of deployments. Previous surveys on air quality monitoring (see Sec. 2.1) either focus on covering specific sensor technologies or dealing with a specific measurement challenge without comprehensively reviewing the entire research landscape. Besides reviewing existing work, we perform a rigorous gap analysis of the field to highlight important open research challenges.

---

[1] https://www.hindustantimes.com/delhi-news/delhi-gets-18-more-monitoring-stations-to-keep-tab-on-air-quality/story-kBtKpMeuPyz0KgOeDB1z9M.html

[2] http://www.baaqmd.gov/about-air-quality/air-quality-measurement/ambient-air-monitoring-network

[3] http://www.chinadaily.com.cn/china/2016-02/22/content_23595631.htm

[4] https://m.economictimes.com/small-biz/startups/newsbuzz/making-sense-of-air-quality-using-sensors/amp_articleshow/69262232.cms

Table 1: Existing related surveys

Scope	Survey
Sensors	Gas sensor technologies [49, 29, 50, 33]
	MOS sensors [51]
	NDIR sensors [52]
	Portable sensors [53, 54]
	Wearable sensors [55]
	Commercial sensors [56, 26]
	Low-cost sensors quality [57, 58]
	Usability of low-cost air quality sensors (AQSs) for atmospheric measurements [59]
Deployment	Cities and projects [33]
Calibration	Adaptation to drift [43]
	Optical PM sensors [44, 45]
	Error sources in calibration [46]
Integration	Air quality sensor networks [60]
	Land-use regression models [47, 48]
	Satellite-based estimations [61, 62]

2 Scope of the Survey

Research on low-cost air quality sensing has been recently gaining momentum as sensor technology has matured to a point where increasingly large-scale deployments are possible. For example, Cheng et al. [41] consider a testbed consisting of 1000 low-cost sensors deployed in Beijing and Motlagh et al. [42] present a testbed with 100+ sensors located in three different districts in Helsinki, Finland. This gain in momentum is also reflected in the number of research papers published on the topic with a query for `low-cost air quality calibration` returning over $180\,000$ results on Google Scholar. Despite this increase in research, limited work has been carried out in surveying and critically assessing the field. Indeed, existing research works mostly focus on individual pollutants and specific parts of the processing pipeline without covering issues surrounding air quality calibration in depth. Our survey addresses this limitation, providing a thorough review of the research field and performing a rigorous gap analysis to identify important open research challenges in the field.

2.1 Related Surveys

Table 1 summarizes previous surveys having partially overlapping scopes with our work. In terms of sensor technology, there has been a number of surveys focusing on specific types of sensors or technology. However, these have not addressed the suitability of different technologies for large-scale air quality monitoring or how technology affects the processing pipeline.

In terms of operations performed on individual sensor devices, only a small number of previous surveys exist. These predominantly focus on a specific research challenge. For example, Morawska et al. [36] provide an overview of deployments of low-cost air quality sensors but do not cover other parts of the processing pipeline. Gama et al. [43] address techniques for dealing with concept drift, but do not address how it relates to calibration. Liu et al. [44] and Zheng et al. [45] focus on calibration of optical particulate matter (PM) sensors. However, these surveys cover calibration broadly without addressing issues related to the use of machine learning specifically. Closest to our work, Maag et al. [46] provide an overview of low-cost sensor calibration, but limit their focus on specific sources of errors without discussing in detail the sensor technologies or the needed machine learning algorithms and their suitability.

Finally, there have been surveys about application areas that require fusing air quality information from several sensors, such as how to use air quality information to generate spatiotemporal pollution maps [47, 48]. These surveys do not address operations needed on individual devices. Our survey complements these works by addressing how to provide better quality data.
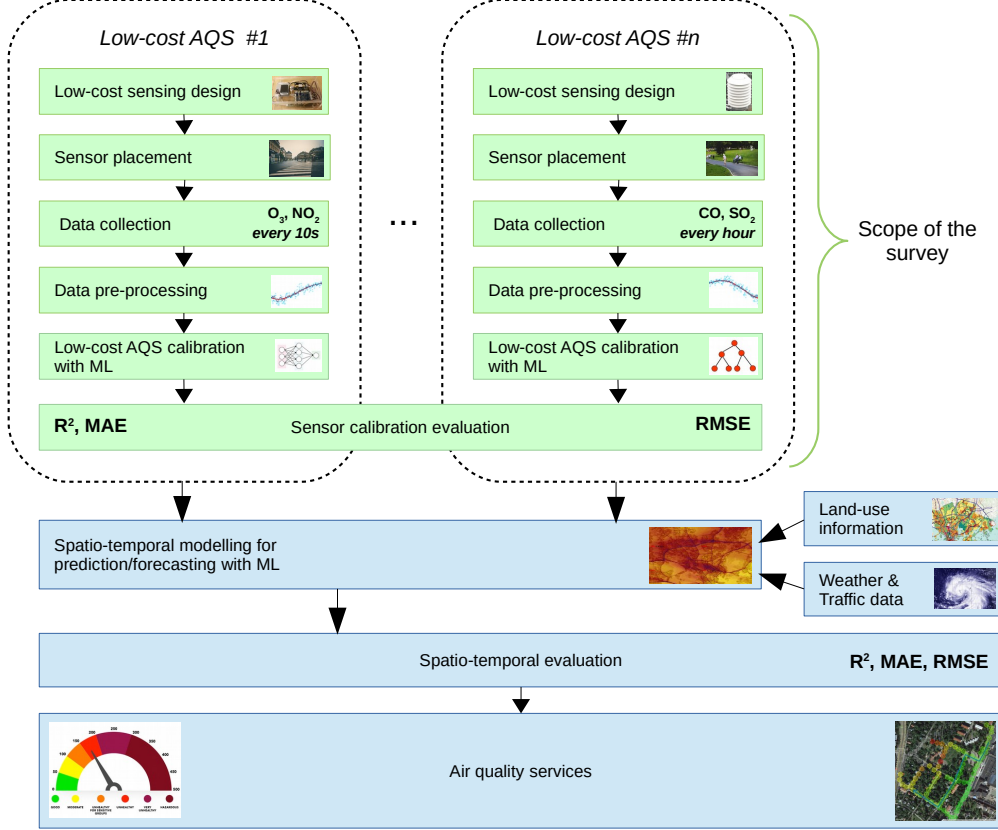
Figure 1: Calibration pipeline

2.2 Selection of Articles

Articles included (and excluded) in the survey were determined through a three-stage process. First, during an identification process, an iterative search strategy was used to determine potentially relevant articles to be included. We first identified an initial set of articles using searches with a small set of keywords on Google Scholar, IEEE Xplore, and ACM Digital Library. The following keywords were used: *air quality*, *sensors*, *low cost*, *machine learning*. The results were complemented with follow-up works from most prominent researchers. Once the initial set was formed, we searched for articles citing them or that were cited in this set. This process was continued until no more articles were found. In line with the interdisciplinary nature of the topic, searches were carried out separately within computer science and atmospheric sciences publications.

Second, during a preliminary screening phase, the articles found by the search engines were pruned by labeling them as potentially relevant or irrelevant by one of the researchers contributing to the survey. Any articles relating to the scope of the survey were preserved.

Third, once the articles were identified, we filtered them, by asking one of the researchers that contributed to the survey to read the articles and to present the main findings for the other authors. A majority decision was then made on whether the article was in the scope of the survey. In this survey, we aimed to focus on applying machine learning for the re-calibration of low-cost sensors, hence we selected only the papers that used machine learning in their re-calibration process. However, several papers specific on low-cost sensors implementations were evaluated to provide insights on the different sensing technologies in air pollution monitoring.

3 Low-Cost Air Quality Sensing Pipeline

Low-cost air quality sensing follows a typical machine learning pipeline for sensor data, illustrated in Figure 1. Within the pipeline, we can separate two types of operations: *per-device* operations that need to be performed separately for each sensor; and *integration* operations that combine data from multiple sensing units. In this survey, our focus is on

per-device operations, as several existing surveys have focused on application domains that are related to the integration operations [60, 47, 48, 61, 62]. Below we briefly give an overview of the different steps in the sensing pipeline.

3.1 Per-Device Operations

The low-cost air-quality sensing pipeline includes six operations that address issues related to individual devices: 1) the design of low-cost sensing units, 2) deciding where the devices are deployed, 3) the actual data collection, 4) pre-processing, 5) machine learning (ML)-based calibration and 6) the evaluation of that calibration. Applications such as prediction and real-time maps can then be built on the results of per-device data.

Low-cost Sensing Design: A sensing unit can be seen as a low-cost monitoring station integrating one to several sensors, each measuring a specific pollutant or environmental variable. In current air quality sensing research, the design of the sensing units is largely overlooked. Indeed, most of the research relies on off-the-shelf sensing units and focuses on other parts of the pipeline. However, the choice of sensor models and the overall design of the sensing unit can have a huge impact on how these sensors produce data. For example, as we discuss in section 4, some sensors need to be heated while others are sensitive to temperature changes. This can result in significant inaccuracies, particularly when sensors for several pollutants are integrated into the same sensing unit. Another concern is related to variations in sampling rates which can make it difficult to synchronize measurements from different sensors or to relate the measurements with real-world events. We discuss the properties of low-cost gas and particle matter sensors in section 4.

Sensor Placement: The sensors can be mobile, mounted onto vehicles, carried as personal sensors (i.e., wearables) or deployed to specific areas of cities. The benefit of mobile sensors is that their measurements can cover a large area. However, this can result in high sparsity, as many areas will only have a small set of measurements. Another challenge with mobile sensors is how to manage them and how to ensure they are operational. Instead of relying on mobile sensors, most real-world deployments rely on sensors placed in fixed positions for an extended time. Examples include the Barcelona Lighting Masterplan[5] and the Chicago Array of Things[6]. In current research, the deployment of sensors is largely driven by practical constraints, such as availability of power and availability of locations where the airflow to the sensor remains unobstructed. We briefly discuss sensor placement in section 4 and refer to the survey by Morawska et al. [36] for a more thorough overview of existing sensor deployments.

Data Collection: This step consists in collecting air pollution data, i.e., the concentration of selected air pollutants and environmental parameters. Pollutants require different collection densities, therefore, not every pollutant needs to be collected by all devices. For example, temperature and humidity have similar patterns within one region, whereas pollutants resulting from vehicles can fluctuate significantly even within a small region. Data collection is discussed in subsection 6.1.

Data Pre-processing: Examples of pre-processing operations include synchronization of different measurements and removal of erroneous measurements from periods where the device is compromised. For instance, the devices could have been operating in extreme conditions which are not supported by the sensors (e.g., high temperatures), the measurement units may be clogged, or power spikes can disrupt the functionality of a sensor. Other common pre-processing techniques include interpolation to achieve consistent sampling rate, and aggregation to reduce sampling rate, e.g., when low-cost sensors produce data at a higher rate than professional-grade reference stations. State-of-the-art low-cost air quality research typically considers measured pollutant concentrations and environmental variables such as temperature and (relative) humidity as features. Pre-processing is discussed in detail in subsection 6.1.

Calibration Based on Machine Learning: This step consists in the application of machine learning techniques to calibrate the measurements of the low-cost sensors and is the main focus of our survey. We critically compare existing machine learning solutions for the calibration of low-cost air quality sensors and identify the main advantages and disadvantages of the methods. Machine-learning-based calibration is discussed in detail in subsection 6.1.

Sensor Calibration Evaluation: The final step related to a single sensing unit is the performance evaluation of their calibration. We discuss this step in detail in section 7, with an emphasis on the selection of performance measures and test data length.

Integration Operations: Air quality data from a single sensor is limited in usefulness, as it captures pollutant concentrations only within a small region. In practice, we need to combine data from numerous sensing units to create high-quality spatiotemporal air quality information with fine granularity. This operation is conducted in the seventh step, which includes spatiotemporal modeling and fusion with additional sources of information, including but not limited

---

[5] http://ajuntament.barcelona.cat/ecologiaurbana/en
[6] https://arrayofthings.github.io

to land-use, weather and traffic data. The eighth step includes the performance evaluation of the models produced by the previous step. Finally, the ninth and last step is the production of air quality services built upon high-quality air pollution models. Such services include more advanced air quality index (AQI) [63] models and green path routing [64] to enhance the quality of life of citizens.

4 Low-cost air quality sensors

Low-cost solutions for air pollution monitoring typically consist of sensing units that package together multiple low-cost air quality sensing components. Besides the components responsible for monitoring pollutants, sensing units can incorporate other components such as power sources, processing units, local data storage, networking interfaces, and atmospheric sensors (e.g., temperature, relative humidity, and wind direction) [65]. Individual components usually cost between \$20 and \$100. However, a complete low-cost sensing unit typically costs upwards of \$500 [66].

Characteristics of components and the overall design of a sensing unit can have a significant impact on resulting measurements, and hence need to be carefully chosen to avoid negatively influencing other parts of the air quality monitoring pipeline. In this section, we survey the most commonly used sensor technologies employed by low-cost air quality sensing units. Before we delve into the details of the sensor technologies, we briefly discuss the two different types of pollutants that are considered in air quality research.

4.1 Types of Pollutants

Air quality sensing research typically categorizes pollutants as gaseous or particulate matter according to the composition of the pollutant [67]. Commonly considered gaseous pollutants include carbon monoxide (CO), carbon dioxide (CO_2), ozone (O_3), nitrogen oxides (NO_x) and sulfur oxides (SO_x). Carbon dioxide monitoring is restricted to indoor environments. For outdoor monitoring, the most common gases are those that belong to prominent air quality indexes, such as EPA in the USA or the Air Quality Index of China. Specifically, these indexes include the following gases: sulfur dioxide (SO_2), ozone (O_3), carbon monoxide (CO) and nitrogen dioxide (NO_2). Particulates (or aerosols), on the other hand, refer to tiny particles of solid or liquid compounds that are suspended in gases. The source of particulates can be natural (e.g., dust or sea spray) or caused by human activity (e.g., burning of fossil fuels, wood, etc., dust from roads and tires, and power plants).

Gas sensors can typically be tailored to support different gases by changing (parts of) the sensing materials or operating parameters of the sensing unit. Particulate matter sensors, in contrast, only monitor the extent of particulates in the air, without being able to identify their exact source or composition. However, particulate matter sensors can be categorized based on the size of the particles they can monitor, with fine ($PM_{2.5}$) and coarse-grained (PM_{10}) being the most common categories in low-cost air quality research and belonging to all major air quality indexes. With some sensor technologies, it is also possible to detect so-called ultra-fine particulates ($PM_{0.1}$). However, these mostly require expensive professional-grade instruments and thus are rarely considered in low-cost sensing research.

4.2 Gas Sensing

Within low-cost sensing units for gaseous pollutants we can identify four main types of sensing technologies: metal oxide semiconductor (MOS), electrochemical (EC), non-dispersive infrared (NDIR) sensors, and photo-ionisation detector (PID) sensors. In this section, we describe the key properties of these technologies. We discuss MOS and EC sensors in more detail as they are the least expensive and most widely used technologies in low-cost sensing units. NDIR and PID sensors have higher costs than MOS and EC sensors (\$500–\$5000) and hence they are rarely used in low-cost sensing units. We note that there are also other technologies for monitoring concentrations of gaseous air pollutants, but these have even higher costs making them unsuitable for low-cost sensing units. For example, gas chromatography (GC) sensors can also be used but they cost between \$15 000 and \$100 000.

Solid-state Metal Oxide Sensors

MOS sensors are a popular sensor technology for monitoring gas concentrations of several pollutants, such as non-methane hydrocarbons (NMHCs), carbon monoxide (CO), carbon dioxide (CO_2), nitrogen dioxide (NO_2), a combination of nitrogen oxide (NO) and NO_2 (NO_x) and ozone (O_3) [68, 40, 69, 70, 71, 72, 73, 25, 55]. MOS sensors consist of a heating element and a semiconducting metal oxide sensing element. The heater warms the surface of the sensing element up to 300–500°C, which is then able to detect gases through a chemical reaction occurring on its surface. This reaction causes a change in the electrical conductivity of the sensing element, which can be monitored using an external circuit to measure the detected gas level [51].

Advantages: MOS sensors are very low-cost and compact. Furthermore, these sensors have a high sensitivity and can even reach sub-parts per billion (ppb) sensitivity for some gases. MOS sensors also have a short response time, i.e., they can produce data at a high sampling frequency. Other advantages of MOS sensors include their long lifespan and resilience. Indeed, MOS sensors can operate in high temperature and humidity environments, making them well-suited for long-term deployments.

Disadvantages: MOS sensors have several disadvantages that can affect other parts of the processing pipeline. First, while they are resilient, their sensitivity is affected by atmospheric conditions and other gases. In terms of calibration, this implies that inputs from temperature and relative humidity sensors are needed while calibrating MOS sensors. Another disadvantage of MOS sensors is that they tend to have low accuracy and are subject to drift [49], which requires them to be re-calibrated often to maintain good quality data outputs [74, 75]. The extent of drift depends on humidity, with higher humidity increasing drift [51]. Finally, MOS sensors require access to a sufficiently large power source due to their need to power an electric heater.

MOS sensors are well-suited for low-cost sensing units due to their low cost and high resilience against environmental conditions. However, their high power requirement and sensitivity to environmental conditions are a concern. High power requirements make MOS sensors better suited for deployments where a fixed power supply is available than for deployments requiring battery power. As an example of the use of MOS sensors for air quality monitoring, Hasenfratz et al. [27] used a `MiCS-OZ-47` sensor, as part of a participatory air quality monitoring campaign. The sensor was linked to a separate battery pack and a smartphone was used to store and transmit measurements. The authors evaluated the impact of the sensor on the battery, which resulted in an estimated operation time of 50 hours when using a separate battery. Burgues et al. [76] state that power consumption of gas sensors can be reduced up to 90% by shutting down the heating elements for a certain time and then heating them in cycles rather than continuously. They showed that this negatively affects the accuracy, therefore, duty cycling cannot be used during the calibration phase. In terms of sensitivity, recently there have been advances in the synthesis of MOS gas sensing materials that can ensure sensitivity remains high even when air humidity increases [77]. However, these materials are not available in mass quantities and hence they are not used by current low-cost sensors.

Electrochemical Sensors

EC sensors are the second most popular technology for monitoring gas concentrations. EC sensors have been used to monitor CO, NO, NO_2, O_3 and sulphur dioxide (SO_2) [78, 72, 71, 73, 25, 79, 80, 39]. These sensors detect gases by oxidation-reduction reactions, employing electrodes separated by an electrolyte substance, such as mineral acid. The working electrode contacts both the electrolyte and the ambient air, which is monitored via a porous membrane. The reaction produces an electrical current between the electrodes, which can be easily measured from the outer pins of the sensor.

Advantages: Similarly to MOS sensors, EC sensors are inexpensive and have high sensitivity (in parts per billion (ppb) levels for some gases). Compared to MOS sensors, the main benefit of EC sensors is their lower power draw as they do not require powering an electric heater. Another advantage of EC sensors is that their sensitivity is less affected by temperature and humidity than with MOS sensors.

Disadvantages: The main drawback of EC sensors is that they have a short lifespan, only lasting between 6 months and a year, depending on the amount of pollution they are exposed to. EC sensors are also less resilient against weather conditions than MOS sensors. A combination of low humidity and high temperature is particularly problematic to EC sensors as it can dry out the sensor's electrolyte and break the sensor. Another drawback with EC sensors is that they are sensitive to weather conditions and other gases may interfere with the measurements, even if they are less sensitive than MOS sensors [53].

EC sensors are well-suited for low-cost deployments spanning a few months, as the sensors are inexpensive and their performance is less affected by temperature and humidity variations than MOS sensors. The accuracy of EC sensors tends to be good, as long as the weather conditions fall within their operational range. Examples of research relying on EC sensors include Nikzad et al. [78] and Wei et al. [81].

Non-dispersive Infrared

NDIR sensors consist of an infrared light source, an atmospheric sampling chamber, an optical filter, and a detector. When a gas passes through the chamber, the light emitted by the infrared source travels through it, and some frequencies get absorbed depending on the gas. The rest of the light hits the optical filter and the detector, which outputs the frequencies through an electrical current. NDIR sensors have been mostly used to detect CO_2 concentrations [70, 73, 25, 39]. However, they can be used to detect also other gases through changes in the wavelength of the light.

Table 2: Advantages and disadvantages of gas sensor technologies with respect to their use within low-cost sensor arrays

Type	Cost	Size	Lifespan	Sensitivity	Drift	Accuracy	Energy	Calibration	Response time	Known issues
MOS	Low	Small	Long	High	Yes	Low	High	Frequent	Fast	Cross-sensitivity to humidity and other gases. Sensitivity reduced in high temperature.
EC	Low	Small	Short	High	2%–15% per year	Good	Low	Reasonable	$\approx 120s$	Sensitive to temperature. Low humidity and high temperatures can cause the electrolyte of the sensor to dry out.
NDIR	High	Small	Long	High	$(0.4 \pm 0.4)\%$	High	Low	Frequent	$\approx 20s$	Spectral interference and high detection limit. Cross-sensitivity at least to water vapour.
PID	High	Small	Long	High	20% in weeks	High	low	Frequent	Fast $\approx 1s$	High sensitivity to high humidity levels or water vapour.

Advantages: They are simple and require little power, and small units also are available. They have a long lifespan (not degraded by exposure to gases) and require only little maintenance.

Disadvantages: They have high detection limits, i.e., cannot measure small pollutant concentrations, and they are susceptible to spectral interference from different gases as well as water [53]. NDIR sensors are also subject to drift [52] and they cost considerably more than MOS or EC sensors (up to a 10-fold increase in price).

Even considering the disadvantages, the long lifespan of NDIR sensors makes them a good choice for long-term deployments in dry areas. The high cost and high detection limit mean that they are better suited for sparser deployments than what most low-cost sensing aims at accomplishing.

Photo-ionisation Detectors

PID sensors operate by illuminating compounds using high energy UV photons. The process results in compounds becoming ionized as they absorb the UV photons and results in an electrical current that can be captured by a detector inside the sensor. The greater the concentration of the measured component in the compound, the more ions are produced, and the greater the current.

Advantages: PIDs are very sensitive and have a short response time. They have a small size, a small weight and low energy requirements [56, 82].

Disadvantages:
> PIDs affect all molecules tss ionization energy than the UV light affecting them, which means PIDs are not specific to a particular pollutant [56]. PIDs are sensitive to high humidity levels or water vapor. They are not very low-cost (between \$500 and \$5000), even if affordable compared to high-end monitoring stations [54]. PIDs are also subject to drift and need to be re-calibrated often (once per month).

The ability of PID sensors to analyze samples of low concentration in ambient temperatures and pressures makes them suitable for testing in a wide range of environments. This has made PID sensors particularly well suited for the analysis of small particles and gases in controlled small-scale experiments. However, the sensitivity to water and the relatively high cost make PID sensors poorly suited for dense long-term deployments.

Spinelle et al. [73, 25] studied the performance of MOS and EC sensors for detecting O_3, NO, NO_2 and CO. The authors concluded that no significant differences in sensor outputs could be observed between the two sensor technologies. In total, they evaluated 15 sensor models for the 5 gases. However, their study included only five months of data, which is within the range of the lifespan of EC sensors, therefore EC sensor degradation effects were not considered within this study.

Sensor mobility is discussed in numerous studies that we review in this survey [83, 84, 27, 85, 38, 55]. For example, Hasenfratz et al. [27] simulated mobility by carrying out experiments in a room with constant O_3 concentration and artificial wind, generated using a table fan. The authors found that, when O_3 concentration is low, the wind does not affect the measurements much. However, when O_3 concentration is high, wind effects produce a measurement offset. Therefore, they recommend shielding the sensor from the wind when moving relatively fast, for example when riding a bicycle.

Generally, the choice between sensor technologies depends on the characteristics of the deployment. MOS and EC sensors are cheapest and thus best suited for large-scale deployments, whereas PID and NDIR sensors are better suited to sparser deployments. In terms of accuracy, MOS and EC sensors have comparable performance, with EC sensors

being more energy-efficient but MOS sensors being more resilient. In particular, EC sensors do not require an electric heater but they can break in conditions with high temperatures and low humidity. MOS sensors also are more durable, being able to operate for years, whereas EC sensors require maintenance or even replacement every half year or so. In terms of calibration, both sensors are somewhat sensitive to weather conditions and concentrations of other gases, suggesting that these variables need to be incorporated as part of the calibration model. In terms of sensor design, MOS sensors can cause cross-interferences to other sensors by heating the air inside the sensing unit. Thus, the sensing unit and the sensor cycles need to be carefully designed to minimize the risk of such effects.

4.3 Particulate Matter Sensors

Particulate matter sensors monitor particle density by assessing changes in the properties of air passing through the sensor unit. Unlike gas sensors that can be tailored to detect different pollutants, particulate matter sensors cannot identify the exact composition of pollutants. However, they can be adapted to recognize particles of different sizes.

Existing low-cost particulate matter sensing technology predominantly is based on Diffusion size classifiers (DiSCs) or light-scattering particle sensorss (LSPs). The former operates by charging air passing through the sensor and estimating particle density from the total electricity charge after applying different filtering operations on the charged air. The latter operates by using light scattering to estimate the density of the particles. Traditional laboratory-grade instruments for particulate matter sensing are based on similar principles, but use additional components to improve detection accuracy. For example, optical particle counter (OPC) are high-quality variants of LSPs whereas condensation particle counters (CPCs) use alcohol or water vapor to change the physical properties of particulates before passing them through an LSP sensor [86, 87, 88, 89]. However, these sensors are typically bulky and more expensive than basic DiSC and LSPs, and hence rarely used in low-cost air quality sensing.

At the top end of the scale, Tapered Element Oscillating Microbalance (TEOM) sensors use changes in the oscillation frequency of a vibrating glass tube, and beta attenuation monitors (BAM) sensors use absorption of beta radiation for estimating particle density. TEOM and BAM sensors are rather expensive, costing over $20 000 [88]. Another high-end option is scanning mobility particle sizer (SMPS), which estimates the size and concentration of particles using electrical mobility sizing to monodisperse the output, which is then monitored with a CPC sensor. In the following subsections, we describe DiSC and LSPs in more detail since these technologies are the most affordable and most widely used in low-cost air quality sensing research.

Diffusion Size Classifiers

DiSCs are used to detect particles by applying electrical signals to identify physical changes on the sensing surface. They contain an air inlet, a corona charger, an induction stage, a diffusion stage, and a backup filter. When air enters the sensor, particles go through a diffusion charger that produces ions using a corona wire. A small fraction of these ions attach to particles in the air, and the remaining ions are captured using an ion trap that is placed between the diffusion charger and the induction stage. The particles then pass through the induction stage where they produce a small electrical current that is proportional to their concentration. After leaving the induction stage, the particles arrive at the diffusion stage, where they are precipitated to produce a small electrical current, which is proportional to their concentration. Since the particles also have an induction effect in the diffusion stage, the current measured in the induction stage is subtracted from the current measured in the diffusion stage to compensate for the induction effect. Larger particles that are not precipitated by the diffusion stage eventually reach the backup filter which produces an electrical current proportional to their concentration [90, 91]. This way DiSC sensors can separate between different particle sizes.

Advantages: The sensitivity of DiSC sensors is extremely high, and they have low power consumption. Manufacturing costs are low when the sensors are developed in large bulks.

Disadvantages: Manufacturing costs of DiSC have a high upfront setup cost as they require clean rooms and other special facilities. Therefore, the production and assembly of low quantities of sensors have a high unit cost ($\approx$ $10 000). Also testing equipment for assessing the quality and performance of DiSC sensors can be expensive. Another problem with DiSC sensors is that the sensing area can become unclean, making it necessary to clean it frequently. Research on how to automatize the cleaning process, e.g., using oscillation and electrostatics is actively pursued [92, 91].

Light-Scattering Particle Sensors

LSPs are small, low-cost sensing units that are widely used to detect particulate matter [83, 85, 93, 94, 95]. They are composed of an air inlet, a light sensor, and a light source, usually infrared or laser. When air enters the sensor, the light

Table 3: Advantages and disadvantages of PM sensor technologies with respect to their use within low-cost sensor arrays

Type	Cost	Size	Lifespan	Sensitivity	Drift	Accuracy	Energy	Calibration	Response time	Known issues
DiSC	High	Small	Good	Good	Yes	Good	High	Yearly (Factory), Hourly (Software)	1s	The instrument can produce wrong results if the incoming aerosol is highly positively charged. Cannot distinguish between narrow and broad particle size distributions [90].
LSP	Ultra-low	Small	Good	Poor	None	Low	Low	Frequent	$\approx 30s$	Mixes all particle sizes, variation of air influx reduces accuracy.

source is focused on a sensing point. An infrared LED is positioned in a forward angle with respect to a photodiode. Particles passing through the light beam scatter light, which generates a measurable signal in the sensor circuitry. The scattered light is focused onto the photodiode by a lens. The sensors may have a light scattering focusing lens and a focusing lens also for the infrared light source. The resolution at which different particle sizes can be detected depends on the configuration of these lenses. Finally, the sensor produces a signal that can be measured to estimate the number of particles in the air [96].

Advantages: LSPs are small and very low-cost. The sensors are mostly very low-power [94].

Disadvantages: Light-scattering based instruments fail to detect very small particles [90]. The sensor readings are also impacted by temperature and relative humidity, which means both temperature and humidity need to be measured. More expensive LSPs can use multi-angular light scattering to reduce the impact of environmental variables [88].

Table 3 summarizes the advantages and disadvantages of the two particulate matter sensing technologies used in low-cost air quality monitoring research. Similarly to gas sensors, the optimal choice of sensing technology depends on the context of the deployment. DiSC sensors have high sensitivity and they are mostly unaffected by weather or other environmental conditions, but they suffer from the need for regular maintenance to ensure the detection surface remains sufficiently clean. LSPs, on the other hand, are susceptible to weather changes but require less maintenance, making them better suited for longer-term deployments. While other technologies for particulate matter sensing exist, currently they are not well suited for low-cost deployments due to the higher cost of sensing equipment and larger size of the sensing units.

Particulate matter sensors generally require information about weather conditions regardless of whether the underlying sensor technology is sensitive to weather or not. This is because weather affects the concentration of particles in the air, and thus the sensor measurements. High winds can disperse particulate concentration, whereas high humidity causes particulates to cluster together, increasing concentration. The effect of temperature, however, is less well understood. Zheng et al. [97] found higher temperatures to lower particulate concentrations when humidity is low, and lower temperature to decrease particle concentrations when humidity is high. Wang et al. [98] studied the effects of temperature and humidity on low-cost PM sensors and found relative humidity to affect the accuracy of sensor technology. For example, as water in the air absorbs infrared radiation, humidity can result in LSPs overestimating particle concentrations as light intensity is reduced. The temperature was not found to directly impact the sensor technology, even if it affects the concentration of particles.

Besides weather, the concentration of particulate matter is affected by the extent of human activity within the area being monitored. The higher the traffic density and the lower the fuel efficiency of the vehicles, the higher the concentration of particulates will be. Note that the density is not solely a result of fuel burning as also tire and road wear produce particulates. In terms of calibration, this implies that the context of the deployment needs to be taken into consideration as locations close to intersections are likely to have an increased particulate matter concentration than other nearby areas [97]. Several research initiatives have explored the possibility of mounting low-cost sensors on vehicles [83, 84, 99]. When the vehicles are moving, the input air flux is constantly changing, which can affect sensor accuracy. In particular, the more air enters into the sensor, the more pollutants will be detected, even if the concentration of particulates would remain constant. Besides vehicles, wind speed can trigger a similar effect. To ensure that an accurate calibration model can be constructed, a possible solution is to include a GPS or another sensor to estimate the velocity of the sensor unit at the time of measurement and to use this information to compensate for the density of pollutants detected by the sensor [84].

5   Data Collection and Pre-Processing

Low-cost outdoor air quality monitoring commonly relies on sensors that have been deployed as part of the urban infrastructure [36]. The most common approaches consist in using fixed infrastructures, such as street lights, or

sensors mounted onto specific types of vehicles, such as trams [29], garbage trucks [100], or even Google Street View vehicles [19]. We next briefly describe the characteristics of the data collected by low-cost sensor deployments and typical pre-processing operations performed on the measurements.

5.1 Measurements

Measurements from low-cost air quality sensors can be interpreted as time-series data consisting of values of different pollutants and environmental variables. The integration of several pollutants is necessary to capture the effects of cross-sensitivities [31] whereas environmental variables are critical for accounting for differences in error as environmental conditions change. As discussed in Section 4.1, the most common pollutants to consider are gaseous pollutants and particulate matter that are part of prominent air quality indexes. In terms of environmental variables, temperature and relative humidity are the most common variables that need to be considered. Wind speed is another variable that can influence pollutant concentrations. However, wind speed is often difficult to measure with low-cost sensors as it requires an unobstructed air intake, whereas other environmental variables can be more accurately and reliably measured with low-cost sensors.

Reference Measurements Calibrating low-cost sensors with machine learning requires access to high-quality reference measurements that can be considered as ground truth.

The most common choice is to deploy low-cost sensors near a high-quality atmospheric station, and use the measurement of the station as ground truth [40, 79, 73, 25]. Another possibility is to use a high-quality mobile measurement laboratory deployed near the low-cost sensors [39]. Generally, the closer the low-cost sensor is to the reference station, the better the appropriate calibration relationship can be established. In cases in which the ground truth is needed in multiple different locations, for example when calibrating mobile sensors, other approaches must be used [42]. For example, public high-quality pollution data from official authorities can be used [27].

Temporal resolution The temporal resolution of measurements also influences the calibration process. The resolution is governed by the sampling frequency of the sensors, which varies across different sensor technologies. For example, as MOS sensors require heating, they have a slower sampling rate than sensors that can operate continuously. In the studies surveyed for this article, the sampling rate in most of the studies is between 5 seconds and 20 seconds [38, 72, 39, 40]. However, there are also studies with sampling intervals as high as 1 hour [58] or as low as 10 milliseconds [73, 25].

5.2 Characteristics of Air Quality Measurements

Low-cost sensor measurements have some characteristics that need to be accounted for while designing calibration solutions. Below we briefly discuss the most important ones.

Autocorrelation. Air pollutant concentrations are known to have a strong spatiotemporal correlation with the weather. Furthermore, seasonal patterns also have a significant influence on them [101, 102, 103]. This means that the used calibration techniques need to be capable of dealing with autocorrelation. The data that is used to test the generality of the model should also be sufficiently long-term to ensure the results are not overfitting to short-term correlations.

Cross-sensitivities. Measurements provided by low-cost sensors suffer from cross-sensitivities between pollutants [31]. Measurements can also be affected by temperature, humidity and wind direction [32]. In terms of calibration models, this means that the used techniques cannot assume the variables to be independent, but instead, they need to consider the values of environmental variables, and potentially also the values of other pollutants, as input.

Concept drift. Low-cost air pollutant measurements are vulnerable to *drift* whereby the relationship between environmental variables and pollutants varies over time [43]. For example, an analysis of metal oxide sensors (MOS; see Section 4) has demonstrated a drift in measurements higher than 200% [104]. The most common reason for drift is wear. For example, metal oxide sensors are vulnerable to oxidation which alters the conductivity of the sensing element and results in drift [105] whereas light-scattering particle sensors are vulnerable to deposits forming on the lens of the optical sensor [106]. This means that the calibration function needs to be re-trained periodically to ensure it accounts for changes in the properties of the sensors. The frequency of re-calibration depends on the characteristics of the deployment. For example, light scattering particle sensor re-calibration frequency depends on the extent of pollutants. The higher the amount of pollutants, the more often re-calibration is required.

Height differences. As discussed, the most common placement of low-cost sensors is near ground level and without isolating them from the urban infrastructure. Professional-grade measurement stations, on the other hand, are typically at least partially isolated from the urban infrastructure, and they have different air intakes located in different parts of the sensor. For example, in Finland, reference stations are either in dedicated containers that can be near the ground

11

or as part of separate measurement towers that are jointly responsible for weather and pollution measurements [17]. Pollutant concentrations can vary significantly also with elevation. For example, seasonality influences the elevation of the atmospheric mixing layer [107], which in turn affects the extent of pollutants that can be captured [108].

5.3 Preprocessing

The measurements from low-cost sensors typically require preprocessing before they can be used to capture a calibration model. Below we briefly describe the most common operations

Resampling. Before training a calibration model, measurements need to be re-sampled to a suitable temporal resolution. A higher resolution implies more samples and a longer model training time, but it can improve the robustness of the model. A common choice is to use a one-minute resolution [73, 25, 79, 84]. Some studies use a much coarser resolution, such as one hour [40, 58], which is the standard resolution for deriving air quality index values [109]. Saukh et al. [72] use a resolution in tens of seconds, while Hasenfratz et al. [27] and Maag et al. [38] use five seconds.

Synchronization. To minimize cross-sensitivities, sensor sampling intervals need to be interleaved. For example, as MOS sensors require heating, they can influence measurements for temperature or other pollutants unless the heating period is sufficiently distant from the sampling period of other sensors. To account for differences in sampling times, the measurements need to be synchronized during the modeling phase. This can be accomplished using aggregation, e.g., using the mean value over a given data window, or interpolating values of different sensor units to have consistent timestamps. In most cases, using a simple linear interpolation is sufficient, especially if the synchronization window is short. However, in some cases more advanced interpolation techniques, such as nearest-neighbor interpolation, complex spline interpolation, or even Gaussian Processes can be used.

Smoothing and Filtering. Pollutant measurements occasionally contain outliers that can significantly decrease the performance of the calibration model, unless the outliers are accounted for. There are several reasons for outliers. For example, sudden wind gusts can result in abnormal measurements, or the air intake of the sensing unit may get temporally obstructed. Common ways to mitigate these issues are to use *smoothing* or *filtering*. In the former case, the measurements are fitted to a model that ensures temporal consistency, whereas in the latter case, values appearing as abnormal or otherwise erroneous are removed. As an example of the former, Cheng et al. [83] smooth the sensor data through a signal reconstruction technique based on a bi-criterion problem with a quadratic smoothing function. As an example of the latter, Hasenfratz et al. [99] used a three-phased filtering process. First, each sensor computes its null-offset and uses it to calibrate the offset of the measured particle concentration. Second, values of internal status variables of each sensing unit are checked for potential malfunctions, and measurements from erroneous periods are removed. Third, measurements with poor location accuracy are removed from consideration. Another example of filtering is proposed by Gao et al. [84] who also use GPS data to filter measurements. Smoothing and filtering are popular techniques for preprocessing and as such are likely to be used in most studies. However, most studies we surveyed for this article do not indicate which kind of preprocessing has been applied [73, 25, 110, 39].

6 Machine Learning Calibration of Low-Cost Air Quality Sensors

Low-cost sensors increasingly rely on machine-learning-based calibration pipelines to improve the accuracy of sensor measurements. Previous surveys have shown that domain characteristics need to be taken into account when building such calibration pipelines [44, 45, 46]. In the following we survey existing calibration pipelines for low-cost air quality sensor data and discuss how they the address requirements stemming from the specifics of air quality measurements. Generally, these approaches learn the relationship between a specific pollutant, as given by the low-cost sensor, and a ground truth value obtained from a reference station. We first discuss issues that affect the choice of machine learning algorithms, after which we survey machine learning algorithms used in previous studies. The different models and how they have been applied are summarized in Table 4.

Figure 2 presents a high-level overview of the (continuous) machine learning calibration process. A high-quality station provides reference data to train calibration models that can correct the measurements of a particular low-cost sensing unit. The performance and usefulness of the models are tested against new measurements, similarly collected from low-cost sensor units and a high-quality station. Once the error of a calibration model is sufficiently small, the corrected measurements can be used by many air quality sensing applications, such as pollution monitoring and prediction, and high definition pollution maps based on spatio-temporal models [48, 112, 113, 114].

Table 4: Summary of calibration studies.

ML model	Reference	Training data length	Test data length	Temporal resolution	Mobility	Exploits temporality	Online training
LR	Hasenfratz et al. [27]	NR	$\leq$ 2 mos.	5 s	X		
	Lin et al. [111]	NR	$\leq$ 2 mos.	5 m			
	Saukh et al. [72]	NR	$\leq$ 6 mos.	10 s, 20 s	X		
MLR	Maag et al. (2016) [38]	4 wks.	1.25 yrs.	5 s	X		
	Maag et al. (2018) [55]	2.1 days	2.7 wks.	NR	X		
	Liu et al. [85]	NR	36 h	1 m	X		
	Cordero et al. [110]	NR	$\leq$ 30 days	NR			
	Zimmerman et al. [39]	NR	1.4–15 wks.	15 m			
SVM	Cordero et al. [110]	NR	$\leq$ 30 days	NR			
RF	Borrego et al. [57, 58]	12.6 days	1.4 days	1 m/1 h			
	Cordero et al. [110]	NR	$\leq$ 1 mo.	NR			
	Zimmerman et al. [39]	5.6 days	1.4–15 wks.	15 m			
FFNN	DeVito et al. (2008) [68]	8 mos.	3 mos.	1 h			
	DeVito et al. (2009) [40]	2 wks.	7 mos.	1 h			
	Spinelle et al. [73, 25]	1 wk.	4.3 mos.	1 m			
	Esposito et al. [79]	1 wk.	3 wks.	1 m		X	
	Borrego et al. [57, 58]	12.6 days	1.4 days	1 m/1 h			
	Maag et al. [55]	2.1 days	2.7 wks.	NR	X		
	Cordero et al. [110]	NR	$\leq$ 30 days	NR			
FFNN + GP	Cheng et al. [83]	3.5 mos.	2 mos.	5 m	X		
	Gao et al. [84]	NR	NR	1 m	X		
NARX	Esposito et al. [79]	1 wk.	3 wks.	1 m		X	
TDNN	Esposito et al. [79]	1 wk.	3 wks.	1 m		X	

6.1 Issues for Calibration

Generality. Calibration models should be capable of operating under different conditions, including different geographic locations, across different seasons, and potentially also across variations in sensing units. In practice, achieving all these goals simultaneously is not feasible. A model might need to be periodically retrained to adjust for variations. Different models might be needed for different seasons, hardware, or even locations.

Distribution of Air Quality Data. Many common machine learning techniques, such as common regression models, have been designed for data that is independent and identically distributed (i.i.d.), or at least close to being. As discussed in the previous section, this rarely is the case for air quality data. This is because the measurements contain significant temporal correlations and the values of different pollutants and environmental variables are dependent on each other. This suggests that calibration models that do not make strong assumptions about the distribution of measurements are likely to perform and generalize better.
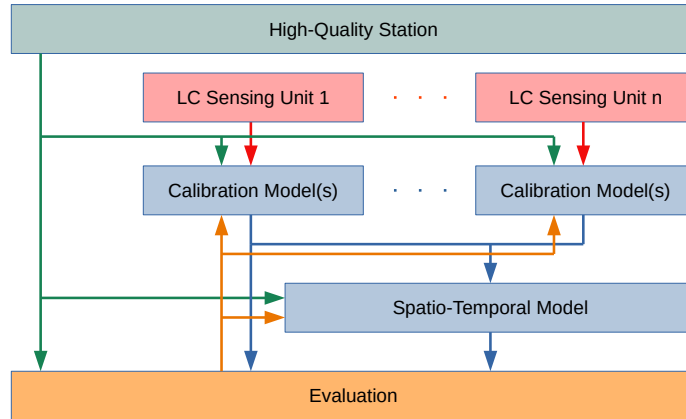

Figure 2: Data flow diagram of machine learning-based calibration.

Interpretability. The most powerful machine learning models often are black boxes that hide their internal logic from the user [115, 116]. This means that while the model may produce accurate estimates, it is difficult or impossible for the user to understand how the model works or why a particular estimate has been reached. If the calibrated measurements will be used to make decisions that have, e.g., economic, legal or safety consequences, then black-box approaches may be unacceptable. Another concern with black-box techniques is that errors in the calibration models they capture may be difficult to rectify. Indeed, a complex learning algorithm may, e.g., use unintentional correlations and features in the training data to obtain the estimates, which may lead to unexpected and counter-intuitive calibration behavior in real-life applications. For a thorough discussion on black-box models and issues with them, we refer to Guidotti et al. [116]. The alternative to black boxes is to use white boxes, models of which the performance can be easily understood and explained. In practice, however, the number of parameters influences the interpretability of models and most common white-box models become gray boxes that can be only partially explained.

Optimization criteria. Most common machine learning algorithms operate by minimizing an objective function that specifies a *loss* between the output of the machine learning model and the desired output. In the case of calibration, the loss function measures differences between the low-cost sensor and the reference station. Generally, the cost function needs to be chosen so that it represents the goal of the application. For example, if we wish to know whether the measured concentration is below or above a certain threshold, mean squared error would be a good choice as the cost function as it penalizes large errors more than small ones. On the other hand, if we wish to optimize the average-case performance of the algorithm, we could use mean absolute error instead. In some cases, we may be interested in other kinds of objective functions. For example, for detecting drastic short bursts in the concentration of a pollutant, the objective function can be mapped into classification error where the different classes represent the severity of the pollutant.

Amount of Data. Choosing the right machine learning model for a calibration problem depends on the available amount of air quality measurements. The amount of air quality measurements, in turn, is linked with the generality and complexity of the model. If the model is simple, it requires relatively few data points for training, but it may not fit the data well. This phenomenon is known as *underfitting*. Conversely, if the model is complex, it will approximate the function to predict the data better, but it will also need a larger training dataset to avoid *overfitting*, i.e., fitting the training dataset well but having poor performance on unseen data. Generally, the more complex the model, the longer the training dataset should be. Training data should also be sufficiently long to ensure that the model has enough information to learn how the cross-sensitivities between the sensors affect their response. A longer period of training data generally implies a longer training time, which results in less data being available for validating and testing the model.

The choice of test data is critical for ensuring the usefulness of the calibration model. Standard machine learning evaluation techniques, such as selecting a subset of all data as test data, or using cross-validation, are not suitable for air quality calibration due to the nature of the measurements. Indeed, these evaluation techniques can result in significant correlations between training and testing data, which would result in the calibration model overfitting on the temporal structure of the measurements. Optimal evaluation of air quality calibration models is currently an open issue as the data should cover a sufficiently long period, different pollutant concentrations, and different environmental conditions. From the studies surveyed for this article, it is difficult to estimate average lengths of training or testing periods since many studies do not report them exactly. In the studies that report the length, the length of training data spans from 2.1 days to 8 months, whereas the length of the test data spans from 36 hours to 1.25 years.

Computational Time and Complexity. Time requirements of machine learning techniques also have some influence on the choice of machine learning models for calibration. Most machine learning models are slow to train and fast to use while predicting the values of new measurements. The time requirements of the prediction phase should be sufficiently low so that the calibration model can be used to correct any new measurements from the low-cost sensing unit. In practice, any machine learning model can achieve this since most low-cost air quality sensors require sufficient air intake before they can take measurements. Indeed, the temporal resolution of measurements is usually in minutes (or once per minute) rather than in seconds. For prediction, a more critical concern is memory and storage complexity. Traditional machine learning techniques, such as linear models, random forests, and support vector regression, have a reasonably small model size, but emerging techniques, such as deep learning, often have large a model size, incorporating tens or even hundreds of thousands of parameters. Storing and loading such models on the sensing units may become a bottleneck, especially on sensing units that have been designed to operate for a long time. For model training, time complexity influences how often re-calibration can be performed, as well as the overall system architecture. Simpler models, such as linear models or support vector regression can be efficiently trained even on low-cost sensing units, but more complex models, such as artificial neural networks or deep models can be too computationally heavy for the sensing units. When the training cannot be performed on the sensing units, sufficient computing and networking infrastructure need to be available.

6.2 Linear Models

Linear regression (LR) is the simplest machine learning regression model and is based on the linear equation.

In the case of a single (input) feature, this model is usually referred to as univariate LR or just LR. In the case of more than one feature, this model is usually referred to as multivariate linear regression (MLR). LR models have widely been used as calibration methods for air quality monitoring [27, 111, 72], or as a baseline for comparing the calibration performances of more complex approaches [73, 25, 39]. LR has also been used as a pre-calibration method, with the output of the model fed to other methods [110]. As discussed earlier, low-cost sensors are vulnerable to cross-sensitivities and meteorological conditions, which renders simple LR models insufficient for most environments [73, 25, 39]. MLR has shown improvement in performance [73, 25, 72, 38, 110], because the model can learn cross-sensitivities between different pollutants and meteorological conditions.

When the relationship between features and the air pollutant being calibrated is not strictly linear, data transformation can be used to improve the accuracy of LR models. For instance, Liu et al. [85] argued that LR using a logarithmic function reacts better to the cross-sensitivity between $PM_{2.5}$ and wind interference. A variation of MLR called geometric mean regression (GMR) has been tested by Saukh et al. [72]. Their comparison with a conventional MLR model shows that GMR is less vulnerable to drift than conventional MLR models.

Conventional linear models assume input data to be independent and identically distributed. As this assumption rarely holds with air pollutants, linear models can only reach modest calibration performance. However, a major benefit of linear models is that they are easily interpretable since the model corresponds to a (hyper)plane fitted on data. Hence the parameters have an intuitive geometric interpretation and linear models can be considered as white-box models. Due to their simplicity, linear models require less training data than more complex models. Furthermore, linear models are very quick, both to train and to predict with. This is true especially in the air quality data scenario: the optimal weights that minimize the cost function can be directly computed, instead of being approximated through iterations, as the number of features is small.

Advantages: Simple to define, and trivial to find the optimum weights.

Disadvantages: The low-cost AQS calibration problem is too complex for this model [73, 25, 39]. LR cannot automatically find all the cross-sensitivities between the various pollutants, and in some cases, the function that models some relations is not linear. In such cases, the features need to be manually transformed to allow LR to fit them. Furthermore, LR and MLR models suffer from drift.

6.3 Support Vector Regression

support vector regression (SVR) is one of the most popular techniques for modeling *non-linear* relationships between input features and the output variable (i.e., calibrated air pollutant). The general idea in SVR is to map the input data to a high-dimensional space using a so-called kernel function on the measurements, and then find a hyperplane in the transformed space that minimizes the distance between the hyperplane and the data points.

In air quality calibration, SVR has mainly been used for comparison against other techniques. For example, Cordero et al. [110] compare SVR, random forest, and Artificial neural networks (ANNs) for calibrating NO_2 measurements. The authors report mixed results, with SVRs outperforming the other models in some cases but having worse performance in other cases.

SVR models are more complex than linear models and hence they require more training data. In terms of runtime, they are fast to compute, even if less efficient than linear models. SVRs, like linear models, assumes that the data is independent and identically distributed. The interpretability of SVR models depends on the kernel that is used with a linear kernel resulting in a white-box model but non-linear kernels effectively turning the model into a black-box that is difficult to interpret.

Advantages: Model training is defined as a convex optimization problem, for which there are efficient solutions to find the optimal parameters. SVR uses a kernel to transform the input data, enabling capturing nonlinear relationships between input features and calibrated pollutants. SVR incorporates a regularization parameter, which makes it possible to control under and overfitting. There are many efficient, mature, and easy-to-use SVR implementations.

Disadvantages: Sensitive to kernel parameters and assumes data to be independently and identically distributed. Poor interpretability for nonlinear kernels.

6.4 Decision Trees and Random Forests

Decision trees (DTs) are another common off-the-shelf machine learning technique. In DTs, every node of the tree has a conditional check, and every branch corresponds to an outcome of the check. While determining the value of a new measurement, we progress through the tree, starting from the root node and following the branches, until a leaf node is reached. The value of the leaf node is then used as the outcome of the calibration. Each check in the DT corresponds to a rule that can be used to subdivide the measurements into an increasingly smaller range of values. Thus, unlike linear models and SVRs, DTs do not fit a model or function to the data and hence they can support both linear and nonlinear relationships. To the best of our knowledge, DTs have found limited use in air quality calibration. They have only been used for calibrating CO, and even then only as a baseline for other methods [117].

When applied to complex problems, DTs can become highly complex and overfit on the training data. These issues can be mitigated using random forests (RFs) which combine multiple simple DTs into a single powerful model. RF is an example of the *bagging* technique, which operates by creating different subsets of the training data, learning separate models for each subset, and aggregating the outputs of all these models together while predicting unseen data. The main disadvantages of RF models are that they can be difficult to interpret and their performance is sensitive to the choice of parameter values. Indeed, while the outputs of individual DTs can be easily interpreted, understanding the joint effect of tens or even hundreds of DTs is much more complex. Similarly, the number of models to integrate influences overall runtime and performance. Having a small number of DTs as part of the RF model is efficient to learn, but easily underfits the data, whereas integrating a large number of trees increases the model size and training data. RFs models have been widely applied to the calibration of sensors for many pollutants, such as CO [39, 58], CO_2 [39], NO_2 and O_3 [39, 110, 58], NO, SO_2, $PM_{2.5}$ and PM_{10} [58]. Lin et al. [118] propose a hybrid model that combines LR and RF to simultaneously learn linear and non-linear relationships.

Advantages: Reduces overfitting by training different models on different artificial datasets generated from the original dataset. The training of the different models can be parallelized. DTs and RFs do not require choosing the function for non-linear problems, which is an advantage in our scenario.

Disadvantages: DTs, when used for regression, can have a very high depth unless properly regularized. The number of models, and therefore of generated datasets adds additional complexity to parameter the training correctly. The computational needs of the RFs are higher than a single DT. RFs can be considered gray-box or even black-box algorithms, depending on the number of models that are integrated.

6.5 Boosting Algorithms

Similarly to bagging, *boosting* algorithms operate by creating subsets of the training data and learning a different model on each subset. The difference to bagging is that boosting trains all the weak models sequentially, aggregating them into a single strong model, instead of running each model separately and aggregating their outputs. Boosting, when training a new weak model, also takes into account the success of the previously trained weak model, and weights the training data accordingly. Example of boosting algorithms are adaptive boosting (AdaBoost) [119], gradient boosting (GB) [120], and extreme gradient boosting (XGBoost) [121]. In the context of air quality monitoring, boosting algorithms have been used for predicting $PM_{2.5}$ [122].

Advantages: Reduces overfitting by training different models on different artificial datasets generated from the original dataset. No need to choose the function for non-linear relationships. When training the new weak learners to combine into the strong learner, the model takes into account the success of the previous weak learner and weighs data accordingly.

Disadvantages: Interpretability and complexity depend on the number of learners that are aggregated together, with a higher number resulting in better performance, but higher complexity and reduced interpretability. Risk of overfitting when the number of learners to aggregate grows large.

6.6 Artificial Neural Networks

ANNs are a popular machine learning technique for modeling time series data. ANNs have been initially motivated by the structure of the human brain [123] and are composed of a set of nodes called *neurons*, that are grouped into *layers*. The first layer of the model is called the input layer, and the last is referred to as the output layer. Intermediate layers between the input and output layers are referred to as hidden layers. Depending on the number of intermediate layers, the network is called either *shallow* or *deep*. For the calibration of low-cost air quality models, the most common approach is to use a shallow feedforward neural network (FFNN), with most models containing one or two hidden layers. This type of ANNs have been applied to the calibration of sensors for measuring NMHCs [68], PM2.5 [83, 84],

CO [40], CO_2 [55], NO [40, 79, 110], NO_x [79], NO_2 [40, 79, 110], and O_3 [55, 110]. Spinelle et al. [73, 25] use an ANN model based on FFNNs applied to the calibration of CO, CO_2, NO, NO_2 and O_3.

In an ANN, each neuron contains a function that is applied to the data that it receives as input, to produce an output value. Such function is called *activation function*. Common activation functions include the rectified linear unit (ReLU) function, the sigmoid function, and the hyperbolic tangent function. Like in LR, the activation functions consist of some weights that need to be set so that the final output of the model is as close as possible to the ground truth.

Just a few of the studies that we survey discuss the used activation function of the neurons. Some report using the hyperbolic tangent function, or a variation of it [68, 40]. Spinelle et al. [73, 25] use different activation functions in different layers in their ANN model. They also report trying the radial basis function (RBF) function but discarding it because it did not yield good results.

In terms of performance, there have been studies that compare ANNs against regression models [73, 25, 55, 110]. The conclusions from these studies have been mixed. Some studies report a better performance from ANNs [73, 25, 55], whereas some studies report the opposite [110]. This might be explained by the fact that the relationship between the pollutants and how they vary in time is a very complex function. Therefore, the distribution of pollutants and characteristics of environmental variables can influence whether LR and MLR models are sufficient for approximating the underlying relationships in data.

The complexity of the relationships that can be captured with ANNs depends on the structure of the network. Standard FFNN structures mostly have been designed for capturing a non-linear function between input and output. However, more complex structures that can incorporate additional considerations as part of the network, such as temporal structure or feature learning, have been proposed. As an example, recurrent neural networks (RNNs) are a class of ANNs designed for incorporating temporal structure and thus well-suited for modeling time series. In an RNN, input values or neuron outputs of the previous time step can influence the state of the ANN in the current time step. For the calibration of low-cost air quality measurements, Esposito et al. [79] tested two more complex RNN architectures: time delay neural network (TDNN) and nonlinear autoregressive exogenous model (NARX). They report that NARXs have a better performance than FFNNs, and TDNNs achieve the best performance. This might be because RNNs take into consideration previous time steps, therefore encoding input value changes over time.

It is also possible to build hybrid models by using layers of different architectures. For example, we can use an RNN layer to learn relationships between different points in the input data, an FFNN layer to learn another level of abstraction, and a final FFNN layer to constrain the size of the output. These hybrid models are called Deep learning (DL) models. The idea behind this is to create a model that can learn multiple levels of abstraction of the input data [124].

Advantages: Very flexible learner that can approximate any function, given enough layers and neurons. It can automatically learn the relationships between the features and learn multiple response values simultaneously, which is very useful for the low-cost AQS calibration problem.

Disadvantages: Its complexity makes it extremely expensive in computing resources to train, requiring dedicated hardware for doing so quickly. It also needs a huge amount of data for avoiding overfitting. ANNs are typically black-box algorithms.

6.7 Gaussian Processes

A Gaussian process (GP) is a model that combines multiple Gaussian random variables into a joint distribution to estimate the function that models data. It is a non-parametric approach, which means that there is no need to specify the number of parameters. However, similarly to SVR, GPs need a kernel function to be specified. In this case, the kernel function is the covariance function that specifies how the Gaussian random variables are shaped. Detailed information about GPs and how to train them can be found in Rasmussen et al. [125].

GPs do not make assumptions on the distribution of input data, and hence are well suited for calibrating low-cost air quality sensors. Another benefit of GPs is that the models can be seen as white boxes as it is possible to plot the probability distributions used to model data. GPs require a training dataset larger than linear models and support vector machines (SVMs) would, but generally smaller than more complex models, such as RFs and ANNs. GPs are lazy learners, meaning they do not need to be trained; instead they approximate the function of the training dataset while predicting. However, this means that the whole training dataset needs to be kept in memory, and every time that it needs to predict the target variable, it needs to compare it to the probability distribution of the features. This means that GPs has high memory complexity, which might be unpractical if the deployed AQS units do not communicate with a central infrastructure and need to run calibration in the field. GPs have been used in the context of low-cost AQS calibration by Cheng et al. [83] and Gao et al. [84] on top of FFNN to improve its performance. In both studies, GPs were able to improve calibration performance compared to standalone FFNN.

17

Advantages: Non-parametric model, no need to specify the number of parameters. Makes no assumptions on the distribution of data. White-box, possible to interpret results.

Disadvantages: High memory complexity, requires a kernel function and can be sensitive to parameters of the kernel function.

7 Measuring performance and comparing models

Machine-learning-based calibration models are only useful if they can consistently improve the accuracy of the sensor measurements produced by a low-cost sensing unit. To ensure this indeed is the case, calibration models need to be validated against measurements collected from high-cost reference stations. We next discuss validation methods and how they can be applied to low-cost calibration, and compare existing low-cost air quality calibration studies by selecting the most commonly used performance measures in them.

7.1 Performance Measures

The performance of calibration models is typically expressed through one or more performance measures, which are functions that characterize the dissimilarity between the output of the calibration model and the ground truth values obtained from a reference station. Existing studies have used differing performance measures, which makes it difficult to compare performance across studies. In the following, we briefly summarize the main performance measures.

Absolute Error Measures: Mean square error (MSE) is the standard error measure for assessing the performance of regression models. MSE is defined as:

$$\text{MSE} = \frac{1}{m} \sum_{i=1}^{m} (y_i - \hat{y}_i)^2 \tag{1}$$

where m is the number of samples, $\hat{y}_i$ is the predicted value and y_i is the actual value of a sample. MSE is useful to evaluate the performance during training and to define a cost function because of its simplicity. A variation of MSE is root-mean-square error (RMSE), defined as:

$$\text{RMSE} = \sqrt{\frac{1}{m} \sum_{i=1}^{m} (y_i - \hat{y}_i)^2} \tag{2}$$

Another absolute error metric is mean absolute error (MAE) which is defined as:

$$\text{MAE} = \frac{1}{m} \sum_{i=1}^{m} |y_i - \hat{y}_i| \tag{3}$$

Finally, mean bias error (MBE) is defined as:

$$\text{MBE} = \frac{1}{m} \sum_{i=1}^{m} (y_i - \hat{y}_i). \tag{4}$$

MSE and RMSE are very similar. MSE can be interpreted geometrically as the average fit of points to a regression model, whereas RMSE is the average distance of points from the regression model. RMSE and MSE weigh errors proportionally to their magnitude, whereas MAE weighs all errors equally. This makes RMSE and MSE more sensitive to outliers [126], suggesting that MAE is a better measure for measuring the average performance, whereas (R)MSE is useful for measuring a model's sensitivity to outliers. In practice, it is recommended to use both measures together as this provides complementary information on the model's performance [126]. MBE, on the other hand, measures whether the average error is positive or negative and can be used to determine whether the model underestimates or overestimates the pollutant values. Spinelle et al. [73, 25] divide RMSE and MBE by the standard deviation of the reference measurements and combine the resulting values into a target diagram. Target diagrams are useful for visualizing these two performance metrics and quickly comparing different models.

Relative Error Measures: The alternative to absolute measures is to rely upon a relative error measure which expresses the error proportionally to the true measurements. The most popular relative error measure is mean relative error (MRE) which is defined as:

$$\text{MRE} = \frac{1}{m} \sum_{i=1}^{m} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \tag{5}$$

Table 5: Comparison of common performance measures.

Method	Metric	Formula	Advantages	Disadvantages
MSE	No	$\frac{1}{m}\sum_{i=1}^{m}(y_i-\hat{y}_i)^2$	Simple measure, can be used as a cost function. Useful for measuring the model's sensitivity to outliers.	Tends to exaggerate errors, especially with noisy data. For very clean data it might overestimate the model performance.
RMSE	Yes	$\sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i-\hat{y}_i)^2}$	Same as MSE, but in the same dimension as the target values.	Same as MSE.
MAE	Yes	$\frac{1}{m}\sum_{i=1}^{m}|y_i-\hat{y}_i|$	Useful for measuring the "average" performance of a model.	Underestimates the outliers.
MBE	No	$\frac{1}{m}\sum_{i=1}^{m}(y_i-\hat{y}_i)$	Useful for measuring the bias, and to see to which value the average error tends.	It takes into account the bias only. It can't be used to evaluate the actual performance of a model.
MRE	Yes	$\frac{1}{m}\sum_{i=1}^{m}\left|\frac{y_i-\hat{y}_i}{y_i}\right|$	Useful for expressing the average error in proportion to the target values.	Tends to exaggerate the error for small values, and underestimate the error for big values.
MAPE	Yes	$\frac{1}{m}\sum_{i=1}^{m}\left|\frac{y_i-\hat{y}_i}{y_i}\right|\cdot 100\%$	Same as above.	Same as above.
R^2	No	$1-\frac{\text{SS}_{\text{res}}}{\text{SS}_{\text{tot}}}$	Useful for measuring how much the variance is accounted for by the model.	Same as MSE and RMSE.

A related measure is mean absolute percentage error (MAPE) which expresses MRE as a percentage:

$$\text{MAPE} = \frac{1}{m}\sum_{i=1}^{m}\left|\frac{y_i-\hat{y}_i}{y_i}\right|\cdot 100\% \tag{6}$$

MRE is useful for expressing how far estimated values are from the reference values, whereas MAPE is useful for characterizing performance when the same model is applied for multiple pollutants.

Coefficient of Determination: The coefficient of determination, or R^2, measures how much a variable influences another variable. For a calibration model, R^2 measures the percentage of variance that the model explains. To compute R^2 we need to compute two variability measures, namely the total sum of squares:

$$\text{SS}_{\text{tot}} = \frac{1}{m}\sum_{i=1}^{m}(y_i-\bar{y})^2 \tag{7}$$

and the sum of squares of residuals:

$$\text{SS}_{\text{res}} = \frac{1}{m}\sum_{i=1}^{m}(y_i-\hat{y}_i)^2 = \text{MSE}. \tag{8}$$

Here $\bar{y}$ is the mean of the target data. The R^2 is now given as:

$$R^2 = 1 - \frac{\text{SS}_{\text{res}}}{\text{SS}_{\text{tot}}} \tag{9}$$

R^2 can be useful in the low-cost AQS calibration scenario to assess how closely the distribution of the predicted values matches the distribution of the ground truth measurements.

Best Practices: MSE, RMSE and R^2 are closely related. This means that, if we rank some models according to one or the other measure separately, the ranking positions for the models in both rankings will be the same. The same holds for MRE and MAPE. However, measures that are not directly related, such as RMSE, MAE, and MRE, do not necessarily result in the same ranking. Hence, in practice, the recommended approach is to use multiple performance measures and take into account how they are affected by the properties of the data. Visual aids, such as target diagrams, should also be used so that different performance measures can be visually compared. A summary of the performance measures can be found in Table 5.

7.2 Evaluation Criteria for Low-Cost Deployments

To evaluate the suitability of air quality sensor calibration methods considered in this survey for long-term deployments of low-cost sensors, we rank them based on four evaluation criteria.

- The reliability of the method, based on the length of the test dataset
- The resolution, determined by the length of the smallest temporal step modeled
- The accuracy of the method as reported by the authors
- The sensor technology, emphasizing technologies are low-cost and that can operate independently for long periods

Each of the four criteria is valued on a scale from 0 to 5 (higher is better). A method can have an evaluation of 0 only when it could not be evaluated based on the information provided. Otherwise, works are evaluated from 1 to 5. The final ranking for the calibration methods depends on a weighted score based on the above evaluation criteria, emphasizing reliability. We introduce how to calculate the evaluation criteria below, followed by the final ranking score.

Reliability This score is based on the length of the test dataset. The minimum score is assigned to one-month long datasets or shorter, the maximum score is assigned to one-year long datasets or longer. Values in between are assigned using the equation:

$$\text{RLB} = L/12 \cdot 4 + 1 \tag{10}$$

where L is the length of the dataset in months.

Resolution This score is based on the temporal resolution of a model. The minimum score is assigned to resolutions of 1 hour or coarser, the maximum score is assigned to resolutions of 1 minute or finer. Values in between are assigned using the equation:

$$\text{RSL} = (1 - R/60) \cdot 4 + 1 \tag{11}$$

where R is the model temporal resolution in minutes.

Accuracy This score is based on the accuracy of the model. This is the most complex to compute since different studies use different similarity measures. It is defined as follows. If a study uses RMSE, or in alternative MAE, it is used as the base value of the performance of the models. Since Spinelle et al. [73, 25] do not provide an exact value, we use a conservative estimate of MAE obtained by analyzing the residual plots that they provide. Models of studies that do not use comparable similarity measures get a score of 0. We group the models by the type of pollutant they predict. In each group, the model with the best performance gets a score of 5, and the model with the worst performance gets a score of 1. The mean μ and the standard deviation σ^2 is then computed. The accuracy score of each model is computed on an equation based on the cumulative distribution function (CDF):

$$\text{ACC} = \left(1 - \Phi_{\mu,\sigma^2}(B)\right) \cdot 4 + 1 \tag{12}$$

where B is the base value of the performance of a model.

Sensor technology This score is based on the sensor technologies used to produce the input values of a model. The score is determined according to sensor technology taking into account the typical cost, reliability, and lifespan of such technology. ECs get a score of 1; DiSCs get a score of 2; NDIRs, PIDs and OPCs get a score of 3; LSPs get a score of 4; and MOSs get a score of 5. The sensor technology (ST) score of each model is computed as the mean of the scores of the sensors, weighted on the number of sensors used for each technology.

Final score The final score is computed using the following equation:

$$\text{RLB}/5 \cdot \frac{\text{RSL} + 2 \cdot \text{ACC} + \text{ST}}{20} \cdot 4 + 1 \tag{13}$$

where RLB is the reliability score, RSL is the resolution score, ACC is the accuracy score, and ST is the sensor technology score. As we can see in the equation, much importance is given to RLB. This is because, as we previously discussed, low-cost AQSs are heavily influenced by seasonal phenomenons, and a dataset of at least a year is required to capture all seasonal phenomenons.

7.3 Comparing Studies

The different studies surveyed for this article are scored and summarized in Table 6. Below we separately discuss the studies according to each of the four criteria.

Reliability The study by Maag et al. [38] is the only one receiving a full mark on the Reliability score. This is because they use a dataset that spans more than a year to test the data. The next highest is the study by De Vito et al. (2009) [40], which uses a test dataset spanning about 7 months. Hence, the results show that most studies use relatively short test datasets, which are unlikely to capture the full extent of seasonal variations.

Table 6: Comparison of the different machine learning approaches for low-cost AQSs calibration, including performance measures and our evaluation criteria. For each pollutant, rows are ranked using our evaluation criteria-based final ranking in descending order.

Target	Method	Reference	CO	CO2	NO	NOx	NO2	O3	SO2	VOCs	PM2.5	PM10	T	RH	AH	WS	RMSE	MAE	MAPE	R²	Reliability	Resolution	Accuracy	Technology	Final Score
CO	MLR	[38]	x				x	x					x	x			0.048 ppm	NR	NR	NR	5.00	5.00	3.90	2.33	4.03
	RF	[39]	x	x			x	x					x	x			NR	7.9 ppb	NR	0.99	2.17	4.00	5.00	2.50	2.43
	FFNN	[25]	x	x			x								x		NR	0.05–0.12	NR	0.367	2.43	5.00	3.38	1.00	2.24
	MLR	[39]	x	x									x	x			NR	39 ppb	NR	0.94	2.17	4.00	4.01	1.00	2.13
	FFNN	[40]	x			x				x							NR	0.31 ppm	27.00 %	NR	3.33	1.00	1.00	5.00	2.07
	RF	[58]	x	x	x		x	x	x			x	x	x			NR	0.07 ppm	280.00 %	0.88	1.00	1.00	3.60	2.57	1.43
	FFNN	[58]	x	x	x		x	x	x			x	x	x			NR	0.09 ppm	34.00 %	0.51	1.00	1.00	3.30	2.57	1.41
	LR	[72]	x										x				0.23 ppm	NR	NR	NR	1.00	5.00	1.50	1.00	1.36
CO2	FFNN	[25]	x	x			x						x	x			NR	2–18	NR	0.787	2.43	5.00	3.89	3.50	2.58
	RF	[39]	x	x			x						x	x			NR	1.7 ppb	NR	0.99	2.17	4.00	5.00	2.50	2.43
	MLR	[39]		x				x					x	x			NR	19 ppb	NR	0.49	2.17	4.00	3.65	3.00	2.24
	FFNN	[55]		x				x		x							NR	44 ppb	NR	0.94	1.21	0.00	2.94	5.00	1.53
	MLR	[55]						x		x							NR	135 ppb	NR	0.49	1.21	0.00	1.00	5.00	1.34
NO	FFNN	[25]	x	x	x		x	x				x	x	x			NR	0.1–0.5	NR	0.208	2.43	5.00	5.00	3.00	2.75
	FFNN	[58]	x	x	x		x	x	x			x	x	x			NR	2.39 ppb	84.00 %	0.39	1.00	1.00	3.07	2.57	1.39
	RF	[58]	x	x	x		x	x	x			x	x	x			NR	4.76 ppb	127.00 %	0.74	1.00	1.00	1.00	2.57	1.22
NO2	MLR	[38]	x				x	x					x	x			5.13 ppb	NR	NR	NR	5.00	5.00	2.22	2.33	3.35
	RF	[39]	x	x			x	x					x	x			NR	0.5 ppb	NR	0.99	2.17	4.00	5.00	2.50	2.43
	FFNN	[73]	x			x	x	x					x	x	x		NR	2–4.5 ppb	NR	0.596	2.43	5.00	3.30	3.00	2.42
	FFNN	[40]	x	x			x	x		x							NR	10.1 ppb	22.00 %	NR	3.33	1.00	1.00	5.00	2.07
	MLR	[39]	x				x	x					x	x			NR	4.6 ppb	NR	0.59	2.17	4.00	2.51	1.00	1.87
	TDNN	[79]			x		x	x					x	x			NR	1.27 ppb	22.00 %	NR	1.23	5.00	4.29	1.00	1.72
	NARX	[79]			x		x	x					x	x			NR	1.30 ppb	21.00 %	NR	1.23	5.00	4.28	1.00	1.72
	FFNN	[79]			x		x	x					x	x			NR	1.50 ppb	25.00 %	NR	1.23	5.00	4.20	1.00	1.71
	RF	[58]	x	x	x		x	x	x			x	x	x			NR	1.97 ppb	17.00 %	0.89	1.00	1.00	3.99	2.57	1.46
	FFNN	[58]	x	x	x		x	x	x			x	x	x			NR	2.00 ppb	25.00 %	0.81	1.00	1.00	3.98	2.57	1.46
	MLR	[110]			x		x	x					x	x			2.92–3.88 ppb	2.34–3.14 ppb	12.3–35.4%	0.81–0.93	0.00	1.00	2.93	1.00	1.27
	RF	[110]			x		x	x					x	x			2.29–4.73 ppb	1.86–3.35 ppb	16.0–32.9%	0.85–0.95	0.00	1.00	2.43	1.00	1.23
	SVM	[110]			x		x	x					x	x			2.87–4.94 ppb	2.07–4.36 ppb	19.8–34.3%	0.79–0.95	0.00	1.00	2.32	1.00	1.23
	FFNN	[110]			x		x	x					x	x			3.24–7.44 ppb	2.76–6.22 ppb	20.2–93.4%	0.62–0.88	0.00	1.00	1.34	1.00	1.15
	LR	[111]					x			x			x				NR	NR	NR	NR	1.00	4.67	0.00	1.00	1.23
NOx	TDNN	[79]			x		x	x					x	x			NR	1.37 ppb	20.00 %	NR	1.23	5.00	5.00	1.00	1.79
	FFNN	[79]			x		x	x					x	x			NR	1.95 ppb	29.00 %	NR	1.23	5.00	1.00	1.00	1.39
O3	MLR	[38]	x				x	x					x	x			2.8 ppb	NR	NR	NR	5.00	5.00	3.82	2.33	4.00
	FFNN	[73]	x			x	x	x					x	x	x		NR	1–4.5 ppb	NR	0.915	2.43	5.00	3.84	2.33	2.46
	RF	[39]	x	x			x	x					x	x			NR	0.7 ppb	NR	0.99	2.17	4.00	5.00	2.50	2.43
	MLR	[39]	x				x	x					x	x			NR	5.1 ppb	NR	0.81	2.17	4.00	2.93	5.00	2.29
	LR	[27]						x					x	x			NR	1.46 ppb	NR	NR	1.00	5.00	4.24	5.00	1.74
	FFNN	[55]		x				x		x							NR	3.5 ppb	NR	0.91	1.21	0.00	3.57	5.00	1.59
	TDNN	[79]			x		x	x					x	x			NR	7.45 ppb	42.00 %	NR	1.23	5.00	2.05	1.00	1.50
	RF	[58]	x	x	x		x	x	x			x	x	x			NR	1.32	NR	0.97	1.00	5.00	4.28	2.57	1.49
	FFNN	[79]			x		x	x					x	x			NR	7.90 ppb	70.00 %	NR	1.23	1.00	1.91	1.00	1.48
	LR	[72]	x					x					x				12.9 ppb	NR	NR	NR	1.00	5.00	1.00	5.00	1.48
	FFNN	[58]	x	x	x		x	x	x			x	x	x			NR	2.60 ppb	18.00 %	0.86	1.00	1.00	3.89	2.57	1.45
	LR	[111]					x			x			x				NR	NR	NR	0.92	1.00	4.67	0.00	5.00	1.39
	MLR	[55]						x		x			x				NR	10.7 ppb	NR	0.16	1.21	0.00	1.29	5.00	1.37
PM2.5	FFNN + GP	[83]									x		x	x		x	96.69 ug/m³	NR	NR	NR	1.67	5.00	5.00	4.00	2.27
	MLR	[44]									x						NR	NR	NR	0.9959	1.02	5.00	0.00	4.00	1.37
	FFNN + GP	[84]									x		x	x			NR	NR	5.40 %	NR	0.00	5.00	0.00	4.00	1.00
SO2	RF	[58]	x	x	x		x	x	x			x	x	x			NR	0.09 ppb	5.00 %	0.95	1.00	1.00	5.00	2.57	1.54
	FFNN	[58]	x	x	x		x	x	x			x	x	x			NR	0.16 ppb	10.00 %	0.86	1.00	1.00	1.00	2.57	1.22

Resolution Many studies have a full mark on the Resolution score because they all use a temporal resolution lower or equal than 1 minute. These studies are Maag et al. (2016) [38], Spinelle et al. (2015, 2017) [73, 25], Cheng et al. [83], Esposito et al. [79], Liu et al. [44], Saukh et al. [72], Hasenfratz et al. (2012) [27], and Gao et al. [84]. Other studies with a high-resolution score are Lin et al. [111] and Zimmerman et al. [39]. Every other study has a resolution of 1 hour or not reported, which makes them get a Resolution score of respectively 1 and 0. This suggests that most studies use a good temporal resolution. Regarding spatial resolution, however, the scores would be significantly lower as most studies use only measurements from a single geographical location. Most surveys do not specify the coverage area of measurements and hence we have not been able to calculate a reliable score for spatial resolution.

Accuracy The accuracy score, by the nature of how it is calculated, is smoothly distributed. The RF model by Zimmerman et al. [39] dominates most of the rankings, namely CO, CO_2, NO_2, and O_3. Notable mentions with good performances are the MLR model by Maag et al. [38] for CO; the FFNN model by Spinelle et al. [25] for CO_2; the TDNN and NARX models by Esposito et al. [79] for NO_2,;and the RF model by Borrego et al. [58] for NO_2 and O_3. Only a few models have been developed for the rest of the pollutants, and a few of the studies that present them do not present meaningful performance measures. Because of these reasons, we will only mention the best model for each. For NO, the best model is the FFNN model by Spinelle et al. (2017) [25]; for NO_x, the best is the TDNN model by Esposito et al. [79]; for SO_2, the best model is the RF model by Borrego et al. (2018) [58]; and for For $PM_{2.5}$, the best is the FFNN model with GP by Cheng et al. [83].

Technology The Technology score, as we have already discussed, evaluates the sensor technologies used in the studies. The studies with the highest score are those by De Vito et al. [40], Hasenfratz et al. [27], Lin et al. [111], Saukh et al. [72], Maag et al. [55] and Zimmerman et al. [39]. All of these studies have in common the fact that they use MOS sensors. Other studies with a good score are Cheng et al. [83], Gao et al. [84] and Liu et al. [44], which all use LSP sensors. The rest of the studies have a score that ranges from middle to low, since they use a combination of sensors including ECs, or use only ECs, except for Zimmerman et al. [39] which uses NDIR. Studies with a higher than average score are Spinelle et al. [73, 25] and Zimmerman et al. [39].

Final Score The final score ranks the studies very differently than a score based on accuracy alone would. There are several reasons for this. Firstly, the final score is to be intended as a score that evaluates the whole methodology of a study, not just the model performance. Secondly, the most important parameter to influence the score is reliability, which is based on the length of the test dataset used. This is because, if a model has a very high performance in a short dataset, it might be overfitting the data. For these reasons, for CO, NO_2 and O_3 we can see that the best-rated model by our evaluation criteria is an MLR model by Maag et al. [38]. Another property that helps their study to be on top is the resolution of their model, which is 5 seconds. According to our evaluation criteria, highly ranked studies on gases are Zimmerman et al. [39] and Spinelle et al. [73, 25], which use advanced models, respectively RF and FFNN. For $PM_{2.5}$, the final scores are low for all studies. This is because in all of them the length of the test dataset is very short, or not reported at all. Also, most of them use relative performance measures or measures based on variance alone. This results in an Accuracy score of 0 since it is difficult to compare them to other studies.

8 Discussion and Roadmap

In this survey, we have critically compared common technologies and methods for machine-learning-based calibration of low-cost sensing units, including the sensing units themselves, the machine learning algorithms used for constructing the calibration models, and the evaluation measures for assessing the usefulness of the models. We next reflect on the current state of the field, highlighting open issues that need addressing, and briefly presenting a research roadmap that fills these gaps.

Combination of Sensors. Considering that sensors have cross-sensitivities between pollutants, an important research problem is to find combinations of sensors that capture as many pollutants as possible without suffering in calibration performance. By best combination, we mean both the combination of sensors for different pollutants, and the use of multiple different sensor technologies for the same pollutants. For example, it may be possible to improve the performance of particulate matter sensing by combining infrared and laser-based LSP sensors.

Life Cycle Management. Massive-scale AQSs deployments are often built out of a heterogeneous base of sensors that are unattended and installed in hard-to-reach locations. Routine tasks such as cleanup or software updates become hard to manage, which can lead to high maintenance costs.

Device life cycle management with minimal need for manual intervention is critical for continuous long-term operation of these deployments. Achieving this with massive deployments is an open issue. Another open problem related to the life cycle management is detecting or predicting when a sensor has failed or is about to fail. This can be potentially

accomplished using ML techniques, but these techniques have not yet been investigated in the context of low-cost AQSs.

Mobility Effects. Mobility can significantly affect the accuracy of sensors. When a sensor is in movement, the quantity of air that enters the sensor increases, which in turn can increase the concentration of the pollutant detected by the sensor. As we have already discussed, there are some ways to measure the mobility, so that it can be taken into account on the accuracy.

Universal Models. Most studies that attempt to calibrate multiple pollutants report mixed performance, with the best model differing for different pollutants. The most evident example can be found in Cordero et al. [110] where no single model performs well for all gases and PM classes. Developing models that perform well for multiple pollutants is currently an open issue.

Deep Learning. Very little work has been done on applying deep learning for calibrating low-cost AQSs. One of the reasons for this is the lack of sufficiently large datasets since deep learning typically requires a large amount of measurements before the models converge. Research in other domains has shown that deep learning can improve generality of models, and hence developing deep learning based calibration models can potentially mitigate effects of cross-sensitives, variations in environmental factors, and other issues with current models.

Dataset length The study by Maag et al. [38, 55], to the best of our knowledge, is the only one that uses a test dataset longer than a year. This is important for capturing seasonal phenomena and effects of different weather patterns. In the future, it would be interesting to see more studies that test the models for periods longer than a year, so that they are tested on different conditions.

Concept Drift and Re-Calibration. Concept drift has been widely reported as an issue for low-cost sensor technologies, but most studies have not been able to assess its effect due to using only short measurement periods. Concept drift can be mitigated using periodic re-calibration or online training where new training data is continuously used to improve model performance. Developing methods for detecting drift and triggering these mitigation techniques are currently open issues.

Use of Multiple Performance Measures. Most studies use only one or two performance measures for evaluating the performance of calibration models. Further work is needed to assess calibration accuracy using multiple, complementary measurements. Also, the practical results of calibration have not been thoroughly assessed with the study of Cheng et al. [83] being the only one to consider calibration performance in practical applications. Specifically, Cheng et al. consider how the results of the calibration model would affect the values of an air quality index.

9 Summary and Conclusion

Low-cost air quality monitoring technology is emerging as a complementary technology to professional-grade air quality stations. The high cost of professional-grade stations limits the granularity at which they can capture pollutant concentrations, whereas low-cost sensors can be deployed densely to increase the spatial granularity of collected information. Unfortunately, the accuracy of low-cost sensors tends to be poor as the sensors are vulnerable to several sources of noise. In this article, we have critically surveyed machine-learning-based calibration of low-cost air quality sensors, the main technique for improving the usefulness of measurements provided by low-cost air quality sensors. Our focus has been on individual sensing units, each of which typically integrates several different sensors (e.g., environmental sensors, particulate matter sensors and sensors for gaseous pollutants). In this survey, we have covered the sensor technology itself, the processing pipeline required for calibration, the machine learning techniques that are used in calibration, and different ways to evaluate the performance of calibration models. Based on our survey, we have highlighted open research issues in the field, with the inconsistency of studies, lack of sufficiently long datasets, and lack of models that perform well across several pollutants being among the most critical research problems.

References

[1] WHO. Ambient air pollution: A global assessment of exposure and burden of disease. Technical report, World Health Organization, 2016.

[2] Robert D. Brook, Sanjay Rajagopalan, C. Arden Pope, Jeffrey R. Brook, Aruni Bhatnagar, Ana V. Diez-Roux, Fernando Holguin, Yuling Hong, Russell V. Luepker, Murray A. Mittleman, Annette Peters, David Siscovick, Sidney C. Smith, Laurie Whitsel, and Joel D. Kaufman. Particulate matter air pollution and cardiovascular disease. *Circulation*, 121(21):2331–2378, 2010.

[3] Mark Goldberg. A systematic review of the relation between long-term exposure to ambient air pollution and chronic diseases. *Reviews on Environmental Health*, 23(4):243–298, 2011.

[4] Zorana J. Andersen, Luise C. Kristiansen, Klaus K. Andersen, Tom S. Olsen, Martin Hvidberg, Steen S. Jensen, Matthias Ketzel, Steffen Loft, Mette Sørensen, Anne Tjønneland, Kim Overvad, and Ole Raaschou-Nielsen. Stroke and long-term exposure to outdoor air pollution from nitrogen dioxide. *Stroke*, 43(2):320–325, 2012.

[5] Ulrike Gehring, Alet H. Wijga, Michael Brauer, Paul Fischer, Johan C. de Jongste, Marjan Kerkhof, Marieke Oldenwening, Henriette A. Smit, and Bert Brunekreef. Traffic-related air pollution and the development of asthma and allergies during the first 8 years of life. *American Journal of Respiratory and Critical Care Medicine*, 181(6):596–603, 2010. PMID: 19965811.

[6] Zorana J. Andersen, Martin Hvidberg, Steen S. Jensen, Matthias Ketzel, Steffen Loft, Mette Sørensen, Anne Tjønneland, Kim Overvad, and Ole Raaschou-Nielsen. Chronic obstructive pulmonary disease and long-term exposure to traffic-related air pollution. *American Journal of Respiratory and Critical Care Medicine*, 183(4):455–461, 2011. PMID: 20870755.

[7] Zorana Jovanovic Andersen, Klaus Bønnelykke, Martin Hvidberg, Steen S Jensen, Matthias Ketzel, Steffen Loft, Mette Sørensen, Anne Tjønneland, Kim Overvad, and Ole Raaschou-Nielsen. Long-term exposure to air pollution and asthma hospitalisations in older adults: a cohort study. *Thorax*, 67(1):6–11, 2012.

[8] Ole Raaschou-Nielsen, Zorana J. Andersen, Martin Hvidberg, Steen S. Jensen, Matthias Ketzel, Mette Sørensen, Johnni Hansen, Steffen Loft, Kim Overvad, and Anne Tjønneland. Air pollution from traffic and cancer incidence: a danish cohort study. *Environmental Health*, 10(1):67, Jul 2011.

[9] EM Saber and G Heydari. Flow patterns and deposition fraction of particles in the range of 0.1–10 μm at trachea and the first third generations under different breathing conditions. *Computers in biology and medicine*, 42(5):631–638, 2012.

[10] Heather Volk, Fred Lurmann, Bryan Penfold, Irva Hertz-Picciotto, and Rob McConnell. Traffic-related air pollution, particulate matter, and autism. *JAMA Psychiatry*, 70(1):71–77, 2013.

[11] Inge Bos, Patrick De Boever, Luc Int Panis, and Romain Meeusen. Physical activity, air pollution and the brain. *Sports Medicine*, 44(11):1505–1518, Nov 2014.

[12] Xin Zhang, Xi Chen, and Xiaobo Zhang. The impact of exposure to air pollution on cognitive performance. *Proceedings of the National Academy of Sciences*, 115(37):9193–9197, 2018.

[13] OECD. The economic consequence of outdoor air pollution, 2016.

[14] Alastair Lewis and Peter Edwards. Validate personal air-pollution sensors. *Nature*, 535:29–31, 2016.

[15] Ruwim Berkowicz, Finn Palmgren, Ole Hertel, and Elisabetta Vignati. Using measurements of air pollution in streets for evaluation of urban air quality—meterological analysis and model calculations. *Science of The Total Environment*, 189-190:259–265, 1996. Highway and Urban Pollution.

[16] Sotiris Vardoulakis, Norbert Gonzalez-Flesca, Bernard E.A. Fisher, and Koulis Pericleous. Spatial variability of air pollution in the vicinity of a permanent monitoring station in central paris. *Atmospheric Environment*, 39(15):2725–2736, 2005. Fourth International Conference on Urban Air Quality: Measurement, Modelling and Management, 25-28 March 2003.

[17] Markku Kulmala. Build a global earth observatory, 2018.

[18] Eemil Lagerspetz, Naser Hossein Motlagh, Martha Arbayani Zaidan, Pak L. Fung, Julien Mineraud, Samu Varjonen, Matti Siekkinen, Petteri Nurmi, Yutaka Matsumi, Sasu Tarkoma, and Tareq Hussein. Megasense: Feasibility of low-cost sensors for pollution hot-spot detection. In *2019 IEEE 17th International Conference on Industrial Informatics (INDIN)*, Helsinki, Finland, July 2019. IEEE.

[19] Joshua S. Apte, Kyle Messier, Shahzad Gani, Michael Brauer, Thomas Kirchstetter, Melissa Lunden, Julian D. Marshall, Christopher J. Portier, Roel Vermeulen, and Steven P. Hamburg. High-resolution air pollution mapping with Google Street View cars: Exploiting big data. *Environmental Science & Technology*, 51:6999–7008, 06 2017.

[20] Adam Moore, Miguel Figliozzi, and Christopher M Monsere. Air quality at bus stops: empirical analysis of exposure to particulate matter at bus stop shelters. *Transportation Research Record*, 2270(1):76–86, 2012.

[21] Michelle L Bell, Keita Ebisu, Roger D Peng, Jonathan M Samet, and Francesca Dominici. Hospital admissions and chemical composition of fine particle air pollution. *American journal of respiratory and critical care medicine*, 179(12):1115–1120, 2009.

[22] Jun Wang and Sundar A. Christopher. Intercomparison between satellite-derived aerosol optical thickness and pm2.5 mass: Implications for air quality studies. *Geophysical Research Letters*, 30(21):4–1–4–4, 2003.

[23] Ying Wang, Guoshun Zhuang, Aohan Tang, Hui Yuan, Yele Sun, Shuang Chen, and Aihua Zheng. The ion chemistry and the source of pm2. 5 aerosol in beijing. *Atmospheric Environment*, 39(21):3771–3784, 2005.

[24] Council of European Union. Directive 2008/50/ec of the european parliament and of the council of 21 may 2008 on ambient air quality and cleaner air for europe, 2008.
`http://data.europa.eu/eli/dir/2008/50/oj`.

[25] Laurent Spinelle, Michel Gerboles, Maria Gabriella Villani, Manuel Aleixandre, and Fausto Bonavitacola. Field calibration of a cluster of low-cost commercially available sensors for air quality monitoring. part b: No, co and co2. *Sensors and Actuators B: Chemical*, 238:706–715, 2017.

[26] Bartosz Szulczyński and Jacek Gębicki. Currently commercially available chemical sensors employed for detection of volatile organic compounds in outdoor and indoor air. *Environments*, 4(1):21, Mar 2017.

[27] David Hasenfratz, Olga Saukh, Silvan Sturzenegger, and Lothar Thiele. Participatory air pollution monitoring using smartphones. In *Proceedings of the 2nd International Workshop on Mobile Sensing*, pages 1–5, Beijing, China, 2012. Academic Press.

[28] Array of Things. Array of Things Expands with Partner Projects Around U.S. `https://medium.com/array-of-things/array-of-things-expands-with-partner-projects-around-u-s-c0ef9171f314`, 2018. [Online; accessed 15-Jan-2019].

[29] Jason Jingshi Li, Boi Faltings, Olga Saukh, David Hasenfratz, and Jan Beutel. Sensing the air we breathe: The opensense zurich dataset. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*, AAAI'12, pages 323–325, Toronto, Ontario, Canada, 2012. AAAI Press.

[30] Mark Bart, David E. Williams, Bruce Ainslie, Ian McKendry, Jennifer Salmond, Stuart K. Grange, Maryam Alavi-Shoshtari, Douw Steyn, and Geoff S. Henshaw. High density ozone monitoring using gas sensitive semi-conductor sensors in the lower fraser valley, british columbia. *Environmental Science and Technology*, 48(7):3970–3977, 2014.

[31] E. S. Cross, L. R. Williams, D. K. Lewis, G. R. Magoon, T. B. Onasch, M. L. Kaminsky, D. R. Worsnop, and J. T. Jayne. Use of electrochemical sensors for measurement of air pollution: correcting interference response and validating measurements. *Atmospheric Measurement Techniques*, 10(9):3575–3588, 2017.

[32] Nicholas Masson, Ricardo Piedrahita, and Michael Hannigan. Quantification method for electrolytic sensors in long-term monitoring of ambient air quality. *Sensors*, 15(10):27283–27302, 2015.

[33] Lidia Morawska, Phong K. Thai, Xiaoting Liu, Akwasi Asumadu-Sakyi, Godwin Ayoko, Alena Bartonova, Andrea Bedini, Fahe Chai, Bryce Christensen, Matthew Dunbabin, Jian Gao, Gayle S.W. Hagler, Rohan Jayaratne, Prashant Kumar, Alexis K.H. Lau, Peter K.K. Louie, Mandana Mazaheri, Zhi Ning, Nunzio Motta, Ben Mullins, Md Mahmudur Rahman, Zoran Ristovski, Mahnaz Shafiei, Dian Tjondronegoro, Dane Westerdahl, and Ron Williams. Applications of low-cost sensing technologies for air quality monitoring and exposure assessment: How far have they gone? *Environment International*, 116:286–299, 2018.

[34] C Borrego, AM Costa, J Ginja, M Amorim, M Coutinho, K Karatzas, Th Sioumis, N Katsifarakis, K Konstantinidis, S De Vito, et al. Assessment of air quality microsensors versus reference methods: The eunetair joint exercise. *Atmospheric Environment*, 147:246–263, 2016.

[35] W. Jiao, G. Hagler, R. Williams, R. Sharpe, R. Brown, D. Garver, R. Judge, M. Caudill, J. Rickard, M. Davis, L. Weinstock, S. Zimmer-Dauphinee, and K. Buckley. Community air sensor network (cairsense) project: evaluation of low-cost sensor performance in a suburban environment in the southeastern united states. *Atmospheric Measurement Techniques*, 9(11):5281–5292, 2016.

[36] Lidia Morawska, Phong K Thai, Xiaoting Liu, Akwasi Asumadu-Sakyi, Godwin Ayoko, Alena Bartonova, Andrea Bedini, Fahe Chai, Bryce Christensen, Matthew Dunbabin, et al. Applications of low-cost sensing technologies for air quality monitoring and exposure assessment: How far have they gone? *Environment international*, 116:286–299, 2018.

[37] Nithya Ramanathan, Laura Balzano, Marci C Burt, Deborah Estrin, Tom Harmon, Colleen K Harvey, Jenny Jay, Eddie Kohler, Sarah E Rothenberg, and Mani Srivastava. Rapid deployment with confidence: Calibration and fault detection in environmental sensor networks. Technical report, University of California, 2006.

[38] Balz Maag, Olga Saukh, David Hasenfratz, and Lothar Thiele. Pre-deployment testing, augmentation and calibration of cross-sensitive sensors. In *Proceedings of the 2016 International Conference on Embedded Wireless Systems and Networks*, EWSN '16, pages 169–180, USA, 2016. Junction Publishing.

[39] N. Zimmerman, A. A. Presto, S. P. N. Kumar, J. Gu, A. Hauryliuk, E. S. Robinson, A. L. Robinson, and R. Subramanian. A machine learning calibration model using random forests to improve sensor performance for lower-cost air quality monitoring. *Atmospheric Measurement Techniques*, 11(1):291–313, 2018.

[40] Saverio De Vito, Marco Piga, Luca Martinotto, and Girolamo Di Francia. Co, no2 and nox urban pollution monitoring with on-field calibrated electronic nose by automatic bayesian regularization. *Sensors and Actuators B: Chemical*, 143(1):182–191, 2009.

[41] Yun Cheng, Xiaoxi He, Zimu Zhou, and Lothar Thiele. Ict: In-field calibration transfer for air quality sensor deployments. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 3(1):6, 2019.

[42] Naser Hossein Motlagh, Eemil Lagerspetz, Petteri Nurmi, Xin Li, Samu Varjonen, Julien Mineraud, Matti Siekkinen, Andrew Rebeiro-Hargrave, Tareq Hussein, Tuukka Petäjä, Markku Kulmala, and Sasu Tarkoma. Towards massive scale air quality monitoring. *IEEE Communications Magazine*, February 2020.

[43] João Gama, Indrė Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. A survey on concept drift adaptation. *ACM Comput. Surv.*, 46(4):44:1–44:37, March 2014.

[44] Di Liu, Qiang Zhang, Jingkun Jiang, and Da-Ren Chen. Performance calibration of low-cost and portable particular matter (pm) sensors. *Journal of Aerosol Science*, 112:1–10, 2017.

[45] T. Zheng, M. H. Bergin, K. K. Johnson, S. N. Tripathi, S. Shirodkar, M. S. Landis, R. Sutaria, and D. E. Carlson. Field evaluation of low-cost particulate matter sensors in high- and low-concentration environments. *Atmospheric Measurement Techniques*, 11(8):4823–4846, 2018.

[46] B. Maag, Z. Zhou, and L. Thiele. A survey on sensor calibration in air pollution monitoring deployments. *IEEE Internet of Things Journal*, 5(6):4857–4870, Dec 2018.

[47] Gerard Hoek, Rob Beelen, Kees de Hoogh, Danielle Vienneau, John Gulliver, Paul Fischer, and David Briggs. A review of land-use regression models to assess spatial variation of outdoor air pollution. *Atmospheric Environment*, 42(33):7561 – 7578, 2008.

[48] Markey Johnson, V. Isakov, J.S. Touma, S. Mukerjee, and H. Özkaynak. Evaluation of land-use regression models used to predict air quality concentrations in an urban area. *Atmospheric Environment*, 44(30):3660–3668, 2010.

[49] Chengxiang Wang, Longwei Yin, Luyuan Zhang, Dong Xiang, and Rui Gao. Metal oxide gas sensors: Sensitivity and influencing factors. *Sensors*, 10(3):2088–2106, 2010.

[50] Ronan Baron and John Saffell. Amperometric gas sensors as a low cost emerging technology platform for air quality monitoring applications: A review. *ACS Sensors*, 2(11):1553–1566, 2017.

[51] Örnek Özgür. An overview of metal oxide semiconducting sensors in electronic nose applications. In *Third International Symposium on Sustainable Development*, pages 506–515, Sarajevo, Bosnia and Herzegovina, Oct 2012. International Burch University.

[52] Trieu-Vuong Dinh, In-Young Choi, Youn-Suk Son, and Jo-Chun Kim. A review on non-dispersive infrared gas sensors: Improvement of sensor detection limit and interference correction. *Sensors and Actuators B: Chemical*, 231, 03 2016.

[53] Jonathan E. Thompson. Crowd-sourced air quality studies: A review of the literature & portable sensors. *Trends in Environmental Analytical Chemistry*, 11:23–34, 2016.

[54] Laurent Spinelle, Michel Gerboles, Gertjan Kok, Stefan Persijn, and Tilman Sauerwald. Review of portable and low-cost sensors for the ambient air monitoring of benzene and other volatile organic compounds. *Sensors*, 17(7):1520, Jun 2017.

[55] Balz Maag, Zimu Zhou, and Lothar Thiele. W-air: Enabling personal air pollution monitoring on wearables. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 2(1):24:1–24:25, March 2018.

[56] Manuel Aleixandre and Michel Gerboles. Review of small commercial sensors for indicative monitoring of ambient gas. *Chemical Engineering Transactions*, 30, 09 2012.

[57] C. Borrego, A.M. Costa, J. Ginja, M. Amorim, M. Coutinho, K. Karatzas, Th. Sioumis, N. Katsifarakis, K. Konstantinidis, S. De Vito, E. Esposito, P. Smith, N. André, P. Gérard, L.A. Francis, N. Castell, P. Schneider, M. Viana, M.C. Minguillón, W. Reimringer, R.P. Otjes, O. von Sicard, R. Pohle, B. Elen, D. Suriano, V. Pfister, M. Prato, S. Dipinto, and M. Penza. Assessment of air quality microsensors versus reference methods: The eunetair joint exercise. *Atmospheric Environment*, 147:246–263, 2016.

[58] C. Borrego, J. Ginja, M. Coutinho, C. Ribeiro, K. Karatzas, Th Sioumis, N. Katsifarakis, K. Konstantinidis, S. De Vito, E. Esposito, M. Salvato, P. Smith, N. André, P. Gérard, L.A. Francis, N. Castell, P. Schneider, M. Viana, M.C. Minguillón, W. Reimringer, R.P. Otjes, O. von Sicard, R. Pohle, B. Elen, D. Suriano, V. Pfister, M. Prato, S. Dipinto, and M. Penza. Assessment of air quality microsensors versus reference methods: The eunetair joint exercise – part ii. *Atmospheric Environment*, 193:127 – 142, 2018.

[59] Alastair C Lewis, Erika von Schneidemesser, and Richard E Peltier. Low-cost sensors for the measurement of atmospheric composition: overview of topic and future applications. Technical Report 1215, World Meteorological Organization, Geneva, Switzerland, May 2018.

[60] Wei Yi, Kin Lo, Terrence Mak, Kwong Leung, Yee Leung, and Mei Meng. A survey of wireless sensor network based air pollution monitoring systems. *Sensors*, 15(12):31392–31427, 2015.

[61] David G. Streets, Timothy Canty, Gregory R. Carmichael, Benjamin de Foy, Russell R. Dickerson, Bryan N. Duncan, David P. Edwards, John A. Haynes, Daven K. Henze, Marc R. Houyoux, Daniel J. Jacob, Nickolay A. Krotkov, Lok N. Lamsal, Yang Liu, Zifeng Lu, Randall V. Martin, Gabriele G. Pfister, Robert W. Pinder, Ross J. Salawitch, and Kevin J. Wecht. Emissions estimation from satellite retrievals: A review of current capability. *Atmospheric Environment*, 77:1011–1042, 2013.

[62] Bryan N. Duncan, Ana I. Prados, Lok N. Lamsal, Yang Liu, David G. Streets, Pawan Gupta, Ernest Hilsenrath, Ralph A. Kahn, J. Eric Nielsen, Andreas J. Beyersdorf, Sharon P. Burton, Arlene M. Fiore, Jack Fishman, Daven K. Henze, Chris A. Hostetler, Nickolay A. Krotkov, Pius Lee, Meiyun Lin, Steven Pawson, Gabriele Pfister, Kenneth E. Pickering, R. Bradley Pierce, Yasuko Yoshida, and Luke D. Ziemba. Satellite data of atmospheric pollution for u.s. air quality applications: Examples of applications, summary of data end-user resources, answers to faqs, and common mistakes to avoid. *Atmospheric Environment*, 94:647–662, 2014.

[63] Henrik Olstrup, Christer Johansson, Bertil Forsberg, Andreas Tornevi, Agneta Ekebom, and Kadri Meister. A multi-pollutant air quality health index (aqhi) based on short-term respiratory effects in stockholm, sweden. *International Journal of Environmental Research and Public Health*, 16(1):105, Jan 2019.

[64] David Hasenfratz, Tabita Arn, Ivo de Concini, Olga Saukh, and Lothar Thiele. Health-optimal routing in urban areas. In *Proceedings of the 14th International Conference on Information Processing in Sensor Networks*, IPSN '15, pages 398–399, New York, NY, USA, 2015. ACM.

[65] Prabal Dutta, Paul M. Aoki, Neil Kumar, Alan Mainwaring, Chris Myers, Wesley Willett, and Allison Woodruff. Common sense: Participatory urban sensing using a network of handheld air quality monitors. In *Proceedings of the 7th ACM Conference on Embedded Networked Sensor Systems*, SenSys '09, pages 349–350, New York, NY, USA, 2009. ACM.

[66] Nuria Castell, Franck R. Dauge, Philipp Schneider, Matthias Vogt, Uri Lerner, Barak Fishbain, David Broday, and Alena Bartonova. Can commercial low-cost sensor platforms contribute to air quality monitoring and exposure estimates? *Environment International*, 99:293–302, 2017.

[67] Seppo T. Rinne and Joel D. Kaufman. Chapter 80 - air pollution. In Stephen G. Spiro, Gerard A. Silvestri, and Alvar Agustí, editors, *Clinical Respiratory Medicine*, pages 937–945. W.B. Saunders, Philadelphia, 3 edition, 2012.

[68] S. De Vito, E. Massera, M. Piga, L. Martinotto, and G. Di Francia. On field calibration of an electronic nose for benzene estimation in an urban pollution monitoring scenario. *Sensors and Actuators B: Chemical*, 129(2):750–757, 2008.

[69] David Hasenfratz, Olga Saukh, and Lothar Thiele. On-the-fly calibration of low-cost gas sensors. In Gian Pietro Picco and Wendi Heinzelman, editors, *Wireless Sensor Networks*, pages 228–244, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.

[70] R. Piedrahita, Y. Xiang, N. Masson, J. Ortega, A. Collier, Y. Jiang, K. Li, R. P. Dick, Q. Lv, M. Hannigan, and L. Shang. The next generation of low-cost personal air quality sensors for quantitative exposure monitoring. *Atmospheric Measurement Techniques*, 7(10):3325–3336, 2014.

[71] D. Oletic and V. Bilas. Design of sensor node for air quality crowdsensing. In *IEEE Sensors Applications Symposium (SAS)*, pages 1–5, Zadar, Croatia, April 2015. IEEE.

[72] Olga Saukh, David Hasenfratz, and Lothar Thiele. Reducing multi-hop calibration errors in large-scale mobile sensor networks. In *Proceedings of the 14th International Conference on Information Processing in Sensor Networks*, IPSN '15, pages 274–285, New York, NY, USA, 2015. ACM.

[73] Laurent Spinelle, Michel Gerboles, Maria Gabriella Villani, Manuel Aleixandre, and Fausto Bonavitacola. Field calibration of a cluster of low-cost available sensors for air quality monitoring. part a: Ozone and nitrogen dioxide. *Sensors and Actuators B: Chemical*, 215:249–257, 2015.

[74] Wataru Tsujita, Akihito Yoshino, Hiroshi Ishida, and Toyosaka Moriizumi. Gas sensor network for air-pollution monitoring. *Sensors and Actuators B: Chemical*, 110(2):304 – 311, 2005.

[75] N. Masson, R. Piedrahita, and M. Hannigan. Approach for quantification of metal oxide type semiconductor gas sensors used for ambient air quality monitoring. *Sensors and Actuators B: Chemical*, 208:339–345, 2015.

[76] Javier Burgués and Santiago Marco. Low power operation of temperature-modulated metal oxide semiconductor gas sensors. *Sensors*, 18(2), 2018.

[77] Alexey Vasiliev, Andrey E. Varfolomeev, Ivan A. Volkov, Nikolay Simonenko, Pavel V. Arsenov, Ivan Vlasov, Victor V. Ivanov, Alexander V. Pislyakov, Alexander Lagutin, Igor Jahatspanian, and Thomas Maeder. Reducing humidity response of gas sensors for medical applications: Use of spark discharge synthesis of metal oxide nanoparticles. *Sensors*, 18, 08 2018.

[78] Nima Nikzad, Nakul Verma, Celal Ziftci, Elizabeth Bales, Nichole Quick, Piero Zappi, Kevin Patrick, Sanjoy Dasgupta, Ingolf Krueger, Tajana Šimunić Rosing, and William G. Griswold. Citisense: Improving geospatial environmental assessment of air quality using a wireless personal exposure monitoring system. In *Proceedings of the Conference on Wireless Health*, WH '12, pages 11:1–11:8, New York, NY, USA, 2012. ACM.

[79] E. Esposito, S. De Vito, M. Salvato, V. Bright, R.L. Jones, and O. Popoola. Dynamic neural network architectures for on field stochastic calibration of indicative low cost air quality sensing systems. *Sensors and Actuators B: Chemical*, 231:701–713, 2016.

[80] K. Hu, V. Sivaraman, H. Bhrugubanda, S. Kang, and A. Rahman. Svr based dense air pollution estimation model using static and wireless sensor network. In *2016 IEEE SENSORS*, pages 1–3, Orlando, FL, USA, Oct 2016. IEEE.

[81] Peng Wei, Zhi Ning, Sheng Ye, Li Sun, Fenhuan Yang, Ka Chun Wong, Dane Westerdahl, and Peter K. K. Louie. Impact analysis of temperature and humidity conditions on electrochemical sensor response in ambient air quality monitoring. *Sensors*, 18, 1 2018.

[82] Samuel O. Agbroko and James Covington. A novel, low-cost, portable pid sensor for the detection of volatile organic compounds. *Sensors and Actuators B: Chemical*, 275:10 – 15, 2018.

[83] Yun Cheng, Xiucheng Li, Zhijun Li, Shouxu Jiang, Yilong Li, Ji Jia, and Xiaofan Jiang. Aircloud: A cloud-based air-quality monitoring system for everyone. In *Proceedings of the 12th ACM Conference on Embedded Network Sensor Systems*, SenSys '14, pages 251–265, New York, NY, USA, 2014. ACM.

[84] Y. Gao, W. Dong, K. Guo, X. Liu, Y. Chen, X. Liu, J. Bu, and C. Chen. Mosaic: A low-cost mobile sensing system for urban air quality monitoring. In *IEEE INFOCOM 2016 - The 35th Annual IEEE International Conference on Computer Communications*, pages 1–9, San Francisco, CA, USA, April 2016. IEEE.

[85] H. Liu, H. Wu, H. Lee, Y. Ho, and L. Chen. A system calibration model for mobile pm2.5 sensing using low-cost sensors. In *2017 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData)*, pages 611–618, Exeter, UK, June 2017. IEEE.

[86] Linda H Schmoll, Thomas M Peters, and Patrick O'Shaughnessy. Use of a condensation particle counter and an optical particle counter to assess the number concentration of engineered nanoparticles. *Journal of occupational and environmental hygiene*, 7:535–45, 09 2010.

[87] Sinan Sousan, Kirsten Koehler, Geb Thomas, Jae Hong Park, Michael Hillman, Andrew Halterman, and Thomas M. Peters. Inter-comparison of low-cost sensors for measuring the mass concentration of occupational aerosols. *Aerosol Science and Technology*, 50(5):462–473, 2016.

[88] Wenjia Shao, Hongjian Zhang, and Hongliang Zhou. Fine particle sensor based on multi-angle light scattering and data fusion. *Sensors (Switzerland)*, 17, 05 2017.

[89] Bean Chen, Diane Schwegler-Berry, Amy Cumpston, Jared Cumpston, Sherri Friend, Samuel Stone, and Michael Keane. Performance of a scanning mobility particle sizer in measuring diverse types of airborne nanoparticles: Multi-walled carbon nanotubes, welding fumes, and titanium dioxide spray. *Journal of Occupational and Environmental Hygiene*, 13, 02 2016.

[90] M. Fierz, C. Houle, P. Steigmeier, and H.Burtscher. Design, calibration, and field performance of a miniature diffusion size classifier. *Aerosol Science and Technology*, 45(1):1–10, 2011.

[91] Reto Meier, Katherine Clark, and Michael Riediker. Comparative testing of a miniature diffusion size classifier to assess airborne ultrafine particles under field conditions. *Aerosol Science and Technology - AEROSOL SCI TECH*, 47, 01 2012.

[92] Min-Geon Kim, Ji-Seob Choi, and Woo-Tae Park. Mems pzt oscillating platform for fine dust particle removal at resonance. *International Journal of Precision Engineering and Manufacturing*, 19(12):1851–1859, Dec 2018.

[93] Liang Liu, Wu Liu, Yu Zheng, Huadong Ma, and Cheng Zhang. Third-eye: A mobilephone-enabled crowdsensing system for air quality monitoring. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 2(1):20:1–20:26, March 2018.

[94] M. R. Chowdhury, S. De, N. K. Shukla, and R. N. Biswas. Energy-efficient air pollution monitoring with optimum duty-cycling on a sensor hub. In *2018 Twenty Fourth National Conference on Communications (NCC)*, pages 1–6, Hyderabad, India, Feb 2018. IEEE.

[95] Joel Kuula, Timo Mäkelä, Risto Hillamo, and Hilkka Timonen. Response characterization of an inexpensive aerosol sensor. *Sensors (Switzerland)*, 17, 11 2017.

[96] Zohir Chowdhury, Rufus D. Edwards, Michael Johnson, Kyra Naumoff Shields, Tracy Allen, Eduardo Canuz, and Kirk R. Smith. An inexpensive light-scattering particle monitor: field validation. *J. Environ. Monit.*, 9:1099–1106, 2007.

[97] Yu Zheng, Furui Liu, and Hsun-Ping Hsieh. U-air: When urban air quality inference meets big data. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, pages 1436–1444, New York, NY, USA, 2013. ACM.

[98] Yang Wang, Jiayu Li, He Jing, Qiang Zhang, Jingkun Jiang, and Pratim Biswas. Laboratory evaluation and calibration of three low-cost particle sensors for particulate matter measurement. *Aerosol Science and Technology*, 49:1063–1077, 11 2015.

[99] D. Hasenfratz, O. Saukh, C. Walser, C. Hueglin, M. Fierz, and L. Thiele. Pushing the spatio-temporal resolution limit of urban air pollution maps. In *2014 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, pages 69–77, Budapest, Hungary, March 2014. IEEE.

[100] Yoshinari Shirai, Yasue Kishino, Futoshi Naya, and Yutaka Yanagisawa. Toward on-demand urban air quality monitoring using public vehicles. In *Proceedings of the 2Nd International Workshop on Smart*, SmartCities '16, pages 1:1–1:6, New York, NY, USA, 2016. ACM.

[101] DP Chock, TR Terrell, and SB Levitt. Time-series analysis of riverside, california air quality data. *Atmospheric Environment (1967)*, 9(11):978–989, 1975.

[102] PH Merz, LJ Painter, and PR Ryason. Aerometric data analysis—time series analysis and forecast and an atmospheric smog diagram. *Atmospheric Environment (1967)*, 6(5):319–342, 1972.

[103] RLR Salcedo, MCM Alvim Ferraz, CA Alves, and FG Martins. Time-series analysis of air pollution data. *Atmospheric Environment*, 33(15):2361–2372, 1999.

[104] Anne-Claude Romain and Jacques Nicolas. Long term stability of metal oxide-based gas sensors for e-nose environmental applications: An overview. *Sensors and Actuators B: Chemical*, 146(2):502–506, 2010.

[105] N Barsan, D Koziej, and U Weimar. Metal oxide-based gas sensor research: How to? *Sensors and Actuators B: Chemical*, 121(1):18–35, 2007.

[106] Elena Austin, Igor Novosselov, Edmund Seto, and Michael G Yost. Laboratory evaluation of the shinyei ppd42ns low-cost particulate matter sensor. *PloS one*, 10(9):e0137789, 2015.

[107] Guiqian Tang, Jinqiang Zhang, Xiaowan Zhu, Tao Song, Christoph Münkel, Bo Hu, Klaus Schäfer, Zirui Liu, Junke Zhang, Lili Wang, et al. Mixing layer height and its implications for air pollution over beijing, china. *Atmospheric Chemistry and Physics*, 16(4):2459–2475, 2016.

[108] Patrick Wagner and Klaus Schäfer. Influence of mixing layer height on air pollutant concentrations in an urban street canyon. *Urban Climate*, 22:64–79, 2017.

[109] David Mintz. *Technical Assistance Document for the Reporting of Daily Air Quality—the Air Quality Index (AQI)*. U.S. Environmental Protection Agency, Research Triangle Park, North Carolina 27711, May 2016.

[110] José María Cordero, Rafael Borge, and Adolfo Narros. Using statistical methods to carry out in field calibrations of low cost air quality sensors. *Sensors and Actuators B: Chemical*, 267:245 – 254, 2018.

[111] C. Lin, J. Gillespie, M.D. Schuder, W. Duberstein, I.J. Beverland, and M.R. Heal. Evaluation and calibration of aeroqual series 500 portable gas sensors for accurate measurement of ambient ozone and nitrogen dioxide. *Atmospheric Environment*, 100:111–116, 2015.

[112] Marloes Eeftens, Rob Beelen, Kees de Hoogh, Tom Bellander, Giulia Cesaroni, Marta Cirach, Christophe Declercq, Audrius Dėdelė, Evi Dons, Audrey de Nazelle, Konstantina Dimakopoulou, Kirsten Eriksen, Grégoire Falq, Paul Fischer, Claudia Galassi, Regina Gražulevičienė, Joachim Heinrich, Barbara Hoffmann, Michael Jerrett, Dirk Keidel, Michal Korek, Timo Lanki, Sarah Lindley, Christian Madsen, Anna Mölter, Gizella Nádor, Mark Nieuwenhuijsen, Michael Nonnemacher, Xanthi Pedeli, Ole Raaschou-Nielsen, Evridiki Patelarou, Ulrich Quass, Andrea Ranzi, Christian Schindler, Morgane Stempfelet, Euripides Stephanou, Dorothea Sugiri, Ming-Yi Tsai, Tarja Yli-Tuomi, Mihály J Varró, Danielle Vienneau, Stephanie von Klot, Kathrin Wolf, Bert Brunekreef, and Gerard Hoek. Development of land use regression models for pm2.5, pm2.5 absorbance, pm10 and pmcoarse in 20 european study areas; results of the escape project. *Environmental Science & Technology*, 46(20):11195–11205, 2012. PMID: 22963366.

[113] Rob Beelen, Gerard Hoek, Danielle Vienneau, Marloes Eeftens, Konstantina Dimakopoulou, Xanthi Pedeli, Ming-Yi Tsai, Nino Künzli, Tamara Schikowski, Alessandro Marcon, Kirsten T. Eriksen, Ole Raaschou-Nielsen, Euripides Stephanou, Evridiki Patelarou, Timo Lanki, Tarja Yli-Tuomi, Christophe Declercq, Grégoire Falq, Morgane Stempfelet, Matthias Birk, Josef Cyrys, Stephanie von Klot, Gizella Nádor, Mihály János Varró, Audrius Dėdelė, Regina Gražulevičienė, Anna Mölter, Sarah Lindley, Christian Madsen, Giulia Cesaroni, Andrea Ranzi, Chiara Badaloni, Barbara Hoffmann, Michael Nonnemacher, Ursula Krämer, Thomas Kuhlbusch, Marta Cirach, Audrey de Nazelle, Mark Nieuwenhuijsen, Tom Bellander, Michal Korek, David Olsson, Magnus Strömgren, Evi Dons, Michael Jerrett, Paul Fischer, Meng Wang, Bert Brunekreef, and Kees de Hoogh. Development of no2 and nox land use regression models for estimating air pollution exposure in 36 study areas in europe – the escape project. *Atmospheric Environment*, 72:10–23, 2013.

[114] Rongrong Wang, Sarah B. Henderson, Hind Sbihi, Ryan W. Allen, and Michael Brauer. Temporal stability of land use regression models for traffic-related air pollution. *Atmospheric Environment*, 64:312–319, 2013.

[115] Zachary C. Lipton. The mythos of model interpretability. *Commun. ACM*, 61(10):36–43, September 2018.

[116] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM Computing Surveys*, 51(5):93, 2018.

[117] K. Hu, A. Rahman, H. Bhrugubanda, and V. Sivaraman. Hazeest: Machine learning based metropolitan air pollution estimation from fixed and mobile sensors. *IEEE Sensors Journal*, 17(11):3517–3525, June 2017.

[118] Yuxiang Lin, Wei Dong, and Yuan Chen. Calibrating low-cost sensors by a two-phase learning approach for urban air quality measurement. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 2(1):18:1–18:18, March 2018.

[119] Yoav Freund and Robert E. Schapire. Experiments with a new boosting algorithm. In *Proceedings of the Thirteenth International Conference on International Conference on Machine Learning*, ICML'96, pages 148–156, San Francisco, CA, USA, 1996. Morgan Kaufmann Publishers Inc.

[120] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, 2001.

[121] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA, 2016. ACM.

[122] Nicholas E. Johnson, Bartosz Bonczak, and Constantine E. Kontokosta. Using a gradient boosting model to improve the performance of low-cost aerosol monitors in a dense, heterogeneous urban environment. *Atmospheric Environment*, 184:9 – 16, 2018.

[123] A. K. Jain, Jianchang Mao, and K. M. Mohiuddin. Artificial neural networks: a tutorial. *Computer*, 29(3):31–44, March 1996.

[124] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521:436, may 2015.

[125] Carl Edward Rasmussen. *Gaussian Processes in Machine Learning*, chapter 4, pages 63–71. Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, August 2004.

[126] T. Chai and R. R. Draxler. Root mean square error (rmse) or mean absolute error (mae)? - arguments against avoiding rmse in the literature. *Geoscientific Model Development*, 7(3):1247–1250, 2014.