Diss. ETH No. 26082

Air Quality Sensor Calibration and its Peculiarities

A thesis submitted to attain the degree of

Doctor of Sciences of ETH Zurich
(Dr. sc. ETH Zurich)

presented by
BALZ MAAG
M.Sc. ETH Zurich

born on 20.02.1990
citizen of
Switzerland
Zurich

accepted on the recommendation of
Prof. Dr. Lothar Thiele, examiner
Prof. Dr. Kay Römer, co-examiner
Prof. Dr. Olga Saukh, co-examiner

2019

Institut für Technische Informatik und Kommunikationsnetze
Computer Engineering and Networks Laboratory

TIK-SCHRIFTENREIHE NR. 178

Balz Maag

Air Quality Sensor Calibration and its Peculiarities

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

A dissertation submitted to
ETH Zurich
for the degree of Doctor of Sciences

DISS. ETH NO. 26082

Prof. Dr. Lothar Thiele, examiner
Prof. Dr. Kay Römer, co-examiner
Prof. Dr. Olga Saukh, co-examiner
Examination date: June 11, 2019

To my family.
Für mini Familie.

Abstract

Air pollution can have devastating effects on human health and the environment and is one of the most challenging environmental burdens we face as a modern society. Monitoring air quality has therefore become more and more important in recent years.

Ongoing monitoring efforts are usually conducted by sparsely distributed and fixed sites equipped with expensive high-end sensing infrastructure. Advances in sensor technology have made it possible to extend these efforts by deploying low-cost air pollution sensors in large-scale deployments that allow data collection with high spatio-temporal resolution. Unfortunately, low-cost air pollution sensors suffer from various limitations and, consequently, the collected data quality is limited and often not fit for meaningful applications.

In this thesis we tackle the different limitations by air pollution sensor calibration. We investigate multiple limiting factors of current sensor technologies and develop calibration strategies that can be applied at various stages of a deployment. The main contributions of this thesis are:

- We propose a new method to perform in-field pre-deployment testing of low-cost air quality sensors. The procedure is able to *(i)* conclude about the usability of a sensor in a given environment, *(ii)* identify fundamental cross-sensitivities and (*iii*) assemble and calibrate a sensor array, which provides accurate measurements with long-term stability.
- We provide a novel algorithm for collaborative multi-hop calibration of sensor arrays in a mobile deployment. Our approach provides a re-calibration framework and tackles cross-sensitivities, meteorological dependencies and signal drift of sensor arrays during their deployment. We theoretically and empirically show that our algorithm is minimizing error accumulation over multiple hops and outperforms related techniques.
- We are the first to uncover and counteract interference from human gas emissions on low-cost metal oxide sensor arrays in wearable and personal air pollution measurement devices. We design a wearable platform that is able to counteract human interference by utilizing non-linear neural network calibration and semi-supervised learning for effortless calibration model updates during deployment.
- We study uncertainty metrics and integrate them into different calibration models. We show that different calibration approaches improve their performance by applying heuristic filtering based on confidence information when trained on noisy data.

Zusammenfassung

Luftverschmutzung kann verheerende Auswirkungen auf die menschliche Gesundheit und die Umwelt haben und ist eine der herausforderndsten Umweltbelastungen, die unsere moderne Gesellschaft beschäftigt. Die Überwachung der Luftqualität wurde in den letzten Jahren deshalb immer wichtiger.

Laufende Überwachungsmassnahmen werden üblicherweise mit wenigen und festen Standorten, die mit teuren High-End Sensoren ausgerüstet sind, ausgeführt. Fortschritte in der Sensortechnologie boten die Möglichkeit diese Bemühungen mit grossangelegten Installationen von kostengünstigen Luftverschmutzungssensoren, die das Sammeln von Daten mit hoher räumlicher und zeitlicher Auflösung ermöglichen, auszudehnen. Leider leiden kostengünstige Luftverschmutzungssensoren unter verschiedenen Einschränkungen, welche dazu führen, dass die erfasste Datenqualität oft unzureichend und nicht für sinnvolle Anwendungen geeignet ist.

In dieser Arbeit bewältigen wir die unterschiedlichen Einschränkungen durch Kalibrierung von Luftverschmutzungssensoren. Wir untersuchen mehrere limitierende Faktoren von aktuellen Sensortechnologien und entwickeln Kalibrierungsstrategien, die zu verschiedenen Zeitpunkten während eines Sensoreinsatzes angewendet werden können. Die Hauptbeiträge dieser Arbeit lauten wie folgt:

- Wir erstellen eine neue Methode zum Testen von kostengünstigen Luftqualitätssensoren bevor deren Installation und unmittelbar im Einsatzgebiet. Das Verfahren erlaubt *(i)* Schlussfolgerung bezüglich der Verwendbarkeit eines Sensors in einer gegebenen Umgebung zu ziehen, *(ii)* grundlegende Quersensitivitäten zu identifizieren und *(iii)* das Erstellen und Kalibrieren von Sensor-Arrays, die genaue Messungen mit langfristiger Stabilität liefern.
- Wir entwickeln einen neuartigen Algorithmus zur kollaborativen Multi-Hop Kalibrierung von Sensor-Arrays in mobilen Anwendungen. Unsere Methode bietet ein Modell für Neukalibrierung und bekämpft Querempfindlichkeiten, meteorologische Abhängigkeiten und Signaldrift von Sensor-Arrays während ihres Einsatzes. Wir zeigen theoretisch und empirisch, dass unser Verfahren die Fehlerakkumulation über mehrere Hops minimiert und verwandte Methoden leistungsmässig übertrifft.
- Wir sind die Ersten, die Interferenzen zwischen menschlich verursachten Gasemissionen und kostengünstigen Metaloxid-Gas-Sensor-Arrays in Wearables und persönlichen Luftverschmutzungs-

messgeräten identifizieren und entgegenwirken. Wir entwickeln dazu eine Wearable-Platform, die die Interferenz mit Hilfe von Kalibrierung anhand nicht-linearen Neuralen Netzwerken bekämpft und durch teil-überwachtes Lernen mühelose Anpassungen des Kalibrierungsmodells während der Installation ermöglicht.

- Wir untersuchen Messunsicherheit-Metriken und integrieren diese in verschiedene Kalibrierungsmodelle. Wir zeigen, dass unterschiedliche Kalibrierungsansätze ihre Performanz durch Verwendung von heuristischen Filtermethoden, die die Messunsicherheit-Werte verwenden, verbessern, falls sie mit fehlerbehafteten Daten trainiert werden.

Acknowledgements

First of all, I would like to express my sincere gratitude to Lothar Thiele for giving me the once-in-a-lifetime opportunity to write this thesis, guiding me through the whole process, all the inspiring discussions and the never-ending support. Further, I would like to thank Kay Römer and Olga Saukh for reviewing my thesis, serving on the committee board and providing me with helpful feedback.

Special thanks go to Olga for all the interesting collaborations we had in the past and might have in the future, for closely helping me in the beginning of my PhD and all the guidance and encouragement. Similarly, I want to thank Zimu Zhou for the great support during the different projects and successful publications we accomplished together. I learned a lot from both of them and I am convinced they will inspire many more aspiring students during their academic careers. This thesis would not have been possible without their help.

Furthermore, I want to thank all my former and current colleagues at the Computer Engineering group for the great time we had together over the past few years. Especially, I would like to show my appreciation to David Hasenfratz, Christoph Walser, Tonio Gsell and Jan Beutel, who all greatly contributed to the OpenSense project and Roman Lim for supervising my master thesis, which initially sparked my interest in pursuing a PhD.

Over the course of my PhD I had the pleasure to supervise numerous students. Their hard work has been included throughout the whole thesis and their help is deeply appreciated.

Finally, I would like to thank my parents Rolf and Esther, my siblings Donat, Basil and Lina, my grandparents, all my close friends and my girlfriend Aresu for their endless encouragement and support.

The work presented in this thesis has been partially supported by NanoTera.ch with Swiss Confederation financing under the project OpenSense II and the Swiss National Science Foundation (SNSF) under the FLAG-ERA CONVERGENCE project. This support is gratefully acknowledged.

Contents

List of Figures

List of Tables

1

Introduction

Air pollution is one of the biggest environmental challenges we face in our modern society. It unquestionably affects the quality of all our lives and public health. According to a study by the World Health Organization (WHO) *"only one person in ten lives in a city that complies with the WHO Air quality guidelines"* [WHO16]. Pollutants such as particulate matter (PM), ozone (O_3), carbon monoxide (CO) or nitrogen dioxide (NO_2) are known to cause respiratory illnesses or cardiovascular diseases. As a result, approximately 3 million people die each year caused by poor air quality. Due to its continuous increase the WHO declared air pollution as a *"public health emergency"* [WHO16]. Heavily polluted air also leads to environmental problems such as acid rain, stratospheric ozone depletion and global climate change. Monitoring air pollution is thus of growing importance to increase public awareness and involvement in human health and sustainable urban environments [TC09].

Traditionally, air pollutants are monitored by fixed sites with expensive high-end sensing infrastructure run by governmental authorities. These monitoring sites are usually distributed sparsely and only suffice to estimate the average pollution affecting large populations. However, air pollution is known to be a complex phenomenon with sophisticated spatial and short-term variations [Mon01]. For instance, in major streets, the pollutant concentrations may vary within tens of meters and over time within minutes [DAS+05]. Therefore, it is desirable to increase the spatio-temporal resolution of available air pollution information for the public to assess their personal health risks and take precaution measures.

A driving factor that enables these increased monitoring efforts is the availability of low-cost portable air pollution sensors. These sensors are usually small, consume low power, cost roughly between 1$ and 1'000$ and are able to measure the concentrations of all the major air pollutants. Compared to bulky high-end solutions (≥ 10'000$), low-cost sensors are particularly convenient for large-scale static and mobile deployments [RKP$^+$17, JSBT$^+$15, SGK$^+$17, BS17]. By now, low-cost air pollution sensors have been successfully integrated into various long-term deployments to provide fine-grained air pollution information for quantitative studies and public services [YLM$^+$15].

Unfortunately, the data provided by these deployments is often lacking sufficient accuracy [FB17a, YLM$^+$15]. Many researchers report about serious inaccuracies when comparing the low-cost sensor measurements to reliable and accurate measurements of conventional monitoring sites [CDS$^+$17, JHW$^+$16]. The reason for this unsatisfying performance can be linked to various limitations of state-of-the-art low-cost sensors, such as low signal-to-noise ratios or interference from environmental factors [SGV$^+$17, SGV$^+$15].

In order to improve the data quality of existing and future air quality monitoring deployments, active research efforts are devoted to counteract these limitations with appropriate *sensor calibration*. By calibrating a low-cost sensor its measurements are transformed in a way that the calibrated measurements are able to closely agree with reference measurements from a high-end device. Sensor calibration is indispensable both before and after the deployments of low-cost air pollution sensors. Pre-deployment calibration is crucial to identify the primary error sources, select and train calibration models for low-cost sensors to properly function in the target deployment. Periodic post-deployment calibration is necessary to maintain consistency among distributed sensors and ensure data quality of long-term deployments.

Although calibration for air pollution sensors dates back to decades ago [SWLL91, Jan92], it has attracted increasing research interest because *(i)* newly available air pollution sensors push the boundaries in terms of power consumption and portability while neglecting sensing accuracy; and *(ii)* air pollution sensors are deployed in new scenarios such as in crowdsourced urban sensing [Tho16] and personal sensing [TDMP16].

1.1 Air Pollution Sensors

Fast advances in technology and strong commercialization efforts are main drivers for an increasing number of low-cost sensors available nowadays [PXM$^+$14]. Compared to high-end monitoring systems low-cost sensors typically require significantly less power and smaller packaging. Although these properties make low-cost sensors favourable for various large-scale monitoring applications, a diverse list of limitations hinders them to achieve a similar level of data quality as more sophisticated sensors. This section summarizes the most common sensing technologies of modern low-cost air pollution sensors.

As highlighted in [RKP$^+$17, YLM$^+$15], common low-cost sensors can roughly be divided in two groups defined by their target pollutant, i.e., particulate matter (Section 1.1.1) and gases (Section 1.1.2).

1.1.1 Particulate Matter Sensors

Particulate matter (PM) describes a mixture of solid and fluid particles, which are typically classified by their size in diameter. PM_{10} describes the mass concentration of particles with a diameter smaller than $10\mu m$, $PM_{2.5}$ smaller than $2.5\mu m$. Ultrafine particles (UFP) are nano-particles with diameters usually below $0.1\mu m$. These particles are known to cause serious effects on environment and human health and, thus, monitoring their concentration, size distribution and composition is of high importance [TC09].

Low-cost PM sensors are almost exclusively based on optical sensing principles. The most prominent principle is based on light scattering, where air is pumped into a small chamber. Inside the chamber a light source, either an LED or a low-power laser, is illuminating the air. Depending on the number of particles in the air mixture, the light is scattered with different intensity, which can be measured by a photodiode. Certain low-cost PM sensors apply more sophisticated optical principles to also differentiate sizes of particles.

1.1.2 Gas Sensors

The most relevant gaseous pollutants in outdoor air with serious negative effects on human beings, animals and the environment are sulphur dioxide (SO_2), oxides of nitrogen (NO, NO_2, NO_x = NO + NO_2), carbon monoxide (CO) and ozone (O_3) [TC09]. In indoor air mainly carbon dioxide (CO_2), volatile organic compounds (VOC) and in some cases also carbon monoxide (CO) are known to be possibly present in harmful concentrations [Jon99].

The majority of commercially available low-cost gas sensors is therefore targeting to measure the concentration of one of these gases. With the exception of CO_2, which is either directly measured with light scattering sensors [PXM+14] or approximated by the presence of VOCs [HHU+10], the most popular sensing principles are based on electrochemical or metal oxide layer reactions.

Electrochemical sensors. An electrochemical cell sensor (EC) consists in its simplest form of two electrodes, a working electrode and a counter electrode. Gases are either oxidized or reduced at the working electrode, which results in electronic charges generated. The generated potential difference at the two electrodes allows a current flow. This current is usually linearly proportional to the gas concentration. More advanced electrochemical sensors incorporate one or two additional electrodes to improve stability and sensitivity [MPS+13, WYZ+10].

Metal oxide sensors. Metal oxide sensors (MOX) use a sensing layer, where gases are either absorbed or desorbed. This reaction causes a change in conductivity of the material. In order to increase sensitivity the sensing layer needs to be heated to temperatures of at least 250°C. State-of-the-art metal oxide sensors are capable of measuring all the major gaseous pollutants [FCAB10].

1.1.3 Black-Box Approach

Based on the above sensing principles, manufacturers produce low-cost sensors and offer different features. Some sensors solely output an analog signal while others offer on-device signal processing, e.g., digitization of the analog signals or internal calibration. In the remainder of this thesis we do not differentiate between these different features. We regard a low-cost air pollution sensor as a black box with a signal output. The sensor is applied out-of-the-box and its output is used for comparison and calibration with references. This is the general approach done in the studies related to this thesis.

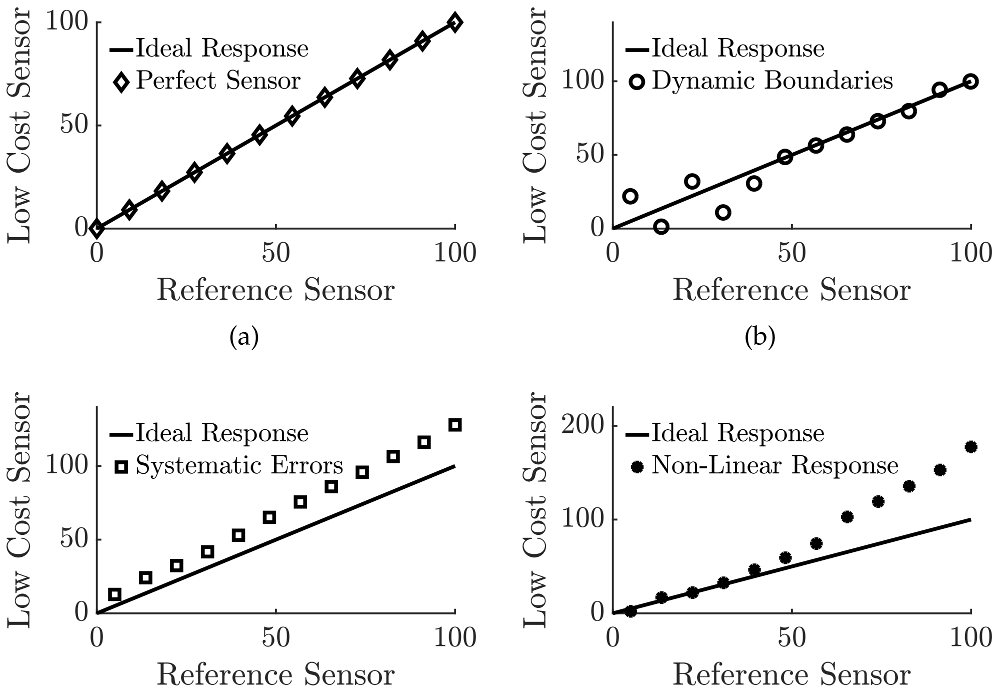
1.2 Error Sources

One of the most essential questions regarding the aforementioned low-cost sensors is how their measurements perform in comparison to high-quality references. An ideal sensor fully agrees with its corresponding reference sensor, i.e., exhibits a perfect linear relationship, as illustrated in Figure 1.1a. Unfortunately, the main reason why low-cost sensors have not yet been established as a trust-worthy air pollution monitoring fashion is their generally poor measurement accuracy [JHW+16, CDS+17, SGV+15, SGV+17]. In an exhaustive test report Jiao et al. [JHW+16] perform a black box testing approach for multiple sensors. Out of 38 tested sensors only 17 correlate well to their corresponding reference sensors. Through extensive sensor testing schemes and signal analysis researchers were able to detect multiple different error sources of state-of-the-art low-cost sensors. As a result, most low-cost sensors significantly deviate from an ideal sensor (Figure 1.1).

In the following we present the most prominent error sources, also summarized in Figure 1.2, that affect low-cost air pollution sensors. Note that we do not include error sources that have not yet been thoroughly tackled by calibration methods, such as slow response time or sensor mobility effects [AMM16a, AMM16b].

1.2.1 Dynamic boundaries

Dynamic boundaries define the range of a pollutant concentration in which a sensor is sensitive to. Especially the lower boundary, the limit of detection (LOD) [SG+11], is important. Below this boundary the noise of a sensor signal starts to dominate and it becomes impossible to differentiate between concentration levels. Low-cost sensors often



(a)

(b)

(c)

(d)

Figure 1.1: Comparison of measurements (arbitrary unit) from a reference sensor (x-axis) versus measurements of different low-cost sensors (y-axis). Figure 1.1a illustrates the response of a perfect low-cost sensor, Figure 1.1b of a low-cost sensor affected by high noise at low concentration due to imperfect dynamic boundaries, Figure 1.1c of a low-cost sensor suffering from systematically overestimated measurements and Figure 1.1d of a sensor with a non-linear response. The ideal response is a perfectly linear relationship between the low-cost and reference sensor.

have a LOD that is close to the range of interest or even surpasses it, however, this also depends on the application. As a result, measurements at low pollution concentration are subject to high noise. An example of a low-cost sensor affected by high noise at low concentration due to imperfect dynamic boundaries is depicted in Figure 1.1b. Especially PM [RKP+17] and electrochemical sensors [HIVF+18] are known to be significantly affected by low signal-to-noise ratios at low concentrations. It is important that calibration procedures are applied with respect to these dynamic boundaries.

1.2.2 Systematic errors

Systematic errors are of non-random nature and typically either characterized by a constant offset over the whole range of concentrations or an under- or overestimation of the concentration in certain ranges [CDS+17, SGV+15, SGV+17]. An example of a sensor response with a constant offset is illustrated in Figure 1.1c. They can often be attributed to imperfect

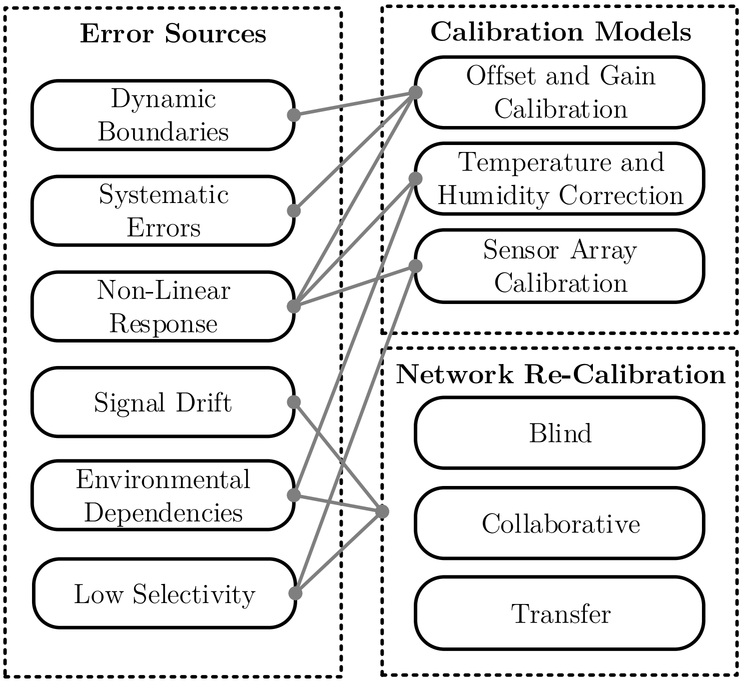


Figure 1.2: Overview of typical low-cost error sources and their corresponding calibration approaches indicated by gray lines.

calibration parameters and are generally not related to the sensing principle. Popular examples where systematic errors pose a challenge are factory calibrated sensors, as elaborated in detail in Section 1.4.

1.2.3 Non-linear response

Due to the nature of certain low-cost sensing techniques non-linear relationships between a sensor's and a reference's response are unavoidable. Sensor manufacturers often already linearise the sensor response, e.g., by internal signal processing, or provide information about typical non-linear behaviour in the datasheet. However, additional factors such as environmental conditions are known to cause non-linear behaviour as well [PSMJ16]. Figure 1.1d shows an example of a non-linear sensor response. A linear relationship is in general favourable because it allows the use of simple calibration models.

1.2.4 Signal drift

Low-cost sensors generally cannot maintain a stable measurement performance over a long time period [KESN+18, MMH17, KSL+18]. This usually happens due to ageing (causing gain drift) and impurity effects (causing offset or baseline drift), and leads to a slow drift of a sensor's sensitivity. Signal drift is one of the most common error sources and seriously impedes long-term deployments with low-cost sensors.

1.2.5 Environmental dependencies

Changing environmental conditions can cause problems that almost any low-cost sensor is facing. Various laboratory reports show that certain *physical* ambient properties, especially temperature and humidity conditions, can have a serious effect on a sensor's response. For instance, increasing humidity is notably decreasing the sensitivity of metal oxide [WYZ+10], electrochemical [PSL+17] and particulate matter sensors [WLJ+15]. As a result, low-cost sensors usually perform significantly worse in field deployments than in a laboratory setup. Further, environmental dependencies can also be responsible for non-linear responses, e.g., for electrochemical sensors [PSMJ16].

1.2.6 Low selectivity

Typical metal oxide and electrochemical sensors suffer from low selectivity. This means they are not exclusively sensitive to their intended target gas but are also cross-sensitive to, sometimes various, interfering substances in the air [LCL+12]. Especially in complex outdoor air these cross-sensitivities impose a fundamental challenge for low-cost gas sensors. Particulate matter sensors are usually not affected by cross-sensitivities because they are designed to detect a composition of different particles. However, in some cases where low-cost particulate matter sensors are either used to detect particles from certain sources like car exhaust or to distinguish different particle sizes, cross-sensitivities are also considered as a fundamental error source [RKP+17]. Compared to environmental dependencies, the low selectivity problem is caused by purely *chemical* interferences and requires more sophisticated calibration efforts.

1.3 Sensor Deployments and Calibration Opportunities

A commonly used solution to reduce the errors of low-cost air pollution sensors is calibration. A typical sensor calibration pipeline is shown in Figure 1.3. The general goal of sensor calibration is to find a calibration model that transforms the raw measurements of a low-cost sensor into a calibrated form. Determining a calibration model is typically done in a way that the measurements of the low-cost sensor are mapped to those of an accurate reference sensor with respect to some optimization goal, e.g., minimizing the calibration error. Sensor calibration is performed both *before* and *after* the deployment of air pollution sensors to deal with different error sources (see Figure 1.2).

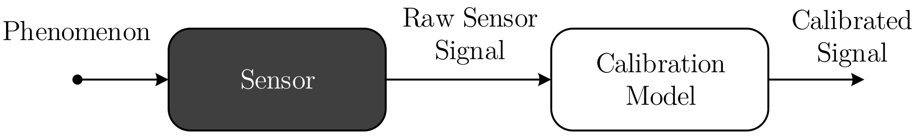


Figure 1.3: Typical sensor calibration pipeline. A sensor is sensing a phenomenon of interest, e.g., a pollution concentration, and produces a sensor signal. We treat the sensor as a black-box and apply a calibration model on the raw sensor signal to produce calibrated measurements. The calibrated measurement ideally describes the true property of the target phenomena, e.g., the true pollution concentration.

1.3.1 Pre-Deployment Calibration

The aim of pre-deployment calibration is to try to identify all possible error sources of a sensor in an observed and/or controlled environment before deploying the sensor in the field. Pre-deployment calibration usually assumes continuous availability of a high-quality reference sensor. One or multiple error sources listed in Figure 1.2 can be detected by comparing the low-cost sensor to the reference sensor. These error sources are then tackled by developing a suited *calibration model* (Section 1.4).

1.3.2 Post-Deployment Calibration

Post-deployment calibration is used for counteracting error sources that impede a consistent performance of a calibration model over time or in the actual deployment environment. These error sources are either heavily deployment-dependent, such as harsh environmental conditions, or due to signal drift, which commonly occurs in long-term deployments. During post-deployment calibration, large numbers of sensors with irregular access to reference measurements need to be calibrated. This is achieved by applying the calibration models extracted from pre-deployment calibration to different *network re-calibration strategies* (Section 1.5).

In Section 1.4 and Section 1.5 we outline the existing calibration approaches, which are found in literature and used in low-cost air pollution sensor deployments.

1.4 Calibration Models

Calibration models are applied in both pre-deployment and post-deployment calibration. We start with the basic and fundamental model, i.e., offset and gain calibration, in Section 1.4.1. Building on this basic model, Section 1.4.2 presents a first extension that corrects for temperature and humidity effects. Finally, Section 1.4.3 summarizes an additional extension of the previous two models by also considering potential interference from other pollutants.

A calibration model takes the raw measurements of a low-cost sensor and transforms them to calibrated measurements, leveraging prior knowledge, e.g., datasheets, or additional information, e.g., measurements from auxiliary sensors. Various mathematical methods can be applied and calibration models may vary for different types of sensors. Calibration parameters can be derived through measurements either in a laboratory setup (controlled environment) or in the field next to reference monitoring sites (observed environment).

1.4.1 Offset and Gain Calibration

Offset and gain calibration tackles calibration errors due to dynamic boundaries and systematic errors and removes potential non-linear responses. It is one of the most essential calibration models that maps the raw sensing measurements to a target pollutant concentration.

Principles. Offset and gain calibration fits a calibration curve, either a linear or a non-linear one, to model relationships between raw sensor readings and pollutant concentrations. The calibration curve is defined by an offset term, i.e., the sensor's response to complete absence of the target pollutant, and a gain term that characterizes the sensor's response to increasing pollutant concentrations. Optimal offset and gain parameters capture the behaviour of a sensor within its sensitivity range, i.e., the dynamic boundaries, and remove systematic errors attributed to poorly fitted calibration parameters.

Methods. The most popular methods to calculate offset and gain terms are ordinary least squares for a linear calibration line and non-linear curve fitting, for instance with an exponential [ANSY15] or power law [DKC+15] gain term. Offset and gain calibration can be performed in both lab and field setups.

Laboratory tests. One way to acquire a calibration curve is to expose a sensor to various target pollutant concentrations in a controlled laboratory setup. Austin et al. [ANSY15] expose a low-cost PM sensor to different aerosol air mixtures in an air-tight enclosure. The gathered measurements are used to calculate a calibration curve defined by an offset and an exponential gain term. Castell et al. [CDS+17] follow a similar approach and calibrate different electrochemical sensors by exposing them to five different gas mixing ratios. Their sensors show high correlation ($R^2 \geq 0.92$) and, thus, a simple linear calibration based on ordinary least squares was used to adapt the offset and gain terms. Similar laboratory calibration can be found in additional works [CLL+14, BL11]. For certain commercially available low-cost sensors an initial laboratory calibration is already performed in the factory. Manufacturers usually follow similar

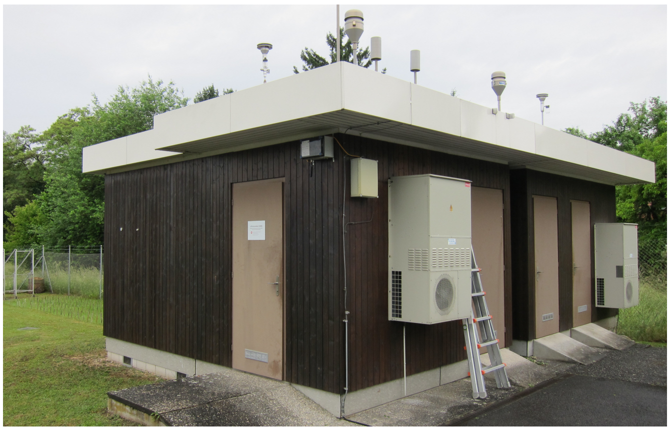


Figure 1.4: Governmental monitoring station located in a suburban area in Switzerland.

approaches as found in the literature and either provide the sensor's response over a range of target pollutant concentrations [Alp13b] or in the form of a calibration curve recorded in a laboratory setup [SGX08].

Field tests. Various recent works propose to directly calibrate their sensors in an environment that is similar to the final deployment. The most prominent way is installing the sensors under test next to high-end sensors. For instance, Dacunto et al. [DKC+15] jointly deploy a low-cost $PM_{2.5}$ sensor with a high-end device in different indoor locations. In outdoor deployments the most prominent approach is to install the sensors under test directly next to governmental monitoring stations that often feature a variety of accurate pollution sensors. For instance, Figure 1.4 shows a monitoring station of the governmental air quality monitoring network NABEL [Nyf01] in Switzerland. Spinelle et al. [SGV+15, SGV+17] deploy 17 different low-cost gas sensors next to high-quality sensors of a air quality monitoring station in a semi-rural area. Carotta et al. [CMC+01] deploy different MOX sensors next to a monitoring station located at a high-traffic road and next to one in a low-traffic intensive area. The highly accurate measurements from these monitoring stations are used to train and evaluate the calibration of the low-cost sensors.

Discussion. While laboratory setups are faster than field tests, many researchers [MPH15, CDS+17, CMC+01, PXM+14] recommend field tests for offset and gain calibration. In a laboratory setup, the environmental conditions during exposure are typically held constant, e.g., at room temperature and moderate relative humidity. Further, the chamber is usually filled with clean air mixed with the target pollutant concentration, i.e., without possible interference from other pollutants. In contrast, field tests allow the sensors to be exposed to situations with realistic environmental conditions, e.g., changing meteorological parameters or interfering gases. Because the sensors are exposed to realistic pollution concentrations the parameters can be optimized to capture the behaviour of the sensor within expected concentration ranges, i.e., with respect to the dynamic boundaries. For instance, Castell et al. [CDS+17] calculate an offset of their calibration curve around 1 ppb (parts-per-billion = 10^{-7}%) in a laboratory calibration and around 166 ppb in a field calibration for a CO sensor. By re-calibrating the CO sensor, i.e., adapting its offset term in the field, they finally reduce the measurement error from 181 ppb by over a factor of 2 to 87 ppb. Zimmermann et al. [ZPK+18] show similar results with four different sensors. Offset and gain calibration models calculated in a laboratory perform poorly in an outdoor deployment and are not in line with re-calibrated models.

As explained in Section 1.1.2, errors of air pollution sensors can be environment-dependent. *In-field* offset and gain calibration *implicitly* mitigates the impact of these errors. However, environmental conditions are complex and subject to short- and long-term changes. As a result, simple offset and gain calibration achieves significantly worse results in field than in laboratory tests. For instance, Castell et al. [CDS+17] observe a drop of $R^2 = 0.99$ to 0.3 of a NO_2 sensor when moving from laboratory to field tests. To *explicitly* account for these environmental conditions information about temperature, relative humidity and interfering gases as well as advanced calibration models are needed, as we will describe in Section 1.4.2 and Section 1.4.3.

1.4.2 Temperature and Humidity Correction

Temperature and humidity correction augments air pollution measurements with concurrently measured temperature and humidity readings to calibrate the low-cost air pollution sensor.

Principles. The motivation of temperature and humidity correction stems from the influence of different temperature or relative humidity settings on sensors observed in laboratory tests. Pang et al. [PSL+17] observe a relative drop in sensitivity of roughly 20% for electrochemical sensors when the relative humidity is increased from 15% to 85%. A similar observation is made by Wang et al. [WYZ+10] for a metal oxide sensor. The sensor almost completely loses its sensitivity when changing from dry air to an extreme relative humidity of 95%. Wang et al. [WLJ+15] demonstrate that increasing humidity can lead to an overestimation of the particle number of typical low-cost light scattering sensors. Similar sensitivity losses are also experienced under changing ambient temperature as summarized by Rai et al. [RKP+17]. These results make it evident that changing environmental conditions such as temperature and humidity need to be incorporated in the calibration process in order to improve the overall measurement accuracy of virtually any low-cost air pollution sensors.

Methods. Temperature and humidity correction is ubiquitous due to the availability of cheap and small but precise low-cost temperature and humidity sensors. Most works include these additional measurements in their calibration methods, and extend the single-variant mathematical models in offset and gain calibration (Section 1.4.1) to the corresponding multi-variant models.

A simple approach found in most of the investigations is to find the linear combination of raw air pollution, temperature and humidity sensor measurements that best captures the target reference concentration. The results in [HPSS14, PXM$^+$14, JHW$^+$16, SWN17, MZK$^+$17, EK12, HIVF$^+$18] all use multiple least-squares to calculate this combination and show beneficial results for any type of low-cost sensor. Different approaches apply more complex methods to model the impact of temperature and humidity. Masson et al. [MPH15] derive a detailed model that captures the physical effect of ambient temperature on their MOX sensor. Popoola et al. [PSMJ16] develop a temperature baseline correction algorithm for electrochemical sensors. They observe notable differences in temperature sensitivity for carbon monoxide (CO) and nitrogen oxide (NO) sensors. While the CO sensor showed a linear relationship to its reference, the NO sensor exhibits a strong exponential relationship. Therefore, they model the reaction to temperature with a linear line fit for the CO sensor and an exponential curve fit for the NO sensor, which is used to correct the corresponding sensor signal. They are able to show a significant improvement for the NO sensor by improving the correlation from $R^2 = 0.02$ to $R^2 = 0.78$. Tsujita et al. [TYIM05] and Sohn et al. [SAZP08] similarly model the relationship of MOX sensors to humidity and temperature with exponential terms and compensate for them by fitting a calibration curve.

Discussion. The extensive list of different sensors that significantly improve their accuracy after temperature and humidity correction underlines the severity of the problem. Temperature and humidity correction needs to be performed for any air pollution sensor regardless of its underlying sensing principles. In rare cases, the impact of ambient conditions can be precisely modelled using chemical process theory. This approach, however, requires deep knowledge of the underlying sensing principle, e.g., physical properties of the metal oxide sensing layer. Therefore, simpler data driven methods dominate the different calibration methods. Due to the popularity of the problem, recent low-cost sensors, especially fully digital sensor solutions, already integrate an internal temperature and humidity correction [SGX08, Aer16]. However, the various field calibration works emphasize the benefit of directly compensating for temperature and humidity dependencies. Thus, it becomes evident that static correction schemes by manufacturers or laboratory calibration may be replaced by in-field calibration for optimal performance.

1.4.3 Sensor Array Calibration

Sensor array calibration is a generic extension of temperature and humidity correction that tackles another environment-dependent factor, interfering gases.

Principles. As described in Section 1.4.1, laboratory tests are usually performed by exposing a sensor to clean air that is mixed with the target pollutant. In most real-world deployments the air mixture is composed of multiple different components [MPS+13]. For instance, multiple pollutants appear concurrently at diverse concentrations in outdoor and common indoor air. These complex air mixtures particularly pose a substantial challenge for gaseous pollutant sensors. Instead of being selective to one single pollutant, low-cost sensors are typically sensitive to multiple pollutants at the same time with different intensities [LCL+12, WYZ+10]. This low-selectivity problem is also referred to as cross-sensitivity and, broadly put, equivalent to the temperature and humidity dependency, i.e., different factors in the environment are influencing a sensors response. Thus, the basic concept is the same as the temperature and humidity correction but often requires more complex methods.

By concurrently measuring all the cross-sensitivities it is possible to compensate for all interfering pollutants. This approach requires a sensor array, i.e., multiple different jointly deployed low-cost sensors. One option to create a sensor array is to install multiple sensors in a box to ensure common air sampling. Note that the majority of sensor arrays also include temperature and humidity sensors and, thus, in this case sensor array calibration is also performing a temperature and humidity correction.

Methods. Popular sensor array calibration methods can be divided in multiple least-squares and neural networks. For certain cross-sensitivity problems a multiple least-squares regression can be successfully used for calibration. One of the most popular examples is the cross-sensitivity of NO_x electrochemical sensors on O_3 concentrations [SGV+15], and vice-versa [PSL+17]. Pang et al. [PSL+17] are compensating for potential influences of ambient NO and NO_2 concentrations on the signal of an O_3 electrochemical sensor. The NO and NO_2 concentrations are, however, measured by a high-end sensing device. The effect of the two cross-sensitivities follows a linear behaviour and, thus, a linear multiple least-squares calibration can be successfully applied. Another investigation [FB17b] follows a similar approach, but compensates for the cross-sensitivity to O_3 of a NO_2 electrochemical sensor. The O_3 measurements are measured by another low-cost oxide sensor.

In more complex cases, linear calibration models do generally not perform well [SGV+15, SGV+17] and, therefore, different authors investigate the feasibility of non-linear calibration models, mostly based on neural networks or related machine learning methods. Spinelle et al.

[SGV+15, SGV+17] show for a wide range of low-cost gas sensors an overall better performance of neural- network-based sensor array calibration compared to multiple least-squares and particularly to an offset and gain calibration based on ordinary least squares. For multiple O_3 and NO_2 sensors the coefficient of determination R^2 is improved from values below 0.3 to at least 0.85 and 0.55, respectively, using neural networks instead of linear models. They also show that for some sensors, in particular metal oxide CO and electrochemical NO sensors, the cross-sensitivity limitation appears to be too severe and could not be solved by calibration with reasonable performance. Similar results are reported by De Vito et al. [DMP+08, DPMF09, DES+18], Esposito et al. [EDS+16, EDS+17, EDS+18], Lewis et al. [LLE+16], Barakeh et al. [BBR+17] and Zimmermann et al. [ZPK+18]. Different types of machine learning techniques, with the majority being neural networks, are able to resolve cross-sensitivities of commercial low-cost sensors with the help of sensor array calibration.

Discussion. Compared to the other two calibration models, sensor array calibration is not a necessity for all sensors. The necessity of sensor array calibration mainly depends on the sensitivity profiles of low-cost sensors and the target pollutant. For instance, O_3 can in general be accurately measured with a single low-cost sensor due to the aggressive nature of ozone, which in return simplifies the development of selective sensing principles. Other pollutants, for instance NO_x, are affected by the presence of aggressive interference factors and complicate the design of selective sensors. These two interacting factors pose a substantial challenge in choosing the optimal sensor array composition, i.e., what low-cost sensors are required to accurately measure the target pollutant. Therefore, various works [DPMF09, ZPK+18, CWL+17] present a thorough analysis on which sensor array composition achieves the best performance in terms of measurement accuracy, precision and stability. Such an analysis requires concurrent data of multiple different low-cost sensors that need to be tested on their feasibility in different sensor arrays. In some cases, the available low-cost sensors may not suffice for a successful array due to unresolved cross-sensitivities [SGV+17]. Thus, finding the optimal sensor array to tackle all cross-sensitivities remains an open problem. Further, similar to the two previous models authors agree that pre-deployment sensor array calibration needs to be performed in the field. The complex composition of pollutants in outdoor air requires the sensors under test to be exposed in their target deployment for a successful calibration.

1.4.4 Comparisons of Calibration Models

In summary, the most essential calibration model that is necessary for all types of sensors is a simple offset and gain calibration, i.e., mapping the raw sensor measurements to a pollutant concentration.

Popular mathematical methods are linear regression or simple curve fitting possibly incorporating a non-linear gain term. Due to the severity of the environmental dependency problem extending the basic model with a temperature and humidity correction becomes indispensable in order to significantly improve the measurement accuracy of any low-cost sensor. The correction can easily be done by concurrently measuring environmental parameters and including them in multi-variable methods, such as multiple least-squares or non-linear curve fitting. Finally, additional environmental influences from interfering gases can be eliminated by incorporating sensor array calibration techniques. Cross-sensitivities are mostly problematic for electrochemical and metal oxide sensors and heavily deployment-dependent. Sensor array calibration requires concurrent measurements from different low-cost sensors and often sophisticated machine learning methods to capture the complex relationship between multiple cross-sensitive sensors and the target pollutant concentration. Overall sensor array calibration has been shown to produce most accurate data. Spinelle et al. [SGV+15, SGV+17] evaluate the performance of the three different calibration steps with different gas sensors. For instance, the NO concentration measured by a calibrated sensor array achieves 15 and 41 times lower measurement errors compared to a single NO sensor with and without temperature correction, respectively. Similar results are shown by Zimmermann et al. [ZPK+18]. Their sensor array calibration based on both linear and non-linear methods achieves an almost one order of magnitude lower error than a simple laboratory offset and gain calibration for four different types of sensors.

The number of additional sensors and the amount of measurements needed to learn the model parameters increase with the complexity of calibration models. Compared to the other two calibration models, sensor array calibration also requires more training samples, i.e., covering a large range of different outdoor situations and, thus, is more time-consuming and complex to perform. De Vito et al. [DMP+08] show a clear positive trend of accuracy and precision with increasing training data. Finally, they achieve a stable calibration with training data collected over 100 days. These long training efforts are, however, justified in order to achieve high data accuracy during long-term deployments possibly spanning multiple years.

Note that a prerequisite to apply calibration models is the access to a highly accurate reference. A reference is usually available in lab or field tests before actual deployment of air pollution sensors. However, the sensors *after deployment* may have irregular access to a reference, which requires additional calibration strategies, as we will discuss in the next section.

1.5 Network Calibration

Low-cost sensors are usually deployed in either a *static* or *mobile* sensor network for long-term air pollution monitoring. Even after pre-deployment calibration, these sensors need periodic re-calibration due to sensor drift over time and changes in the target environment. Some works report a significant drift after already one month of deployment [MMH17]. Thus, re-calibrating sensors appears to be an absolute necessity in any long-term deployment.

An important commonality of post-deployment calibration is the lack of reference sensors to verify and potentially re-calibrate low-cost sensors. This section summarizes existing *network re-calibration* methods, which calibrate a network of sensors with irregular or even no access to a highly accurate reference. We group the existing literature into three fundamental network calibration approaches, i.e., blind (Section 1.5.1), collaborative (Section 1.5.2) and transfer (Section 1.5.3) calibration, based on their assumptions or usage of virtual references. Note that calibration in sensor networks is a general problem and, thus, some of the presented methods can also be directly applied or adapted to other types of sensor network applications consisting of temperature and relative humidity sensors [WYL+16], microphones [SITN17] or barometers [YGTL14].

1.5.1 Blind Calibration

The concept of *blind calibration* [BN07] or *macro calibration*, is originally designed for general sensor networks and has also been applied to temperature and relative humidity sensor networks [BN07, WYL+16]. The idea is to achieve a high similarity between measurements of all sensors in a network. A key assumption is that neighbouring sensors measure almost identical values, or are at least correlated. This assumption is often not true for air pollution monitoring deployments. First, air pollution is known to be a highly complex system with large spatio-temporal gradients. Second, typical inter-device differences of low-cost air pollution sensors hinder equal measurements even in a dense small-scale network. As a result measurements of air pollution sensors in a large-scale deployment are in general neither identical nor necessarily correlated. A more practical assumption is to exploit situations in space and time where we can safely assume that all sensors within the given deployment measure the same pollution concentrations.

Tsujita et al. [TYIM05] installed a low-cost NO_2 sensor in the city of Tokyo, Japan. They recognize that the major error source of their sensor appears to be a baseline drift of the calibration parameters over time. Because they continuously install their sensor at different locations where no accurate governmental stations are deployed, they propose an auto-calibration method. The sensor can be calibrated to reference stations that are not necessarily in their spatial vicinity when one can

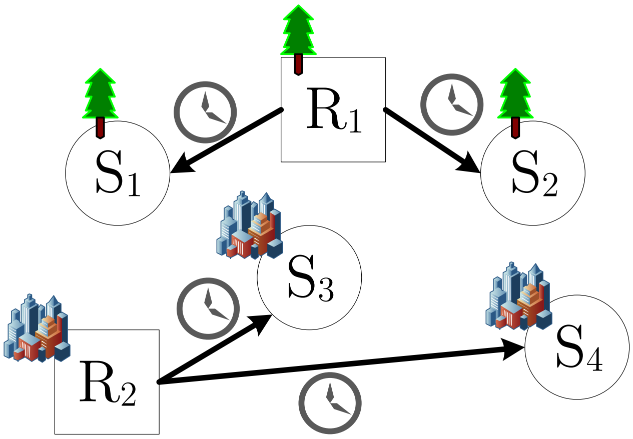


Figure 1.5: Blind calibration scenario with rurally located sensors S_1, S_2, rural reference R_1, urban sensors S_3, S_4 and urban reference R_2. Sensors that are located in similar areas (rural or urban) are calibrated to references in similar areas during times when it is safe to assume that all sensor measurements are identical.

safely assume that the NO_2 concentration is almost identical at any point in the deployment region. To check these circumstances they use NO_2 measurements from four different monitoring stations and re-calibrate the offset term of their low-cost sensor as soon as all four stations report a NO_2 concentration below 10 ppb. A similar method is also applied by Pieri et al. [PM16]. A slightly adapted approach is presented by Moltchanov et al. [MLE+15]. Instead of assessing the possibility of a uniform concentration with reference measurements, they use specific time periods. In order to calibrate low-cost O_3 sensors they assume that the O_3 concentration is uniform during night time (01:00-04:00 AM), when local emissions of precursors, e.g., NO_2 traffic emissions, are negligible. During these time periods they calibrate six O_3 sensors to the reference measurements of one monitoring station. Because O_3 usually reaches concentrations close to zero during night, this approach again only allows for an offset re-calibration. Finally, Mueller et al. [MMH17] also divide their low-cost sensors in two groups, i.e., sensors that measure traffic-related pollution variations deployed in inner city areas and background pollution sensors in outer city areas. This scenario is also illustrated in Figure 1.5. They assume that at inner city locations O_3 and NO_2 concentrations are usually uniform during night and at outer city locations during the afternoon. Individual sensors installed in the inner city are then calibrated to a remote monitoring station in the inner city during night-time and correspondingly for sensors located in the outer parts of the city in the afternoon.

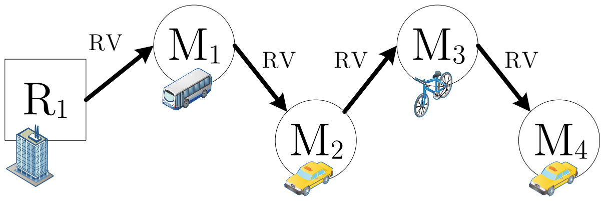


Figure 1.6: Collaborative multi-hop calibration scenario exploiting sensor rendezvous (RV) between static reference R_1 and mobile sensors $M_{\{1,2,3,4\}}$. Whenever sensor M_1 is in the vicinity of the reference R_1 the low-cost sensor can be calibrated. In return, the freshly calibrated M_1 is calibrating M_2 during rendezvous, and so forth.

1.5.2 Collaborative Calibration

Collaborative calibration extends blind calibration by creating virtual references where two mobile sensors meet in space and time such that they should measure the same physical phenomena. The basic idea of collaborative calibration is to exploit situations where two or more *mobile* sensors meet in space and time, i.e., referred to as sensor rendezvous. The notion of sensor rendezvous can also be found in other sensor network problems, such as energy efficient data collection [XWXJ08] or sensor fault detection [SHWT14]. Further, collaborative calibration exploiting sensor rendezvous is also used in other sensor networks, e.g., crowdsensing applications using microphones [SITN17] or barometers [YGTL14].

Sensor rendezvous can be utilized as references for calibrating mobile air pollution sensors. Sensors in a rendezvous are assumed to sense the same physical air and the range of a rendezvous can be empirically determined. For instance, Xiang et al. [XBP+12] define a distance of at most 2 m between two sensors to constitute a rendezvous in an indoor air pollution monitoring deployment. Saukh et al. [SHT15] show that a distance of 50 m in urban outdoor deployments is a reasonable upper limit. Whenever a mobile low-cost sensor is in a sensor rendezvous with a highly accurate sensor, e.g., from a governmental monitoring site, the low-cost sensor can use the reference measurement for calibration [SHT15].

Arfire et al. [AMM15] apply a non-linear temperature correction for mobile electrochemical sensors in a collaborative fashion with a reference sensor. Hasenfratz et al. [HST12] present three different calibration methods based on weighted least squares that also incorporate the age of measurement at the time of the calibration parameter calculations. The methods in [HST12] are also applied by Budde et al. [BEMRB13] to calibrate PM sensors in a participatory sensing scenario. These methods assume that a sensor is in rendezvous with one or more reference sensors multiple times under different conditions so that the sensor can collect a calibration dataset with high variance for calibration.

Unfortunately, not all sensors are necessarily in rendezvous with

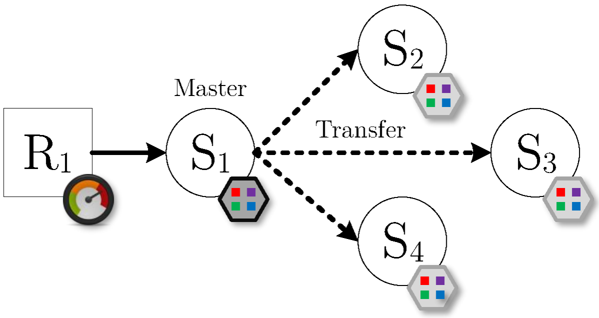


Figure 1.7: Transfer calibration scenario between a reference sensor R_1 and sensors $S_{\{1,2,3,4\}}$. In a first step, sensors $S_{\{2,3,4\}}$ are standardized to a master sensor S_1 in order to achieve high similarity of raw measurements. In a second step, a calibration model acquired by the master S_1 with reference R_1 is transferred to all other sensors $S_{\{2,3,4\}}$.

reference sensors frequently enough. As a consequence, some sensors in the network cannot be re-calibrated. Therefore, some works additionally exploit rendezvous between a freshly calibrated and an uncalibrated low-cost sensor. In this case, a sensor that has been freshly calibrated is used to calibrate an uncalibrated one, e.g., a sensor that has no rendezvous with references. In return, the second freshly calibrated sensor can also be used to calibrate others, and so on. Calibration is therefore performed in a chain-like fashion and, thus, this concept is also known as multi-hop calibration. A typical multi-hop calibration chain is illustrated in Figure 1.6. Although multi-hop calibration allows to calibrate more sensors compared to calibration exclusively with references, it also poses multiple challenges. The most severe challenge is error accumulation over multiple hops, first reported by Hasenfratz et al. [HST12] and in detail evaluated by Saukh et al. [SHT15] and Kizel et al. [KESN+18]. Due to the nature of least-squares-based calibration models at every hop of the calibration chain calibration errors are accumulated. To counteract this error accumulation, Saukh et al. [SHT15] propose to use an alternative method, i.e., the geometric mean regression. It is not suffering from error accumulation, theoretically and practically proven, and is successfully used for offset and gain calibration of a real-world air pollution network. Additional challenges of multi-hop calibration are tackled in [FRD17, MBS+18]. Fu et al. [FRD17] study the effect of reference sensor placement on the performance of multi-hop calibration and present an algorithm to optimally design a practical deployment of static reference and mobile low-cost sensors. A privacy-reserving multi-hop calibration scheme for participatory and crowd sensing deployments is introduced by Markert et al. [MBS+18].

1.5.3 Transfer Calibration

The third group of network calibration methods is known as transfer calibration. It has its origins mainly in industrial deployments using electronic noses (e-noses), i.e., metal oxide sensor arrays for hazardous odour detection. Although the related work mainly focuses on e-nose calibration, transfer calibration can be applied to any sensor model. E-noses are typically calibrated by neural networks to detect multiple different odours or gases with one calibration model. Training such a neural network requires a lot of effort mainly due to training sample collection and model optimization. Metal oxide sensor arrays do typically not produce identical responses compared to similar arrays, even coming from the same production batch [ZTK+11], i.e., there are significant inter-device differences for e-noses. Therefore, each e-nose needs to be calibrated independently and mass production becomes an almost impossible task. Transfer calibration tackles this problem by applying a two-step calibration process. Assuming multiple e-noses, one e-nose acts as a master sensor. In a first step, all non-master e-noses standardize their raw sensor array signals individually to the raw ones of the master. This step is usually performed by linear regression methods, such as robust regression [DKJ+14], ridge regression [YZ16], direct standardization [FFGG+16] or weighted least squares [ZTK+11], and counteracts the inter-device differences. In a second step, the master node calibrates its response to the target gas or odour concentrations, e.g., by training a neural network calibration model [ZTK+11, DKJ+14]. This model is now transferred to all non-master nodes, as illustrated in Figure 1.7. Other popular methods used in the second step are support vector machines regression [YZ16, FFGG+16] or classification-based methods to classify the presence of a certain gas using support vector machines [YZ16, FFGG+16] or logistic regression [YZ16]. Some works also combine the two steps using a global training framework, such as auto encoders by Zhang et al. [ZGY17] or a mixture of multi-task and transfer learning by Yan et al. [YZ15]. Bruins et al. [BGvdS+13] show that the standardization in the first step can also be performed by applying an elaborate heating temperature control of the MOX sensor array.

Since transfer learning only requires one complex calibration process for the master sensor array, it is clearly able to minimize calibration efforts in large-scale deployments. Unfortunately, transfer learning approaches have mainly been evaluated in lab setups and not yet intensively in real-world deployments. One of the few transfer calibration adaptations using a real-world large-scale PM sensor deployment is presented by Cheng et al. [CLL+14]. In a first laboratory calibration step the PM sensors are standardized to a master sensor using second degree curve fitting. In the second step a neural network is used to perform a temperature and humidity correction. The neural network is constantly updated throughout the deployment. Overall they achieve an increase in approximately 8% measurement accuracy compared to uncalibrated situations.

1.5.4 Comparisons of Network Re-Calibration Strategies

The three network calibration approaches all rely on different assumptions and fundamental design choices and, thus, also have different advantages. The least complex method based on blind calibration exploits time periods and locations of reference and low-cost sensors for calibration to assure that all sensors generate identical measurements. While this approach can be applied to almost any type of sensor in almost any deployment, the opportunities for calibration are generally sparse and, hence, only offset and gain calibration can be successfully performed.

In order to increase the opportunities for calibration, collaborative calibration exploits meeting points or rendezvous between sensors. Consequently, collaborative calibration can only be applied to mobile sensor deployments. Depending on the mobility of the sensors it might not be possible to calibrate all sensors within the network, e.g., a sensor with no rendezvous can not be calibrated. So far it is unclear how collaborative calibration scales with the network size. This is not a substantial problem for the other two methods.

Finally, transfer calibration uses a two-step approach by first standardizing all deployed sensors to a master sensor and then transferring calibration parameters acquired by the master to all sensors. Transfer calibration has no restrictions on the possible calibration models or the mobility of sensor, with the exception of the static master sensor next to a reference. However, transfer calibration assumes that all sensors in the network *(i)* drift in an equal way as the master node and *(ii)* are equally affected by environmental conditions. These two assumptions are in general not true in typical air quality monitoring networks [ZTK+11]. Therefore, up to now transfer calibration has not achieved satisfactory performance. Further, there is only little experience in real-world deployments.

Overall, all the methods have been proven to be successful in counteracting decreasing accuracy in their specific long-term deployments. In general the average measurement accuracy is increased after re-calibrating a sensor network and, thus, the existing results point out the necessity of re-calibration. However, the different strengths and weaknesses of the three methods present the need for an universal network calibration method. Currently, there is no one-size-fits-all network calibration solution available. Recent research efforts investigate the possibility of a generally applicable network calibration method, e.g., by combining different aspects from the three methods. Some theoretical investigations already provide mixtures of different models. For instance, Dorffer et al. [DPDR15, DPDR16a, DPDR16b] combine the two ideas of blind and collaborative network calibration to increase the possibilities for sensor re-calibration. A key benefit of enhancing and mixing different network calibration aspects will thus help to assure that all sensors in a network can be calibrated.

1.6 Thesis Outline and Contributions

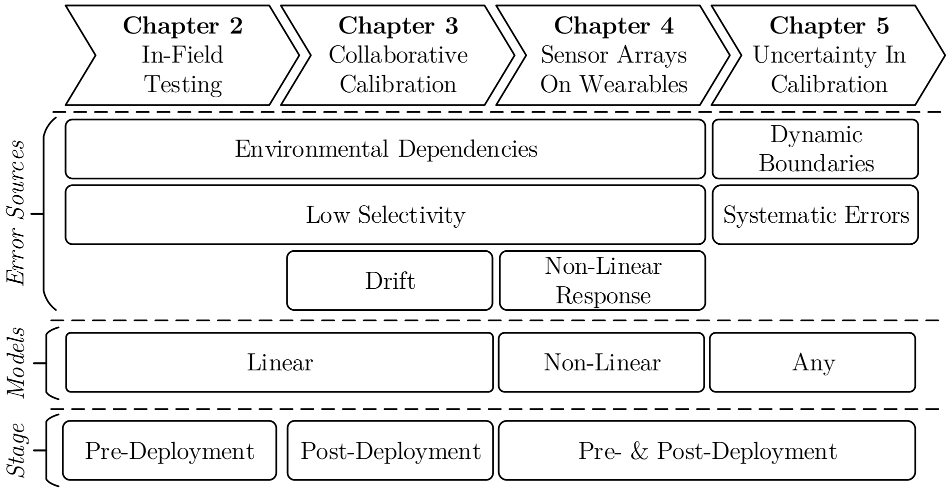


Figure 1.8: Overview of the error sources, the types of calibration models and the stage of the deployment discussed in each chapter of this thesis.

This thesis presents novel calibration techniques and aims to push the boundaries towards accurate measurements from low-cost air pollution sensors. As summarized in Chapter 1, low-cost air pollution sensor deployments are becoming more and more popular, however the collected data is not yet of sufficient quality for beneficial applications. Different error sources impact low-cost sensors and lead to a substantial body of related work discussing different calibration techniques counteracting the errors. We extend these related works and propose novel methodologies and techniques by tackling the different error sources presented in Section 1.2 and designing tailored calibration models, which can be applied at different stages of a low-cost air quality sensor deployment. The overview of this thesis and the specific topics are displayed in Figure 1.8. Specifically, we present a pre-deployment testing method to model sensor cross-sensitivities and to augment optimal sensor arrays (Chapter 2), design a collaborative multi-hop calibration algorithm for sensor arrays (Chapter 3), uncover and tackle non-linear error sources by human emissions in wearable use-cases (Chapter 4) and devise a scheme to estimate uncertainties of calibrated measurements, which is used to further improve collaborative calibration (Chapter 5).

In the following, we present the main contributions of each individual chapter.

Chapter 2: Pre-Deployment Testing, Augmentation and Calibration of Cross-Sensitive Sensors. In this chapter we tackle the error sources *low selectivity* and *environmental dependency* and show the importance of performing *in-field* and *pre-deployment* testing of low-cost air pollution sensors. As shown in Section 1.4.3, cross-sensitivities and environmental dependency are two major limiting factors of low-cost air pollution sensors. Existing works successfully show how different calibration models can be used to calibrate sensor arrays and eventually improve measurement accuracy. However, in order to counteract these error sources one needs to build an appropriate sensor array. Unfortunately, information on which cross-sensitivities and to what extent they are affecting a given sensor is often missing or is not relevant for the target deployment.

- We design a novel approach to perform in-field and pre-deployment testing of low-cost air quality sensors. The approach is able to uncover substantial cross-sensitivities and conclude about the usability of a sensor for the target deployment. We use the gained information to construct a sensor array that counteracts substantial cross-sensitivities and environmental dependencies by calibration using *linear* multiple least-squares.

- We demonstrate the feasibility of our approach and the importance of sensor arrays to resolve cross-sensitivities on real-world low-cost sensor data. We are able to uncover different sensitivity profiles of the sensors under test and improve the data quality of our constructed sensor arrays.

Chapter 3: Multi-Hop Calibration for Mobile Sensor Arrays. This chapter investigates an additional error source, *drift*, or in general, changing sensor behaviour over time. As summarized in Section 1.5, there exist different methodologies that are able to frequently re-calibrate sensors during their deployment. Unfortunately, they all have different strengths and weaknesses. While collaborative multi-hop calibration has been proven to be able to calibrate large numbers of mobile sensors in a deployment, the state-of-the-art calibration model designed for multi-hop calibration only allows a simple offset and gain calibration.

- We design the first algorithm for collaborative multi-hop calibration in a mobile sensor array deployment. Our method is able to counteract cross-sensitivities, environmental dependencies and signal drift while minimizing error accumulation over multiple hops.

- Using different datasets we show that our method outperforms various other calibration models and is able to calibrate substantially more sensors compared to one-hop calibration.

Chapter 4: Enabling Personal Air Pollution Monitoring on Wearables. This chapter focuses on personal air pollution monitoring using wearables. We integrate a sensor array into a personal air quality monitor to quantify the immediate exposure to air pollution of a user. This special use case of air pollution sensor arrays is causing another error source: *non-linear response*.

- We show that natural human gas emissions can impact low-cost metal oxide sensors when equipped in wearables. We investigate this human interference and its effect on the sensors and highlight a non-linear response.

- We build a wearable prototype featuring a *pre-* and *post-deployment* calibration model facilitating *non-linear* neural networks and semi-supervised learning. Our approach allows to tackle the human interference and to recover accurate ambient air pollution values while reducing training and updating efforts of the calibration model before and during deployment.

Chapter 5: Enhancing Sensor Calibration with Uncertainty Estimates. In the last chapter we present a way to determine how confident we can be in calibrated measurements from any calibration model. A calibration is never error-free, especially error sources like *dynamic boundaries* or *systematic errors* can introduce large inaccuracies. Creating a notion on how severe these inaccuracies are, i.e., how certain we can be about a measurements accuracy, is thus of great interest.

- We develop a scheme that is able to estimate two major uncertainties sources, aleatoric and epistemic, for *any calibration model*.

- We integrate the uncertainties into multi-hop calibration by developing an uncertainty-based data filtering at each hop in the network. Finally, we are able to improve the measurement accuracy of a real-world multi-hop calibration setup by selecting measurements for calibration based on our uncertainty metrics.

2

Pre-Deployment Testing, Augmentation and Calibration of Cross-Sensitive Sensors

Over the past few years, many low-cost air pollution sensors have been integrated into measurement platforms for air quality monitoring. In Section 1.2, we list various different limitations and error sources that pose a significant challenge on using these sensors in real-world applications. In this chapter, we specifically focus on the following three limitations: concentrations of dangerous pollutants in ambient air often lie at the boundaries of a sensor's dynamic range, environmental conditions affect the sensor signal and the sensors are cross-sensitive to multiple pollutants. As highlighted by different authors, see Section 1.4.1 for a short summary, these error sources are responsible that typical low-cost sensors do not perform well when they are deployed in the field without any post-processing, e.g., calibration, of their signal. Unfortunately, datasheet information on these effects is often scarce or may not cover deployment conditions. Consequently, the sensors need to undergo extensive pre-deployment testing to examine their feasibility for a given application and to find the optimal measurement setup, i.e., a sensor array, that allows accurate data collection. By creating an optimized sensor array we are able to apply the two important calibration models, temperature and humidity correction and sensor array calibration presented in Section 1.4.2 and Section 1.4.3 and, consequently, counteract the error sources. This procedure enables accurate measurements of low-cost air pollution sensors and provides an important basis for effective post-deployment calibration methods in long-term deployments, which we will discuss in Chapter 3.

In this chapter, we propose a novel method to conduct in-field testing of low-cost sensors. The proposed algorithm is based on multiple least-squares and leverages the variation of urban air pollution to quantify the amount of explained and unexplained sensor signals. We verify (i) whether a sensor is feasible for air quality monitoring in a given environment, (ii) model sensor cross-sensitivities to interfering gases and environmental effects and (iii) use the acquired information to compute an optimized array and its calibration parameters for stable and accurate sensor measurements over long-time periods. Finally, we apply our testing approach on 5 off-the-shelf low-cost sensors and 12 reference signals using over 9 million measurements collected in an urban area. We propose an optimized sensor array and show—compared to a simple offset and gain calibration using ordinary least-squares—an up to 45% lower calibration error with better long-time stability of the calibration parameters.

2.1 Introduction

Breakthroughs in sensor technology made a new generation of small, cheap and portable air quality sensors available on the market. As summarized in detail in Section 1.1, they are usually based on optical sensing principles[1], electrochemical cells[2] or semiconductor technologies[3], which allow a compact and inexpensive design. Researchers and start-ups frequently integrate these sensors in their measurement platforms to monitor air pollution. Over the past years, numerous research projects, e.g., CommonSense [AHM$^+$09], MESSAGE [MPS$^+$13], OpenSense [HSW$^+$14], and public initiatives, such as, Air Quality Egg [Wic18] and Data Canvas [Dat14], were launched to explore opportunities of these new technologies and raise awareness in the society. A comprehensive list of successfully realized projects can be found in [YLM$^+$15, Tho16]. However, results of laboratory tests and comparison against precise analytical instruments often report insufficient sensor accuracy, sensor drift and low correlation with reference measurements when trying to measure pollutant concentrations in ambient air [JHW$^+$16, PXM$^+$14, EPC12].

In this chapter, we present a novel method to quantify the real-world usability of a low-cost sensor for monitoring urban air pollutants by splitting the sensor's measurements into explained, unexplained and noise components. We show that many low-cost sensors can be used to monitor air quality, if deployed as part of a sensor array and jointly calibrated in combination with other low-cost sensors of the sensor array. Furthermore, we can provide a more stable calibration accuracy for a sensor array compared to a simple offset and gain calibration and, hence, need to calibrate the sensors less often.

[1] e.g., Shinyei PPD42NS Particle Sensor [ANSY15]

[2] e.g., Alphasense CO-B4 [Alp13a]

[3] e.g., MICS-OZ-47 O_3 Module [SGX13]

Challenges. Using low-cost sensors to monitor pollutants in ambient air is challenging due to various error sources, see Section 1.2 for a detailed list of reasons that may lead to the degradation of a sensor's measurement quality. In this chapter, we focus on the following limiting error sources:

1. Measured concentrations are low and often lie at a sensor's sensitivity boundaries, especially at the lower boundary (limit of detection—LOD) of the dynamic range, because many sensors are primarily designed to sense higher pollution concentrations, e.g., in the automotive industry [PPS+99].
2. Environmental conditions—typically temperature and humidity—impact the speed of chemical reactions and, thus, the sensor output.
3. Low-cost air quality sensors often suffer from low selectivity and their response is affected by a wide range of substances in the air, referred to as *cross-sensitivities*.

These three challenges are the most prominent error-sources of state-of-the-art low-cost air quality sensors. Unfortunately, datasheet information on a sensor's specific sensitivity profile is often scarce. Sensor producers may list suggested operating temperature ranges or measured cross-sensitivities per interfering pollutant based on laboratory experiments. These laboratory experiments are typically conducted in a certain fixed setting [Alp13a] and do not reflect the final deployment environment of the sensor. More often manufacturers provide no quantitative sensitivity evaluation at all [Dov, Shi10, Amp19]. For instance, Austin et al. [ARG06] test different electrochemical sensor and uncover 17 different cross-sensitivities. Unfortunately, these cross-sensitivities are not specified or described in a conclusive way in the datasheets of the tested sensors. Consequently, considering only information from datasheets can limit the sensor performance during deployment. Another example is presented by Eugster and Kling [EK12]. They use a methane (CH_4) sensor for rural air monitoring in Alaska and derive a temperature and humidity correction model from information in the datasheet. Due to sparse testing at only three different humidity levels, the appliance of their model is limited to situations with a relative humidity larger than 40%.

Unless the sensors undergo extensive pre-deployment testing, using such sensors to monitor air quality is difficult. Outdoor air composition is complex and can exhibit notable variations over time, both daily and seasonally. Sensor testing and calibration in a laboratory requires exposing a sensor to a wide range of artificially created but feasible gas mixtures. This is labour intensive, requires complicated setup, e.g., environmental chambers and gas mixtures, and assumes a thorough knowledge of the air composition in the target environment and cross-sensitivities of the sensor [PSL+17, CLL+14]. In contrast, testing and calibrating sensors in-field leverages existing high-quality measurement stations deployed outdoors and, thus, the results are more relevant for an outdoor deployment and no complex setup is needed [CDS+17, ZPK+18].

Contributions and road-map. In this chapter, we test and calibrate low-cost sensors in the field by conducting parallel measurements at a high-quality reference station. By analysing obtained data, we can (i) verify whether air pollution monitoring with a given sensor in a given environment is feasible, (ii) characterize sensor cross-sensitivities and construct the sensor array, which can optimally monitor the specific pollutant, and (iii) compute calibration parameters for the sensors in the system. The contributions of this chapter are organized as follows:

- After introducing our assumptions and models in Section 2.3 and summarizing sensor calibration using multiple least-squares [VG08] in Section 2.4, Section 2.5 describes our sensor testing methodology to uncover sensor cross-sensitivities. Given measurements of different reference sensors, we design an indicator that allows quantifying the amounts of captured and uncaptured cross-sensitivities and sensor noise.

- Section 2.6 reports the results of the proposed testing methodology on real data. We use several low-cost sensors available on the market and previously used by other researchers [PXM+14, EPC12] to show that our approach allows to conclude on the feasibility of using a sensor in a given setting. We give positive and negative examples of sensors used for air quality monitoring in current and past projects. Furthermore, we leverage the approach to test and calibrate our urban air quality measurement system [LFS+12] equipped with low-cost cross-sensitive sensors.

2.2 Related Work

This section summarizes existing work on sensor selectivity, testing and calibration.

2.2.1 Testing Cross-Sensitive Sensors

As discussed in detail in Section 1.4.1, sensors are traditionally tested in a fully controlled environment in a laboratory [ARG06, CDS+17, Mor07, MSVA99, VDR+10]. This type of testing allows for (i) fast sensor characterization (ii) for a given interval of interest and (iii) in a controlled environment, e.g., fixed temperature and humidity and no interfering pollutants [LMM14, STP08]. However, different researchers [MPH15, CDS+17, CMC+01, PXM+14, JHW+16] motivate the need of testing the feasibility of a specific sensor in a given scenario. A laboratory setup is not reflecting the typical deployment environment of the sensor and, consequently, the sensors perform poorly once they are deployed and need immediate re-calibration [CDS$^+$17, JHW$^+$16]. We need to uncover the sensor's sensitivity, environmental dependency and operating range and test whether these match the application requirements. Consequently, latest works on air quality monitoring favour in-field sensor testing against reference sensors over laboratory tests [SGV$^+$14, PXM$^+$14]. In-field sensor testing allows sensor evaluation at a wide range of varying environmental conditions and pollutant concentrations under deployment conditions. Therefore, various results on large-scale in-field tests [SGV$^+$14, JHW$^+$16, PXM$^+$14] have been described and typically report identical results: Only a fraction of all the different sensors shows high correlation to their corresponding references. However, while most investigations already perform a temperature and humidity correction, a deeper evaluation of the potential reasons, e.g., cross-sensitivities, is often missing.

This chapter advocates in-field sensor evaluation at reference stations located directly at or close to the deployment site, which assures similar environmental conditions and air composition during testing and deployment of a sensor. We use multiple least-squares (MLS) to decide on sensor qualification for the environment of interest. Further, we draw conclusions about the unexplained part of the sensor measurements by leveraging frequency characteristics of atmospheric phenomena.

2.2.2 Calibrating Cross-Sensitive Sensors

Low-cost sensors usually suffer from substantial deviation when compared to highly accurate references. Sensors need to be calibrated to minimize this deviation. Under the assumption that there is a linear relationship between sensor and reference measurements a widely used approach is to apply univariate linear regression techniques, such as ordinary least-squares (OLS) [HST12]. These techniques, however, only perform well if the low-cost sensor is highly selective to the target gas. If the sensor is affected by cross-sensitivities to interfering gases or depends on meteorological conditions, a multi-variable model should be used instead [Bro03]. These models, however, require collocated measurements of additional sensors—often referred to as sensor array—to compensate for sensor cross-sensitivities. As highlighted in Section 1.4.3, popular approaches to calibrate a sensor array to reference measurements can be classified into multiple linear regression techniques, such as multiple least-squares (MLS) [PSL$^+$17, KBP06, Mor08, ZW09, PXM$^+$14, VG08, LDC18], and artificial neural networks (ANN) based techniques [DPMF09, EDS$^+$16, SGV$^+$15, BBR$^+$17]. While ANNs can be powerful to resolve complex cross-sensitivities, linear calibration methods like MLS are usually less prone to overfitting and in general easier to train and interpret. Due to the typically linear or polynomial response of low-cost air pollution sensors to references, we utilize MLS for testing and calibration in this chapter.

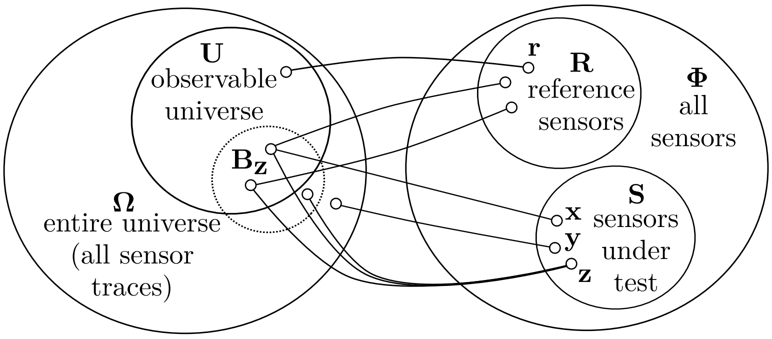


Figure 2.1: Relationship between different sets introduced in this chapter.

2.3 Assumptions and Models

This section introduces the basic terminology and discusses assumptions and models used throughout this chapter.

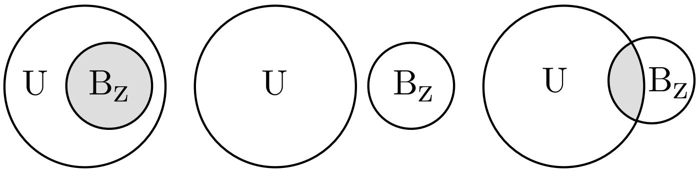
2.3.1 Observable Universe

Let Φ be a set of all sensors one can possibly build. Let $R \subset \Phi$ be a set of available sensors that accurately measure some phenomena of interest. That is, a sensor $r \in R$ accurately measures a single phenomenon, e.g., ambient temperature. The sensors in R can be used as *reference sensors* to test the quality of other sensors, such as low-cost sensors, measuring the same phenomena.

A time-ordered sequence of discrete measurements $m = \{m(t_j)\}$ taken by a sensor at times t_j for $j \in \{1, 2, ..., n\}$ within a time interval $[t_1, t_n]$ is referred to as a *trace*. We consider a measurement as a point measurement, that is, it has no duration. The application scenario of a sensor limits the number of possible traces of that sensor, e.g., an air quality sensor reports different measurements in an automotive industry application than in monitoring outdoor air quality. Consider some fixed scenario of interest, let Ω and $U \subset \Omega$ be sets of corresponding traces produced by the sensors Φ and $R \subset \Phi$, respectively. Ω can be understood as the *entire universe* of all sensor traces and U as the *observable universe* determined by a set of references R, see Figure 2.1.

2.3.2 Low-cost Sensors

Let $S \subset \Phi$ be a set of low-cost sensors under test. We assume no prior knowledge about the sensors. In order to explain a trace of a low-cost sensor y, we relate it to the traces of reference sensors in the observable universe U, i.e., we represent $m_y = \{m_y(t_j)\}$ as a function of traces in U and an *unexplained*—or residual—trace in $\Omega \setminus U$. If m_y is solely representable as an unexplained trace in $\Omega \setminus U$, there is no possibility to explain measurements of sensor y with references in the given scenario. If a trace $m_x \in \Omega$ of a sensor $x \in S$ can be completely explained with a single



(a) Inclusion. (b) Exclusion. (c) Intersection.

Figure 2.2: Scenarios to quantify the amount of the observable universe U captured by measurements B_z.

reference trace m_r in U, one can compare the readings of sensor $x \in S$ to the response of a corresponding reference sensor $r \in R$ and calibrate it if needed.

2.3.3 Cross-sensitivity

Low-cost sensors show in general a close to *linear* or *polynomial response* to reference traces[CDS+17, DKC+15], but typically suffer from offsets and drifts that result in a substantial deviation of their trace from a reference trace. In order to minimize this deviation, the trace m_x of a sensor x needs to be *calibrated* to a reference trace m_r. Calibration is usually performed by representing the calibrated sensor trace $\hat{m}_x$ as a function *cal* of the raw trace, i.e., $\hat{m}_x = cal(m_x)$. An optimal calibration function *cal* minimizes some norm of the difference between the calibrated sensor trace $\hat{m}_x$ and the reference sensor trace m_r.

An interesting sensor testing and calibration challenge arises if the trace of sensor $z \in S$ is a function of multiple traces $B_z \subset \Omega$, also known as sensor *cross-sensitivities*. In this case, calibrating a cross-sensitive sensor z to a reference r requires to describe the calibrated trace as a function of multiple traces, i.e., $\hat{m}_z = cal(m_z, m_{s_1}, m_{s_2}, ...)$ given multiple sensors $s_i \in \Phi$ such that some norm of the difference between $\hat{m}_z$ and m_r is minimized. Given an observable universe U formed by some reference sensors R and a cross-sensitive sensor z, we distinguish three types of relationships between U and B_z as illustrated in Figure 2.2:

Inclusion: $B_z \subseteq U$. The set of available reference sensors R can measure all cross-sensitivities of sensor z. Hence, the sensor response can be fully *explained* by the set of available references R, see also sensor x in Figure 2.1. Moreover, R can be used to calibrate the trace of sensor z to any trace in B_z as is detailed later. In this case, we can unambiguously conclude on sensor quality and perform best-possible sensor calibration.

Exclusion: $B_z \cap U = \emptyset$. There is no relation between the sensor response and the observable universe U. In this case, the sensor response *can not be explained* by means of reference sensors $R \in \Omega$ and hence also not calibrated, see sensor y in Figure 2.1.

Intersection: $B_z \cap U \neq B_z$. The most common case is that only a part of B_z can be explained by U, see sensor z in Figure 2.1. We refer to $B_z \cap U$ as to the *explained* part of z's response and to $B_z \setminus U$ as to its *unexplained* part. The usability of sensor z in a given scenario depends on whether the explained part of the trace dominates its unexplained part.

Low-cost gas sensors are often cross-sensitive, because their small sensing surface area and low power consumption requirements limit the selectivity [THMA03]. Moreover, environmental parameters, such as ambient temperature and humidity, influence chemical reactions [KBBS07] or falsify optical sensing techniques [JLT$^+$18] and, thus, often affect the sensor response. We assume that a measurement of a cross-sensitive sensor is an *additive* combination of different, possibly non-linear effects describing the impact of different *phenomena*, e.g., interfering gases or meteorological effects [Bro03].

2.3.4 Sensor Array

Ignoring sensor cross-sensitivity or environmental parameters leads to poor sensor calibration. Cross-sensitive sensors are usually *augmented* with collocated sensors to a set of sensors $M \subseteq \Phi$, called *sensor array*. Sensor arrays are used to compensate for cross-sensitivities. A cross-sensitive sensor z can be perfectly calibrated to a reference sensor r using a *multiple regression* method, given low-cost sensors in M that cover *all* phenomena in B_z. These multiple regression methods can find the function of pre-processed and aggregated measurements from sensors in M that minimizes its deviation from a reference r. However, the knowledge of B_z is often incomplete or unknown due to scarce datasheet information obtained through basic tests conducted in the laboratory, or there is no information at all. Even if B_z is known (it might consist of multiple relevant phenomena [ARG06]), the quality of sensor tests and calibration of sensor z is limited by the set of available reference sensors R. In this chapter, we give answers to the following questions: (i) Given a cross-sensitive sensor z and references R, how can we identify cross-sensitivities $B_z \cap U$ of sensor z? (ii) How should z be augmented to a sensor array M when using it in a measurement system to improve the measurement quality? (iii) If sensor z is sensitive to phenomena not covered by U i.e., $B_z \cap U \neq B_z$, can z still be reasonably calibrated and used in a given scenario? Since the list of cross-sensitivities is typically long, $B_z \cap U \neq B_z$ presents a common case when dealing with gas sensors.

2.3.5 Test Deployment Conditions

Testing a cross-sensitive sensor z in a laboratory requires simulating common deployment conditions and varying concentrations of every substance in B_z. This is expensive, time and labour-intensive if the list of sensor cross-sensitivities is long. For instance, the datasheet of the

Alphasense NO_2-B4 Nitrogen Dioxide 4-Electrode (2013) sensor [Alp13a] lists 11 possible cross-sensitivities, whereof at least three can have a considerable impact on the sensor response depending on the scenario. In contrast, in-field sensor calibration with parallel measurements with reference sensors R is an alternative and gives the advantage that sensor packaging and deployment conditions are similar to those of the target deployment and environment. The latter is crucial when testing and calibrating low-cost sensors, which often measure at their sensitivity boundaries. In this chapter, we assume that the data collected for sensor testing and calibration is gathered under similar conditions as in the target deployment.

2.4 Sensor Calibration

Sensor calibration is necessary to establish and maintain high sensor data quality. Since the sensor response can be affected by multiple factors, a large body of work tackles the task to compensate for these factors. This section recapitulates sensor calibration based on least-squares [VG08]. Further, we discuss an application example for a cross-sensitive gas sensor.

2.4.1 Ordinary Least-Squares

Regression analysis is often used to calibrate sensor measurements according to a reference trace [CDS+17, HST12, BN07], i.e., to conduct a simple offset and gain calibration (see Section 1.4.1). The common approach is to calibrate a raw sensor measurement $m(t)$ at time t to a given reference sample $m_r(t)$ such that

$$m_r(t) = b_0 + b_1 \cdot m(t) + \varepsilon(t), \tag{2.1}$$

where b_0 and b_1 are calibration parameters describing offset and gain of a *calibration line*, and ε is a regression error component. *Ordinary Least Squares* (OLS) [RTSH08b] regression is typically used to compute estimates of the calibration parameters $\hat{b}_0$ and $\hat{b}_1$ such that the error ε is minimized according to the $L_2 = ||\varepsilon||_2$ norm. A raw sensor measurement $m(t)$ can then be converted to its calibrated version $\hat{m}(t)$ as follows

$$\hat{m}(t) = \hat{b}_0 + \hat{b}_1 \cdot m(t). \tag{2.2}$$

2.4.2 Multiple Least-Squares

The measurements of cross-sensitive sensors are aggregated measurements of multiple phenomena. Consequently, the measurements correlate poorly to measurements of a single reference, i.e., measurements of a single phenomenon. Using the one-dimensional regression model described above to calibrate cross-sensitive sensors consequently also leads to poor calibration accuracy. The standard solution—also known as *multiple regression* [RTSH08a, VG08]—is to include additional regressors $m_l \in M, l \in \{1, 2, ..., k\}$ into the model. The goal of multiple regression is to find coefficients b_i with $i \in \{0, 1, ..., k\}$ of the linear combination of different sensor measurements and compositions thereof $m_l(t)$, which best fits a reference measurement $m_r(t)$ as follows

$$m_r(t) = b_0 + b_1 \cdot m_1(t) + ... + b_k \cdot m_k(t) + \varepsilon(t) \tag{2.3}$$

and accordingly in matrix form

$$m_r = M \cdot b + \varepsilon, \tag{2.4}$$

where $m_r \in \mathbb{R}^{n\times 1}$, $M \in \mathbb{R}^{n\times(k+1)}$, $b \in \mathbb{R}^{(k+1)\times 1}$, $\varepsilon \in \mathbb{R}^{n\times 1}$ and n is the number of samples at times t_j with $j \in \{1, 2, ..., n\}$. The estimates of the regression parameters $\hat{b} \in \mathbb{R}^{(k+1)\times 1}$ are calculated by *multiple least-squares* (MLS) [VG08] and the raw sensor traces M are calibrated by applying

$$\hat{m} = M \cdot \hat{b}. \tag{2.5}$$

2.4.3 Calibration Quality

In the following we list three important metrics to assess the quality of our calibration approaches used in this chapter but also in the remainder of this thesis.

Root-Mean-Square-Error (RMSE). The goal of least-squares based regressions is to minimize some norm of the regression error ε with

$$\varepsilon = \hat{m} - m_r. \tag{2.6}$$

Ordinary and multiple least-squares both minimize the $L_2 = ||\varepsilon||_2$ norm [RTSH08a]. The main metric we use is the *root-mean-square error* (RMSE) between m_r and $\hat{m}$ to evaluate the calibration accuracy of calibrated trace $\hat{m}$ and its corresponding reference trace m_r. RMSE is a standard metric [SHT15, BN07, CBKL10, CLL+14] to quantify calibration quality and is computed as follows

$$\text{RMSE}(\hat{m}, m_r) = \left(\frac{1}{n} \sum_{j=1}^{n} (\hat{m}(t_j) - m_r(t_j))^2 \right)^{\frac{1}{2}} = \text{RMS}(\varepsilon) = \left(\frac{1}{n} ||\varepsilon||_2^2 \right)^{\frac{1}{2}}. \tag{2.7}$$

Standardized RMSE. In order to assess if our calibrated air quality measurements are of sufficient quality for a given application we also adopt a standardized version of the RMSE [TPP12] defined as

$$RMSE_\sigma = \frac{\left(\frac{1}{n} \sum_{j=1}^{n} (\hat{m}(t_j) - m_r(t_j))^2 \right)^{\frac{1}{2}}}{\sigma \cdot \left(\frac{1}{n} \sum_{j=1}^{n} m_r^2(t_j) \right)^{\frac{1}{2}}} = \frac{RMSE}{\sigma \cdot RMS_r}, \tag{2.8}$$

where RMS_r is the root-mean-square value of the ground-truth trace m_r and σ is defined as a relative uncertainty measure for a specific pollutant. For instance, according to the air quality directive by the European Parliament [PEA08], $\sigma = 0.15$ is required for O_3, CO and NO_2 measurements. The $RMSE_\sigma$ acts as a statistical parameter to assess the quality of the measurement. If $RMSE_\sigma \leq 1$ the measurements fulfil the air quality directive and can be used for accurate air quality measurements. In the case where $1 < RMSE_\sigma \leq 2$ the measurement quality is not in line with the air quality directive but might still be used for certain applications. For instance, the data quality allows indicative measurements, i.e., assessing the air quality in terms of pollution levels or an air quality index (AQI) [Val14]. The quality of the measurements is not sufficient for any application if $RMSE_\sigma > 2$. The investigations in [TPP12] point out several flaws of the $RMSE_\sigma$, e.g., it is not concentration level-dependent, and suggest that $RMSE_\sigma < 2$ indicate measurements with adequate quality.

Goodness of fit. Finally, another important metric we apply is the coefficient of determination $R^2 \in [0, 1]$, given by

$$R^2 = \left(\frac{\sum_{j=1}^{n} \left(m_r(t_j) - \mu_r\right)\left(m(t_j) - \mu_{\hat{m}}\right)}{\left(\sum_{j=1}^{n} \left(m_r(t_j) - \mu_r\right)\right)^{\frac{1}{2}} \left(\sum_{j=1}^{n} \left(m(t_j) - \mu_{\hat{m}}\right)\right)^{\frac{1}{2}}} \right)^2, \tag{2.9}$$

where μ_r and $\mu_{\hat{m}}$ are the mean values of the reference trace m_r and the calibrated measurement trace $\hat{m}$, respectively. The R^2 value is a widely used metric to asses the amount of variance in the calibrated measurements that can be explained by the calibration model [RTSH08a]. Values close to 1 indicate a well-fitted calibration model, values close to 0 indicate that there is no correlation between the calibrated and ground-truth measurements.

2.4.4 Application Example: NO_2 Sensor

When calibrating a cross-sensitive sensor, a simple offset and gain calibration by OLS is usually not suited. For instance, when calibrating a nitrogen dioxide (*Alphasense NO_2-B4 Nitrogen Dioxide 4-Electrode (2013)*) sensor from AlphaSense [Alp13a], we use OLS to calculate the calibration parameters based on a training trace of two weeks gathered in February 2014. As reference we use NO_2 measurements from a static, high-quality NABEL reference station (see Figure 1.4 and Figure 2.3) in an urban area in Duebendorf, Switzerland, where our sensors are installed on the roof of the station to ensure collocated measurements.

The outcome of the calibration for a test dataset of two weeks during March 2014 is presented in Figure 2.4, which shows the calibrated measurements and the corresponding NO_2 reference over time.

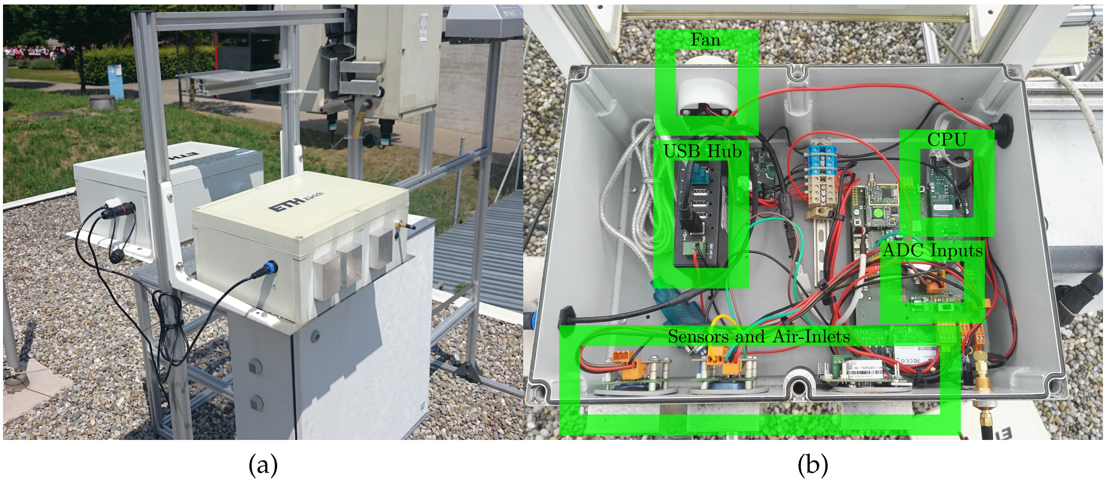


Figure 2.3: Sensor box (Figure 2.3a) deployment at the NABEL measurement station and the hardware (Figure 2.3b) used to collect measurements.

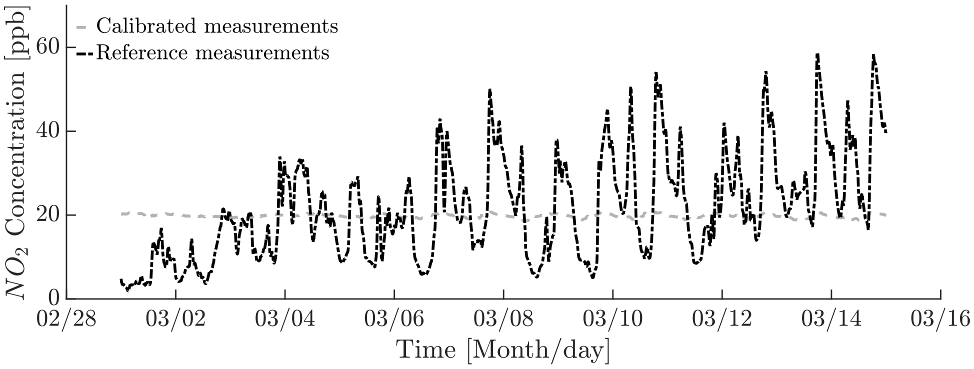


Figure 2.4: Ordinary least-squares (OLS) calibrated m = NO_2^* measurements and m_r = NO_2 reference measurements over time. The calibrated measurements do not correlate to the reference.

The calibrated measurements remain nearly constant over the whole calibration period. Due to sensor cross-sensitivities there is no correlation between uncalibrated sensor measurements and the reference, in fact the R^2 value equals 0.003. As a result the slope of the calibration gain (i.e., $\hat{b}_1$) component has a strong bias towards zero and the overall RMSE of the calibration is 12.4 ppb, while the average true NO_2 concentration is 21.5 ppb during this period. The calibrated measurements are clearly not of sufficient quality to make any conclusions about the actual NO_2 concentration, which is also reflected in a high value of the standardized error RMSE_σ = 3.4.

In order to improve the calibration quality, we apply MLS on measurements from multiple sensors M, measuring phenomena to which our sensor to be calibrated is cross-sensitive to. However, the calibration quality heavily depends on the choice of the sensors in M. We choose

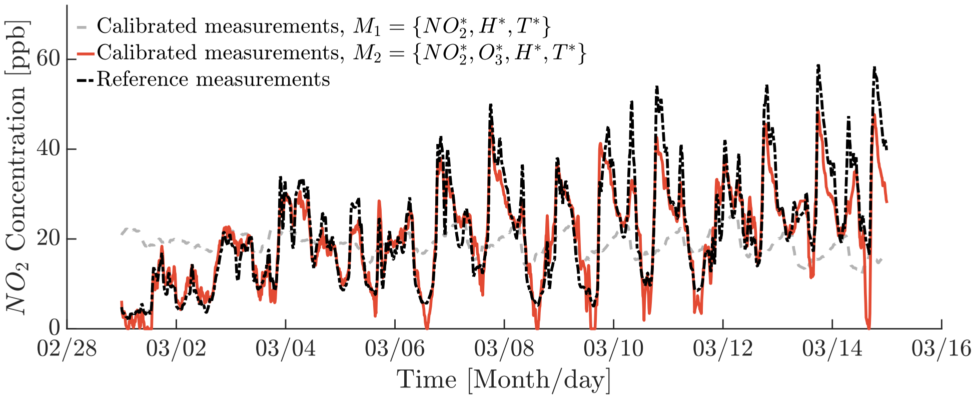


Figure 2.5: Multiple least-squares (MLS) calibration with different sensor arrays M and m_r = NO_2 reference measurements over time. The calibration quality heavily depends on the choice of sensor traces in M.

a sensor array consisting of $M_1 = \{NO_2^*, H^*, T^*\}$, where NO_2^*, H^* and T^* are low-cost nitrogen dioxide, humidity and temperature sensors. This means, we perform a temperature and humidity correction, see Section 1.4.2, for our electrochemical sensors [PSMJ16] using a sensor array M_1. In the remainder of this chapter, we denote all low-cost sensors that need calibration with an asterisk (*). Figure 2.5 shows the calibration outcome of MLS during the same two weeks in March. The calibrated measurements still correlate poorly to the reference and have a notable RMSE of 12.8 ppb. The reason for the poor performance is the cross-sensitivity of the NO_2 sensor to ozone (O_3), as we will show in Section 2.6.1.

The calibrated measurements when a collocated O_3 sensor is added to M_1, i.e., we construct a new array $M_2 = \{NO_2^*, O_3^*, H^*, T^*\}$, is shown in Figure 2.5. The calibration quality is improved significantly. There is a clear correlation between calibrated measurements and reference with $R^2 = 0.88$ and an almost 3 times smaller RMSE of 4.6 ppb. Finally, a $RMSE_\sigma = 1.2$ indicates that the measurements can be used to determine the current air quality level. We conclude that MLS is able to calibrate a cross-sensitive sensor when *augmented* with appropriate sensors to a sensor array M. However, due to scarce or no information about cross-sensitivities, the set of sensors in M and their respective impact on the measurements of the cross-sensitive sensor to be calibrated is often unknown.

2.4.5 Discussion

The example of the low-cost NO_2 sensor in Section 2.4.4 emphasizes the need of a pre-deployment testing methodology. The calibration accuracy of a cross-sensitive sensor z is limited without the thorough knowledge of the phenomena B_z to which the sensor is sensitive. If a sensor trace is

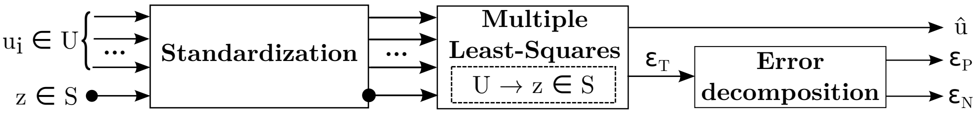


Figure 2.6: Sensor testing methodology.

an additive combination of multiple phenomena, it correlates poorly to a single reference trace. In order to calibrate the sensor measurements to a reference, the phenomenon of interest needs to be *segregated* from the measurements. This is only possible, if all remaining phenomena in the measurements can be compensated for. A sensor may not be sensitive to the same extent on the individual phenomena and it is possible—depending on the scenario—that some cross-sensitivities have minor impact on the sensor behaviour. It is therefore important to identify which phenomena need to be measured and their individual impact on the sensor trace under application-related circumstances.

The sensitivity list of a cross-sensitive sensor allows to augment it with additional sensors to a sensor array M. Having measurements from sensors in array M, which measures all phenomena in Bz, it is possible to accurately calibrate the measurements using multiple least-squares to the corresponding reference trace, as described in Section 2.4.2. The fit of the linear combination in (2.3) can then be seen as the segregation of the part in a cross-sensitive sensor trace $m_z \in M$ that is induced by the phenomenon of interest by compensating for the other measured phenomena and fitting it to a reference.

2.5 Testing Methodology

Datasheet information on sensors' cross-sensitivities and their dependency on meteorological parameters is often scarce. Even though some sensors undergo laboratory testing and calibration, these test settings typically only cover a few points in the sensing range. Given the usually long list of sensor cross-sensitivities, extensive tests and sensor calibration is highly time-consuming and is, therefore, hardly possible. To solve the problem, we propose a novel method that uncovers sensor dependencies under deployment-related conditions. We ignore any prior knowledge about the sensor, i.e., do not rely on any information given in the datasheet, and treat the sensor as a black-box, see also Figure 1.3. We choose the observable universe U as a set of all relevant phenomena for a given application. Our testing methodology consists of three steps depicted in Figure 2.6:

Standardization. All input traces are converted to a standardized representation with zero mean and unit variance in order to get scale-invariant sensor traces and thereby scale-invariant cross-sensitivity factors.

Inverse calibration. Multiple least-squares is used to regress the standardized measurements from the phenomena in the observable universe U on the measurements of low-cost sensor z. The resulting regression parameters are used to generate insights about the composition of the sensor measurements, i.e., they identify cross-sensitivities of the sensor.

Error decomposition. The regression error ε_T of the inverse calibration step is used as an indicator for missing phenomena in U and substantial sensor noise of z. We distinguish the latter on the frequency characteristics of typical atmospheric phenomena and, therefore, decompose the error by applying a low-pass filter.
We explain the three steps in more detail below.

2.5.1 Standardization

The measurements we use for testing can be any pollutant concentration, temperature or relative humidity and, thus, all these measurements have different scales and units. In order to get scale-invariant results, all variables need to be standardized, i.e., they need to be centred and have unit variance. We denote in the remainder of this chapter the standardized form of a trace $m \in \mathbb{R}^{n \times 1}$ as $\tilde{m} = \frac{m - \mu_m}{\sigma_m}$, where μ_m and σ_m are mean and standard deviation of m, respectively.

2.5.2 Inverse Calibration

The primary goal of the testing procedure is to uncover the explained part $B_z \cap U$ of a low-cost sensor $z \in S$, i.e., expose the phenomena B_z the sensor is sensitive to (see Figure 2.1). This is achieved by decomposing the sensor measurements into single phenomena of the observable universe U. In contrast to calibration, where usually multiple sensors are regressed on a reference, we reverse the process. Hence, similar to (2.4) and given collocated measurements $z \in \mathbb{R}^{n \times 1}$ of sensor z and references traces $U \in \mathbb{R}^{n \times |U|}$ of multiple reference sensors in U, the standardized regression equation is

$$\tilde{z} = \tilde{U} \cdot b + \varepsilon_T. \tag{2.10}$$

This means, we describe the measurements of sensor z as a linear combination of different references given in U with parameters b and some residual error term ε_T, as described in Section 2.3. The estimation $\hat{b}$ of the true regression parameters b calculated by MLS (see Section 2.4.2) will give insights about the extent of any cross-sensitivities or dependency on meteorological effects of sensor z. We calculate the regression estimation of the inverse calibration to determine how well we can describe the trace of sensor z as a combination of the different references in U , i.e.,

$$\hat{u} = \tilde{U} \cdot \hat{b}. \tag{2.11}$$

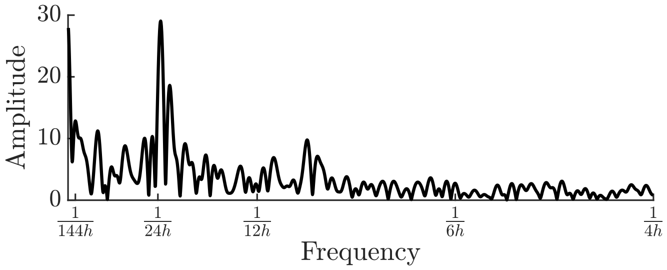


Figure 2.7: Frequency spectrum of O_3 reference measurements with a peak at frequency $\frac{1}{24h}$.

In Section 2.5.3 we show how we decompose the error component $\varepsilon_T = \tilde{z} - \hat{u}$ of our testing procedure into two separate parts to uncover the cross-sensitivities of sensor z. With this knowledge it is possible to determine, whether additional sensors need to augment z forming a sensor array to accurately measure the target phenomenon. For instance, whether we need temperature and humidity values to compensate for meteorological dependencies.

2.5.3 Error Decomposition

It is possible that a sensor measures a phenomenon not captured by the observable universe U. In this case, we are not able to explain a certain part of the sensor trace with U, limiting the benefit of using the sensor in the given application. Hence, it is important to determine the fraction of the explained and unexplained parts of z given U.

The unexplained part of z depends on the performance of the inverse calibration, defined as regression error ε_T, i.e., the root-mean-square error

$$\varepsilon_T = RMSE(\tilde{z}, \hat{u}), \tag{2.12}$$

where $\hat{u} = \tilde{U} \cdot \hat{b}$ is the regression estimation solved by MLS in (2.11). The larger the RMSE of ε_T, the larger is the unexplained part of the sensor measurements.

The contribution of the unexplained part is often two-fold, *(i)* the sensor is impacted by chemical or physical phenomena, such as interfering gases or meteorological effects, which cannot be explained with the current universe U, and *(ii)* the measurements suffer from sensor noise.

In order to distinguish between these two causes, we exploit that the underlying phenomena and noise differ in their frequency representations. Sensor noise is often a high-frequent signal, whereas phenomena like pollution concentrations or ambient temperature show distinct low-frequent variation patterns. For instance, the concentrations of primary air pollutants usually reach their maxima during the day and drop in the night [RP04], which is based on the increased activity of pollutant sources such as traffic or industrial plants during daytime. We exploit this daily periodicity by using it as indicator for any possible missing phenomena in U, indicated by a substantial low-frequent part (e.g., with frequency $\geq \frac{1}{24h}$) in the error ε_T induced by the missing phenomenon. For instance, Figure 2.7 shows the frequency spectrum of O_3 reference measurements recorded during April 2014. We observe substantial frequency components at frequencies larger or equal than $\frac{1}{24h}$. Consequently, the unexplained part of any sensor that is sensitive to O_3 will contain a notable low-frequent part, if O_3 is not included in universe U during the inverse calibration step. We use a low-pass filter to decompose ε_T, see (2.12) and Figure 2.6, in a low-frequency part ε_P, which represents *uncaptured* periodic phenomena. Similarly we use a high-pass filter to create a high-frequent part ε_N, which we treat as the *noise* component of the sensor. We apply two 3rd order *Butterworth* filters [Str04] and for simplicity use $\frac{1}{24h}$ as cut-off frequency.

To quantify the impact of each error component, we compute the root-mean-square of the different error components, see also (2.7). The low-frequent component

$$\mathrm{RMS}(\varepsilon_P) \in [0,1] \tag{2.13}$$

serves as a measure for uncaptured phenomena in our model. High $\mathrm{RMS}(\varepsilon_P)$ indicates that it is likely that the sensor is cross-sensitive to a phenomenon not included in U but still related to a relevant phenomenon. By contrast, large values of the high-frequent component

$$\mathrm{RMS}(\varepsilon_N) \in [0,1] \tag{2.14}$$

is attributed to high sensor noise. Finally, the amount that can be explained with references in U is measured with

$$\mathrm{RMS}(\hat{u}) \in [0,1]\,. \tag{2.15}$$

Depending on the decomposed errors, we can draw conclusions about the feasibility of deploying the sensor under test in a given environment. Assuming a sensor z can be fully explained with reference variables in U and is not affected by noise, i.e., $B_z \cap U = B_z$, the regression estimation equals to the sensor measurements, i.e., $\hat{u} = \tilde{z}$. Consequently, both $\mathrm{RMS}(\varepsilon_P)$ and $\mathrm{RMS}(\varepsilon_N)$ are zero and $\mathrm{RMS}(\hat{u})$ corresponds to the standard deviation of $\tilde{z}$, i.e., equals one. We can expect values close to zero and one, respectively, for any good low-cost sensor given an adequate observable universe.

2.5.4 Sensor Signature

The determined error components $\mathrm{RMS}(\varepsilon_P)$ and $\mathrm{RMS}(\varepsilon_N)$, which we calculate by decomposing the testing error ε_T (see (2.12)) in Section 2.5.3, and the explained part RMS($\hat{u}$), see (2.11) in Section 2.5.2, of a sensor z with a given observable universe U describe a *sensor signature*. Based on the sensor signature it is possible to determine the sensor array that can be used for compensating cross-sensitivities. If the testing methodology for sensor z is conducted multiple times with different compositions of universe U, the universe U that optimizes the sensor signature, i.e., minimizes RMS(ε_P) and RMS(ε_N) and maximizes RMS($\hat{u}$), best describes the set of phenomena B_z the sensor under test is sensitive to. If $h \in B_z$ is a phenomenon of interest, then the necessary sensor array M to measure h is created by augmenting sensor z with low-cost sensors that measure phenomena $B_z \setminus h$.

2.6 Experimental Evaluation

In this section, we apply our testing methodology to different types of low-cost sensors. In Section 2.6.1 we test three sensors, which have different sensitivity profiles, i.e., have different cross-sensitivities with different magnitudes. We analyse their cross-sensitivities and meteorological dependencies and show how collocated measurements of multiple sensors in a sensor array can be used to accurately calibrate the measurements to reference gases in Section 2.6.2. Further, we investigate the stability of the calibration accuracy over time in Section 2.6.3. Finally, in Section 2.6.4, we present results from two sensors, which are not suitable for air quality monitoring in our setting. For these sensors we were not able to explain the sensor measurements with reference variables to an adequate extent.

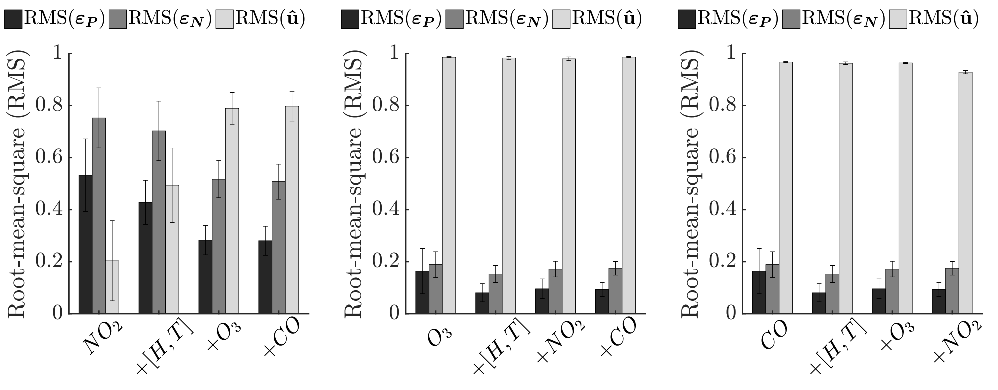
2.6.1 Sensor Testing

Setup. Our goal is to use low-cost sensors for monitoring major pollutants, namely O_3, CO and NO_2, in an urban environment. To achieve this goal, we build a measurement system consisting of multiple low-cost sensors. In order to find the optimal sensor array that ensures accurate measurements, we apply the testing methodology presented in Section 2.5. We deploy three sensors at a station of the Swiss National Air Pollution Monitoring Network (NABEL) in Duebendorf, Switzerland, depicted in Figure 2.3. The sensors are an electrochemical-based NO_2 sensor[4], a metal oxide-based O_3 sensor[5] and an electrochemical-based CO sensor[6]. They are placed inside a box, see Figure 2.3b, which is mounted on the roof of the station next to the air inlets of the highly accurate devices.

[4] *Alphasense NO_2-B4 Nitrogen Dioxide 4-Electrode (2013)* [Alp13a]

[5] *SGX Sensortech (formerly e2v) MiCS-OZ-47 Ozone Sensing Head with Smart Transmitter PCB O_3* [SGX13]

[6] *Alphasense CO-B4 Carbon Monoxide 4-Electrode (2014)* [Alp13a]



(a) NO_2 sensor

(b) O_3 sensor

(c) CO sensor

Figure 2.8: Average root-mean-square value of the periodic (low-frequent) part ε_P, noise (high-frequent) part ε_N of the regression error and the regression estimation $\hat{u}$ for three low-cost sensors. The x-axis shows the evaluation for different observable universes U.

Therewith we ensure collocated measurements of low-cost sensors and reference devices.

The following evaluations are based on over 9 million measurement samples gathered during 15 months from January 2014 to March 2015. We use all sensing modalities, i.e., 10 different pollutant concentrations, measured by the official air quality measurement station as reference variables to build our observable universe. Further, we use temperature and humidity reference measurements to gain insights about the meteorological dependencies of the low-cost sensors. Because low-cost sensors usually do not show a completely linear dependency to phenomena, the samples of all references in U have additionally been included in quadratic and cubic form in the regression (2.10).

Testing procedure. In order to show the feasibility of our testing methodology, we start with an observable universe U, which consists only of the reference corresponding to each sensor. For example, the universe $U = \{NO_2\}$ is used to initially perform the testing methodology for the low-cost sensor $z = \{NO_2^*\}$. The universe is then gradually extended with further references and the testing methodology repeated to highlight the impact of adding references to universe U.

Results. Figure 2.8a shows the evolution of RMS(ε_P) (see (2.13)), RMS(ε_N) (see (2.14)) and RMS($\hat{u}$) (see (2.15)) for the NO_2^* low-cost sensor, where each set of bars corresponds to a different U. The values are calculated in steps of two weeks over the whole measurement period and the height of the bars and the whiskers indicate the average and standard deviation, respectively, over all tests.

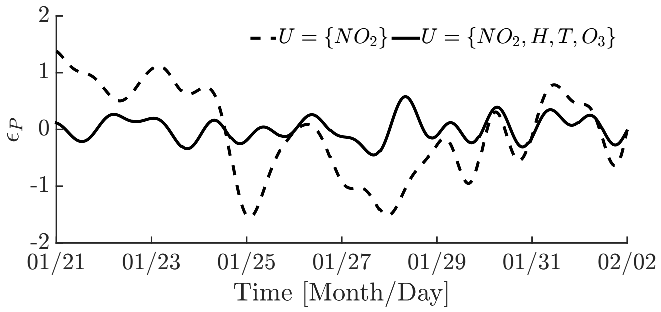


Figure 2.9: Low-frequent error component ε_P for different observable universes U and $z = \{NO_2^*\}$ over time. Including O_3 in U lowers the amplitude of ε_P, which highlights the sensor's cross-sensitivity to O_3.

We make the following observations:

- The results for the initial $U = \{NO_2\}$ show a remarkably larger periodic error RMS(ε_P) and noise component RMS(ε_N) than the explained part RMS($\hat{u}$). This finding points out that the sensor is to a large extent cross-sensitive to some phenomena beyond NO_2 and is affected by noise.

- The second set of bars shows the results when humidity and temperature measurements including their squared and cubed versions are added to $U = \{NO_2, H, T\}$. Although we observe an increase of the explained part of the sensor measurements, the unexplained low-frequency estimate RMS(ε_P) remains dominant. Though the sensor is influenced by meteorological effects, it still is sensitive to other phenomena, which are not included in U yet.

- Extending the observable universe $U = \{\mathrm{NO_2, H, T, O_3}\}$ with the O_3 modality decreases the unexplained part of the sensor measurements by roughly 35% and confirms the sensor's strong cross-sensitivity to O_3. This result is reflected in Figure 2.9, where we observe a smaller amplitude of ε_P compared to the initial universe $U = \{NO_2\}$.

- We fail to improve the sensor performance with the addition of CO reference samples to the universe $U = \{\mathrm{NO_2, H, T, O_3, CO}\}$ and, thus, conclude that our NO_2^* sensor is not impacted by a change of CO concentration in the ambient air. Finally, we were not able to find any notable changes of the test results by adding the remaining references.

In contrast to the cross-sensitive NO_2^* sensor, the results of the O_3^* (Figure 2.8b) and CO^* sensor (Figure 2.8c) are different. Both sensors are highly sensitive to their target gases and, hence, the initial U suffices already to explain sensor measurements to a great extent. In fact, only including temperature and humidity references in the regression lessens the unexplained part of the error. Introducing additional references does not improve the outcome of the testing procedure, because both sensors are not cross-sensitive to any of the tested interfering gases.

The above findings show that the NO_2^* sensor can be used to monitor air quality in our setting only if the measurement system additionally acquires collocated O_3 measurements to compensate for the sensor's cross-sensitivity to O_3. Moreover, all gas sensors (O_3^*, NO_2^* and CO^*) depend on meteorological conditions and, thus, should be augmented with temperature and relative humidity sensors to achieve accurate calibration.

2.6.2 Sensor Array Calibration

Our novel testing methodology allows revealing cross-sensitivities and their extent for every sensor under test. These results immediately suggest the necessary measurement system augmentation to segregate mutual dependencies between cross-sensitive sensors. In the next step, the measurement system can be calibrated using MLS to measure the desired phenomena. Therefore, we used the acquired sensor signatures and construct three different sensor arrays, that can be calibrated to the three corresponding pollutant reference concentrations.

Setup. We augment an array $M_{NO_2} = \{NO_2^*, O_3^*, H^*, T^*\}$, which is calibrated to the NO_2 reference, and $M_{O_3} = \{O_3^*, H^*, T^*\}$ and $M_{CO} = \{CO^*, H^*, T^*\}$, which are calibrated to O_3 and CO reference measurements, respectively. In this section, we investigate the calibration accuracy of these sensor arrays and compare the performance of the multiple least-squares (MLS, see Section 2.4.2) and ordinary least-squares (OLS, see Section 2.4.1) approaches. The measurements from the sensors as well as the quadratic form of the O_3^* sensor in the corresponding sensor arrays are calibrated to O_3, CO and NO_2 references provided by the high quality sensors using MLS[7]. For the OLS approach we use the sensor measurements corresponding to the reference, i.e., the O_3^* sensor measurements are calibrated to the O_3 reference using OLS. The calibration parameters have been repeatedly trained with data over four weeks and used to calibrate the consecutive four weeks. The average RMSE for the OLS and MLS calibration over the whole deployment period of 15 months is summarized in Table 2.1.

Results. As already seen in Section 2.4.4, MLS manages to achieve accurate calibration of the cross-sensitive NO_2^* sensor and outperforms the OLS approach by up to 45% in terms of calibration error. Note that the measurements still have a relatively high normalized error $RMSE_\sigma > 2$, which is due to the overall low concentration of NO_2 with a median of only

[7] In contrast to the sensor testing, we do not standardize the variables for calibration.

		$NO_2\,[ppb]$	$O_3\,[ppb]$	$CO\,[ppm]$
$[25, 50, 75]^{th}$ perc.		[6.8, 13.1, 21.8]	[2.2, 18, 31.8]	[0.17, 0.22, 0.32]
RMSE	OLS	9.72 ± 2.97	4.88 ± 1.67	0.051 ± 0.020
	MLS	5.42 ± 1.20	3.04 ± 1.15	0.050 ± 0.021
$RMSE_\sigma$	OLS	3.84 ± 0.84	1.37 ± 0.22	1.20 ± 0.26
	MLS	2.21 ± 0.53	0.85 ± 0.21	1.17 ± 0.28
R^2	OLS	0.03 ± 0.02	0.89 ± 0.04	0.83 ± 0.07
	MLS	0.67 ± 0.11	0.96 ± 0.01	0.83 ± 0.07

Table 2.1: Comparison of calibration performance (mean ± standard deviation) between OLS and MLS and the $[25,50,75]^{th}$ percentiles of the actual pollutant concentrations. MLS improves the calibration performance over OLS, especially for the NO_2 and O_3 measurements. The difference for CO measurements are negligible.

13 ppb. Nevertheless, the improvement of the R^2 value from 0.03 to 0.67 in average shows that the MLS calibration is able to capture significantly more variability of the actual NO_2 concentration than the basic OLS calibration. Further, the sensor array calibration to O_3 and CO references is beneficial as well. Although both sensors have no significant cross-sensitivities to interfering gases, MLS compensates for the meteorological influences resulting in a lower calibration error compared to OLS.

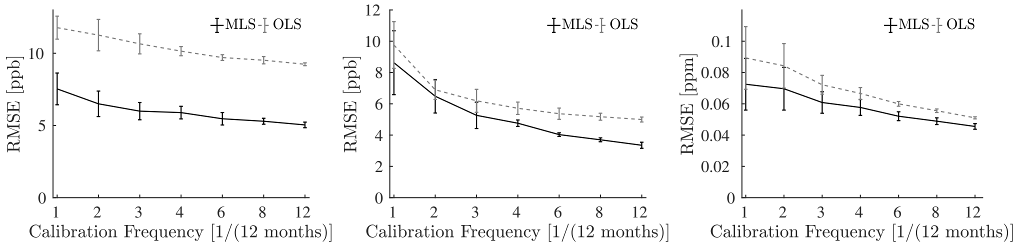
These results emphasize the necessity of uncovering the sensitivity profiles of low-cost sensors by our pre-testing methodology and then augmenting the measurement system with appropriate sensors.

2.6.3 Calibration Stability

Various works, which use an univariate calibration approach such as OLS [CDS+17, SGV+15, SGV+17] or calibration techniques based on artificial neural networks (ANN) [ESV+18, DES+18] emphasize that low-cost sensors need frequent re-calibration. Some works already notice a need for re-calibration within four weeks [MMH17, CDS+17]. We show that if our measurement system is augmented with appropriate sensors and then calibrated with MLS, it needs less-frequent re-calibration than reported above.

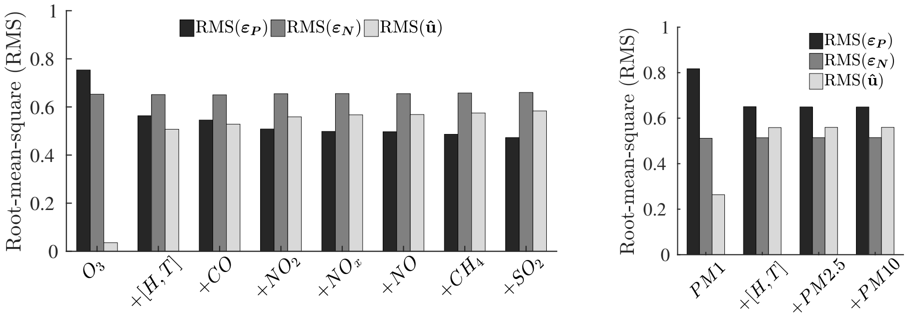
Setup. We compare the calibration error of MLS and OLS with different re-calibration frequencies over a period of 12 months for the same three sensor arrays as in Section 2.6.2. We calculate the regression parameters for both techniques using a training dataset of four weeks. The resulting parameters are then used to calibrate and evaluate a testing dataset between the end of the current training dataset and the end of the consecutive one. The individual training intervals are uniformly spread over a period of 12 months, i.e., a calibration frequency of four re-calculates the calibration parameters every three months. The procedure is performed 26 times for each re-calibration frequency setting, where each time the start of the initial training dataset is increased by one week starting at January 10*th* 2014.

Results. Figure 2.10 shows the average RMSE for the MLS and OLS calibration to NO_2, O_3 and CO references. We observe a decreasing error with an increasing calibration frequency for all three references and both techniques. As shown before, OLS is not suited to calibrate the cross-sensitive NO_2^* sensor and consequently MLS clearly outperforms OLS for all calibration frequency settings in Figure 2.10a. Of more interest are the results of the sensor array calibration to O_3 and CO references. In Figure 2.10b we observe that using a sensor array and MLS to calibrate to the O_3 references between three and four times a year performs better than using OLS on a monthly basis. A similar trend is shown in Figure 2.10c, where the CO sensor array always performs better than the simple OLS



(a) NO_2 (b) O_3 (c) CO

Figure 2.10: Average RMSE over 12 months with different calibration frequencies of multiple least-squares and ordinary least-squares. Calibrating a sensor array with multiple least-squares needs less frequent re-calibration than ordinary least-squares.



(a) *Air quality* sensor [Dov] (b) *Particle* sensor [Shi10]

Figure 2.11: Average root-mean-square value of the periodic (low-frequent) part ε_P, noise (high-frequent) part ε_N of the regression error and the regression estimation $\hat{u}$ for two unqualified low-cost sensors. The x-axis shows the evaluation for different observable universes U.

calibration. Calibrating the array every six to eight weeks using MLS achieves a lower error than using OLS every four weeks.

In conclusion, we can see that calibrating our sensor array with multiple least-squares needs less frequent re-calibration to achieve the same performance when compared to the state-of-the art ordinary least-squares based calibration.

2.6.4 Unqualified Sensors

Low-cost sensors available on the market may fail under certain application conditions. Fig. 2.11 shows test results for two sensors that can not be used to monitor air quality in our setting. The *TP-401A Indoor air quality gas sensor* [Dov] from Shenzhen Dovelet Sensors Technology is a metal oxide sensor, which is—according to its datasheet—sensitive to a long list of pollutants, including carbon monoxide, even at low concentrations. Unfortunately, no information about the extent of each sensitivity is provided. The *Particle Sensor Model PPD42NS* [Shi10] from Shinyei measures the concentration of particle matter, i.e., dust particles with diameter larger than 1 μm (*PM1*).

In Figure 2.11a, we observe for both sensors and each observable universe that the error components dominate and the explained part RMS($\hat{u}$) does not exceed values above 0.6. Besides temperature and humidity references we used all pollutants measured by the static measurement station to construct the universe for testing the *TP-401A* sensor. We observe that the sensor is sensitive to meteorological influences but not to any pollutants, which are helpful to explain sensor measurements. We suspect that reasons for the negative result might be the sensor's low accuracy and its design for higher pollution concentration.

On the contrary, even though the measurements of the *PPD42NS* sensor can be partly explained with *PM1*, temperature and humidity, it has still a remarkable periodic error component with a RMS(ε_P) ≈ 0.65 as shown in Figure 2.11b. The result indicates that the sensor is impacted by a phenomenon that we failed to identify. The sensor may be useful to measure PM1 if one can find and compensate for that phenomenon in question.

2.7 Summary

Nowadays, low-cost air quality sensors are integrated in an increasing number of measurement platforms for air quality monitoring. Calibrating these sensors to reference measurements is however challenging. They typically suffer from cross-sensitivities, poor stability and sensor noise. Information about all the limiting effects is often not provided by the manufacturers. Even if the information is given in a datasheet, it is often scarce and reflects sensor performance under laboratory conditions. Neglecting sensor cross-sensitivities and deployment settings usually results in poor sensor performance, frequent calibration necessity and calibration failures. This arises the need for pre-deployment sensor testing under application conditions.

In this chapter, we present an in-field sensor testing methodology for low-cost and possibly cross-sensitive sensors. Our novel algorithm is based on multiple least-squares and uses collocated measurements of low-cost sensors and various reference sensors to quantify the amount of captured and uncaptured cross-sensitivities, and substantial sensor noise. With the obtained testing results we are able (i) to conclude about the usability of a given sensor under test in the given setting, (ii) identify fundamental cross-sensitivities and compute the sensor array, which can optimally measure a specific pollutant and (iii) compute calibration parameters that provide accurate measurements with long-term stability. We extensively evaluate our algorithm with various low-cost sensors using a dataset of 9 million sensor measurements and show the improved accuracy and long-term parameter stability when calibrating an augmented sensor array to reference measurements. We believe the proposed algorithm can be an essential step in the design of measurement platforms with low-cost cross-sensitive sensors.

3

Multi-Hop Calibration for Mobile Sensor Arrays

Urban air pollution monitoring with mobile, portable, low-cost sensors has attracted increasing research interest for their wide spatial coverage and affordable expenses to the general public. However, low-cost air quality sensors not only suffer from cross-sensitivities and dependency on meteorological effects, as we discuss in Chapter 2, but also from *drift over time*, an additional error source (see Section 1.2) we tackle in this chapter. Therefore, frequent calibration of measurements from low-cost sensors is indispensable to guarantee data accuracy and consistency to be fit for quantitative studies on air pollution. In Chapter 2, we present a pre-deployment testing and calibration approach, which is performed on sensor arrays that have continuous access to reference measurements. However, constant access to references can usually not be guaranteed in real-world deployments and, therefore, frequent re-calibration becomes a challenging task. In this chapter, we propose *sensor array network calibration* (SCAN), a multi-hop calibration technique for low-cost sensors in mobile deployments. SCAN is a *collaborative* calibration framework to counteract drift effects, see Section 1.5.2, and applicable to sensor arrays to compensate for cross-sensitivities and dependencies on meteorological influences. SCAN minimizes error accumulation over multiple hops of sensor arrays, which is unattainable with existing multi-hop calibration techniques. We formulate SCAN as a novel constrained least-squares regression, provide a closed-form expression of its regression parameters and theoretically prove that SCAN is free from regression dilution even in presence of measurement noise. In-depth simulations demonstrate that SCAN outperforms various calibration techniques. Evaluations on two real-world low-cost air pollution sensor datasets comprising 66 million samples collected over three years show that SCAN yields 16% to 60% lower error than state-of-the-art calibration techniques.

3.1 Introduction

The availability of portable and low-cost air quality sensors has made them promising not only for qualitative air pollution monitoring to raise public awareness, but also for quantitative analysis to facilitate public policies, infrastructure control and health studies. Installed on vehicles [LFS+12, SHT15, AMM16b, SGG+13] or in wearable devices [ZLYX15, TDMP16, BKB15, BEMRB13, KPG10], these sensor nodes travel citywide and their users, both professional and amateur, upload air quality measurements with time and location stamps. If collected in long-term deployments, these air quality measurements can provide valuable insights on personal air pollution exposure and validation of high-resolution air pollution models.

To fully unlock the potential of the big urban air quality data collected by mobile, low-cost sensors, it is essential to calibrate the measurements to obtain a *consistent* dataset. However, the quality of measurements from low-cost air quality sensors are affected by multiple factors and can vary dramatically across sensors and over time. In Chapter 2, we investigate the following three potential error sources. *(i)* Low-cost air quality sensors suffer from low selectivity, i.e., they are cross-sensitive to various substances in the air. *(ii)* Changing environmental conditions, such as temperature and humidity, impact the sensor output. *(iii)* Many air quality sensors operate at their sensitivity boundaries when measuring pollution in ambient air, which leads to high noise in sensed data. In this chapter, we investigate a fourth important limiting factor of low-cost air quality sensors: *(iv)* Sensor sensitivity *degrades over time* due to sensor ageing effects.

In Chapter 2, we show that a powerful solution to compensate for the dependencies of *(i)* and *(ii)* is to augment a measurement system with additional sensors to form a *sensor array*. A sensor array consists of co-located sensors that measure, in addition to the target air pollutants, a set of correlated pollutants and environmental parameters e.g., temperature. In Section 2.6, we demonstrate the feasibility of resolving cross-sensitive dependencies of low-cost air quality sensors by jointly calibrating a set of measurements collected by sensor arrays. In fact, an increasing number of customized [ZLYX15, TDMP16, SGV+15] and commercial [SGX08] air quality sensing nodes is integrated with multiple correlated sensors and report measurements of pollutants and environmental parameters simultaneously.

Challenges. Adopting sensor arrays alone is insufficient to ensure consistent data quality of the measurements. As we show in Section 2.6.2, the less frequent sensor arrays are calibrated, the more inaccurate the data becomes. Especially in long-term deployments spanning over multiple months or even years *frequent re-calibration* is an important task to maintain consistent data quality. Due to noise effects, see *(iii)*, and sensitivity drift, see *(iv)*, measurements of sensor arrays need to be re-calibrated to the accurate, static governmental-run air monitoring stations, which are sparsely deployed in cities. However, sensors may have infrequent or no access to the static stations, making it unreliable or infeasible to calibrate all sensors to the static references. As we have summarized in Section 1.5.2 in order to improve the opportunities for calibration, some studies [MLCOS08, HST12, SHT15] propose to calibrate noisy sensors during deployment by exploiting *rendezvous* [SHWT14]. During a rendezvous two sensors are in each other's spatial and temporal vicinity and thus sense the same phenomena and should have similar outputs. Rendezvous-based calibration allows newly calibrated sensors to recursively calibrate other sensors, known as *multi-hop calibration*. By constructing a calibration path starting from a static station and performing calibration for each pair of sensors that frequently meet, calibration can reliably propagate through all sensor arrays within a large-scale deployment. Compared to calibration solely based on rendezvous between mobile nodes and reference stations multi-hop calibration is proven to frequently calibrate a significantly larger number of nodes [SHT15, MBS$^+$16, YGTL14]. Thus, multi-hop calibration is a powerful tool for maintaining high data quality in large scale deployments over long time periods.

However, it is non-trivial to apply multi-hop calibration on sensor arrays. Multi-hop calibration is only effective and guarantees data consistency if the calibration error at each hop does not accumulate. Otherwise a newly calibrated sensor array after multiple hops will still be too noisy as a reference to calibrate its successor sensor array. State-of-the-art multi-hop calibration schemes [HST12] based on *ordinary least-squares regression* (OLS) suffer from error-accumulation over multiple hops due to regression dilution [Woo41]. The results in [SHT15] propose to use *geometric mean regression* (GMR) to minimize multi-hop error-accumulation. The scheme is efficient with single sensor calibration but inapplicable to sensor array calibration. Although there are several generalizations of GMR to higher dimensions [DY97, Tof02], they all perform poorly when calibrating sensor arrays over multiple hops, as shown in Sec. 3.5. Existing sensor array calibration schemes leverage multiple least-squares (MLS), see Chapter 2, or artificial neural networks (ANN) [SGV$^+$15], but are primarily designed for one-hop calibration, i.e., calibration with highly accurate references.

Contributions and road-map. In this chapter, we propose *sensor array network calibration* (SCAN), a low-error multi-hop calibration scheme for sensor arrays. We formulate multi-hop sensor array calibration as a novel *constrained least-squares regression*, and come up with a closed-form solution that minimizes error accumulation over multiple hops for multi-dimensional calibrations. We theoretically prove that SCAN minimizes error accumulation with practical array composition and measurement settings. Experiments with a one-year measurement comprising 10 million samples from a 11-hop metal oxide gas sensor array chain show that SCAN yields up to 38% lower error than multiple least-squares (MLS) and geometric mean regression (GMR). Evaluations on a public air pollution dataset [LFS+12] of 56 million samples collected over two years demonstrate the benefits of multi-hop calibration over one-hop calibration. Additionally, we show that SCAN outperforms MLS by 60%, which is a significant improvement in enabling low-cost sensors for quantitative analysis such as validation of air pollution models and personal exposure studies. The contributions of this chapter are summarized as follows:

- To the best of our knowledge, we are the first to theoretically tackle the error accumulation problem of multi-hop sensor array calibration. While we primarily target low-cost gas sensors, SCAN is a generic multi-hop multi-dimensional calibration scheme and is widely applicable to various heterogeneous, correlated sensors working in an ad-hoc manner.

- We formulate the problem of multi-hop sensor array calibration as a novel constrained least-squares regression and propose a closed-form solution that minimizes error accumulation. We prove that SCAN works with practical settings such as non-zero mean measurements and non-squared sensor array composition.

- We evaluate SCAN on two real-world low-cost air pollution sensor datasets consisting of 10 and 56 million samples, respectively, collected over three years. Experiments show that SCAN achieves 16% to 60% lower error than state-of-the-art calibration techniques, which significantly improves the quality of measurements for qualitative studies such as validation of high-resolution air pollution models and personal air pollution exposure.

In the rest of this chapter, we summarize related work in Section 3.2, formulate SCAN in Section 3.4.2 and come up with a closed-form solution in Section 3.4.3. We conduct extensive simulations and real-world experiments to evaluate the performance of SCAN in Section 3.5 and summarize the chapter in Section 3.6.

3.2 Related Work

Exploiting rendezvous between sensors in a mobile network is a popular approach for sensor calibration [SHT15, SHWT14, HST12, YGTL14, MBS+16]. Saukh et al. [SHWT14] introduce the concept of sensor rendezvous and its potential for applications such as sensor fault detection and calibration. In a further work [SHT15] they present multi-hop calibration based on a rendezvous path. They show that ordinary least-squares (OLS) suffers from error accumulation when applied to multi-hop calibration and propose the use of the geometric mean regression (GMR). GMR does not suffer from error accumulation and, thus, improves network-wide data accuracy. However, GMR can only be applied to single sensors by performing a simple offset and gain calibration (see Section 1.4.1) and, hence, can not compensate for limiting effects of common low-cost sensors, such as low selectivity. Our proposed method can be applied to sensor arrays while also minimizing error accumulation and, thus, improves data accuracy of low-cost sensor deployments even more.

Ye et al. [YGTL14] present a multi-hop calibration approach for a large network of mobile barometric sensors embedded into commercial smartphones. The work shows two major challenges; *(i)* common low-cost barometric sensors show a considerable cross-sensitivity to temperature and wind regimes and *(ii)* errors accumulate along rendezvous paths. In order to reduce error accumulation over multiple hops they propose an approach that finds paths with minimal length and calibrates all sensors in the network. They show that this is a NP-complete problem and propose a heuristic solution. We are convinced that our SCAN approach can *(i)* compensate for the barometric sensors cross-sensitivities when augmented to an appropriate sensor array and *(ii)* does not require to find short rendezvous paths for improved measurement quality.

Markert et al. [MBS+16] present a privacy-protecting multi-hop calibration which enables its appliance for participatory environmental sensing. They demonstrate the benefits of multi-hop over one-hop calibration while protecting the location privacy of participating users.

3.3 Assumptions and Models

This section recapitulates the basic assumptions on sensor arrays and defines the models for multi-hop sensor array calibration.

3.3.1 Sensor Array Calibration

We refer to a *trace* as a time-ordered sequence of n discrete and instantaneous measurements $y_i = (y_{ij}) \in \mathbb{R}^n$ taken by sensor s_i at times t_j and spatial locations l_j for $j \in \{1, 2, \ldots, n\}$ within a time interval $[t_1, t_n]$. A trace measured by a low-cost gas sensor is usually inaccurate due to *(i)* low selectivity, i.e., low-cost sensors are cross-sensitive to multiple substances in the air, and *(ii)* the response of low-cost gas sensors is affected by meteorological conditions. To compensate for these dependencies, low-cost sensors are usually augmented by a set of heterogeneous sensors to form a *sensor array* (for instance see Figure 2.3b). Specifically, instead of measuring the target phenomenon using one low-cost sensor, an array $A = [s_1, s_2, \ldots, s_k]$ of sensors $s_1, \ldots, s_k$ is employed to simultaneously measure k phenomena. Prominent examples are gas sensors based on electrochemical cells designed to measure nitrogen dioxide (NO_2) concentrations, whose sensitivities are generally affected by ambient temperature and interfering gases, such as ozone (O_3) [SGV+15]. We show in detail that we are able to notably improve the accuracy of an $\mathrm{NO_2}$ sensor by concurrently acquiring ambient temperature, relative humidity and O_3 measurements to compensate for the sensor's dependency in Section 2.4.4. Calibrating a sensor array consisting of these multiple low-cost sensors to a NO_2 reference improves the overall measurement accuracy by up to 45% compared to simple offset and gain calibration using only the NO_2 low-cost sensor as presented in detail in Section 2.6.2. Similar cross-sensitivities are in fact a well-known limitation for the majority of state-of-the-art low-cost environmental sensors [JHW+16, SGV+15, EK12, BKW07].

We now formulate the basic model for sensor array calibration. We distinguish between *reference arrays*, which consist of precise and selective sensors, i.e., they accurately measure their target phenomenon, and mobile *low-cost sensor arrays*, which suffer from mutual dependencies and drifts stated above. Let $Y = (y_{ij}) \in \mathbb{R}^{k \times n}$ be a matrix describing traces of k possibly correlated phenomena, such as pollutant concentrations and ambient temperature, where y_{ij} describes the j-th measurement of phenomenon y taken by a reference sensor s_i at the given time instance t_j and location l_j. Denote $\overline{X} = (\overline{x}_{ij}) \in \mathbb{R}^{k \times n}$ as the uncalibrated traces of a mobile low-cost sensor array A taken at the same time instances and locations. We assume that the uncalibrated trace $\overline{x}_i = (\overline{x}_{ij}) \in \mathbb{R}^n$ of a single cross-sensitive sensor is a linear combination of different effects describing the impact of different phenomena and sensor noise, see also Section 2.3. Further, we assume that all sensors s_i for $i \in \{1, 2, \ldots, k\}$ of an array A are sampled at the same time instance and at the same location, i.e., a sensor array forms one physical unit. Then, the uncalibrated traces of a low-cost sensor array can be defined as

$$\overline{X} = B^{-1} \left(Y + N_1 \right), \tag{3.1}$$

where $N_1 \in \mathbb{R}^{k \times n}$ are k noise components (zero mean and uncorrelated to phenomena Y) and $B^{-1} \in \mathbb{R}^{k \times k}$ is an unknown matrix capturing the linear combination of phenomena and their magnitude for each individual sensor in the array, see Figure 3.1.

The aim of sensor array calibration is to find a calibration matrix $\overline{B}$ such that the calibrated measurements

$$\overline{Y} = \overline{BX} = \overline{B} \left(B^{-1}(Y + N_1) \right) + C \tag{3.2}$$

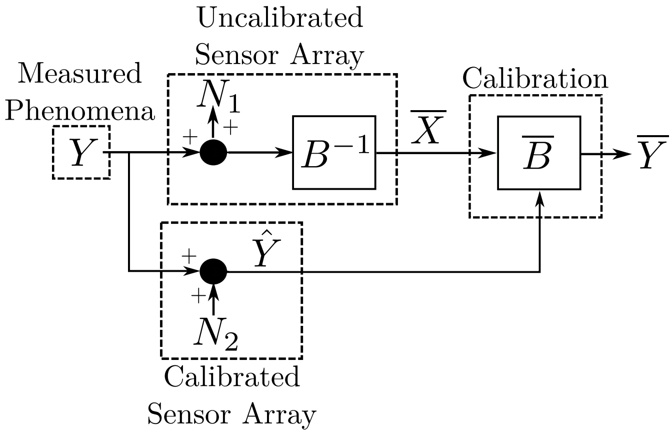


Figure 3.1: Basic sensor array calibration model. The traces $\hat{Y}$ of an already calibrated sensor array can be used as reference to calibrate the traces $\overline{X}$ of an uncalibrated sensor array. If the calibrated sensor array is a reference array it measures the physical phenomena with perfect accuracy, i.e., we assume the measurement noise $N_2 = \mathbf{0}_{k \times n}$.

minimize the calibration error, defined as a distance to the phenomena traces Y in some metric. Matrix $C \in \mathbb{R}^{k \times n}$ describes a constant additive offset. For simplicity of presentation we assume $C = \mathbf{0}_{k \times n}$ and all the rows in $\overline{X}$ and Y have zero mean. We relax this assumption and present the calibration including a non-zero constant matrix C in Section 3.4.4. Further, we assume for a general low-cost sensor array that B is non-singular.

3.3.2 Multi-hop Sensor Array Calibration

Highly accurate reference arrays are often static and limited in a real-world deployment and not all mobile sensor arrays have access to these reference arrays. To calibrate measurements of all sensor arrays, a practical strategy is to use calibrated sensor arrays as *virtual references* to calibrate another mobile sensor array. The rationale is that when two mobile sensor arrays are close in space and time, known as *rendezvous*, they are exposed to the same physical processes, and thus provide an opportunity for calibration. Such a rendezvous based calibration scheme can propagate along multiple hops and ultimately can cover all the deployed mobile sensor arrays.

One-hop calibration. Initially, we calibrate a mobile low-cost sensor array to a high quality reference array, which measures physical phenomena Y with perfect accuracy. Hence, we calibrate a sensor array $A_{(h=0)}$ using its traces $\overline{X}_{(0)} = B_{(0)}^{-1}(Y + N_1)$ by estimating the optimal $\overline{B}_{(0)}$. We refer to the calibration between reference and low-cost array as *one-hop calibration* with $h = 0$.

Multi-hop calibration. In the next hop, $h = 1$, given some rendezvous between an already calibrated array $A_{(0)}$ and an uncalibrated array $A_{(1)}$, we

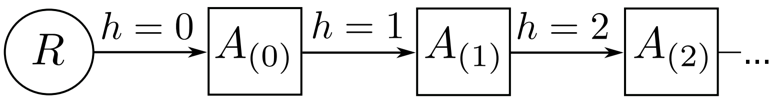


Figure 3.2: In a first hop $h = 0$, an uncalibrated array $A_{(0)}$ is calibrated with traces from an accurate reference array R. In all consecutive hops $h > 0$ an uncalibrated array $A_{(q)}$ is calibrated with traces from an already calibrated array $A_{(q-1)}$.

take $A_{(0)}$ as a *virtual reference*. We use its measurement $\overline{X}_{(0)}$ at the current rendezvous to determine a set of calibrated measurements $\overline{Y}_{(0)} = \overline{B}_{(0)}\overline{X}_{(0)}$ with $\overline{B}_{(0)}$ from the one-hop calibration. The measured values $\overline{X}_{(1)}$ of array $A_{(1)}$ at the current rendezvous can now be calibrated with the calibrated measurements $\hat{Y}_{(1)} = \overline{Y}_{(0)}$ from array $A_{(0)}$, see Figure 3.1. As illustrated in Figure 3.2, this concept can be applied to consecutive hops $h > 1$ resulting in a *rendezvous path*. We refer to calibration of sensor arrays along rendezvous paths with $\max(h) > 0$ as *multi-hop* calibration. Specifically, at hop $h = q$ with $q > 0$, we calibrate the sensor array traces as $\overline{Y}_{(q)} = \overline{B}_{(q)}\overline{X}_{(q)}$, where $\overline{B}_{(q)}$ is obtained by the raw traces $\overline{X}_{(q)}$ and virtual reference traces $\hat{Y}_{(q)} = \overline{Y}_{(q-1)}$ measured by a calibrated sensor array $A_{(q-1)}$.

Thus, given uncalibrated measurements $\overline{X}_{(q)}$ from array $A_{(q)}$ and calibrated measurements $\hat{Y}_{(q)} = \overline{Y}_{(q-1)}$ from array $A_{(q-1)}$ we intend to calculate calibration parameters $\overline{B}_{(q)}$ such that

$$\left\|\overline{Y}_{(q-1)} - \overline{B}_{(q)}\overline{X}_{(q)}\right\|, \tag{3.3}$$

where some norm $\|\cdot\|$ is minimized.

Discussion. *(i)* Our multi-hop sensor array calibration scheme is performed along a *static*, *pre-defined* rendezvous path. For example, the path can be determined by first selecting the mobile sensor array that most frequently meets the static reference array as the first hop, i.e. the mobile sensor array with the most rendezvous with a reference station, and recursively selecting the most frequently met array as the next hop. It is possible that a sensor array A meets a sensor array B frequently and the reference array R infrequently. In this case, we omit all the rendezvous between A and R. Determining the optimal rendezvous paths is out of the scope of this chapter. We present multi-hop calibration based on a calibration graph with multiple parents per node in Chapter 5 and refer interested readers to [SHT15] for calibration parent selection strategies and [FRD17] on reference placement schemes to ensure network-wide calibrability. *(ii)* Our calibration scheme mainly serves as an important data cleaning technique to guarantee data accuracy and consistency in big air quality data collected by mobile sensor arrays. We assume the dataset has been pre-processed to filter incomplete samples due to, e.g., GPS failures as in [SHWT14].

3.4 SCAN: Multi-hop Calibration for Sensor Arrays

In this section, we present in detail our novel multi-hop calibration method. In Section 3.4.1, we show that state-of-the-art *multiple least-squares* (MLS) is not suitable for multi-hop sensor array calibration due to its error accumulation over multiple hops. We tackle this problem by modifying MLS in order to make it suitable for multi-hop calibration. We formulate our *sensor array network calibration* (SCAN) solution as a new constrained least-squares regression in Section 3.4.2 and present a closed-form solution in Section 3.4.3.

3.4.1 Limitations of Multiple Least-Squares

Multiple least-squares (MLS) minimizes the squared calibration error in each individual hop $h = q$, that is the deviation of virtual references $\hat{Y}_{(q)} = \overline{Y}_{(q-1)}$ taken by array $A_{(q-1)}$ and calibrated signals $\overline{Y}_{(q)} = \overline{B}_{(q)}\overline{X}_{(q)}$ taken by array $A_{(q)}$. For simplicity of presentation we set $\overline{X} = \overline{X}_{(q)}$, $\overline{B} = \overline{B}_{(q)}$ and $\hat{Y} = \hat{Y}_{(q)} = \overline{Y}_{(q-1)}$ in all the following sections. Hence, the minimization problem looks as follows [BR14]

$$\min_{\overline{B}} \operatorname{tr}\left(\left(\hat{Y} - \overline{B}\,\overline{X}\right)\left(\hat{Y} - \overline{B}\,\overline{X}\right)^T\right) \tag{3.4}$$

where *tr* is the *trace* operator [BR14]. The solution to (3.4) [RTSH08a] is given by

$$\overline{B} = \hat{Y}\overline{X}^T\left(\overline{X}\,\overline{X}^T\right)^{-1}. \tag{3.5}$$

Since the traces of a virtual reference array are usually imperfect, i.e.,

$$\hat{Y} = Y + N_2, \tag{3.6}$$

where N_2 are k noise components related to the sensor readings of the virtual reference sensor, e.g., due to calibration error, a major challenge in multi-hop calibration is *error accumulation* over the rendezvous path [SHT15]. We apply MLS to multi-hop sensor array calibration at hop $h = q$ with $q > 0$, where N_1 denotes the noise when reading the sensors of the array $A_{(q)}$ that needs to be calibrated, i.e., $\overline{X} = B^{-1}(Y + N_1)$. The calibration matrix $\overline{B}$ calculated by MLS becomes

$$\begin{aligned}\overline{B} &= (Y+N_2)(Y+N_1)^T B^{-T} B^T \left[(Y+N_1)(Y+N_1)^T\right]^{-1} B\\ &= \left[YY^T + YN_1^T + N_2Y^T + N_2N_1^T\right]\left[YY^T + YN_1^T + N_1Y^T + N_1N_1^T\right]^{-1} B.\end{aligned} \tag{3.7}$$

Because we assume independent noise and no correlation between noise and phenomena, it holds that $YN_1^T = N_1Y^T = N_2Y^T = N_2N_1^T = \mathbf{0}_{k\times k}$ and finally

$$\overline{B} = \left[YY^T\right]\left[YY^T + N_1N_1^T\right]^{-1} B. \tag{3.8}$$

From (3.8) we see that MLS underestimates $\overline{B}$ under the presence of noise N_1 with $N_1 N_1^T \neq \mathbf{0}_{k\times k}$, better known as *regression dilution* or *bias towards zero* [FT00]. This effect grows with increasing sensor noise variance $N_1 N_1^T$. Thus, calibration parameters $\overline{B}$ estimated by MLS depend on sensor noise. The bias towards zero of the parameters of a calibrated array affects the uncalibrated array of the consecutive hop. As a result calibration error is accumulated over the whole rendezvous path, as we will show in the simulations and real-world experiments in Section 3.5.

3.4.2 Sensor Array Network Calibration

To reduce error accumulation in multi-hop sensor array calibration, we propose a solution formulated as a constrained least-squares regression referred to as *sensor array network calibration* (SCAN). As far as we are aware of, this is the first formulation of the multi-hop sensor array calibration problem as a novel constrained least-squares regression with a closed-form solution.

We first define the symmetric matrices $\hat{Y}\hat{Y}^T = YY^T + N_2 N_2^T$ and $\overline{X}\,\overline{X}^T = B^{-1}(YY^T + N_1 N_1^T)B^{-T}$. For the following results, suppose $\overline{X}\,\overline{X}^T$ and $\hat{Y}\,\hat{Y}^T$ are non-singular. The SCAN regression problem can be formulated as follows:

$$\boxed{\begin{array}{ll} \underset{\overline{B}}{\text{minimize}} & \operatorname{tr}\left(\left(\hat{Y} - \overline{B}\,\overline{X}\right)\left(\hat{Y} - \overline{B}\,\overline{X}\right)^T\right) \\ \text{subject to} & \overline{B}\,\overline{X}\,\overline{X}^T\overline{B}^T = \hat{Y}\hat{Y}^T \end{array}} \qquad \begin{array}{r} (3.9) \\ (3.10) \end{array}$$

Thus, SCAN minimizes the least-squares error (3.9) with the constraint on the regression parameters stated in (3.10). As we will show later, (3.10) reduces the bias towards zero and eliminates it under certain realistic assumptions. To the best of our knowledge, we are the first to formulate this constraint regression problem.

3.4.3 Closed-form Solution to SCAN

Degrees of freedom. Let $\overline{X} = U_X D_X V_X^T$ be the singular value composition of the matrix $\overline{X} \in \mathbb{R}^{k\times n}$, where $U_X \in \mathbb{R}^{k\times k}$ and $V_X \in \mathbb{R}^{n\times k}$ are orthogonal matrices and $D_X \in \mathbb{R}^{k\times k}$ is a diagonal matrix holding the singular values of $\overline{X}$ on the diagonal. Similarly, let $\hat{Y} = U_Y D_Y V_Y^T$ be the singular value decomposition of $\hat{Y} \in \mathbb{R}^{k\times n}$. The constraint in (3.10) can be formulated as the following equivalent equation

$$\begin{aligned} \overline{B}\,\overline{X}\,\overline{X}^T\overline{B}^T &= \hat{Y}\hat{Y}^T \\ \overline{B}U_X D_X V_X^T V_X D_X^T U_X^T \overline{B}^T &= U_Y D_Y V_Y^T V_Y D_Y^T U_Y^T \\ \overline{B}U_X D_X^2 U_X^T \overline{B}^T &= U_Y D_Y^2 U_Y^T \end{aligned}$$

$$
\begin{aligned}
D_Y^{-1}U_Y^T\overline{B}U_XD_XD_XU_X^T\overline{B}^T U_YD_Y^{-1} = I& \\
\Leftrightarrow F = D_Y^{-1}U_Y^T\overline{B}U_XD_X \wedge FF^T = I,& \qquad (3.11)
\end{aligned}
$$

where $F \in \mathbb{R}^{k\times k}$ is an orthogonal matrix. We can re-formulate the constraint (3.10) to

$$\overline{B} = U_YD_YFD_X^{-1}U_X^T, \qquad (3.12)$$

where the orthogonal matrix F defines the *degrees of freedom* of the calibration matrix $\overline{B}$.

Minimizing least-squares error. We now present a method to determine F that minimizes the least-squares error (3.9). Recall the minimization problem (3.9), we observe

$$
\begin{aligned}
\min_{\overline{B}} \operatorname{tr}\left(\left(\hat{Y}-\overline{B}\,\overline{X}\right)\left(\hat{Y}-\overline{B}\,\overline{X}\right)^T\right) &= \min_{\overline{B}} \operatorname{tr}\left(\hat{Y}\hat{Y}^T + \overline{B}\,\overline{X}\,\overline{X}^T\overline{B}^T - \overline{B}\,\overline{X}\hat{Y}^T - (\overline{B}\,\overline{X}\hat{Y}^T)^T\right) \\
&= \max_{\overline{B}} \operatorname{tr}\left(\overline{B}\,\overline{X}\hat{Y}^T\right). \qquad (3.13)
\end{aligned}
$$

Thus, we simplify the original minimization problem to a maximization problem stated in (3.13). Recall constraint (3.12) on $\overline{B}$, it follows

$$
\begin{aligned}
\max_{\overline{B}} \operatorname{tr}\left(\overline{B}\,\overline{X}\hat{Y}^T\right) &= \max_F \operatorname{tr}(U_YD_YFD_X^{-1}U_X^TU_XD_XV_X^TV_YD_YU_Y^T) \\
&= \max_F \operatorname{tr}\left(U_YD_YFV_X^TV_YD_YU_Y^T\right) \\
&= \max_F \operatorname{tr}\left(FV_X^TV_YD_Y^2\right). \qquad (3.14)
\end{aligned}
$$

We define $Q = V_X^TV_YD_Y^2 \in \mathbb{R}^{k\times k}$ and its singular value decomposition $Q = U_QD_QV_Q^T$ and followingly

$$\max_F \operatorname{tr}\left(FV_X^TV_YD_Y^2\right) = \max_F (FQ) = \max_F \left(FU_QD_QV_Q^T\right). \qquad (3.15)$$

The solution for F that maximizes (3.15) and (3.14), respectively, is given by

$$F = V_QU_Q^T. \qquad (3.16)$$

From (3.12) and (3.16) it follows that

$$\boxed{\overline{B} = U_YD_YV_QU_Q^TD_X^{-1}U_X^T} \qquad (3.17)$$

minimizes the least-squares error (3.9) under the constraint (3.10).

Proof. The proof for the solution to (3.14) is similar to the *orthogonal Procrustes problem* [Sch66] as well as a consequence of the Courant-Fischer theorem [BR14]. It holds

$$\operatorname{tr}(FQ) = \operatorname{tr}\left(FU_QD_QV_Q^T\right) = \operatorname{tr}\left(\left[FU_QD_Q^{\frac{1}{2}}\right]\left[V_QD_Q^{\frac{1}{2}}\right]^T\right) = \left\langle FU_QD_Q^{\frac{1}{2}}, V_QD_Q^{\frac{1}{2}}\right\rangle, \qquad (3.18)$$

where $\langle \cdot, \cdot \rangle$ is the inner product [BR14]. According to the *Cauchy-Schwarz* theorem [BR14] it follows

$$\left\langle FU_Q D_Q^{\frac{1}{2}}, V_Q D_Q^{\frac{1}{2}} \right\rangle \leq \left\| FU_Q D_Q^{\frac{1}{2}} \right\|_2 \left\| V_Q D_Q^{\frac{1}{2}} \right\|_2, \tag{3.19}$$

where the equality holds if $F = V_Q U_Q^T$ and thus maximizes $\mathrm{tr}(FQ)$, i.e.,

$$\begin{aligned} \mathrm{tr}(FQ) &= \left\langle FU_Q D_Q^{\frac{1}{2}}, V_Q D_Q^{\frac{1}{2}} \right\rangle \\ &= \left\langle V_Q D_Q^{\frac{1}{2}}, V_Q D_Q^{\frac{1}{2}} \right\rangle \\ &= \left\| D_Q^{\frac{1}{2}} \right\|_2 \left\| D_Q^{\frac{1}{2}} \right\|_2 \\ &= \mathrm{tr}(D_Q). \end{aligned} \tag{3.20}$$

Thus, $F = V_Q U_Q^T$ is an orthonormal transformation that maximizes $\mathrm{tr}(\overline{B}\,\overline{X}\hat{Y}^T)$ and consequently minimizes the least-squares error (3.9) under constraint (3.10). □

3.4.4 Discussions

Existence. There always exists a solution to the SCAN regression. The closed-form expression of the regression parameters $\overline{B}$ stated in (3.17) involves the singular value decomposition of $\overline{X} \in \mathbb{R}^{k\times n}$, $\hat{Y} \in \mathbb{R}^{k\times n}$ and $Q \in \mathbb{R}^{k\times k}$. Because there always exists a singular value decomposition for any general real matrix [BR14], there also exists a regression parameter matrix $\overline{B} \in \mathbb{R}^{k\times k}$ according to (3.17).

Relationship to GMR. For the two-dimensional calibration problem where we calibrate a single sensor on a single reference sensor to perform an offset and gain calibration, i.e., $k = 1$, the constraint (3.10) of SCAN reduces to $\overline{B} = \pm\sqrt{\frac{\overline{YY}^T}{\overline{XX}^T}}$. This is in fact the solution for the regression gain parameter according to GMR [Woo41].

No bias towards zero property. We now show that the calibration matrix obtained from our SCAN regression is free from regression dilution even in presence of noise in the measurements. Let $\overline{B}$ a solution that minimizes the least-squares error (3.9) and satisfies constraint (3.10), then it follows by (3.1) and (3.6)

$$\overline{B}B^{-1}\left(YY^T + N_1N_1^T\right)B^{-T}\overline{B}^T = YY^T + N_2N_2^T. \tag{3.21}$$

In the case where the variance of the noise of the two sensor arrays are equal and N_1 and N_2 are uncorrelated, we assume

$$N_1N_1^T = N_2N_2^T = NN^T = \lambda I, \tag{3.22}$$

where $\frac{\lambda}{n} \in \mathbb{R}$ describes the variance of noise components in N and $I \in \mathbb{R}^{k \times k}$ is the unit matrix. Further, let $YY^T + NN^T = UD^2U^T$ be the eigenvalue decomposition of the symmetric matrix $YY^T + NN^T = YY^T + \lambda I$ and accordingly it holds $YY^T = U(D^2 - \lambda I)U^T$. Hence, (3.21) simplifies to

$$\begin{aligned} \overline{B}B^{-1}\left(YY^T + NN^T\right)B^{-T}\overline{B}^T &= YY^T + NN^T \\ \overline{B}B^{-1}(UD^2U^T)B^{-T}\overline{B}^T &= UD^2U^T \\ (D^{-1}U^T\overline{B}B^{-1}UD)(D^{-1}U^T\overline{B}B^{-1}UD)^T &= I \\ \Leftrightarrow G = D^{-1}U^T\overline{B}B^{-1}UD \, \wedge \, GG^T &= I, \end{aligned} \tag{3.23}$$

where G is an orthogonal matrix. Recall the maximization problem stated in (3.13), it follows

$$\begin{aligned} \max_{\overline{B}} \operatorname{tr}\left(\overline{B}\,\overline{X}\hat{Y}^T\right) &= \max_{\overline{B}} \operatorname{tr}\left(\overline{B}\,B^{-1}(Y + N_1)(Y^T + N_2^T)\right) \\ &= \max_{\overline{B}} \operatorname{tr}\left(\overline{B}\,B^{-1}YY^T\right) \\ &= \max_{G} \operatorname{tr}\left(UDGD^{-1}U^TYY^T\right) \\ &= \max_{G} \operatorname{tr}\left(UDGD^{-1}U^TU(D^2 - \lambda I)U^T\right) \\ &= \max_{G} \operatorname{tr}\left(G\left(D^2 - \lambda I\right)\right), \end{aligned} \tag{3.24}$$

where due to the diagonal form of $(D^2 - \lambda I)$ and (3.23) the maximum is achieved if $G = I$ [Sch66]. In this case the calculated regression parameters $\overline{B} = B$. That means, even if $NN^T \neq \mathbf{0}_{k \times k}$, SCAN estimates the true underlying calibration parameters B. Hence, the calibration parameters $\overline{B}$ calculated by SCAN are not affected by regression dilution, whereas MLS underestimates the calibration parameters due to its bias towards zero.

In a real-world deployment we cannot assume that (3.22) ideally holds. In this case $\overline{B}$ does depend on a relation between N_1N_1 and N_2N_2 as stated in (3.21). In Section 3.5 we discuss this assumption in detail and experimentally show that SCAN reduces error accumulation over multiple hops and outperforms various state-of-the-art calibration techniques.

Relaxing the zero-mean variable assumption. Both $\hat{Y}$ and $\overline{X}$ contain in general rows with a non-zero mean. In this case, the calibration also needs to compensate for an offset term C, see (3.2). This is done similarly to MLS [BR14] and basic GMR [Woo41]. Let $\tilde{Y} = \hat{Y} + \operatorname{mean}(\hat{Y})$ and $\tilde{X} = \overline{X} + \operatorname{mean}(\overline{X})$ be non-zero mean representations of $\hat{Y}$ and $\overline{X}$, where $\operatorname{mean}(\hat{Y}) \in \mathbb{R}^{k \times n}$ and $\operatorname{mean}(\overline{X}) \in \mathbb{R}^{k \times n}$ are mean values of each row in $\hat{Y}$ and $\overline{X}$, respectively. It follows

$$\tilde{Y} = \overline{B}\tilde{X} + C$$

$$\hat{Y} + \text{mean}(\hat{Y}) = \overline{B}\left(\overline{X} + \text{mean}(\overline{X})\right) + C. \tag{3.25}$$

Let $C = \text{mean}(\hat{Y}) - \overline{B}\text{mean}(\overline{X})$, then (3.25) equals to the calibrated measurements by SCAN with zero-mean variables. The calibrated measurements are calculated as $\overline{Y} = \overline{BX} + \text{mean}(\hat{Y}) - \overline{B} \cdot \text{mean}(\overline{X})$. Accordingly we can reformulate the constraint (3.10) using non zero-mean variables and the calculated offset to

$$\begin{aligned}&\left(\hat{Y}+\text{mean}(\hat{Y})\right)\left(\hat{Y}+\text{mean}(\hat{Y})\right)^T=\left(\overline{B}\left(\overline{X}+\text{mean}(\overline{X})\right)+C\right)\left(\overline{B}\left(\overline{X}+\text{mean}(\overline{X})\right)+C\right)^T\\ &\Leftrightarrow \hat{Y}\hat{Y}^T+\text{mean}(\hat{Y})\cdot\text{mean}(\hat{Y})^T=\overline{BXX}^T\overline{B}^T+\text{mean}(\hat{Y})\cdot\text{mean}(\hat{Y})^T, \end{aligned}\tag{3.26}$$

which is equal to constraint (3.10) of our SCAN regression with zero-mean variables. Thus, the offset calculation does not affect the *no bias towards zero* property, i.e., error accumulation.

Relaxing the squared calibration matrix assumption. So far we assumed that a sensor array consisting of k different sensors is always calibrated to an array with k sensors as well. In some situations this setup is not suitable. For instance, although a sensor of an array is cross-sensitive to a certain phenomenon it might not be possible to calibrate the array to said phenomenon due to the lack of highly accurate reference measurements. Therefore, a calibration of k sensors to l references with $1 \leq l < k$ is in certain cases required.

Let $\tilde{Y} = S(Y + N_2)$ be l phenomena measured by a calibrated sensor array. $S \in \mathbb{R}^{l \times k}$ is a matrix with $S_{ij} = 1$ if phenomenon $j \in \{1, \dots, k\}$ equals the phenomenon $i \in \{1, \dots, l\}$ measured by the calibrated array and all other elements equal 0. Further, it holds $SS^T = I$. Let $\tilde{B} \in \mathbb{R}^{l \times k}$ be the calibration matrix of an uncalibrated sensor array to the l phenomena. Accordingly the constraint (3.10) on $\tilde{B}$ looks as follows

$$\tilde{B}B^{-1}(YY^T + N_1N_1^T)B^{-T}\tilde{B}^T = S(YY^T + N_2N_2^T)S^T. \tag{3.27}$$

In a similar way to (3.21), we define the matrix

$$\tilde{G} = D^{-1}U^TS^T\tilde{B}B^{-1}UD, \tag{3.28}$$

with $\tilde{G}\tilde{G}^T = I$. We can use the same result in (3.24), i.e.,

$$\max_{\tilde{B}} \text{tr}\left(\tilde{B}\overline{X}\tilde{Y}^T\right) = \max_{\tilde{G}} \text{tr}\left(\tilde{G}\left(D^2 - \lambda I\right)\right), \tag{3.29}$$

and see that again the maximum is achieved with $\tilde{G} = I$ and, hence, it follows from (3.28) that we find the true calibration parameters $\tilde{B} = SB$. Hence, the *no bias towards zero* property is preserved.

In the case of non-equal noise variances, the constraints of the calibration with l and k phenomena differ. Unlike MLS, SCAN calculates different calibration parameters depending on the number of phenomena in $\hat{Y}$, which is elaborated in more detail in Section 3.5.

3.5 Experimental Evaluation

This section presents the evaluations of SCAN on both simulated and real-world datasets.

3.5.1 Simulation

We first show through simulations the advantages of SCAN over multiple least-squares (MLS, see Section 3.4.1), which motivates our multi-hop sensor array calibration scheme, as well as performance comparison with the state-of-the-art calibration schemes. We then investigate the robustness of SCAN with various impacting factors.

Setup. We artificially generate 20 different reference sensor arrays $Y_{(h)} \in \mathbb{R}^{k \times n} \sim \mathcal{N}(\ln Y_{(h)}; 0, \gamma)$ representing n ground-truth measurements of k different phenomena with standard deviation $\gamma \in [0.2, 0.45]$. The log-normal distribution models typical air pollution and meteorological measurements in an urban area [SZW+04, Kah73]. These reference arrays serve as basis for a rendezvous path consisting of 20 noisy sensor arrays. In all experiments the number of physical phenomena k is set to 4. The measurements of each array at hop h with $h \in \{0, 1, \dots, 19\}$ are defined by $\overline{X}_{(h)} = B_{(h)}^{-1}(Y_{(h)} + N_1)$, where $B_{(h)}$ is a general matrix with randomly chosen entries $B_{ij} \in [0.2, 2]$ and $N_1 \in \mathbb{R}^{k \times n} \sim \mathcal{N}(0, \sigma^2)$ describes sensor noise. These sensor arrays are in line with our findings in Section 3.5.2 and Section 3.5.3, where we use real-world sensors that exhibit similar cross-sensitivity intensities, as well as with existing literature [SGV+15, JHW+16, SHT15]. In the first hop $h = 0$ the low cost sensor array is calibrated with the reference array. In all consecutive hops the measurement of the previously calibrated array $\hat{Y}_{(h)} = \overline{B}_{(h-1)} \left(B_{(h-1)}^{-1}(Y_{(h)} + N_1) \right)$ are used as reference array. We use 500 samples for training the calibration parameters and 500 samples for evaluation. The 500 samples for evaluating the performance of an array are also used to train the calibration parameters of the consecutive sensor array. Each experiment is run 100 times with re-sampled reference measurements and ground-truth calibration parameters $B_{(h)}$.

Evaluation metrics. We calculate the difference between the calculated calibration parameters $\overline{B}$ and the ground-truth B using the Froebinus norm, defined as

$$\left\| B - \overline{B} \right\|_F = \left(\operatorname{tr} \left(\left(B - \overline{B} \right) \left(B - \overline{B} \right)^T \right) \right)^{\frac{1}{2}}. \tag{3.30}$$

Additionally, we investigate the root-mean-square value of the calibration

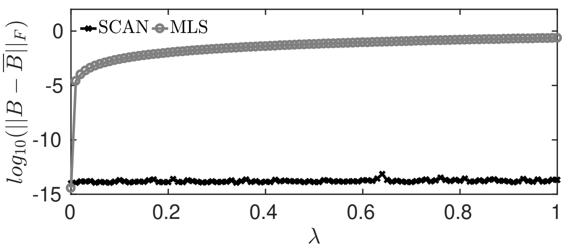


Figure 3.3: Difference between estimated $\overline{B}$ and ground-truth calibration parameters B for different sensor noise variance λ.

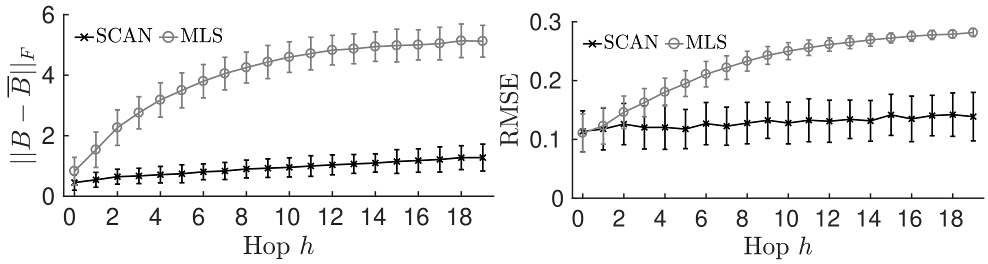
error (RMSE) of the calibrated sensor array to one phenomena $y_4 \in Y$, i.e.,

$$RMSE = \left(\frac{1}{n} \sum_{j=1}^{n} \left(y_{4j} - \overline{y}_{4j} \right)^2 \right)^{\frac{1}{2}}. \tag{3.31}$$

Overall performance. In this set of experiments, we first verify the correctness of our theoretical findings by comparing with MLS, and then compare SCAN with various state-of-the-art multi-hop calibration schemes.

True calibration matrix estimation. As theoretically shown in Section 3.4.4, SCAN is able to calculate the true underlying calibration parameters $\overline{B} = B$ if the virtual reference and uncalibrated array suffer from noise with equal variance, i.e., $N_1 N_1^T = N_2 N_2^T = \lambda I$. We first investigate this finding and simulate the assumption for an arbitrary hop in the rendezvous path by enforcing $\hat{Y}_{(h)} \hat{Y}_{(h)}^T = Y_{(h)} Y_{(h)}^T + \lambda I$ and $\overline{X}_{(h)} \overline{X}_{(h)}^T = B_{(h)}^{-1} (Y_{(h)} Y_{(h)}^T + \lambda I) B_{(h)}^{-T}$. Figure 3.3 reflects our theoretical findings in Section 3.4.4. We observe that SCAN is able to calculate the true B up to machine precision independent of λ. Further, we observe the estimation by MLS is dependent on $\lambda \neq 0$.

Error accumulation. Figure 3.4 shows the results of the calibration at each hop when the standard deviation σ of the noise for all sensor arrays is randomly chosen from $[0.05, 0.2]$. Due to the small number of samples, the noise components are in general correlated, i.e., $N_1 N_1^T$ has a general form with dominating diagonal elements, which relates to real-world sensor arrays as we will show later in Section 3.5.2. In Figure 3.4a we observe that our SCAN approach clearly outperforms MLS with respect to how accurate the underlying $B_{(h)}$ is estimated in each hop. This finding is reflected in the calibration error over multiple hops in Figure 3.4b, where SCAN achieves a reduced error accumulation over 20 hops compared to MLS. In fact, already after 5 hops MLS shows a 49% higher average error than SCAN. Although SCAN outperforms MLS, we also observe an error accumulation for SCAN over 20 hops. Relative to the first hop the error increased by less than 22% at the last hop. The reasons are the different



(a)

(b)

Figure 3.4: SCAN calculates calibration parameters $\overline{B}$ that deviate over all hops to a less extent from the ground-truth B compared to MLS, depicted in Figure 3.4a. Hence, SCAN calibrates in Figure 3.4b all sensor arrays of a 20-hop calibration path with clearly reduced error accumulation.

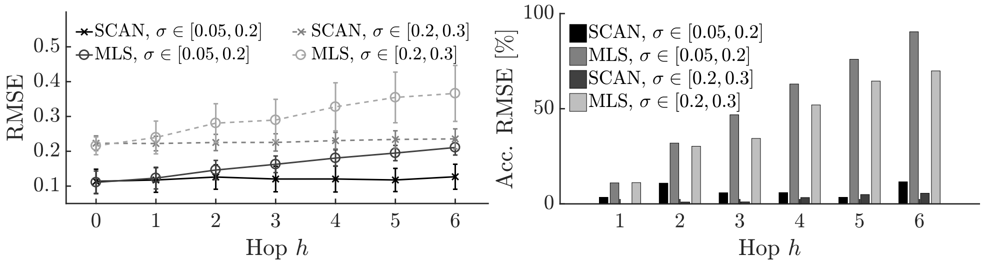
Method	$h = 0$	$h = 5$	$h = 19$
SCAN	0.12 ± 0.04	0.12 + ±0.04	0.14 + ±0.04
MLS	0.11 ± 0.04	0.21 ± 0.05	0.34 ± 0.07
GMR	0.4 ± 0.06	0.54 ± 0.18	0.54 ± 0.15
TLS	0.14 ± 0.07	0.62 ± 0.3	4.2 ± 2.7
ANN	0.19 ± 0.3	> 100	> 100
Draper	0.13 ± 0.05	0.24 + ±0.05	0.34 + ±0.07
Tofallis	0.14 ± 0.06	0.25 + ±0.06	0.35 + ±0.07

Table 3.1: Calibration error of different techniques for one- and multi-hop calibration.

sensor noise variances σ^2 of each individual sensor, i.e., assumption (3.22) is not satisfied, and the different measurement ranges the sensor samples of each array lie in. The impact of these two reasons is elaborated in detail in the following section.

Comparison with other techniques. Table 3.1 summarizes the calibration error for seven different techniques, namely our sensor array network calibration regression (SCAN), 2-dimensional geometric mean regression (GMR [SHT15]), multiple least-squares (MLS, see Chapter 2), total least-squares (TLS [GL80]), artificial neural networks (ANN [SGV+15]) and two different generalizations of the basic geometric mean regression to multiple dimension by Draper et al. [DY97] and Tofallis [Tof02]. The comparison in Table 3.1 is based on 100 runs with a 20-hop rendezvous path and $\sigma \in [0.05, 0.2]$.

As previously shown SCAN clearly outperforms MLS over multiple hops due to the reduction of error accumulation. For one-hop calibration MLS is more accurate. The 2-dimensional GMR calibration is not able to compensate for cross-sensitivities. Thus, the calibration error on all hops is severely larger compared to the other multi-dimensional regression techniques. TLS is a multi-dimensional error-in-variables regression, i.e., it assumes that both X and Y are affected by errors, which relates to our



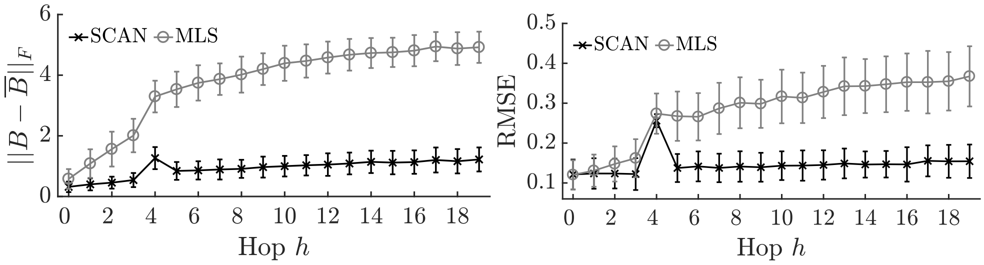
(a) (b)

Figure 3.5: Calibration error over seven hops (Figure 3.5a) and total error accumulation (Figure 3.5b) after each hop relative to $h = 0$ when the sensor signals are affected by low ($\sigma \in [0.05, 0.2]$) and high noise ($\sigma \in [0.2, 0.3]$).

multi-hop calibration problem. However as experimentally shown, TLS is also suffering from an error accumulation that is even more severe compared to MLS over multiple hops. ANNs have been widely used for sensor array calibration due to their ability of modelling complex and possibly non-linear relationships between sensor and reference measurements. In our setup we use a simple network with 1 hidden layer and 10 neurons [DE92]. We show that already after 5 hops the accuracy of the calibration exceeds all other techniques. This can be traced back to the fact that neural networks tend to overfit and are usually a prominent choice for compensating for non-linear cross-sensitivities. Finally, we compare SCAN to two different generalizations of the geometric mean regression to multiple dimensions introduced by Draper et al. [DY97] and Tofallis [Tof02]. From our results we observe that both generalizations suffer from considerable error accumulation over multiple hops and, thus, are not suited for multi-hop sensor array calibration. Overall, we conclude that our SCAN approach outperforms all other tested techniques for multi-hop calibration and MLS is the best choice for one-hop calibration.

Robustness. In the following set of experiments, we investigate the impact of various factors that may violate the assumptions of our theoretical analysis to assess the robustness of SCAN.

Impact of noise variance. The amount of sensor noise has an important effect on the multi-hop calibration performance. Figure 3.5a shows the comparison between different intervals for σ, i.e., $\sigma \in [0.05, 0.2]$ (bold lines) and $\sigma \in [0.2, 0.3]$ (dashed lines). For both SCAN and MLS the average error over all hops is increased with increasing variance in sensor noise. As shown before, SCAN reduces the error accumulation compared to MLS. In fact, the accumulated error over 7 hops is below 13% with SCAN as shown in Figure 3.5b. When calibrating with MLS the amount of noise affects the error accumulation per hop. After $h = 5$ the accumulated error per hop decreases and the calibration error stabilizes with MLS under high sensor noise. This is because the calculated calibration



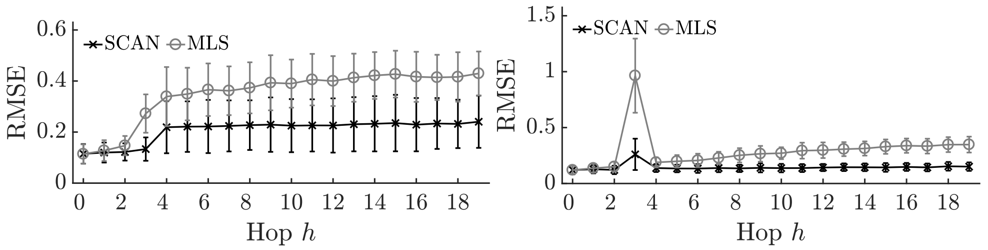
(a) (b)

Figure 3.6: Impact on calibration parameter estimation (Figure 3.6a) and error (Figure 3.6b) of increased sensor noise on a single sensor array at hop $h = 4$. SCAN is able to accurately calculate calibration parameters independent of the sensor noise of its parent.

parameters $\overline{B}$ already strongly deviate from the ground-truth due to the bias towards zero of $\overline{B}$ with MLS, as theoretically proven in (3.8). With lower noise MLS achieves a lower average error compared to SCAN and MLS under higher noise conditions, but accumulated over 95% error over 7 hops.

Impact of increased noise. So far we affected all sensors with noise whose variance lied in a narrow interval. In a real-world deployment this assumption may not hold. Noise of individual arrays might be significantly higher compared to others, especially in networks with heterogeneous nodes. In Figure 3.6 the standard deviation of the noise of all sensors in the uncalibrated array at hop $h = 4$ is set to $\sigma = 0.3$ and the one of its neighbours $h = 3$ and $h = 5$ to $\sigma = 0.12$. Despite the increased calibration error of the array at hop $h = 4$, arrays at hops $h = [5, 19]$ are not notably affected when calibrating with SCAN. In contrast the increased noise at $h = 4$ has a strong impact on the parameter estimation by MLS and, hence, sensor arrays at hops $h = [5, 19]$ suffer from a high calibration error. This result shows that the calibration parameter estimation of a child sensor array by SCAN depends on the child sensor array noise but barely on the noise of its parent.

Impact of variable measurement range. In the previous experiments we always assumed that the measurements of all sensor arrays along the rendezvous path include samples over the whole range of the underlying phenomena, i.e., the interval defined by the smallest and highest absolute value of the phenomena measurements. For instance, the ambient temperature in a deployment ranges from 5 °C to 30 °C, then so far we assumed that all arrays along a rendezvous path contain temperature samples within the whole range. This assumption is difficult to enforce in a real-world deployment. Physical phenomena typically show large variance depending on time and location. Therefore, capturing samples over the whole range of all target phenomena is a difficult task without



(a)

(b)

Figure 3.7: A lower measurement range between arrays at $h = 4$ introduces an increased calibration error in Figure 3.7a for both SCAN and MLS in all hops $h > 3$. In Figure 3.7b an increased measurement range only affects sensor array at hop $h = 3$ without effect on the subsequent sensor arrays.

enforcing rendezvous between sensor arrays.

In Figure 3.7 we show the effect of a rendezvous between array at hop $h = 3$ and $h = 4$ with a different measurement range compared to all other sensor arrays in the path. We decrease the range by a factor of 3 in Figure 3.7a and observe that both SCAN and MLS are affected. The lower measurement range introduces an increase of the calibration error for sensor array at $h = 4$ and all its subsequent arrays. The reason for this effect is that the calibration parameter estimation for all sensor arrays with hop $h > 4$ is only valid for the smaller range of array at $h = 4$ and, thus, the array can only be used to calibrate subsequent arrays with equal measurement range. In Figure 3.7b we increase the measurement range of the rendezvous between arrays at $h = 3$ and $h = 4$ by a factor of 3. This only affects the sensor array at $h = 3$ but not the ones at hops $h = [4, 19]$. The sensor array at $h = 3$ is not able to estimate parameters for its whole range, only for the smaller range of its parent array. This does not affect sensor arrays at hops $h = [4, 19]$ because their measurement range is equal to the one at hop $h = 2$.

We conclude that it is important to use calibration data including a large measurement range of equal size for all sensor arrays in a rendezvous path. In a real-world deployment the density of the network and, hence, the number of rendezvous between nodes, can therefore have an impact on the accuracy of the calibration. In Chapter 5 we also extend our SCAN method with uncertainty metrics for the calibrated measurements. This approach helps to filter samples which are potentially error-prone and tackles the variable range effect.

Impact of number of reference signals. Unlike MLS, SCAN calculates different calibration parameters of an array to a certain reference depending on the number of other references in the regression setup. As shown in Section 3.4.3 this does however not violate the *no bias towards zero* property under the assumption of equal sensor noise of parent and

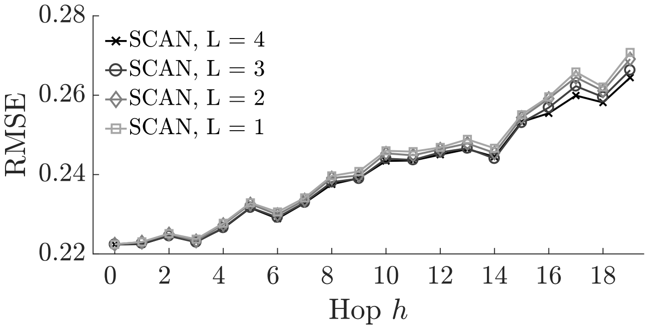


Figure 3.8: Calibration error of our SCAN method with different numbers l of reference signals. The effects on the error accumulation are negligible.

child array. We now show that even if this assumption does not hold, the impact on the calibration error with different number of reference signals is negligible. We investigate the calibration error to phenomena $y_4 \in Y$ with different number of reference signals $l \in \{1, 2, 3, 4\}$. That means, if $l = 1$ each sensor array is only calibrated to y_4. Accordingly, if $l = 2$ the array is calibrated to two phenomena $[y_3; y_4]$, and so forth. The standard deviation of the noise is set to $\sigma \in [0.2, 0.3]$ for all sensor arrays. Figure 3.8 shows the calibration error $y_4 \in Y$ over multiple hops for different l. We observe, that the number of references does not have a notable impact on the error. The differences only become clear after multiple hops, where SCAN with $l = 4$ achieves the lowest average error. In fact, at $h = 19$ the calibration errors for all l values differ by less than 3% .

Summary. Through extensive simulations, we demonstrate that our SCAN scheme outperforms the state-of-the-arts in multi-hop calibration for sensor arrays. SCAN still accumulates mild errors in presence of numerous practical factors including noise variance, noise dependency, variable measurement range, etc., but still significantly outperforms MLS, which becomes unusable in these situations. We further extend SCAN by calculating different uncertainty metrics for each calibrated samples in Chapter 5. This uncertainty information is used to further reduce the calibration error at each hop due the different unavoidable practical factors investigated in this section.

In the next two sections, we evaluate the performance of SCAN on two real-world air quality datasets, showing both the advantages of our novel SCAN scheme and the lessons learned in calibrating large-scale air quality measurements collected by low-cost mobile sensor arrays. We mainly compare our SCAN scheme with MLS, which we show is the best choice for one-hop sensor array calibration, and GMR, the state-of-the-art for multi-hop sensor calibration.

3.5.2 Metal Oxide Sensor Array

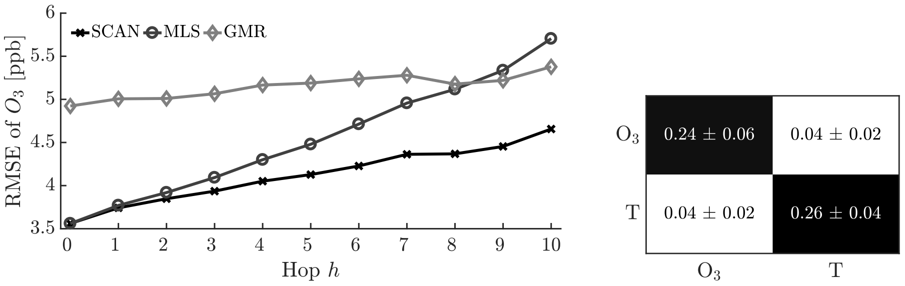
In this set of experiments, we evaluate the performance of SCAN on measurements collected by low-cost metal oxide based gas sensors. The sensor is a prototype featuring an array of sensing layers, whereof each individual layer in the array exhibits a different sensitivity to certain environmental gases. The sensitivity of these layers can be controlled by setting the temperature of the sensing layer to a specific value, which is a common technique for metal oxide based sensors [BKW07]. Due to its small size and low power consumption, the sensor is suitable for a large variety of Internet-of-Things, wearable devices and crowdsensing applications.

Setup. We deploy the low-cost metal oxide based gas sensors next to the static and highly accurate air monitoring station in Duebendorf, Switzerland (see Figure 2.3) to monitor the ambient ozone (O_3) concentration. We heat the sensing layers of the gas sensors to different temperatures to simulate heterogeneous sensor nodes with different noise levels. Further, we deploy a temperature sensor [Sen16] for ambient temperature measurements. The sensors are sampled in an interval of 30 sec. As sensor array we use measurements from two of the gas sensor layers and from the temperature sensor, i.e., $k = 3$. We collect measurements from July 2015 to July 2016.

Rendezvous path. We construct a rendezvous path of $h = 11$ hops because there were at most eleven sensor arrays measuring at the same time. Each calibration is trained on 200 samples within a time frame of at most two weeks. The calibration is tested on 200 different samples within the consecutive two weeks. The whole evaluation over one year of data is executed 500 times with a randomly re-sampled calibration path for every execution. For the calibration in the first hop to measurements of the reference station we use MLS instead of our SCAN method. As shown in Table 3.1 and also confirmed by evaluating the dataset, MLS performs better for one-hop calibration and, thus, the overall multi-hop calibration achieves better accuracy.

Ground-truth. For training the calibration parameters of the array at the initial hop and evaluating the calibration of all arrays in the path we use highly accurate ozone and temperature measurements from the station as reference.

Performance. We compare the multi-hop sensor array calibration performance of our SCAN approach to MLS. 2-Dimensional GMR is used to compare sensor array and simple sensor calibration by solely calibrating the measurements from a single sensing layer in the array that exhibits the highest sensitivity to ozone. Again, we use three different metrics (see Section 2.4.3), i.e., the RMSE (2.7), its normalized form RMSE_σ (2.8) and the R^2 (2.9) value, to benchmark the performance of the calibration.



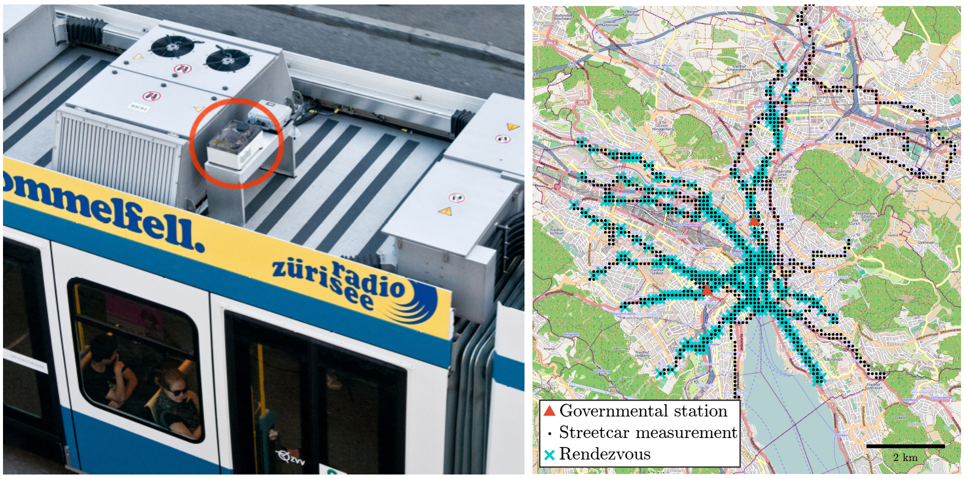
	O_3	T
O_3	0.24 ± 0.06	0.04 ± 0.02
T	0.04 ± 0.02	0.26 ± 0.04

(a) (b)

Figure 3.9: (a) Sensor array calibration error over multiple hops when calibrated to ambient ozone (O_3) concentration. SCAN clearly outperforms MLS and GMR. (b) Covariance matrix of sensor noise after sensor calibration to O_3 and temperature.

Error accumulation. The average RMSE over the whole time period for each calibrated array per hop is depicted in Figure 3.9a. The results highlight the two major advantages of our SCAN sensor array calibration method; *(i)* SCAN reduces error accumulation over multiple hops compared to MLS and *(ii)* improves the overall calibration accuracy compared to the simple offset and gain calibration based on GMR. Over eleven hops MLS increases the average calibration error from 3.5 ppb to 5.85 ppb which is equal to a relative increase of over 60% error. SCAN considerably reduces this error accumulation to a relative increase of 30%. Compared to the simple sensor calibration with GMR, sensor array calibration based on SCAN improves the calibration error by up to 1.36 ppb O_3 concentration. Over all hops SCAN achieves a 16% to 38% smaller error than GMR and an up to 23% smaller error than MLS. Due to the large error accumulation of MLS the normalized RMSE$_\sigma$ increases from 1.17 to 1.87, while it stayed between 1.17 and 1.54 for SCAN and 1.61 and 1.72 for GMR. The R^2 value stayed constant around 0.8 for GMR, and reduced slightly from 0.93 to 0.87 for both MLS and SCAN. Overall these results indicate accurate ozone measurements, however SCAN achieves the best performance in all aspects and is able to provide the most accurate measurements in a large-scale deployment with long calibration paths.

Investigation on noise characteristics. As shown in Section 3.4.4, our SCAN approach is able to completely remove the bias towards zero of its regression parameters if the noise components of the sensor array are uncorrelated, see assumption (3.22). We therefore investigate if this assumption holds for our sensor arrays. Because N_1 is not directly measurable without the knowledge of the true calibration matrix $\overline{B}$, we calculate the covariance matrix $N_2 N_2^T$ of the calibration error N_2 from the one-hop sensor array calibration using MLS, i.e., the best possible calibration for each array. The reference traces for ozone and temperature have different units and ranges. Thus, we scale them to assure that the noise components of the low-cost sensor arrays have equal impact, i.e.,



(a) (b)

Figure 3.10: (a) Measurement box mounted on top of a streetcar. (b) Locations of sensor measurements and rendezvous between two or more sensor arrays.

equal variance. Figure 3.9b shows the average covariance matrix for all eleven sensor arrays. We observe that the two noise components, i.e., the calibration error of ozone O_3 and temperature T, are not completely uncorrelated. However, the diagonal elements dominate the off-diagonal elements, i.e., the $N_2 N_2^T$ matrix resembles a diagonal matrix. This result relates to our assumptions on the noise components in Section 3.2 and Section 3.5.1. In conclusion, our SCAN approach considerably improves the calibration accuracy even if the noise components are correlated, i.e., (3.22) is not satisfied.

3.5.3 Mobile Air Pollution Sensor Network

In this set of experiments, we evaluate the performance of SCAN on a large dataset from a real-world mobile air pollution sensor network deployment [LFS+12]. We first demonstrate the benefits of using multi-hop calibration over one-hop calibration in a real-world deployment. Further, we show that our sensor array network calibration outperforms MLS and GMR for arrays of sensors with low selectivity.

Setup. The dataset was collected by air quality measurement boxes mounted on top of ten streetcars of the public transport network in the city of Zurich, Switzerland, depicted in Figure 3.10a. Each measurement box includes an *MICS-OZ47* ozone (O_3) [SGX08], an *Alphasense CO-B4* carbon monoxide (CO) [Alp15] and a *Sensirion SHTC1* temperature sensor [Sen16]. These sensors have been previously tested by the method presented in Chapter 2. The sampling interval of the sensor array is set to 30 sec. Each box is equipped with a GPS receiver to record location and time of each measurement. All the sampled data is transferred to a global database via GSM. We filter incomplete samples due to network connection or sensor failures and only focus on calibrating errors induced by sensor dependencies as mentioned in Section 3.1. In Section 2.6 we uncover a substantial dependency on ambient temperature of both the deployed low-cost O_3 and CO sensor. Therefore, we augment the gas sensors with the temperature sensor to an array and calibrate the concurrent measurements to the corresponding high-quality reference signals.

Rendezvous path. As shown in Figure 3.10b, the streetcars meet occasionally at different times and locations. We exploit these rendezvous between streetcars to construct rendezvous paths. In the initial hop of a rendezvous path one sensor array is calibrated with measurements from the governmental station. In all following hops an already calibrated sensor array calibrates an uncalibrated one using measurements at rendezvous between two streetcars. We define a rendezvous between two sensor arrays as within a time interval of $5\,min$ and spatial closeness of $50\,m$, which has been validated through extensive testing in [SHT15]. Each rendezvous path consists of at least 200 samples per sensor array pair within a time window of at most four weeks. The average length of all rendezvous paths is 3 hops.

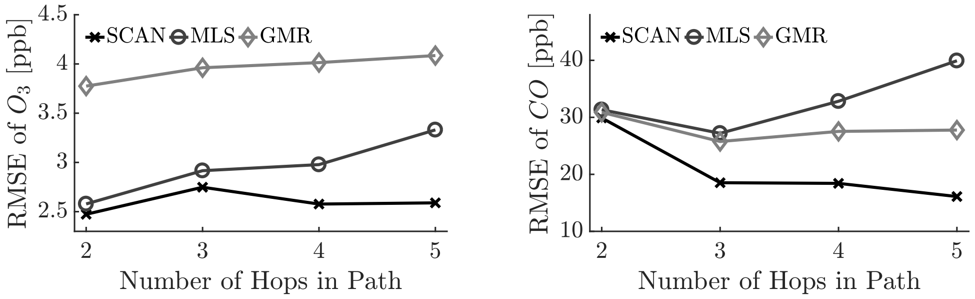
Ground-truth. As reference signals we use O_3, CO and temperature measurements from two static monitoring stations within the deployment area. We use data from both stations to train the calibration of any initial hop. This assures that we use sufficient measurements, i.e., at least 200 samples, for one-hop calibration and removes potential error sources, such as variable measurement range as described in Section 3.5.1. Each calibration is calculated on a training dataset within four weeks. The performance of each sensor array calibration is evaluated on a dataset from the consecutive four weeks. These two datasets do not overlap in time. The location of these two monitoring stations is depicted in Figure 3.10b. The streetcars operate daily on various routes in the city and, thus, achieve a high spatial and temporal measurement coverage, illustrated in Figure 3.10b.

Performance. To benchmark the multi-hop calibration performance we compare a particular array at the last hop of a rendezvous path and its baseline calibration. The baseline calibration is obtained by calibrating the array to reference measurements from the governmental stations using MLS, i.e., the best possible calibration for the array in question. This approach removes effects on the overall measurements accuracy of each array that cannot be compensated by calibration, such as mobility influences [AMM16b, AMM16a].

Overall, we use roughly $2.2 \cdot 10^5$ rendezvous between seven of the ten streetcars and over 550 different rendezvous paths out of 56 million sensor samples recorded from March 2014 to March 2016 for evaluation.

	Max. calibrated arrays	Avg. calibrated arrays
One-hop	3	32.7%
Multi-hop	7	94.3%

Table 3.2: Number of calibrated arrays with one-hop and multi-hop calibration



(a) (b)

Figure 3.11: Calibration error of sensor arrays compared to baseline depending on the number of hops in the rendezvous path. SCAN achieves best results for both target pollutants, ozone (O_3: Figure 3.11a) and carbon monoxide (CO: Figure 3.11b).

One-hop versus multi-hop calibration. This evaluation aims to show the necessity of multi-hop calibration when calibrating datasets collected by a mobile sensor network. Depending on the availability of reference stations one-hop calibration is able to calibrate only a fraction of all the sensor arrays in the network. In the streetcar deployment there are two highly accurate reference stations that can be used to calibrate sensor arrays on streetcars that pass by these two stations. All the remaining sensors can be calibrated using multi-hop calibration. Table 3.2 shows that at most three out of seven arrays can be calibrated with one-hop calibration whereas it is possible to calibrate all arrays with multi-hop calibration. Over the two years the multi-hop approach calibrates in average over 94% of all arrays every month which is an improvement by a factor of 2.88 compared to one-hop calibration. This result clearly shows the benefit of using multi-hop calibration over one-hop calibration. Reasons for not calibrating 100% of all sensors are the irregular schedules and routes of the streetcars over two years or missing sensor data. As a consequence certain streetcars did not meet other streetcars often enough for a successful calibration on a monthly basis. An interesting possibility for future work is to dynamically optimize the rendezvous path selection to ensure accurate and network-wide calibration. We expect that in a more dense network the rate of calibrated sensors is even higher.

Sensor array versus simple sensor calibration. This evaluation aims to validate the effectiveness of our SCAN scheme on large-scale real-world mobile sensor networks. Figure 3.11 shows the difference in calibration error compared to the baseline depending on the number of hops in the rendezvous path for ozone (O_3, Figure 3.11a) and carbon monoxide (CO, Figure 3.11b). We show again the two important contributions of our proposed SCAN approach. SCAN achieves the lowest error accumulation over multiple hops and sensor array calibration improves the overall accuracy compared to the simple sensor calibration based on GMR for both target pollutants. Overall SCAN achieves an up to 42% lower calibration error than GMR and up to 60% compared to MLS over 5 hops, which can dramatically benefit further quantitative analysis such as validation on air pollution models and health studies.

In conclusion, multi-hop calibration considerably increases the number of calibrated sensor arrays in the deployment. Additionally, our proposed SCAN method improves the calibration accuracy, especially when compared to the simple sensor calibration based on GMR.

Discussions. While evaluations on the mobile sensor network deployment show that our SCAN scheme outperforms the state-of-the-arts, we also observe that the benefits are not as distinct as in the simulations and the metal oxide sensor arrays. Here we discuss multiple issues to further clarify the applicability of SCAN.

- **SCAN works in practice even if the assumptions on noise may not hold.** As shown in section Section 3.5.2, noise from real-world sensors can correlate but exhibit typically low cross-correlation, therefore SCAN still works and notably outperforms MLS.

- **SCAN shows more notable benefits for long rendezvous paths.** The current deployment consists only of seven nodes with 5 hops at most. SCAN reduces calibration errors by 42% to 60% after 5 hops, and we expect more notable gain in minimizing error accumulation over multiple hops for larger networks with many low-cost sensor arrays.

- **SCAN is mainly designed to tackle cross-sensitivities in multi-hop calibration.** There are possibly additional error sources in a mobile sensor deployment, such as network faults and mobility. For instance, recent studies [AMM16b] tackle the impact of slow sensor dynamics in a mobile deployment. SCAN can be combined with these schemes to further reduce errors in sensor measurements. In Chapter 5 we present a way to tackle other error sources that are investigated in Section 3.5.1, i.e., high sensor noise and variable measurement ranges, by enhancing SCAN with uncertainty metrics.

3.6 Summary

Monitoring air pollution with mobile wireless sensor networks has received increasing research interest in recent years. Low-cost, portable air quality sensors on the market introduced the opportunity for large scale deployments with high spatial coverage. However, maintaining high-quality measurements from low-cost air pollution sensors is challenging. Low-cost air pollution sensors not only drift over time, but are also cross-sensitive to interfering gases and depend on meteorological conditions. Therefore, calibrating the air quality measurements is vital if the dataset is used for quantitative analysis such as air pollution modelling and health studies. In Chapter 2, we explore constructing sensor arrays to compensate for cross-sensitivities and meteorological dependencies, yet existing multi-hop calibration techniques lead to dramatic error accumulation when applied to sensor arrays, making multi-hop sensor array calibration an open question.

In this chapter, we propose *sensor array network calibration* (SCAN), a novel constrained multi-dimensional linear regression technique, that *(i)* calibrates sensor arrays and *(ii)* reduces error accumulation over multiple hops. We theoretically prove that SCAN is free from regression dilution, the root cause of error accumulation, even in presence of measurement noise. Extensive evaluations on two datasets of 56 million samples collected over three years demonstrate the benefits of SCAN over the state-of-the-art calibration techniques. SCAN compensates for all major limiting factors to maintain high-quality measurements from low-cost air pollution sensors, thus improving reliability of large-scale air quality datasets. We envision SCAN as a general calibration technique for not only air pollution monitoring, but also a range of mobile sensor network applications with dependent sensors, especially in participatory and crowdsourcing sensing.

4

Enabling Personal Air Pollution Monitoring on Wearables

Accurate, portable and personal air pollution sensing devices enable quantification of individual exposure to air pollution, personalized health advice and assistance applications. Wearables are promising (e.g., on wristbands, attached to belts or backpacks) to integrate commercial off-the-shelf gas sensors for personal air pollution sensing. Yet previous research lacks comprehensive investigations on the accuracies of air pollution sensing on wearables. In this chapter, we propose W-Air, an accurate personal multi-pollutant monitoring platform for wearables, which we prototype on a wristband with two low-cost metal oxide gas sensors. Through an extensive measurement study we discover an additional limiting factor of low-cost air quality sensors when used in wearables: human-generated emissions. These emissions pose a substantial challenge for our system by causing non-linear interference, another error source presented in Section 1.2. We observe that the linear regression methods presented in Chapter 2 and Chapter 3 are not able to successfully counteract this interference. Thus, we tackle the non-linear response of our low-cost sensors by applying a neural network with complex modelling capabilities. As summarized in Section 1.2, neural networks are powerful tools to resolve complex and non-linear cross-sensitivities of sensor arrays. In particular, W-Air adopts a sensor array calibration scheme to recover high-fidelity ambient pollutant concentrations from the human interference and leverages a tailored neural network with shared hidden layers to boost calibration parameter training with fewer measurements. W-Air also utilizes semi-supervised regression to facilitate post-deployment calibration model updating with little user intervention. Evaluations demonstrate that W-Air reports accurate measurements both with and without human interference and is able to automatically learn and adapt to new environments.

4.1 Introduction

Air pollution affects human health, productivity and comfort. A prominent problem is caused by ambient ozone (O_3), which contributes to respiratory symptoms when people engage in outdoor exercises and activities [Lip89]. Continuous ozone monitoring is not only important due to its toxicity but also due to its strong impact on various other major pollutants such as nitrogen oxides (NO_x) [FPPJ99]. Similarly, popular pollutants in indoor environments such as carbon dioxide (CO_2) and volatile organic compounds (VOC) may cause discomfort, headache or the sick building syndrome [Jon99, RSC97]. Thus, providing air pollution information to individuals enables them to understand and improve the air quality of their living environments.

There is a growing demand to increase the spatio-temporal resolution of air pollution monitoring. Governmental institutions deploy expensive high-end air pollution sensors in a few stations across a city, like the one in Figure 1.4. The measurements only suffice to estimate the average pollution exposure experienced by the majority of the population for urban planning and policy making. In recent years low-cost sensors have been deployed by researchers and agencies in static stations [ZLH13] or on mobile vehicles [HSW+14] to build air pollution maps for citizens [HSW+15]. However, these maps have in general low accuracy and low spatio-temporal resolution and, therefore, can be misleading when assessing personal exposure for quantitative studies and applications. Pollutants in indoor environments are known to cause various health related problems similar to outdoor air pollution [Jon99] and, thus, monitoring indoor air quality with high resolution is of equal importance. However, similar to outdoor environments, expensive sensors are typically equipped at only a few locations in large buildings and are not able to provide high spatio-temporal resolution data [KSB+16]. Due to the complex heterogeneity of air pollutants [Mon01] and the diverse moving patterns of individuals [JLT+11], *personal* air pollution sampling is therefore necessary for meaningful personal exposure analysis [OKR+00, PXM+14].

Motivation. Personal air pollution monitoring for quantitative health and well-being applications requires accurate, convenient, quasi-continuous collection of heterogeneous data. For instance, researchers record both individual micro-environment (e.g., temperature, wind speed, noise, air pollution) and psychological states (e.g., skin temperature, heart rate) via a set of wearable sensors to investigate the impact of urban environments on citizen's health and quality of life [Eur17, SD17, NKSdD15]. Simultaneous sensing of biological and environmental data enables personalized advice and assistance applications that promote a healthier lifestyle and improved health-care prevention.

To design air pollution sensing devices for the above studies and applications, we primarily focus on integrating low-cost gas sensors into wrist-worn devices. *(i)* Many users wear wristbands or smart-watches most of the day, providing long-term closeness to the body and exposure to the ambient air. *(ii)* Biological parameters are commonly measured by wearables such as wristbands [Eur17, SD17]. Integrating environmental sensors into wearables can improve the compactness and usability of the infrastructure for environment-related physiological applications.

Challenges. Despite various portable air pollution sensing devices [BEMRB13, DAK+09, JLT+11, OB15, ZLYX15], there is a void in air pollution sensing on wearables. Due to their small size, metal oxide (MOX) sensors have been widely adopted to measure a wide spectrum of important air pollutants [KCS14] on portable devices [NVZ+12, PXM+14]. However, similar to other low-cost gas sensors, MOX sensors suffer from low selectivity, i.e., they are *cross-sensitive* to various substances in the air, i.e., the most common error source we tackle throughout this thesis. Since human beings can emit multiple gases through natural skin oils [WW10], cosmetics [Wes16], textiles [RLC14] or respiration [FP99], these gases may interfere with the measurements of the *ambient* atmospheric pollutants when the sensors are placed close to the human body (e.g., on wristbands, attached to belts or backpacks). As a result, MOX sensors equipped in wearables are not only cross-sensitive to the natural ambient pollutants, which we intend to monitor, but also to human-generated substances, which distort the measurements. Hence, it is crucial to investigate and filter the *human interference* to acquire accurate measurements of ambient air pollution.

In this chapter, we design and implement *W-Air, the first-of-its-kind personal air pollutant monitoring platform for wearables with low-cost gas sensors.* As a proof-of-concept, we focus on monitoring concentrations of ambient O_3 and CO_2, two important outdoor and indoor gas pollutants. We aim to answer the following questions.

- *Does the wearable setting introduce new challenges for accurate air pollution sensing?* Through measurement studies, we observe that the measurements of low-cost MOX gas sensors significantly deviate from the ground truth when there is close human presence. This effect introduces another error source (see Section 1.2): *non-linear response*. A linear calibration model, which we use in Chapter 2 and Chapter 3, yields measurement errors that make it impossible to draw any conclusions about the actual O_3 and CO_2 concentration when there is human interference. We also note that such human interference is not restricted to the wrist-worn setting. Attaching the gas sensors to other popular wearable settings, such as belts or backpacks, still yield notable interference from human emissions.

- *How to enable accurate air pollution monitoring on wearables?* A key finding is that the human interference can be characterized by increasing VOC and H_2 concentrations as well as increasing temperature and humidity conditions. Different MOX gas sensors react differently to these interfering concentrations due to their individual specific cross-sensitivities. Therefore, we design a neural network based calibration framework by jointly considering measurements from two different low-cost MOX sensors to eliminate the non-linear interference of emissions caused by human beings. Since W-Air is expected to provide ambient O_3 and CO_2 concentration measurements accurately both with and without human interference, large amounts of measurements in both cases are needed for training the calibration parameters. To improve the usability of W-Air, we apply a neural network architecture with shared layers and a semi-supervised regression technique to boost the training process with *fewer samples* and *little user intervention* compared to traditional neural network architectures.

Contributions and road-map. The main contributions of this chapter are summarized as follows:

- We discover and characterize the human interference on the measurements of low-cost MOX sensors, a crucial yet largely overlooked issue to obtain accurate ambient gas measurements on wearables (e.g., wrist-worn, attached to a belt or a backpack). This human interference is causing another error source affecting the sensors, which we tackle in this chapter: *non-linear response*.

- We propose an effective sensor array calibration scheme to recover ambient outdoor O_3 and indoor CO_2 concentrations from low-cost gas sensor readings during human interference situations. Our calibration method is based on a neural network that uses measurements from two different MOX gas sensors and a temperature sensor to accurately estimate O_3 and CO_2 concentrations. We further utilize a shared layer architecture to bootstrap the supervised training process of our neural network with fewer samples compared to typical neural network architectures. Additionally, we apply semi-supervised regression for parameter updating with little human intervention. W-Air sets a new standard for portable air pollution monitoring with easy maintenance.

- We prototype the above design using COTS MOX gas sensors integrated on a wristband platform. To the best of our knowledge, W-Air is the first working air pollutant monitoring platform using COTS gas sensors for wristbands and smart-watches. Evaluations show that W-Air is able to measure ambient O_3 and CO_2 concentrations with an error of around 4.3 ppb (parts-per-billion

= 10^{-7}%) and 64 ppm (parts-per-million = 10^{-4}%), respectively. Overall, W-Air achieves a data quality that is sufficient for personal air pollution monitoring [TPP12, PEA08].

In the rest of this chapter, we first review relevant literature (Section 4.2), present a measurement study (Section 4.3) and an exploratory calibration (Section 4.4) on the human interference. Then we elaborate on the design (Section 4.5) and evaluation (Section 4.6) of W-Air. We discuss limitations and future work in Section 4.7 and finally conclude this work in Section 4.8.

4.2 Related Work

W-Air is proposed to meet the need for applications in *crowdsourced environment sensing* and *personal environmental sensing*. The design of W-Air is built upon previous research on *portable air pollution sensing devices* and *non-linear sensor array calibration methods*. We review the closely relevant works as follows.

4.2.1 Crowdsourced Environmental Sensing

Mobile crowdsourcing, or participatory sensing, has been widely adopted for environmental sensing. In crowdsourced environmental sensing, unprofessional users take measurements of the environment *"in the form of an open call"* [How06] with their smartphones or other portable devices to cover a large spatio-temporal range. For example, Ear-Phone [RCK+10] is an end-to-end crowdsourced noise sensing and mapping system that builds an urban noise map by measuring noise from smartphones. It leverages compressive sensing to construct accurate noise maps from the sparse and random crowdsourced noise measurements. Overeem et al. [ORL+13] design techniques to infer ambient temperature from crowdsourced smartphone battery temperature and air temperature reported by meteorological stations. Combined with location and time information, they achieve an average estimation accuracy of 1.45°C. Atmos [NVL15, NVL17] is a crowdsourced weather data application that not only automatically samples smartphone sensors (GPS, temperature, light, pressure), but also allows manual input for current and future weather condition estimation. The Atmos application shows an average accuracy of less than 2.7°C for ambient temperature estimation using such a hybrid (automatic sensing and manual user input) crowdsourcing approach.

Crowdsourced air pollution sensing is also of growing research interest because static air pollution monitoring stations are sparsely deployed in cities and limited in spatial resolution [HSW+14]. Compared with other environmental data, air pollution (e.g., gases and dust) can hardly be measured by smartphone sensors. Hence, the first step for crowdsourced air pollution sensing is to design portable and accurate sensing devices suitable for crowdsourced users, which partially motivates our work.

4.2.2 Personal Environmental Sensing

Due to the complex spatial heterogeneity of pollutants [Mon01], personal environmental sensing is important for quantitative studies such as personal exposure assessment. For example, Oglesby et al. [OKR+00] report that personal air pollution sampling is necessary for short-term analysis of personal ultrafine particle exposure. In many personal environmental sensing applications, the sensing device needs to measure both biological responses and environmental data. Nakayoshi et al. [NKSdD15] investigate outdoor thermal physiology by deploying wearable sensors to record individual microclimate (temperature, humidity, wind speed and radiation) and psychological states (skin temperature, heart rate). Project ESUM [SD17] studies the impact of urban morphology on citizen's social potential (e.g., perception). Environmental conditions (e.g., noise, temperature, illumination, air pollution) and skin conductance responses are recorded when participants walk around the city. Project CONVERGENCE [Eur17] integrates low-power environmental sensors and biological sensors on wearables for new generations of human-machine interfaces and health-care and lifestyle applications.

In these personal environmental sensing studies and applications, participants have to carry a set of sensors to collect environmental and biological data. The complexity and bulkiness of the sensing infrastructure may discourage user engagement and even induce bias for physiological studies. Note that biological or physiological parameters are commonly measured by wearables such as wristbands [SD17]. Therefore, our work explores environmental sensing on wearables to improve the compactness and usability of the sensing devices for environment-related physiological studies and personalized health advice and assistance applications.

4.2.3 Portable Air Pollution Sensing Devices

The need for crowdsourced and personal air pollution monitoring has fostered various portable air pollution sensing devices [BEMRB13, DAK+09, JLT+11, NVZ+12, OB15, PXM+14, TDMP16, ZLYX15]. Table 4.1 summarizes the target pollutants, sensor technologies, device usage, scenarios and applications of existing portable air pollution sensing devices. As summarized in Section 1.1, mainstream COTS air pollution sensor technologies include metal oxide (MOX), electrochemical or optical approaches, where MOX sensors are the most popular for their small sizes [KCS14]. MOX sensors are primarily used for monitoring gaseous

System	Pollutant	Technology	Usage	Scenario	Application
Budde et al. [BEMRB13]	Dust ($PM_{2.5}$, PM_{10})	Optical	Carriable	O	C
AirSense [ZLYX15]	Dust ($PM_{2.5}$)	Optical	Carriable	I + O	C + P
MyPart [TDMP16]	Dust (PM_{10})	Optical	Wearable	I + O	C + P
MAQS [JLT+11]	Gas (CO_2)	Optical	Carriable	I	P
CitiSense [NVZ+12]	Gas (CO, NO_2, O_3)	MOX	Carriable	I + O	C
Common Sense [DAK+09]	Gas (CO, NO_x, O_3)	N/A	Handheld	O	C
Oletic et al. [OB15]	Gas (CO, NO_2, SO_2)	Electrochemical	Handheld	O	C
Piedrahita et al. [PXM+14]	Gas (CO, CO_2, NO_2, O_3)	MOX, Optical	Carriable	I + O	C + P
W-Air	**Gas (CO_2, O_3)**	**MOX**	**Wearable**	**I + O**	**C + P**

Note 1: I - indoor; O - outdoor; C - crowdsourced air pollution sensing; P - personal exposure monitoring.

Note 2: The usage of the device is partially determined by its compactness. Wrist-worn devices are smaller in size than carriable and handheld designs.

Table 4.1: Comparison of portable air pollution sensing devices. W-Air is designed for wearable settings, evaluated in both indoor and outdoor scenarios, and targets both crowdsourced and personal environmental sensing applications.

pollutants such as CO, NO_2, O_3 and VOCs [DAK+09, NVZ+12, PXM+14], while optical sensors are preferable for dust monitoring (e.g., $PM_{2.5}$, where PM stands for particulate matter, and $PM_{2.5}$ means fine particles with a diameter of 2.5 micrometers or less) [BEMRB13, ZLYX15]. Sensing devices with MOX sensors vary in size depending on mechanical designs and functionalities, while platforms integrated with COTS electrochemical or optical sensors can only be handheld or attached to backpacks limited by the form factor of the sensors.

Some works focus on platform integration and qualitative validation. Common Sense [DAK+09] and CitiSense [NVZ+12] design sensor nodes with wireless connection and interfaces for crowdsourced gas pollutant monitoring. AirSense [ZLYX15] integrates dust sensors for personal $PM_{2.5}$ monitoring and conduct feasibility studies in various contexts. Other works investigate the reliability and accuracy of sensors and emphasize more on quantitative evaluation. Oletic et al. [OB15] conduct outdoor field tests for gas pollutant sensing. MyPart [TDMP16] is a wrist-worn PM_{10} sensing device that works both indoors and outdoors with validated accuracy.

The closest to our work is [PXM+14], where the authors measure a set of gas pollutants (CO, CO_2, NO_2, O_3) using low-cost MOX and NDIR sensors, calibrate and validate the sensor readings in both indoor and outdoor scenarios. W-Air is also an accurate, multi-gas sensing device that works both indoors and outdoors. However, W-Air primarily focuses on a wrist-worn setting and identifies the human interference problem. We demonstrate that the human interference problem is generic for other sensor placement (e.g., attached to a belt or backpack), and propose effective calibration schemes to filter such interference. W-Air can also deal with insufficient and imbalanced data, which is important for sensor deployments in the wild, yet largely overlooked in previous studies.

4.2.4 Non-Linear Calibration of Gas Sensors

In Chapter 2 and Chapter 3 we present linear methods for sensor array calibration, e.g., multiple least-squares (MLS) and our SCAN method. We show that these linear methods are able to provide accurate data of both static and mobile low-cost sensor arrays. However, in situations where the cross-sensitivities and relationships between the sensors in an array are complex and exhibit non-linear behaviour, the linear regression methods fail to provide accurate calibrated measurements. Different works tackle this challenge by applying non-linear regression methods. The most prominent approach is to train a neural network that finds complex relationships between all the sensors in the array and a reference sensor [DMP+08, DPMF09, EDS+16, EDS+17, SGV+15, SGV+17, BBR+17]. Multiple investigations show that neural networks, typically in the form of multilayer perceptrons [SGV+15, SGV+17], are able to generate more accurate measurements of complex sensor arrays than linear regression methods [SGV+15, SGV+17]. However, there are multiple downsides when using neural networks or similar machine learning methods in general. Typically, neural networks require abundant training samples [DMP+08], they are prone to overfitting [TLL95], which may lead to poor generalization of the calibration function, and require multiple parameters (e.g., learning rate, number of neurons etc.) to be optimized [GD98]. W-Air also features a neural network to allow accurate calibrated measurements even under the presence of human interference. We tackle the high data amount requirements and overfitting problem by using a flat but tailored network structure coupled with semi-supervised learning as described in detail in Section 4.5.

4.3 Measurement Study

In Chapter 2, we show that despite laboratory and in-field calibration from manufacturers, it is still indispensable to conduct pre-deployment testing and calibration of low-cost gas sensors in the target environment of the final deployment. Many existing portable air pollution monitoring devices aim to measure coarse-grained air quality indices, and their measurements are not validated by highly accurate gas sensors [DAK+09, JLT+11, NVZ+12]. Other works either perform data validation in chambers [BEMRB13] or directly adopt calibration parameters from manufacturers [OB15]. MyPart [TDMP16] conducts in-field validation for *customized* dust sensors. In [PXM+14], the authors report low correlation of O_3 measurements between the metal oxide (MOX) gas sensors and the ground-truth, yet without in-depth investigations.

In this section, we show through field studies the human interference on the measurements of COTS MOX sensors. Such human interference imposes enormous errors when using compact gas sensors to monitor ambient atmospheric pollutants. We mainly focus our measurements study on O_3 for outdoor environments and CO_2 for indoor environments due to *(i)* their importance for air quality assessments and *(ii)* the availability of promising low-cost sensor technology.

4.3.1 Measurement Settings

We deploy different sensors, depicted in Figure 4.1, in different indoor and outdoor settings.

Locations. Outdoors we deploy the sensors on the roof of our academic building in the city-centre and on a balcony in a residential area. Indoors the sensors are deployed either in an office environment, a living-room or a bed-room.

Low-cost sensors. The sensors are two low-cost MOX sensors, namely a *MICS-OZ-47* O_3 sensor [SGX13] (also used in Section 2.6.1 and Section 3.5.3) and a *CCS811* VOC sensor [ams17]. The two sensors are connected to the *Thunderboard Sense* [Sil17], a compact multi-sensor development platform, that acts as a wearable device. Note that the *MICS-OZ-47* O_3 sensor is only providing raw and uncalibrated analog-digital-converter (ADC) values and the *CCS811* sensor already calibrated VOC concentration values within [0,1187] ppb. Additionally, we use the *Si7021* temperature and relative humidity sensor to monitor environmental conditions.

Ground-truth. As ground-truth reference devices we use a *SM-50* O_3 measurement unit [Aer16] in outdoor environments and a *Telaire 6713* CO_2 measurement unit [Amp17] in indoor environments. The *SM-50* sensor provides highly accurate ozone measurements within [0,150] ppb, i.e., typical ambient outdoor concentrations. The *Telaire 6713* measures typical indoor CO_2 concentrations within [400,5000] ppm with high accuracy. Although it is conceivable to also integrate the *Telaire 6713* into wearables due to its form factor, its power consumption of 125 mW is almost 10 times higher than the one of a MOX sensor.

Approach. We use the readings from the VOC sensor to approximate the CO_2 concentration because the *CCS811* sensor is not directly measuring CO_2. This is a widely used approach of available low-cost MOX sensors [RHB18, HHU+10]. Furthermore, CO_2 is mainly used to assess indoor air quality [Jon99]. The sensors on the wristband are sampled at 2 Hz, the O_3 reference sensor every 1 min and the CO_2 reference every 5 sec. Additionally we smooth the wristband sensor readings over a sliding window of 3 sec. We power all the sensors and collect their measurements using a laptop to ensure long-term measurements.

We use the measurements from the MOX sensors to reveal a major problem of state-of-the-art MOX gas sensors: human beings act as a source of interference due to emissions from the skin, clothes or respiration [Wes16]. These emissions can be detected by state-of-the-art MOX sensors and, therefore, interfere with the assessment of the ambient air quality. Related work has shown that the emissions are detectable up to 1 m away of a human being [GS01] and, thus, potentially affect typical wearable air quality monitors. Therefore, during the measurements with human interference a user is wearing the device, depicted as S_H in Figure 4.1, while staying between 1.5 m and 2 m away from the reference sensors. We consider three popular ways of utilization of wearable air quality monitors, namely *(i)* a wrist-band [TDMP16], *(ii)* attached to a belt [FMT+99] and *(iii)* attached to a backpack [BEMRB13, JLT+11], as shown in Figure 4.2. Further, we place a second wearable device (S_N) next to the reference sensor to highlight that the human interference problem is in particular severe when the sensors are close to a human being.

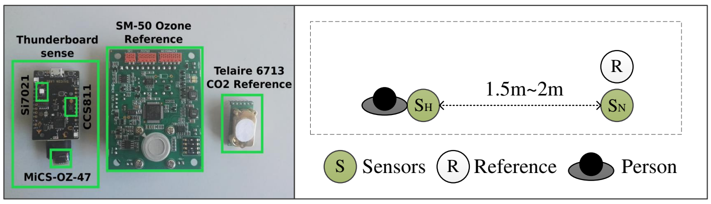
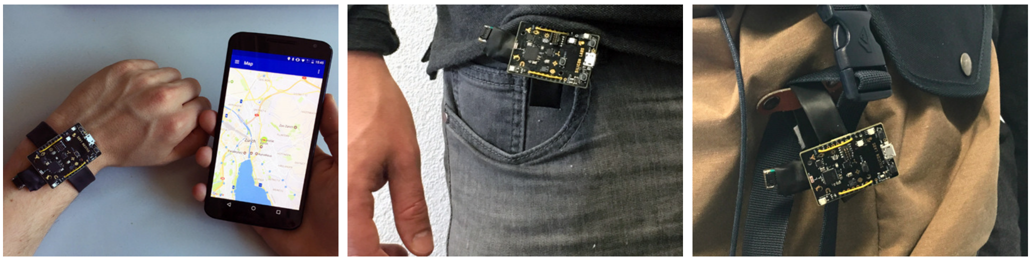


Figure 4.1: Settings of measurement studies: sensors used for the measurement study (left); deployment of the MOX sensors and the reference (right).



(a) (b) (c)

Figure 4.2: Different usages of W-Air: (a) worn as wrist-band, (b) attached to a belt and (c) a backpack

4.3.2 Data Collection

We collected data consisting of approximately 100 hours each in indoor and outdoor environments distributed over 21 days between April and October, 2017. The measurements were conducted at different times during the day, e.g., morning, midday, afternoon, evening and night, and in various weather conditions, e.g., heavy and light rain, cloudy, sunny and windy. We collected both measurements with and without human presence. During human presence situations a user is wearing the wearable device in one of the three different usages shown in Figure 4.2 and sitting or standing next to the corresponding reference device as indicated in Figure 4.1. The user is mainly reading or working with a laptop in both environments. We investigate the impact of additional human activities in Section 4.7.1. The episodes of measurements with human presence last between 15 and 120 minutes and account for a total of approximately 20 hours for each environment. During situations without human presence the sensors are placed next to each other without any human being present in the close-by vicinity. We only use O_3 reference measurements outdoors due to usually very low and short-lived concentrations indoors [Jon99]. In fact, our O_3 reference sensor always reports 0 ppb when used indoors. The CO_2 references are only used indoors because CO_2 is not considered as a major air pollutant with direct and immediate effect on human beings in outdoor settings [Val14].

4.3.3 Observations

This subsection presents the key observations through measurements, which motivate the design of W-Air.

There is a clear impact on the measurements of the two MOX sensors caused by human interference, see Figure 4.3. Figure 4.3a, Figure 4.3b and Figure 4.3c show the measurements of the two MOX sensors outdoors when W-Air is used in three different settings, i.e., as a wrist band, attached to a belt and to a backpack, respectively. The same impact is shown when the sensors are placed indoors in Figure 4.3d, Figure 4.3e and Figure 4.3f. For both environments and all three W-Air usages we observe the same behaviour of the measurements. In the absence of humans, the O_3 measurements (raw and uncalibrated ADC values within $[0, 1]$) and the VOC measurements (within $[0, 1187]$ ppb) of all the sensors vary moderately. This result is expected because the ground-truth O_3 (Figure 4.3a: 30 ppb, Figure 4.3b: 25 ppb, Figure 4.3c: 25 ppb) and CO_2 (Figure 4.3d: 700 ppm, Figure 4.3e: 675ppm, Figure 4.3f: 700ppm) remain constant within the short time periods. As soon as one person is equipping W-Air the readings of both MOX sensors (O_3 and VOC sensor S_H) close to the user immediately increase. In fact, in all cases the O_3 values increase between 15% and 40% relative to the situation without any human presence. Note an increasing raw value is indicating a decreasing O_3 concentration [WYZ+10]. The VOC values show peaks over 600 ppb outdoors and over 200 ppb indoors. Yet the usual VOC concentration measured during non human presence situations is below 100 ppb.

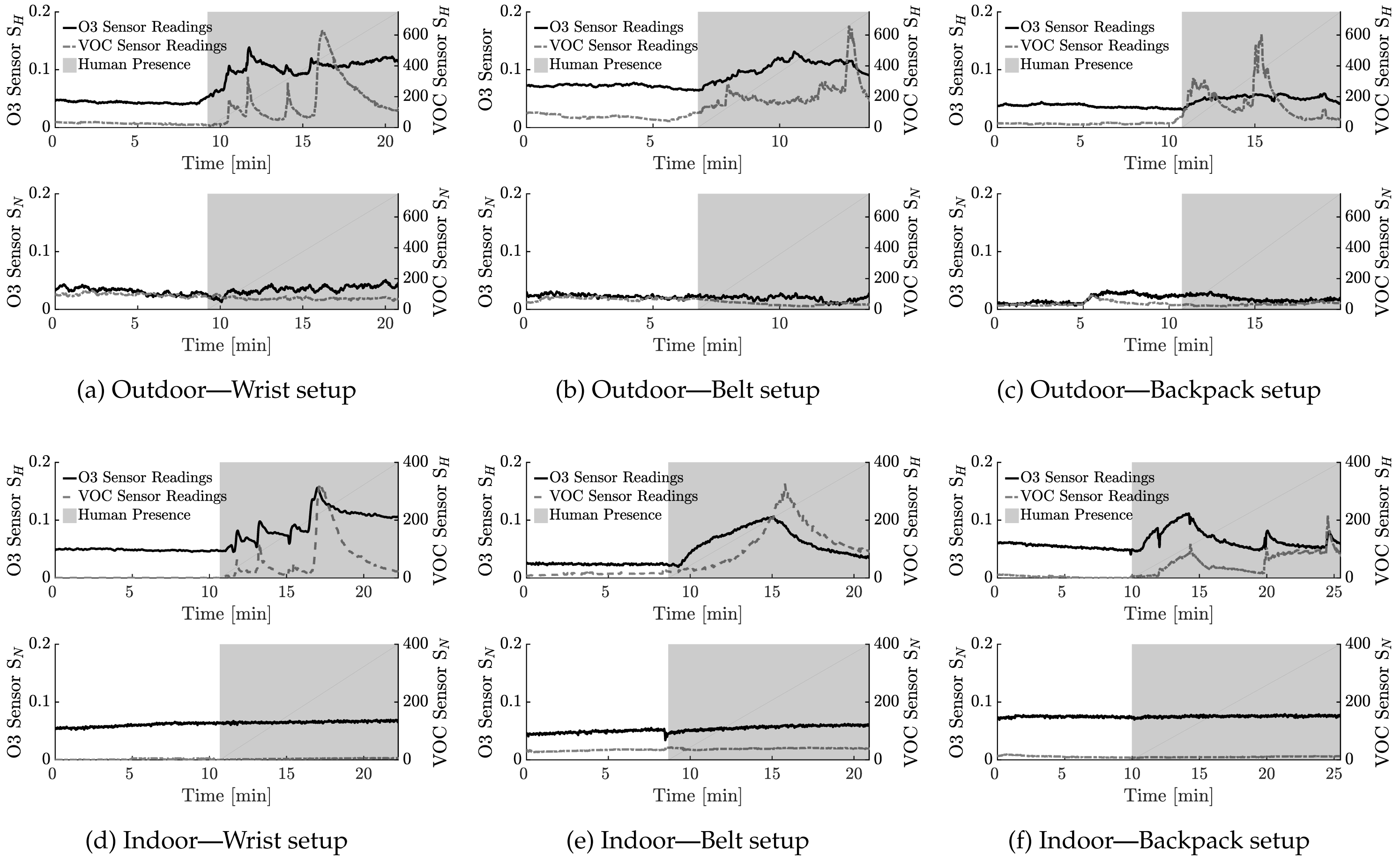
One explanation on such abnormal ambient O_3 and VOC measurements may be different human-generated emissions:

- **Human skin emissions:** Human skin can emanate different VOCs, both from natural skin oils or from ingredients in cosmetic products [GWL+08, Wes16]. These emissions can further be magnified by their reaction with ambient O_3 [WW10]. Previous research [GWL+08] has identified the depletion of ozone while increasing certain organic components in simulated office environments due to the chemical reactions with skin oils.

- **Textile emissions:** The different skins oils and cosmetics are usually also found in clothing [RLC14, Wes16], which constitute as a substantial VOC source. Again O_3 can also react with textiles and increases the VOC concentration while being decreased.

- **Human breath emissions:** Exhaled breath of human beings contains various different VOCs [FP99, Wes16] and hydrogen (H_2) [TSE80]. In indoor environments these emissions can have an important impact on the overall air quality.

As a result during situations with human presence our two sensors generally indicate an *(i)* increased VOC concentration and a *(ii)* decreased



(a) Outdoor—Wrist setup

(b) Outdoor—Belt setup

(c) Outdoor—Backpack setup

(d) Indoor—Wrist setup

(e) Indoor—Belt setup

(f) Indoor—Backpack setup

Figure 4.3: Outdoor (a)-c)) and indoor (d)-f)) measurements of the human interference when W-Air is worn as (a) & (d) a wristband, (b) & (e) attached to a belt and (c) & (f) attached to a backpack.

O_3 concentration. The O_3 sensor is not only measuring the decreasing O_3 concentration in outdoor environments, but also the increasing VOC and H_2 concentrations indoors where background ozone is usually very low. This is due to the typical cross-sensitivity of MOX sensors. Although the O_3 sensor is designed to be most sensitive to O_3 it is also sensitive to VOCs and H_2. Vice-versa the same behaviour holds for the VOC sensor. This cross-sensitivity during human presence situations poses a substantial challenge to capture accurate air quality measurements.

Finally, another possible interfering factor is the increased temperature and humidity levels when the sensors are placed close to a human body. Due to the general dependency of MOX sensors on these factors, the sensors can be additionally influenced. We expect this effect to play a notable role when the user is performing different activities like running or biking, which we elaborate more in Section 4.7.

The two MOX sensors approximately 2 m away are not affected and remain in the reasonable range, indicating no notable human interference. This is because the convective boundary layer of a standing person in quiet ambient air is within $1m$ in diameter [GS01]. Human emissions from more than $1m$ away can be ignored.

We conclude that state-of-the-art MOX sensors are prone to human interference due to their typical cross-sensitivities to VOCs [WYZ$^+$10] and H_2 [RHB18] when used in typical wearable devices. In Section 4.7.1 we also discuss the impact of additional factors on this human interference, such as different human activities or weather situations.

4.4 Exploratory Calibration of Human Interference

Human interference severely downgrades the accuracy of MOX gas sensors. Therefore, we investigate in this section how we can tackle the human interference problem by exploratory sensor calibration. As shown in Chapter 2 and Chapter 3 as well as by a large body of existing research [SGV$^+$15, KBP06, HST12] the calibration of MOX-based air quality sensors is able to provide accurate and stable measurements during situations without human presence. Based on the measurements presented in Section 4.3.1 we investigate in this section the effectiveness of calibration methods during human presence situations.

4.4.1 Methods

Most of the state-of-the-art calibration methods are either based on linear models, e.g., *Multiple Linear Regression* (MLS, Chapter 2) or our SCAN method (Chapter 3), or non-linear models, such as *Artificial Neural Networks* (ANN) [SGV$^+$15]. In general linear calibration models are preferred over non-linear ones due to *(i)* less vulnerability to overfitting, *(ii)* lower computational complexity and *(iii)* reduced training dataset requirements. Non-linear models on the other hand are more suited for complex calibration problems and can in general, if trained successfully, achieve higher data quality. Therefore, we investigate the performance of MLS and ANNs when applied to our collected data. The ANN is based on a multilayer perceptron using two hidden layers with 100 and 10 neurons, respectively.

In order to assess the performance of the models we use different metrics that we already used throughout this thesis, see Section 2.4.3, and are also widely used in assessing the performance of air quality measurements [SGV+15, TPP12].

Root-mean-square-error (RMSE). The RMSE [TPP12] value, which we introduce in Chapter 2: (2.7), is a standard metric to assess the calibration error and is defined as follows

$$RMSE = \left(\frac{1}{n} \sum_{i=1}^{n} (\hat{m}_i - m_i)^2 \right)^{\frac{1}{2}}, \tag{4.1}$$

where $\hat{m}_i$ are calibrated sensor measurements and m_i ground-truth measurements over a window of $i = 1, ..., n$ samples.

Standardized RMSE (RMSE$_\sigma$). We use a standardized RMSE [TPP12], see also Chapter 2: (2.8), to evaluate if the measurements from our W-Air platform suffice for certain applications. The metrics is defined as

$$RMSE_\sigma = \frac{\left(\frac{1}{n} \sum_{i=1}^{n} (\hat{m}_i - m_i)^2 \right)^{\frac{1}{2}}}{\sigma \cdot \left(\frac{1}{n} \sum_{i=1}^{n} (m_i)^2 \right)^{\frac{1}{2}}} = \frac{RMSE}{\sigma \cdot RMS_r}, \tag{4.2}$$

where RMS_r is the root-mean-square value of the reference concentration, i.e., O_3 or CO_2 in our case, and σ is defined as a relative uncertainty measure for a specific pollutant. The European Parliament [PEA08] defines $\sigma = 0.15$ for O_3 measurements. To the best of our knowledge there is no σ defined for CO_2, for simplicity we assume it is also $\sigma = 0.15$. RMSE$_\sigma$ values below 1 indicate accurate measurements fit for any applications and $1 < RMSE_\sigma \leq 2$ indicates data quality that suffices for indicative measurements. Finally, if $RMSE_\sigma > 2$ the data quality is not sufficient for any application.

Coefficient of determination R^2. The $R^2 \in [0, 1]$ value, which is defined in Chapter 2: (2.9), indicates a well-fitted calibration model if the metric is close to 1 and a poorly fitted model for values close to 0.

Prediction confidence. A perfect calibration model is able to estimate the exact ground-truth measurements, i.e., $m_i = \hat{m}_i$. In reality this is rarely the case and the relationship between ground-truth and estimation can be formulated as $m_i = \beta_0 + b_1 \cdot \hat{m}_i$, where $b_0 = 0$ and $b_1 = 1$ for a perfect model.

Target	Method	Inputs	RMSE	RMSE$_\sigma$	R^2	$b_0 \pm c_{95\%}$	$b_1 \pm c_{95\%}$
O_3	Linear	{O_3, T}	7.4 ppb	1.5	0.78	0 ± 2.2	1 ± 0.08
CO_2	Linear	{VOC, T}	81 ppm	0.55	0.88	0 ± 13	1 ± 0.01

Table 4.2: Calibration results for the two target pollutants applied to a linear calibration model (multiple least-squares) with different inputs when sensors are not affected by human interference.

An intercept b_0 far from 0 indicates that predictions are systematically too low or too high and a slope b_1 far from 1 indicates overfitting [SVC+10]. We calculate b_0 and b_1 as well as a 95% confidence intervals $c_{95\%}$ by fitting a simple least squares regression on the calibration prediction and the ground-truth measurements. The confidence intervals $c_{95\%}$ are calculated by bootstrapping the parameters with 10^5 iterations and using the 2.5 and 97.5 percentiles as interval bounds [ET94]. This approach provides a statistical confidence on the model predictions.

Cross-validation. Finally, all the results are based on a 10-fold cross-validation on 2000 samples from our dataset described in Section 4.3.1. As input features we use different combinations of the two MOX sensor readings (O_3 and VOC) and the temperature measurements (T).

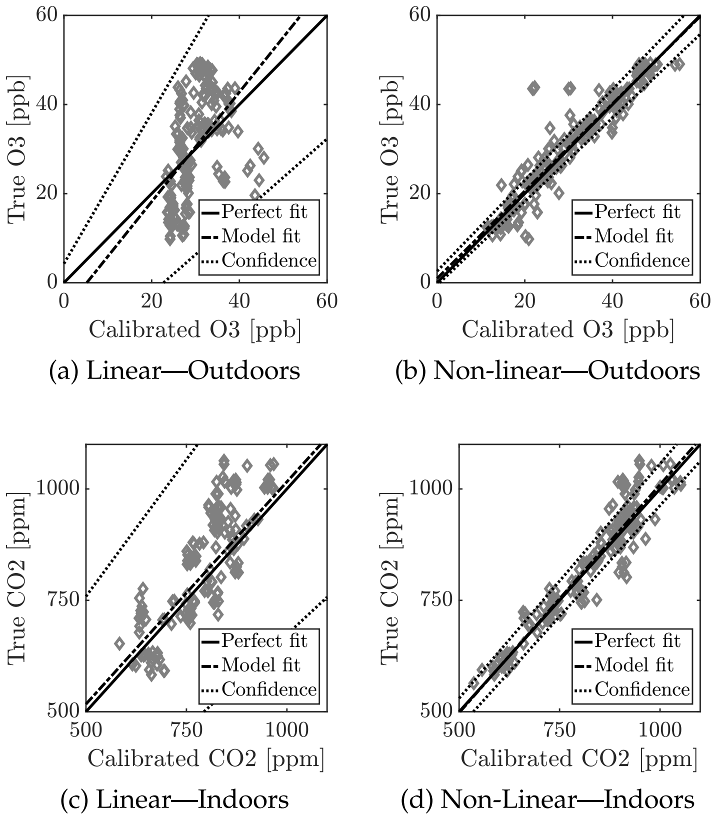
4.4.2 Observation

We first obtain a baseline calibration using only data from situations without any human presence. Such a baseline calibration is sometimes provided by the manufacturer for integrated circuit sensor solutions [SGX08]. Table 4.2 summarizes the average metrics over all 10 folds when we use the linear MLS calibration technique in both environments. Outdoors we predict the O_3 concentration using the O_3 MOX and temperature sensor and indoors we predict the CO_2 concentration using the VOC MOX and temperature sensor. We observe that already a linear model achieves accurate results in both environments. The RMSE$_\sigma$ of approximately 1.5 for the O_3 calibration indicates that the measurements can be used for indicative air quality results. Recent studies also show that the quality of low-cost O_3 sensor measurements can be improved to fulfil the air quality directive using more complex calibration models [SGV+15]. The CO_2 calibration performs even better with a $RMSE_\sigma$ of 0.55. This overall outcome is expected, since similar results are described in Chapter 2 but have also been report in related work, see Section 1.4.3.

In a next step we investigate the calibration during human presence situations. The coherent changes in the measurements of co-located O_3 and VOC sensors when there is human interference lead us to explore joint calibration of O_3 and VOC measurements. Since we aim to perform calibration on resource-constrained wearable and mobile devices, a

Target	Method	Inputs	RMSE	RMSE$_\sigma$	R^2	$b_0 \pm c_{95\%}$	$b_1 \pm c_{95\%}$
O_3	Linear	{O_3, T}	10.8 ppb	2.24	0.15	0 ± 10.8	1 ± 0.36
O_3	Linear	{O_3, VOC, T}	10.7 ppb	2.23	0.16	0.16 ± 11.1	0.99 ± 0.38
O_3	Non-Linear	{O_3, T}	3.9 ppb	0.81	0.89	0.7 ± 1.6	0.98 ± 0.05
O_3	Non-Linear	{O_3, VOC, T}	3.5 ppb	0.72	0.91	0.8 ± 1.4	0.98 ± 0.04
CO_2	Linear	{VOC, T}	182 ppm	1.56	0.1	−395 ± 1680	1.5 ± 2.25
CO_2	Linear	{VOC, O_3, T}	135 ppm	1.17	0.49	1.1 ± 152	0.99 ± 0.19
CO_2	Non-Linear	{VOC, T}	68 ppm	0.6	0.86	33.9 ± 69	0.95 ± 0.07
CO_2	Non-Linear	{VOC, O_3, T}	44 ppm	0.38	0.94	16.8 ± 42	0.98 ± 0.04

Table 4.3: Calibration results for the two target pollutants applied to different methods and inputs during human interference.



(a) Linear—Outdoors

(b) Non-linear—Outdoors

(c) Linear—Indoors

(d) Non-Linear—Indoors

Figure 4.4: Exploratory sensor calibration during human interference using O_3, VOC and temperature sensor readings with: (a) linear calibration (MLS, Section 2.4.2) outdoors; (b) non-linear calibration (artificial neural network [SGV+15]) outdoors; (c) linear calibration indoors; (d) non-linear calibration indoors.

natural question arises *which sensor measurements* we use as inputs and what *model complexity*, e.g., linear versus non-linear, is necessary.

Table 4.3 presents the result of the calibration when using data during human presence situations. We compare the non-linear ANN calibration to the linear MLS one with different inputs. We observe that the linear model for both indoor and outdoor environments performs notably worse than during situations without human interference. The RMSE_σ is in all cases above 1 and therefore not satisfying the data quality goals. Further, the R^2 values are in all cases below 0.5, indicating a poorly fitted model. This result is reflected in the scatter plots in Figure 4.4a and Figure 4.4c where we plot the predictions of the testing dataset of the 10th fold against the corresponding ground-truth. The X-axis and the Y-axis denote the calibrated and the ground-truth measurements, respectively. The fitted model exhibits poor confidence, i.e., the regression line defined by $b_0 \pm c_{95\%}$ and $b_1 \pm c_{95\%}$. This result changes if we use a non-linear ANN calibration model. All metrics drastically improve and perform best when we use all three sensor measurements as inputs. The RMSE_σ is below 1 and, thus,

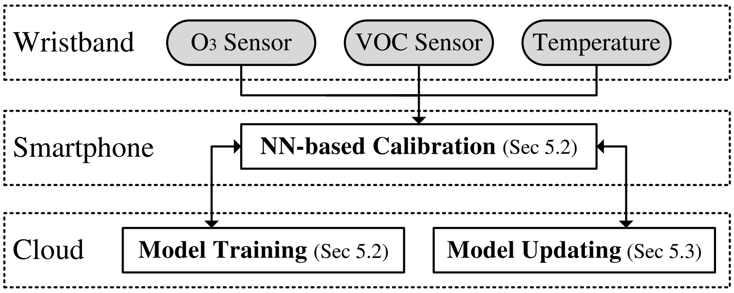


Figure 4.5: Work flow of W-Air. A sensor array consisting of a low-cost O_3, VOC and temperature sensor is utilized to calibrate O_3 and CO_2 (estimated by VOC readings) concentrations. Calibration is based on neural networks and performed on a smartphone. The cloud server performs calibration parameter training and updating for different individuals to provide personalized calibration parameters and adapt to new environments.

implying sufficient data quality. In Figure 4.4b and Figure 4.4d we also observe the improvement of the prediction confidence. The fitted model is close to the perfect fit with a significantly higher confidence compared to the linear model.

We conclude, that a non-linear neural network is able to compensate for the human interference and recover the true O_3 and CO_2 concentrations with low error and high confidence.

Summary. Emissions of human beings impose non-linear interference on ambient O_3 and VOC concentrations. The non-linearity might come from the fact that low-cost MOX sensors are optimized to work linearly within a limited concentration range. Sensors designed and calibrated for lightly polluted areas may suffer severe non-linearity in the high concentration range (due to human interference). At minimal, a non-linear calibration scheme that combines raw O_3 measurements, VOC measurements and environmental factors is indispensable to accurately calibrate the O_3 readings outdoors and CO_2 readings indoors. Despite previous work on non-linear calibration [DPMF09, SGV+15] for static sensor arrays, an effective and efficient sensor array calibration for wearables is missing, which motivates the design of W-Air.

4.5 System Design

This section first presents the overview of W-Air and then elaborates on the calibration scheme with a focus on the techniques to improve the usability. Finally, we present the implementation of W-Air in Section 4.5.4.

4.5.1 Overview

As illustrated in Figure 4.5, W-Air consists of *(i)* a wristband featuring a temperature and two COTS air pollutant sensors; *(ii)* a smartphone; and *(iii)* a cloud server. Users wear W-Air on wristbands, attach it to belts or to backpacks to measure raw O_3 and VOC concentrations as well as environmental factors such as temperature. They also carry smartphones, which calibrate the raw measurements to eliminate the human interference and other environmental factors to output accurate ambient O_3 concentration (outdoor) and CO_2 concentration (indoor). The cloud server communicates with W-Air clients and is responsible for calibration model training and updating.

There are two major functional components in W-Air: *(i)* neural network based calibration, and *(ii)* calibration training and updating. The core technique to filter the human interference and environmental factors (e.g., temperature) is the **neural network (NN) based calibration scheme** (Section 4.5.2). We explicitly distinguish the cases with and without the human interference for more effective calibration and design two separate neural networks for outdoor O_3 calibration and indoor CO_2 concentration calibration, respectively. For easy maintenance and usage, we apply a NN architecture with shared hidden layers for **model training** (Section 4.5.2) with fewer samples than traditional NN architectures and utilize **semi-supervised updating** (Section 4.5.3) to adjust model parameters with little human intervention.

W-Air stores two neural network models to calibrate O_3 and CO_2, respectively, but only runs one to output accurate O_3 concentration *outdoors* and CO_2 concentration *indoors*. This is because O_3 is one of the major air pollutant outdoors and the O_3 concentration is expected to be zero indoors [WWS+00]. Conversely, CO_2 is an important indicator for indoor air quality [SFM99]. To detect whether a W-Air user is indoor or outdoor, any indoor/outdoor detection scheme applies [RKSM14, ZZL+12].

For brevity, the rest of this section mainly focuses on O_3 calibration. Model structure, training and updating for CO_2 calibration all follow the same principle and only differ in specific parameters.

4.5.2 Calibration Methods

As depicted in Section 4.3.3, a non-linear model e.g., a neural network (NN) that inputs O_3, VOC and temperature measurements is sufficient for accurate compensation of human interference. However, we argue that a naive neural network is challenging to apply for the following reasons. *(i)* A complete calibration scheme should operate both *with* and *without* human interference. *(ii)* An easy-to-use calibration scheme should involve little training effort. *(iii)* The calibration performance with and without human interference should be relatively accurate despite

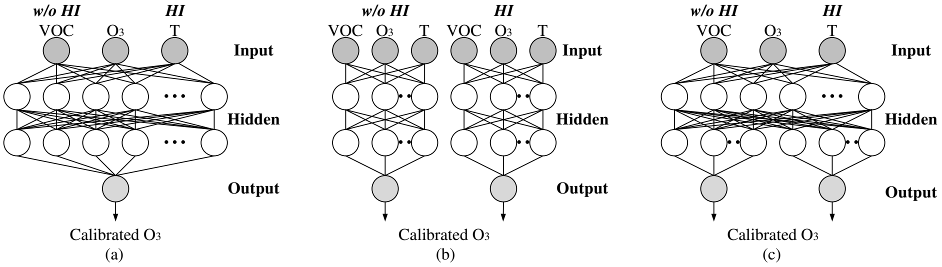


Figure 4.6: Architectures of neural networks (NNs) for O_3 calibration: (a) a single NN; (b) two independent NNs; (c) two NNs with shared hidden layers. Here HI is short for human interference.

imbalanced training dataset sizes for the two cases. In the following, we assume the labels of the two cases (with or without human interference) are known a priori, and discuss the potential for automatic detection of human interference in Section 4.7.2.

There are three candidate NN architectures for calibration. Figure 4.6 illustrates the architectures for O_3 calibration. The same architectures also apply for CO_2 calibration. An intuitive proposal is to adopt a unified NN that calibrates O_3 measurements for both cases, with human interference and without human interference (Figure 4.6(a)). However, a unified, fully-connected NN model is vulnerable to imbalanced training data.

Imbalanced training data are common in practice because most sensor manufacturers only perform generic calibration in laboratories. That is, the sensors are calibrated without human interference. We highlight in Chapter 2 that field calibration is necessary before the sensors can achieve the claimed accuracies. However, end users are reluctant to collect sufficient measurements for field calibration (with human interference in our context). In this case, W-Air needs to train calibration parameters with limited samples with human interference.

Since the O_3 measurements behave differently with and without human interference, the calibration parameters for the two cases also differ notably. With an extremely imbalanced training dataset, a unified NN model tends to train the calibration parameters for human interference largely on the samples collected without human interference. Alternatively, we can create two independent NNs for the cases with human interference and without human interference and train each NN separately (Figure 4.6(b)). Nevertheless, this scheme may involve substantial training data for the two cases since the samples for human interference are of no help to train the NN for the case without human interference. Note that the calibration models for human interference and without human interference may share certain common features. Hence, a more efficient architecture is to allow shared hidden layers for the two cases (Figure 4.6(c)). By enabling data sharing between the two cases, samples for human interference will contribute to extracting the common features useful for calibration when there is no human interference. Consequently, the amount of training data necessary to train the calibration models for both cases will decrease. We evaluate the efficiency of the three NN architectures in detail in Section 4.6.2.

By default, we assign two hidden layers for architecture (a) and (b) with 100 neurons in the first layer and 10 neurons in the second layer. Architecture (c) is composed of 1 shared layer with 100 neurons and two additional separate layers with 10 neurons each. The cost functions of all architectures are to minimize the mean-squared-error (MSE) between the calibrated and ground-truth O_3 readings. We observed through extensive testing that these architecture setups achieve accurate and robust output values. Especially less than 10 neurons in the second layer has a negative effect on the calibration error. Also, we chose a relatively flat architecture because it is easier to train and less prone to overfitting than architectures with more hidden layers and neurons.

4.5.3 Parameter Updating

It is essential to train the NNs on abundant and diverse data covering different environments for accurate calibration. The optimal calibration parameters may differ for different environmental conditions (e.g., sunny versus rainy weather). Additionally, the optimal calibration parameters for each user can also differ because the living environments of users may notably differ. Therefore, it is important to tune the NNs to adapt to the actual target environments. Given sufficient *labeled* measurements from the target environments, the calibration parameters can be trivially updated by adding the newly labeled measurements and retraining the NNs. Here a *labeled* measurement refers to a tuple of VOC, O_3 and temperature together with the true O_3 value (i.e., the label). However, it is cumbersome, if not impossible for users to label the new O_3 measurements collected from his/her own living environments for W-Air to learn and update the calibration parameters. This is because the calibration is a regression problem, and users need to carry a highly reliable reference O_3 sensor to label the ground-truth O_3 concentration, which is impractical in our application scenarios.

To allow W-Air to adapt to new environments, we harness the paradigm of semi-supervised learning, where *unlabeled* (without ground-truth) data are used to improve the accuracy of learning models [ZG09]. In W-Air, we apply COREG [ZL07], a co-training style semi-supervised regression framework to update the calibration models with unlabeled gas measurements. Co-training operates by running two regressors iteratively [ZG09]. The two regressors assign pseudo-labels to each unlabeled sample, and the pseudo-label with higher confidence is used to retrain and improve the performances of both regressors. COREG [ZL07] utilizes two *k Nearest Neighbour* (kNN [Das91]) regressors for co-training.

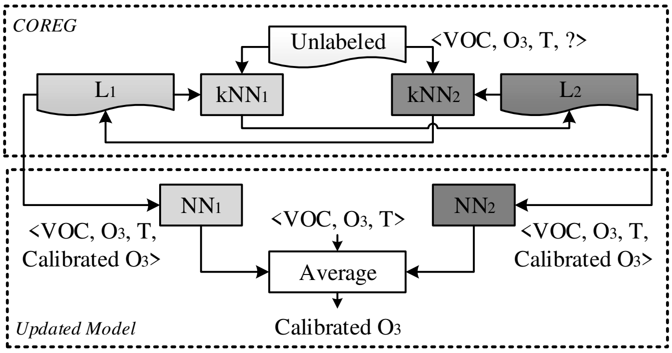


Figure 4.7: An illustration of the COREG work flow for O_3 calibration in W-Air.

The effectiveness of the co-training is fulfilled by using different distance metrics and the number of nearest neighbours (i.e., k) for the two kNNs. After assigning the unlabeled samples using the kNNs, any regressor can be constructed based on the extended (containing both labeled and pseudo-labeled samples) training set to improve regression accuracy.

Figure 4.7 illustrates the work flow of COREG in the context of W-Air. Initially, two kNNs (kNN_1 and kNN_2) are trained using the labeled sets L_1 and L_2, respectively. In each iteration, an unlabeled sample (<VOC, O_3, T>) is assigned a pseudo-label (calibrated O_3) by kNN_1 (kNN_2), and the most confident pseudo-labeled sample (calculated based on MSE, see [ZL07] for details) is moved to L_2 (L_1) to retrain kNN_2 (kNN_1). The outputs of COREG are two extended training sets L_1 and L_2, where two NNs are constructed. A new testing sample will be calibrated by the two NNs, and the final calibration result is the average of the outputs of the two NNs. Note that the co-training framework is applied for W-Air to adapt to different environments and is performed for both human interference and without human interference.

4.5.4 Implementation

This subsection describes the implementation of W-Air.

Hardware. We integrate two mainstream COTS MOX sensors, a *MICS-OZ-47* and a *CCS811*, into a *Thunderboard Sense* development platform (see Figure 4.2). Note that W-Air serves as a proof-of-concept for demonstrating and eliminating the human interference problem. More compact mechanical designs are out of the scope of this thesis. Further, we use the on-board *Si7210* temperature and relative humidity sensors. By default, the MOX gas sensors and the temperature sensor sample at 2 Hz when turned on. The *Thunderboard Sense* provides a GATT-Server data architecture to provide raw sensor measurements via Bluetooth LE. We use a *Motorola Nexus 6* smartphone featuring a 2.7 GHz quad-core

Environment	Human Interference	Baseline	NN-based
Outdoor (O_3)	Without	10 ppb (2.0)	4.3 ppb (0.86)
	With	16.8 ppb (3.7)	4.3 ppb (0.94)
Indoor (CO_2)	Without	177 ppm (1.5)	64 ppm (0.57)
	With	325 ppm (2.3)	38 ppm (0.29)

Table 4.4: Overall calibration errors denoted as RMSE (RMSE_σ) of a linear baseline calibration using MLS (Section 2.4.2) and our NN-based approach.

CPU, 3 GB of RAM and a 3220 mAh battery, running Android 7.0 OS for calibration. The smartphone communicates with the wristband via Bluetooth. A *Lenovo Thinkpad T440p* featuring a 2.5 GHz octa-core CPU, 16 GB RAM and running Ubuntu 16.04 serves as the cloud server for calibration model training and updating.

Software. We apply an indoor/outdoor inference scheme similar to [ZZL+12] to trigger the corresponding neural network for outdoor O_3 calibration and indoor CO_2 calibration. We implement the indoor/outdoor inference and the calibration scheme in *python*. The neural networks are implemented using *Tensorflow*, an open-source machine learning library by Google [Goo17]. We train the indoor/outdoor classifier and the calibration model on the cloud server. The final classifiers and calibration networks are exported and integrated into an Android application. Tensorflow facilitates updating of the neural networks after the initial application installation. Thus, re-training can be done on the cloud server and seamlessly exchanged in the application of a user without recompilation.

4.6 Evaluation

In this section, we thoroughly evaluate the overall calibration performance of our NN-based approach (Section 4.6.1), the benefits of using a NN structure with a shared layer (Section 4.6.2) and the effectiveness of semi-supervised model updating (Section 4.6.3).

4.6.1 Overall Calibration Performance

We first show the overall performance of our NN-based calibration and its advantage over a naive baseline, i.e., a typical calibration performed during pre-deployment tests presented in Chapter 2 or provided by sensor manufacturers [SGX08]. The dataset is based on the measurements described in Section 4.3.1, i.e., we calibrate the W-Air measurements to the ground-truth measurements provided by the two reference sensors. The RMSE and its standardized version RMSE_σ as described in Section 4.4.1 are used to assess the performance of the calibration. As a baseline we construct a linear model using MLS (see (2.3) in Section 2.4.2) that calibrates the raw O_3 (VOC) sensor and temperature measurements to the actual ground-truth O_3 (CO_2) concentration when there is no human interference. This baseline is first trained on 10^3 samples when there is no human presence. Finally, the baseline is tested on 10^4 samples each for both cases, i.e., with and without human interference, and the errors shown in Table 4.4. The testing dataset includes measurements from various locations and times as described in Section 4.3. The baseline performs well for measurements without human interference with an error that is acceptable for indicative measurements. Note that we achieve a slightly higher error for the calibrated O_3 measurements of the same sensor type than in Section 2.6. This is mostly due to the long testing period of half a year without any frequent recalibration and a different mechanical design, i.e., there is no constant airflow guaranteed like in the measurement setup in Figure 2.3a.

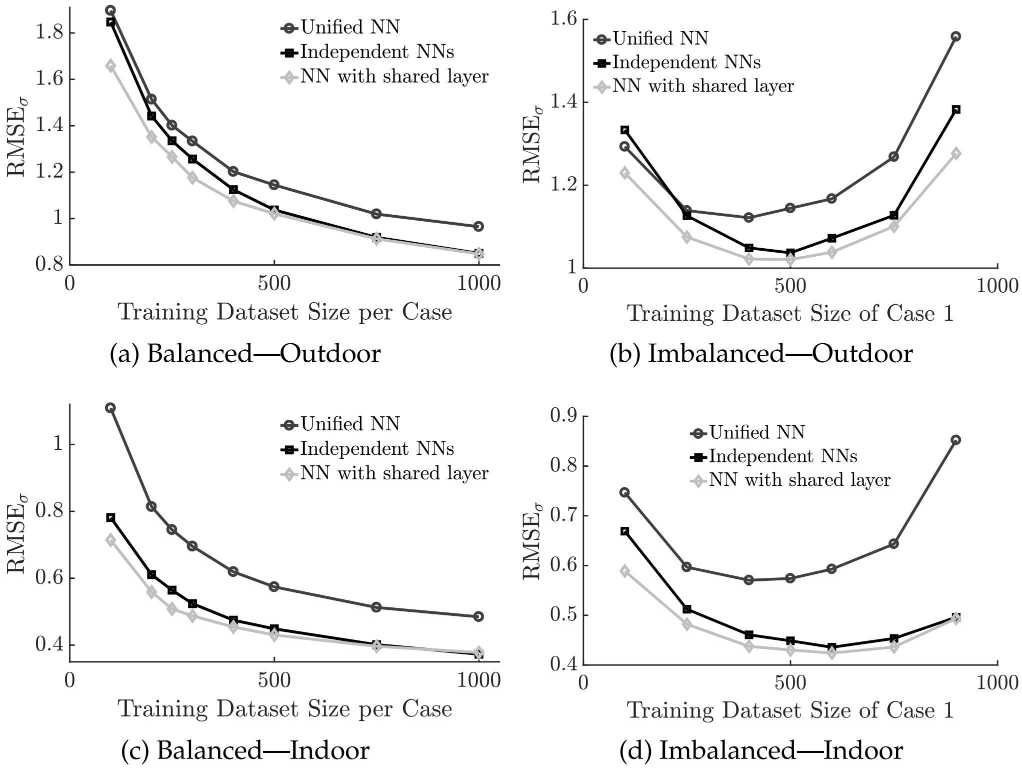
During situations with human interference the error is notably higher and the RMSE_σ is larger than 2 for both CO_2 and O_3. Thus, these calibrated measurements are not usable for any conclusions about the actual air pollution. When applying our non-linear NN-based calibration (shared layer architecture trained with 1000 samples for each case) we achieve the best performance for both cases and both pollutants with a RMSE of 4.3 ppb for O_3 measurements and 64 ppm for CO_2 measurements. In addition, we achieve a $\text{RMSE}_\sigma < 1$ for both pollutants.

In conclusion, our NN-based approach is able to infer the true ambient O_3 and CO_2 concentration during situations with and without human interference with an accuracy that is sufficient for personal air pollution monitoring.

4.6.2 Effectiveness of the NN Architecture

Settings. In this section, we evaluate the effectiveness of our NN-based calibration. We make again use of the measurements described in Section 4.3.1. The following results are based on a 10-fold cross-validation and evaluated on the $RMSE_\sigma$. We evaluate the calibration accuracy of the three NN architectures (see Section 4.5.2) using both balanced and imbalanced training sets in sequel.

Calibration performance using balanced training sets. We first investigate the calibration performance if we have the equal amount of training samples (balanced training sets) for both cases, i.e., case 1 with human interference and case 2 without. Figure 4.8a and Figure 4.8c depicts the calibration performance averaged for human interference and without human interference using balanced training sets for O_3 outdoors and CO_2 indoors, respectively. As expected with more training samples, the calibration errors for all the three NN architectures gradually



(a) Balanced—Outdoor

(b) Imbalanced—Outdoor

(c) Balanced—Indoor

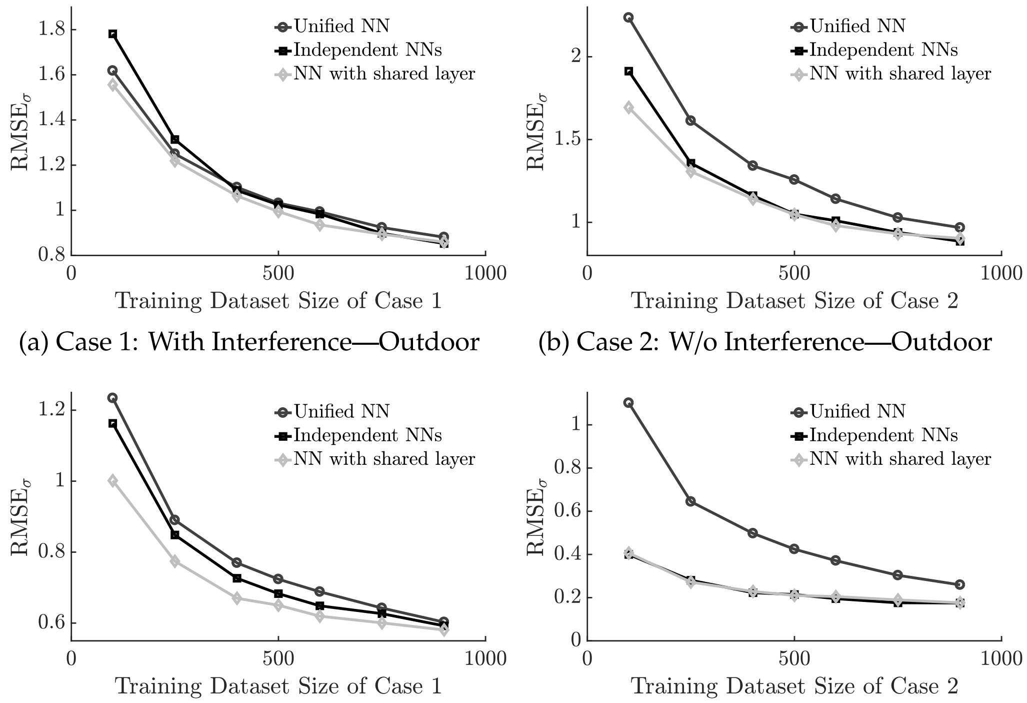
(d) Imbalanced—Indoor

Figure 4.8: Average calibration errors denoted as RMSE$_\sigma$ for both human interference and without human interference using (a) increasing amounts of balanced training samples and (b) imbalanced training sets in an outdoor environment. The same evaluation for indoor environments is presented in (c) and (d), respectively.

decreases. We observe that two independent NNs perform in average better than a single unified NN when trained with few samples, indicating the necessity to distinguish the two cases. By allowing data sharing between the two cases, i.e., using a NN with a shared layer, the calibration error can be additionally decreased. W-Air achieves an up to 10 % lower error for both O_3 and for CO_2 than the two independent NNs using the same amount of training data. For training datasets with more than 500 samples both structures perform equally. Thus, our NN approach with a shared layer is particularly helpful when available training data is limited.

Calibration performance using imbalanced training sets. Due to the diversity in user lifestyles, users are likely to collect different proportions of measurements with and without human interference. Note that we do not assume users always wear W-Air to take measurements. Users may utilize W-Air as a static air pollution sensor when not wearing it. Hence, it is crucial that W-Air still works both with and without human interference, even when trained by imbalanced data.

Figure 4.8b and Figure 4.8d plot the average calibration errors using a fixed 1000 training samples with varying numbers of measurements



(a) Case 1: With Interference—Outdoor

(b) Case 2: W/o Interference—Outdoor

(c) Case 1: With Interference—Indoor

(d) Case 2: W/o Interference—Indoor

Figure 4.9: Calibration errors with (a) human interference and (b) without human interference using imbalanced training sets for outdoor O_3 measurements. The same evaluation for indoor CO_2 measurements is presented in (c) and (d), respectively.

with human interference for outdoors and indoors, respectively. We choose a total of 1000 samples for training because roughly 500 samples are needed to successfully train a well-performing NN for one case (see Figure 4.8a and Figure 4.8c). The results show that the shared layer architecture always yields lower calibration errors, up to 8% outdoors and 12% indoors. This finding highlights the benefit of using a shared layer. We can exploit training samples from one case to train the other case due to common features modelled within the shared layer.

Figure 4.9 further depicts the calibration errors using imbalanced training sets for the two cases separately. We again observe that the shared layer architecture is usually most efficient when using small amounts of training data. With larger amounts of training data the three different NN architectures perform similarly. An interesting observation is the almost identical performance of the shared layer and the independent NN architecture for case 2 indoors. Already a few samples suffice to provide accurate data with an RMSE$_\sigma$ < 0.5. This is due to the strong correlation between the VOC measurements and the ground-truth CO_2 concentration during non-human interference situations.

Robustness to different users. The effect of human interference depends on the amount of VOC a user emits. Thus, we investigate the calibration

	O_3 outdoors	CO_2 indoors
User 1	6 ppb (0.83)	28.8 ppm (0.29)
User 2	4.5 ppb (0.81)	115 ppm (1.5)
User 3	14.3 ppb (3.8)	32 ppm (0.37)
User 4	12 ppb (2.85)	40 ppm (0.47)
User 5	7.3 ppb (1.86)	38 ppm (0.41)

Table 4.5: Calibration errors denoted as RMSE ($RMSE_\sigma$) for different users.

accuracy of different users. Five different users, 4 male (user 1 to 4) and 1 female, additionally recorded between 10 and 15 min of data each outdoors and indoors. The users are wearing W-Air on their wrist while standing or sitting close to a reference sensor as illustrated in Figure 4.1. We train a calibration NN with the measurements from Section 4.3 recorded by user 1 with 1000 samples for each case and finally test on all the data from the four remaining users. Table 4.5 summarizes the resulting errors for each user. For outdoor measurements we observe notable differences between the users. Especially the measurements of user 3 and 4 suffer from a substantial error that makes it impossible to draw reliable conclusions about their personal exposure to air pollution. A similar behaviour we observe indoors for user 2. While all other users achieve accurate results, the results of user 2 suffer from a substantial error that is almost three times higher than the one of the other users. The reason for this behaviour are notable differences of the measurements between the different users. This indicates that the human interference can differ between multiple users. We believe that the individual calibration accuracy can be improved by performing per-person training for each user.

4.6.3 Effectiveness of Semi-supervised Updating

Settings. The two NNs (i.e., NN_1 and NN_2 in Figure 4.7) are based on the shared layer architecture with 100 and 10 neurons each in the hidden layers as used before. For each environment, i.e., indoor and outdoor, we use separate NNs. In order to evaluate the performance of the semi-supervised updating on adapting to new situations, we split our dataset into 12 different distinct situations. Outdoors we distinguish eight different situations based on time and weather conditions because these two factors are the main influences that affect the ambient O_3 concentration. For the remaining 4 indoor situations we distinguish between different air quality levels based on the CO_2 concentration. Table 4.6 summarizes the different conditions for the twelve situations (a) to (l). We apply a leave-one-out validation with these situations by performing supervised learning on seven outdoor situations. Semi-supervised learning and testing is performed on the remaining situation

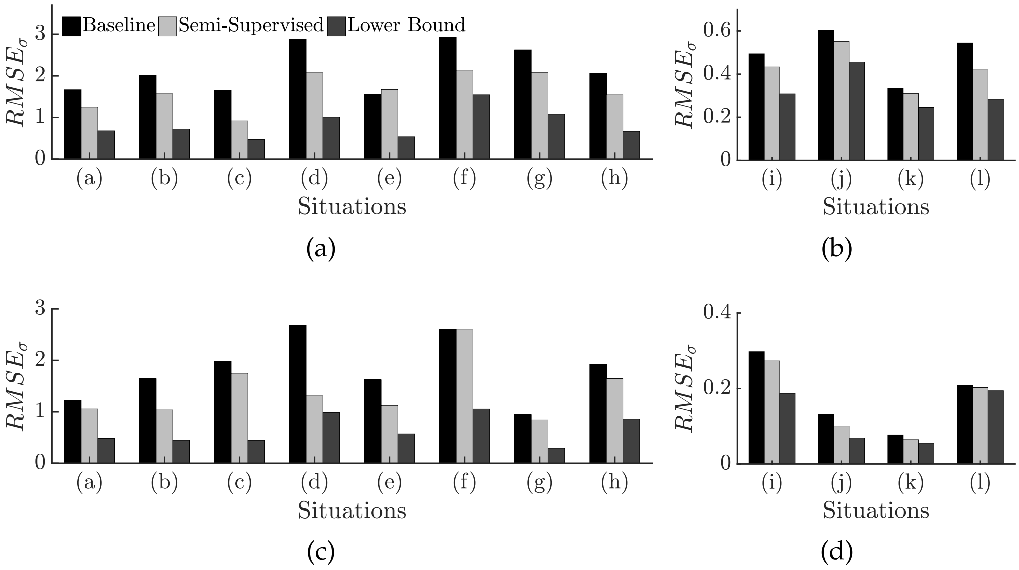


Figure 4.10: Semi-supervised learning performance for situations (a) to (l) for case 1 with human interference outdoors (Figure 4.10a) and indoors (Figure 4.10b) and case 2 without human interference outdoors (Figure 4.10c) and indoors (Figure 4.10d).

for all eight possible combinations for the outdoor situations. In an equal way, we apply the leave-one-out validation with the four indoor situations, but added 10% of data from the left-out situation to the training data. This approach shows that the COREG algorithm is able to improve a model that is trained on a dataset with lack of data for certain ground-truth data ranges.

We compare the performance of our semi-supervised learning to a baseline approach, i.e., the testing error using only the supervised model. Further, we compute a lower bound for the error by adding the unlabeled data to the training set including the true labels. We use 750 samples for each case, with and without human interference, as labeled data for training, and a pool of 500 unlabeled samples for the semi-supervised COREG algorithm. The error is tested on 300 different samples of testing data. The evaluation is finally performed on 10 different pools of labeled and unlabeled data.

Performance. Semi-supervised learning is able to improve the calibration error in all but one situation. Figure 4.10a and Figure 4.10c shows the RMSE$_\sigma$ for the eight outdoor situations for case 1, with human interference, and case 2, without human interference, respectively. The same results are shown in Figure 4.10b and Figure 4.10d for the four indoor situations. The improvement is between 0.5% and 51% and in average about 19.6% and 18.4% for case 1 and case 2, respectively, as summarized in Table 4.6. Only in situation (e) the semi-supervised updating resulted in a degradation of the accuracy of 7.5% for the case

		Improvement		# Samples added	
Situation (outdoor)	Time, Weather	Case 1	Case 2	Case 1	Case 2
(a)	April, afternoon, rain	25.1%	13.5%	58	47
(b)	April, night, cloudy	22.1%	36.1%	68	50
(c)	April, afternoon, sunny and windy	44.4%	11.5%	86	2
(d)	April, morning, sunny	27.8%	51.2%	20	49
(e)	April, morning, light rain	-7.5%	30.7%	25	42
(f)	June, noon, sunny and hot	26.9%	0.5%	96	20
(g)	June, evening, sunny and hot	20.1%	11.0%	56	53
(h)	August, afternoon , hot and windy	25.4%	14.4%	55	57
Situation (indoor)	CO_2 concentration				
(i)	low (400-600 ppm)	12.3%	8.2%	96	93
(j)	low-medium (600-750 ppm)	23.3%	3.1%	85	86
(k)	medium (750-850 ppm)	7.2%	16.4%	90	102
(l)	high (850-1000 ppm)	22.9%	2.8%	94	90
Average		19.6%	18.4%	69	58

Table 4.6: Different situations used for semi-supervised updating with their relative improvement compared to a baseline and the average number of unlabeled samples added based on the COREG results.

with human interference. The reason might be that this situation is too different from the other situations in terms of measurements. The COREG algorithm was in fact able to label only 25 unlabeled samples, which is notably below the average of 69 unlabeled samples. Further, we see that for certain situations the lower error bound is significantly lower than the error of the semi-supervised approach. Especially in outdoor situations we observe the need to adapt our calibration model to new situations. For indoor measurements we observe smaller improvements than for outdoor measurements, in particular for case 2 without human interference. This is due to an already relatively well performing calibration model before the semi-supervised updating, i.e., the difference of the baseline and the lower bound are notably smaller compared to the outdoor situations.

It is in general challenging to adapt a model to an unfamiliar situation. Especially in outdoor situations the model can be further improved. Thus, it will be important to also exploit supervised calibration methods, e.g., by exploiting opportunistic calibration with other devices [SHT15] or static reference sensors [HSST12, HST12], which we already present in detail in Chapter 3. One future direction is the improvement of our neural network based calibration approach to be fit for collaborative and multi-hop calibration scenarios, e.g., by also considering error accumulation like our linear SCAN method (see Section 3.4). In Chapter 5 we present a first direction by including different uncertainty metrics into the collaborative calibration process. We believe that a combination of semi-supervised and supervised calibration methods, such as the three different network calibration approaches presented in Section 1.5, will improve the overall accuracy of W-Air.

4.7 Discussions and Future Works

This section discusses the limitations and future improvements of W-Air.

4.7.1 Impact of User Context

Environmental conditions. Environmental conditions have not only an effect on the air pollution but also on the intensity of the human interference. Especially for different weather conditions in outdoor environments we observed different human interference behaviours. For instance, our W-Air device measured an average VOC concentration of 130 ppb during human interference situations in July and only 35 ppb in October. While the average ground-truth O_3 concentration during both months barely differed with 44 ppb in July and 42 ppb in October, the weather conditions showed notable differences. The temperature was around 28° C in July without any rain and 18° C in October with multiple rain periods.

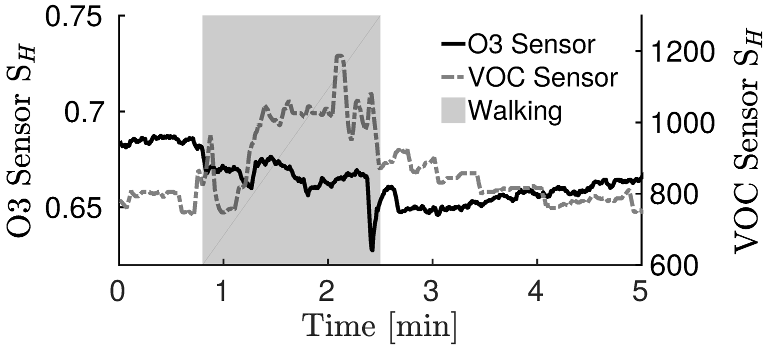


Figure 4.11: VOC (in ppb) and O_3 (raw ADC values) MOX sensor measurements of W-Air attached to a backpack of a user standing and walking indoors. As soon as the user starts to walk we observe an increase of the VOC sensor signal and a decrease of the O_3 sensor signal, indicating an increased intensity of the interference problem.

This dependency of the human interference on environmental conditions poses a substantial challenge for our W-Air system. We need to train the calibration model with data from various environmental conditions in order to provide accurate measurements during these conditions. This is possible by exploiting semi-supervised updating as we show in Section 4.6.3. To further improve the measurement accuracy it will also be important to incorporate supervised techniques such as opportunistic calibration, like the multi-hop calibration presented in Chapter 3 or network calibration strategies, see Section 1.5.

Activities. The current activity of a user is known to be an interfering factor for high data quality in wearable sensing [LZW+17, HSST12]. In order to exclude those influences in our evaluation we only considered situations with the user either sitting, standing or doing office-work. Other daily activities, such as walking, running or exercising can in fact have an additional impact on the sensor readings, i.e., an increased intensity of the human interference problem. In Figure 4.11 we investigate the impact of walking on the sensor data. We observe that as soon as the user starts to walk the VOC sensor measurements increase and the O_3 sensor measurements decrease. As we show in the measurement study in Section 4.3, this is the typical behaviour of the human interference problem. The impact of walking diminishes again as soon as the user stops walking around 2:30 min. The reason for this behaviour might be the consequence of the increased air flow over the sensing layer and, thus, the two gas sensors react differently. We also observe this behaviour when using W-Air in the wrist and belt setting, as well as in outdoor environments.

Consequently, not only the VOC emissions of a user but also its different activities, especially mobility [HSST12], can interfere with the sensor readings. In a first step of possible future work it will be important to thoroughly study the impact of human activities on the air quality measurements of our W-Air system. In a second step it is equally important to adapt and expand our sensor calibration by also incorporating different human activities.

4.7.2 Automatic Detection of Human Interference

The distinctive characteristics of the human interference (Section 4.3.3) lead us to distinguish between the cases with and without human interference. In this work, we assume the knowledge of whether there is human interference is known in advance. Autonomous human interference detection is feasible by leveraging additional sensors of the wearable device. We can use accelerometers to detect movements or a switch that closes a circuit when one clips on the wristband.

Some pioneer studies have integrated proximity sensors into watchbands for wrist gesture input [GYI16], which can also be applied to detect whether a wristband or a smart-watch is worn by the user. Autonomous human interference detection will further reduce the overhead for users to label new measurements to update the calibration parameters.

4.7.3 System Performance

Energy and delay are two crucial factors for practical wearable systems. The most energy-hungry components of W-Air are the two MOX based gas sensors. We measure an average current draw of 50 mA when turning both gas sensors on. Continuously powering the gas sensors will drain a standard 250 mAh coin cell battery within 5 hours and, thus, would limit usability to a great extent. One future direction to improve the energy efficiency of W-Air is to duty-cycle the gas sensors in undesired situations. For instance, the average current draw of the VOC sensor integrated in W-Air drops to 0.7 mA with a duty-cycle of 60 s. However, long duty-cycles also cause slower response time and lower sensitivity of the sensor [ams17].

In terms of delay, the most computational and time intensive task of W-Air is the NN-based calibration. The majority of the delay is induced by the feature extraction, namely a 3 sec smoothing window for the sensor measurements. Running the neural network to output the final calibrated concentrations takes negligible time (around 5 ms). However, there is a delay of around one minute when first turning on the gas sensors before they work in a stable state. During this warm-up time W-Air is not providing accurate measurements. In a duty-cycling scenario the trade-off between energy efficiency and warm-up delay needs to be carefully evaluated. We also envision the one-shot delay due to the warm-up time of the gas sensors will be notably reduced by the rapid development of sensor technologies.

4.8 Summary

In this chapter, we propose W-Air, an accurate personal multi-pollutant monitoring platform for wearable devices. We identify the human interference problem when integrating low-cost MOX gas sensors into wearable platforms such as wrist-worn, attached to belts or backpacks, which is largely overlooked in previous research. We propose a partitioned neural network based sensor array calibration scheme to eliminate the non-linear human interference on ambient outdoor O_3 measurements and indoor CO_2 measurements. The architecture of the calibration model is carefully devised to reduce training efforts and facilitate parameter updating with little human intervene. We prototype W-Air on a wristband with low-cost COTS gas sensors. Evaluations show that W-Air is able to yield accurate ambient O_3 and CO_2 measurements whether there is human interference. It can also learn from unlabeled measurements and adapt to unfamiliar circumstances by exploiting semi-supervised learning.

Future directions include *(i)* more thorough investigations how user context and activities affect the human interference problem, *(ii)* combining W-Air with state-of-the-art health and fitness trackers to explore correlations between air pollution and well-being of a user and *(iii)* large-scale deployments for extensive user-studies.

5

Enhancing Sensor Calibration with Uncertainty Estimates

Throughout this thesis we presented different calibration models, for instance based on linear regression in Chapter 2 and Chapter 3 or non-linear neural networks in Chapter 4. All these models have been proven to be powerful tools to improve the overall data quality of low-cost air pollution sensors. However, the performance of these models is in general not perfect. In fact, various different error sources can affect a calibration model and, therefore, its output is always subject to some uncertainty, i.e., the amount of trust we can have in the accuracy of the measurement. Therefore, in a first part of this chapter we investigate typical uncertainties of calibration models. In particular, we analyse the impact of different data-induced uncertainties on calibration errors and devise a scheme to estimate these uncertainties of calibrated model outputs.

In a second part, we integrate these uncertainties into the multi-hop calibration approach presented in Chapter 3 by proposing an uncertainty-based metric for data filtering at each hop. By filtering samples with high uncertainty, we can build an improved dataset for calibration and minimize the impact of different error sources, in particular dynamic boundaries and systematic errors (see Section 1.2), that accumulate over multiple hops. We evaluate the effectiveness of our method in a real-world ozone sensor deployment. Experimental results show that our method works with both linear and non-linear calibration models and reduces calibration errors in multi-hop setups by up to 25% compared with existing techniques.

5.1 Introduction

In this thesis, we describe multiple calibration models with different characteristics and use-cases. Chapter 2 presents a pre-deployment testing approach of low-cost air pollution sensors based on linear multiple least-squares to resolve cross-sensitivities and meteorological dependency. In Chapter 3 we develop SCAN, a linear regression model, which can be applied in a multi-hop calibration setup during the deployment of a sensor network to tackle sensor drift over time. Finally, in Chapter 4 a non-linear neural network is used to resolve complex cross-sensitivities of low-cost sensors equipped in wearables and a semi-supervised learning algorithm allows the update of the neural network during deployment without any user involvement. All these models are used to minimize the impact of different error sources, see Section 1.2 and Figure 1.8, which are either induced by certain sensor technology limitations, e.g., dependency on meteorological effects, or are application-dependent, e.g., cross-sensitivities to human generated emissions. In this chapter, we focus on error effects, which are introduced by the way these calibration models are *trained*. For instance, the distribution of the training samples may introduce high errors in concentration regions where only few training samples are available, i.e., *dynamic boundaries* (see Section 1.2). Inadequate calibration model capabilities, e.g., using a linear model to resolve non-linear cross-sensitivities like in Section 4.4, may lead to *systematic errors* of the calibrated measurements.

Challenges. Especially during post-deployment calibration, whose goal is to ensure data consistency of the sensor measurements by periodic re-calibration, these additional error sources may impose new challenges. In Chapter 3, we describe multi-hop calibration, an effective concept to calibrate low-cost sensors in a mobile deployment with limited references [XBP+12, SHT15, FRD17, LDC18]. While multi-hop calibration allows to calibrate substantially more sensors in a mobile deployment, it usually encounters severe error accumulation [SHT15] over multiple hops, see also Section 3.4.1. For multi-hop calibration to function, measurements of the virtual reference need to be as accurate as possible at every hop in order not to affect later calibration procedures. However, both the *data* and the *model* for calibration are not perfectly accurate in practice. Hence, calibration at each hop is never error-free. Consequently, the error of multi-hop calibration tends to accumulate over hops if not explicitly controlled. In Chapter 3 we mitigate this error accumulation by designing SCAN, a novel calibration model. However, SCAN and other multi-hop calibration methods like the geometric mean regression (GMR) [SHT15], have certain short-comings.

- They only apply to linear calibration models. In some cases non-linear models are necessary to generate accurate measurements, such as for the wearable use-case of our W-Air system in Chapter 4 or to calibrate other important air pollutants such as particulate matters [FRD17, LDC18, CLL+14].

- They rely on certain assumptions concerning the sensor noise that can be reduced over multiple hops. As we show in Section 3.5.1, SCAN can also suffer from error accumulation due to other effects and error sources, for instance different dynamic boundaries or systematic errors (see Section 1.2) of the sensors in the calibration path.

Contributions and road-map. In this chapter, we first investigate how effects such as high sensor noise, unequal training sample distributions and inadequate calibration models can lead to calibration errors, which can not be minimized by SCAN or similar methods. Next, we present a scheme to estimate the impact of these error sources by *uncertainty* metrics. Finally, we propose to reduce error accumulation in multi-hop calibration by also considering these uncertainties during the calibration process. Rather than designing error-resilient calibration models like SCAN in Chapter 3, we quantify the data-induced uncertainties and conduct explicit data filtering at each hop such that only reliable data are utilized for later calibration. To enable data filtering in multi-hop calibration, two problems need to be solved. *(i)* What uncertainties are introduced by sensor data into calibration and how to estimate them during data processing in the calibration pipeline? *(ii)* How to define a unified metric to quantify the uncertainties for data filtering at each hop?

We address the above problems and make the following contributions.

- We design a scheme to estimate two types of uncertainties: *epistemic* and *aleatoric*, which contribute to data-induced errors in sensor calibration. Our approach is *agnostic to the underlying calibration models* (linear or non-linear) and can be plugged into the standard calibration procedure. To the best of our knowledge, this is the first work to augment sensor calibration with uncertainty estimates.

- We comprehensively analyse the impact of different uncertainties on potential calibration errors and propose an uncertainty-based metric for data filtering in multi-hop calibration.

- We evaluate our method in real-world ozone (O3) sensor deployments. Our results show that we are able to reduce the calibration error in a multi-hop setup by up to 25% when using uncertainty-based data filtering compared to traditional calibration.

In the rest of this chapter, we review related work in Section 5.2, explain calibration uncertainties in Section 5.3, introduce methods to estimate them in Section 5.4 and propose a data filtering metric in Section 5.5. We conduct simulations in Section 5.6 and real-world evaluations in Section 5.7 and finally summarize the chapter in Section 5.8.

5.2 Related Work

This chapter is related to research on multi-hop calibration, which we present in detail in Chapter 3 and Section 1.5.2, and measuring uncertainty, which is summarized below.

Uncertainty metrics. Quantifying the expected performance of statistical models, for instance by estimating major uncertainty sources, is an important problem. Various works [KD09, KG17, PR97] differentiate between different uncertainties, the two most important ones being *epistemic* and *aleatoric*.

Epistemic uncertainty characterizes the knowledge of a model based on the data it has been trained on, i.e., missing knowledge may lead to ignorant predictions. The most prominent method to estimate epistemic uncertainty is ensemble learning, where multiple models are trained to learn the same underlying function but on a different view of the dataset. These different views are often created by using bootstrapping [ET94] or bagging [Bre96] and have been applied in general modelling frameworks [Par13], neural networks [OBPVR16] or Bayesian learning [KG17].

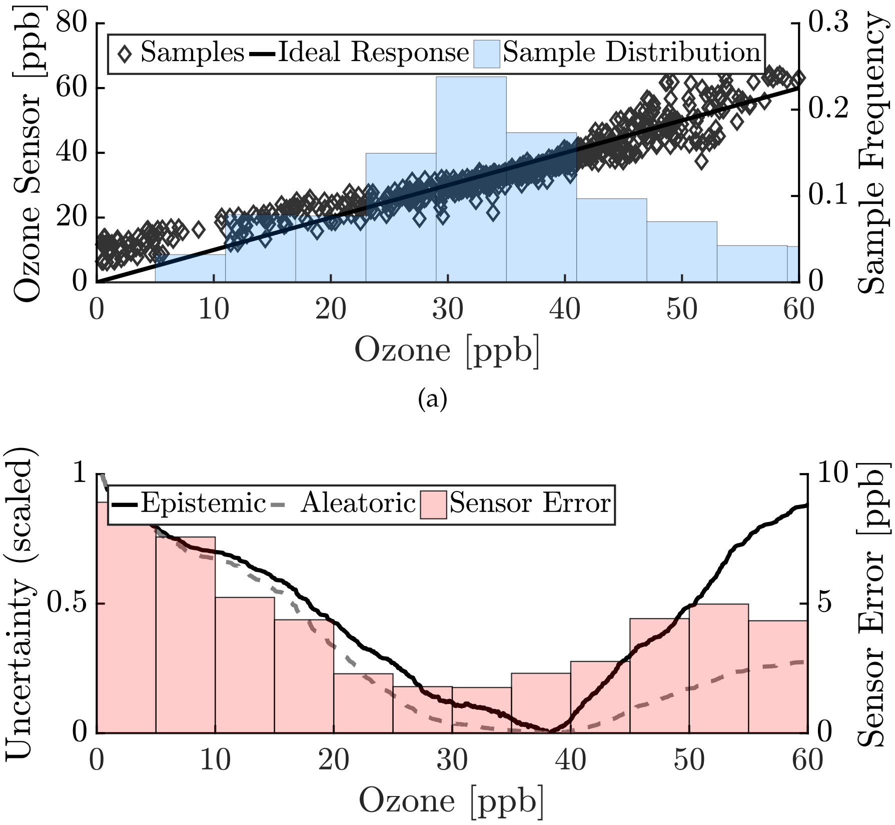
Aleatoric uncertainty, which captures the noise inherent in the measurements, can be modelled by auxiliary models [NW95] that learn the relationship between the input of a model and its corresponding expected error. Recent methods, especially tailored for complex machine learning methods, model the output as a probabilistic predictive distribution. Examples of such methods are Bayesian neural networks [GG16] or deep ensemble learners [LPB17, YZS+18].

Our work is inspired by all these works. In particular, we combine the concepts presented in [NW95, ET94, OBPVR16] into one concise model and apply it in the context of sensor calibration, which is presented in detail in Section 5.4.

5.3 Uncertainties in Sensor Calibration

In this section, we quickly recapitulate the basis of sensor calibration and then explain two types of uncertainties in the context of air pollution sensor calibration.

The general goal of sensor calibration is to find a calibration function *cal* (i.e., calibration model) that transforms some raw sensor measurements M into a calibrated form $\hat{m} = cal(M)$. An optimal calibration model minimizes some norm between the calibrated measurements $\hat{m}$ and some reference or ground-truth measurements m_r. We use a real-world example of a low-cost ozone (O_3) sensor, MICS-OZ-47 [SGX13], which has been deployed next to a highly accurate governmental monitoring station, see Figure 1.4. We use the sensor array calibration method presented



(a)

(b)

Figure 5.1: Calibration results of an ozone sensor. Figure 5.1a shows the calibrated measurements versus the actual ozone concentrations. These samples are also used to train the calibration model. The distribution across the ozone range is highlighted by the blue area. The black solid line indicates the ideal response, i.e., the case when the calibrated measurements are identical to the actual ozone concentration. Figure 5.1b shows the behaviour of the epistemic, see equation (5.2), and aleatoric uncertainty, see equation (5.3), as well as the sensor measurement error defined as the absolute difference between calibrated and actual ozone measurements.

in Chapter 2 to calibrate the sensor. Specifically, the ozone sensor is augmented with a temperature and humidity sensor (i.e., a *sensor array*) and multiple least-squares is applied to find a function *cal* that calibrates n measurements $X \in \mathbb{R}^{n\times 3}$ of the three sensors to the ozone reference $m_r \in \mathbb{R}^{n\times 1}$. The three sensors were placed inside a ventilated box, see Figure 2.3, and deployed next to the governmentally maintained ozone sensor in a suburban area in Switzerland [LFS+12] during 2 weeks in May, 2014.

We are interested in the distributions of the calibration errors across the sensing range (the bars in Figure 5.1b). We observe for the calibrated sensor, the calibration errors vary at different sensor readings. The calibration error increases where the measurements to train the calibration model are limited or have high variation. For example, at low ozone concentrations below 10 ppb the training measurements are scarce. As a result the calibration model is overestimating the ozone concentration. Similarly, at concentrations above 45 ppb the measurements exhibit higher variance than at lower concentrations and, hence, the sensor error is also growing. The consistent overestimation and lower precision in certain concentration ranges can be described as systematic errors and limited dynamic boundaries of the sensor, two error sources described in Section 1.2. We aim to explain the non-uniform calibration errors across the sensing range of a sensor using *uncertainties*.

Epistemic uncertainty. This uncertainty captures the *general ignorance* of our calibration model and is typically caused by lack of knowledge, for instance by missing data or insufficient modelling power [KD09, KG17]. It is in general difficult, if not impossible, to generate confident calibrated measurements in ranges where the calibration model is trained on little or no data at all. This is in particular true for complex models, such as non-linear methods, when they have to excessively extra- or interpolate. The same holds when a model, which lacks appropriate capabilities to capture a complex calibration function, is applied, e.g., when a linear model is used to learn a non-linear function. However, epistemic uncertainty can typically be reduced by collecting and adding more data into the calibration process or applying more appropriate models with adequate capabilities for the problem. Epistemic uncertainty can point out potentially inaccurate sensor readings. In Figure 5.1a, we can observe that most of the samples for training the calibration model are distributed between [20,40] ppb. Outside this range fewer samples have been used to train the calibration model. Consequently, from 20 ppb to 0 ppb and above 40 ppb, both the calibration error and the epistemic uncertainty begin to grow, as shown in Figure 5.1b. Note that above 40 ppb the epistemic uncertainty is equally high as for measurements below 20 ppb, however the sensor error is notably smaller. This additional error at low concentrations is due to a constant overestimation, which may be caused by missing modelling power. Note that there is in general no direct causation between sensor error and epistemic uncertainty. It is for instance still possible for a calibration model to learn an accurate calibration function even with lower amounts of samples. We defer the method to estimate the epistemic uncertainty to Section 5.4.

Aleatoric uncertainty. This uncertainty captures the *natural noise*, i.e., measurement error, in low-cost sensor measurements [NW95, KD09, KG17]. In Figure 5.1a we can observe the scatter of the measurements is larger for higher concentrations. Especially between [45,50] ppb we can observe various outliers, which are at least 10 ppb from the ideal response. Potential reasons for this effect may be that the linear model is not capable to capture the underlying calibration function of the sensor or the uncalibrated measurements exhibit a generally higher noise in this high concentration region. As indicated in Figure 5.1b, the aleatoric uncertainty increases in regions where the deviation of the measurements is increasing as well. Similarly to the epistemic uncertainty, note that high variance of measurements in the training data does not necessarily imply high measurement error. A calibration model may still be able to perfectly capture the ideal underlying function between sensor and reference measurements, e.g., if the noise is following a Gaussian distribution. Further, we can also observe an increased aleatoric uncertainty at lower concentrations. As we have discussed before, the calibration model is not able to perfectly capture the underlying function due to either missing data or modelling power and, thus, the aleatoric uncertainty also captures these high inaccuracies below 20 ppb. In Section 5.4, we lay out a method to learn the relationship between the sensor measurements and the aleatoric uncertainty.

Summary. Not all outputs of a calibrated sensor have the same accuracy due to the distributions of the raw sensor measurements to train the calibration model. This motivates us to associate auxiliary uncertainties to each output of a calibrated sensor to indicate the potential errors. We identify two types of uncertainties: *epistemic* and *aleatoric*. The former characterizes the uncertainty in output ranges where the measurements for training the calibration model are sparse or the model is lacking sufficient modelling capabilities, while the latter represents the uncertainty in output ranges where the calibrated measurements exhibit large inaccuracies. These two types of uncertainties provide extra information about the reliability of the outputs of a calibrated sensor at different sensing ranges, which helps to filter unreliable outputs for further use (multi-hop calibration in our case). In the next two sections, we design methods to estimate the two types of uncertainties, investigate their relationship and integrate them into multi-hop sensor calibration for better accuracy.

5.4 Estimating Uncertainties

This section describes our scheme that can be plugged into any calibration model to output calibrated sensor measurements and the corresponding uncertainties.

5.4.1 General Overview

Given a dataset (M,m_r), where $M \in \mathbb{R}^{n\times k}$ are the uncalibrated sensor measurements that consist of k input feature vectors, i.e., a sensor array consisting of k sensors such as the one in Section 5.3, and $m_r \in \mathbb{R}^{n\times 1}$ is the reference sensor measurements, our scheme works in two phases.

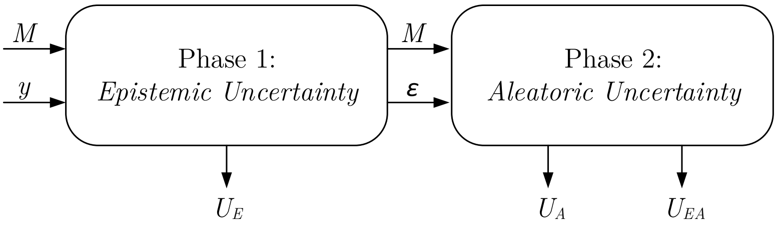


Figure 5.2: General overview of our model, which is divided into two separate phases. In phase 1 sensor calibration is performed and epistemic uncertainty U_E information is retrieved. Phase 2 predicts the aleatoric uncertainty U_A as well as the corresponding epistemic uncertainty U_{EA} of this process.

1. An ensemble of calibration models is trained to output, for an uncalibrated measurement $m_i \in M \in \mathbb{R}^{1 \times k}$, a calibrated measurement $\hat{m}_i = cal(m_i) \in \mathbb{R}^{1 \times 1}$ and a corresponding epistemic uncertainty $U_E(\hat{m}_i) \in \mathbb{R}^{1 \times 1}$. Any calibration model that performs a regression task can be applied to find *cal*. We use bootstrapping [ET94] to train our ensemble, i.e., we artificially generate different training datasets to create diverse models in our ensemble.

2. A neural network ensemble is trained to learn the relationship between uncalibrated sensor measurements M and calibration errors $\varepsilon = (\hat{m} - m_r)^2 \in \mathbb{R}^{n \times 1}$ to quantify the aleatoric uncertainty $U_A(\hat{m})$, also known as local error bars [NW95], for each calibrated measurement $\hat{m}_i \in \hat{m}$ of the sensor. This approach helps us to quantify the variance of our measurements around the ground-truth, i.e., the calibration error as a function of the inputs $f(M) = \varepsilon$. Since the estimation of the aleatoric uncertainty is also subject to some potential error sources, we also estimate the epistemic uncertainty $U_{EA}(\hat{m}_i) \in \mathbb{R}^{1 \times 1}$ of the aleatoric uncertainty estimation.

Figure 5.2 illustrates the overall procedure of our scheme. We explain the two phases in detail below.

5.4.2 Estimating Epistemic Uncertainty

In the first phase we train multiple calibration models to output the calibrated sensor measurements, which then are used to generate epistemic uncertainty. To achieve this goal, we apply an ensemble of sensor calibration models via bootstrapping [ET94]. In particular, given the dataset (M,m_r), p different calibration models are trained. This is done via the following process:

1. Generate p bootstrapped datasets, each consisting of $(\overline{M}, \overline{m}_r)$. This is achieved by the standard bootstrapping methodology, where

each pair of bootstrapped samples $(\overline{M_i}, \overline{m}_{r,i})$ is randomly sampled with replacement from the input datasets (M,m_r). As a result, p different datasets with the same cardinality as the input dataset are artificially generated. According to the .632 bootstrapping rule [ET94], approximately 63.2% unique samples form one newly bootstrapped input dataset.

2. Train p calibration models using the p generated datasets. A typical calibration model minimizes the mean squared error $\frac{1}{n}\sum_{i=1}^{n}(\hat{m}_i^{(j)} - \overline{m}_{r,i})^2$, where $\hat{y}^{(j)}$ is the output the j-th ($j \in \{1, 2..., p\}$) model of the ensemble for a given $m_i \in \overline{X}$ and $\overline{m}_{r,i}$ is the i-th sample of the bootstrapped reference measurements $\overline{m}_r$.

As a result, we get p different model outputs. Therefore, for a given $m_i \in M$ the final output is the mean over all p ensemble model outputs, i.e.,

$$\hat{m}_i = \frac{1}{p}\sum_{j=1}^{p}\hat{m}_i^{(j)} \tag{5.1}$$

Finally, the epistemic uncertainty is calculated by the standard deviation over all the p model outputs, i.e.,

$$U_E(\hat{m}_i) = \left(\frac{1}{p}\sum_{j=1}^{p}(\hat{m}_i^{(j)} - \hat{m}_i)^2\right)^{\frac{1}{2}}. \tag{5.2}$$

Since we use the mean over all outputs as final calibrated measurement, the epistemic uncertainty $U_E(\hat{m}_i)$ gives a notion of how much confidence we can have in this mean $\hat{m}_i$. The higher the disagreement of the p calibration models, the higher the epistemic uncertainty U_E, the less confidence we have in our calibrated measurement. If we add more knowledge, e.g., in the form of data, to the calibration model we might be able to reduce this disagreement and consequently the epistemic uncertainty, see also Section 5.3.

5.4.3 Estimating Aleatoric Uncertainty

In the second phase the goal is the estimation of the aleatoric uncertainty inherent in the output of our ensemble of calibration models in phase 1. This is achieved by a second ensemble of models, which individually learn the relationship between the input vectors M and the calibration error $(\hat{m}_i - m_{r,i})^2$. The calibration error measures the distance of our calibrated measurements to the ground-truth values and, therefore, the aleatoric uncertainty, see Section 5.3. Similarly to phase 1, this task could be performed by any type of regression technique. However in order to facilitate powerful modelling capabilities of non-linear functions we solely use non-linear neural networks. The following procedure performs the estimation of the aleatoric uncertainty.

1. Generate q new bootstrapped datasets $(\overline{M}, \overline{m}_r)$, which are not identical to the ones as in phase 1.

2. For each dataset and each sample $m_i \in \overline{M}$ calculate the squared calibration error $\varepsilon_i = (\hat{m}_i - m_{r,i})^2$, where $\hat{m}_i$ is the calibrated measurement from (5.1).

3. Train q models that approximate the function f with $f(M) = \varepsilon$.

In order to reduce the training efforts, we apply an optimized neural network ensemble structure [OBPVR16]. Instead of q individual neural networks, one neural network with q outputs and a shared hidden structure is used. During each training run, the outputs are individually trained using the bootstrapped datasets and the hidden structure is updated in every training step for every output. The optimization functions for each output are also applying a L2-regularization [WRH91] to assure smooth uncertainty estimations and to avoid over-fitting. Finally, to assure positive outputs (due to $\varepsilon \geq 0$), each output uses the soft-plus activation function, i.e., $softplus(x) = log(1 + e^x)$.

Similar to the ensemble in phase 1, the final output, in this case the estimated aleatoric uncertainty U_A for a given $m_i \in M$ is the mean over all q ensemble model outputs, i.e.,

$$U_A(\hat{m}_i) = \hat{\varepsilon}_i = \frac{1}{q} \sum_{j=1}^{q} \hat{\varepsilon}_i^{(j)}, \tag{5.3}$$

where $\hat{\varepsilon}_i^{(j)}$ is the j-th ($j \in \{1, 2, ..., q\}$) output of the neural network.

Finally, we also estimate the epistemic uncertainty of the aleatoric uncertainty estimation, given by,

$$U_{EA}(\hat{m}_i) = \left(\frac{1}{q} \sum_{j=1}^{q} (\hat{\varepsilon}_i^{(j)} - \hat{\varepsilon}_i)^2 \right)^{\frac{1}{2}}. \tag{5.4}$$

This process allows us to model the calibration error, or the variance of our calibrated measurements, as a function of the input and consequently of a single calibrated measurement. Similarly to U_E, which we calculate in phase 1, the epistemic uncertainty U_{EA} serves as a measure of confidence in our aleatoric uncertainty estimation.

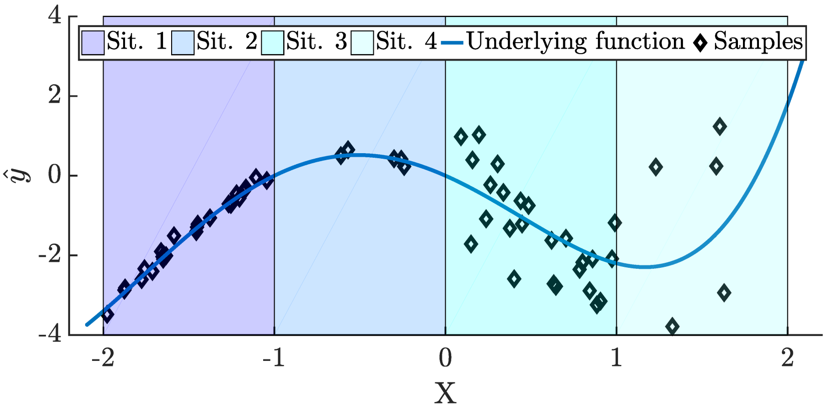


Figure 5.3: Four different uncertainty situations regarding epistemic U_E/U_{EA} versus aleatoric U_A. From left ro right, situation 1) low—low, 2) high—low, 3) low—high, 4) high—high.

5.5 Integrating Uncertainties in Calibration

In this section, we first interpret different situations we may experience when applying our scheme and then propose a data filtering metric for sensor calibration.

5.5.1 Interpretation: Epistemic versus Aleatoric

Before explaining different situations, we first discuss the relationship between the two epistemic uncertainties U_E and U_{EA}. Both U_E and U_{EA} capture the epistemic behaviour of the two ensembles we use. However, U_E captures the uncertainty we have in our calibration model and U_{EA} in our calibration error estimation. Basically, both metrics decrease as the number of samples, i.e., the knowledge we feed into our model, increases, because the ensembles converge to a common output. For simplicity and due to the same basic interpretation we treat U_E and U_{EA} as similar in the following description of the different situations but differentiate them during data filtering (Section 5.5.2).

- **Situation 1: Low U_E/U_{EA}—Low U_A.** In a situation where we have low epistemic and low aleatoric uncertainty at the same time we have high confidence in our calibrated measurement. This ideal case is shown in situation 1 in Figure 5.3. The low epistemic uncertainty U_E and U_{EA} suggests that all the individual models in the two ensembles agree on their outputs and, thus, have been trained with a sufficient amount of data. A low aleatoric uncertainty U_A at the same time, also points out that the calibration error of the measurements during training is low.

- **Situation 2: High U_E/U_{EA}—Low U_A.** The next situation, see also situation 2 in Figure 5.3, appears when we have only a few samples, therefore higher epistemic uncertainty, but these are not affected by large noise, thus low U_A. Although the low aleatoric uncertainty points out that the calibration model is able to approximate an accurate calibration function, there is only little proof. In order to gain more confidence the calibration model should be trained with more samples and re-evaluated, especially if the model has to extra or interpolate large areas of the input space given by X.

- **Situation 3: Low U_E/U_{EA}—High U_A.** In the third situation the calibrated measurements show a large variation and, thus, also lead to a large calibration error. This situation may appear if the uncalibrated input measurements are affected by noise or the calibration model is not able to find an optimal calibration function, for instance due to missing features or insufficient complexity of the model.

- **Situation 4: High U_E/U_{EA}—High U_A.** This is the most undesirable situation (situation 4 in Figure 5.3). It occurs if the calibration model has been trained on only a few measurements and the corresponding outputs are far from the true measurements. In this situation we have no confidence in our calibration model due to missing knowledge and a potentially poorly calibrated model due to the observation of high calibration errors during training.

Summary. All situations except the ideal case with overall low uncertainty point out potential problems with the calibration model. In particular, we face the potential of inaccurate calibrated measurements. In order to further process the calibrated measurements for calibration in subsequent hops we need to pinpoint confident measurements by exploiting our uncertainty estimates. We show how to integrate these uncertainties into multi-hop sensor calibration below.

5.5.2 Uncertainty Based Data Filtering

Multi-hop calibration. The idea of multi-hop calibration, which we describe in Section 3.3.2, is to exploit rendezvous between sensors, i.e., situations when two or more sensors meet in time and space. During rendezvous, sensors are exposed to the same environment and sense the same phenomena, thus, creating a calibration opportunity. Whenever a low-cost sensor is in rendezvous with a reference sensor it can use the reference measurement for calibration. To further increase calibration opportunities, multi-hop calibration also exploits freshly calibrated sensors. That is, a freshly calibrated sensor is providing its calibrated measurements as a virtual reference to an uncalibrated sensor. This process can be repeated until all sensors are calibrated, forming

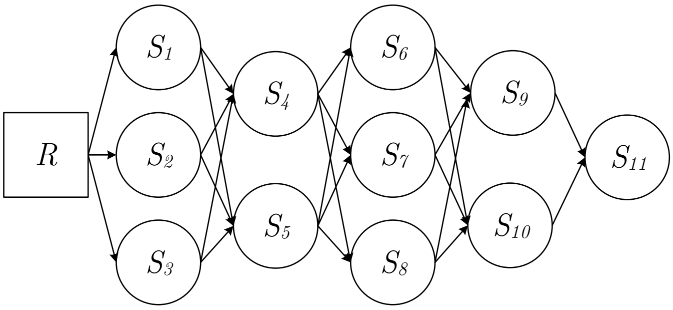


Figure 5.4: Example of a calibration graph consisting of one reference and 11 low-cost sensors. A reference sensor R is calibrating three low cost sensors $S_{\{1,2,3\}}$ in a first hop. These freshly calibrated sensors then provide their calibrated measurements as references to calibrate two additional sensors $S_{\{4,5\}}$. This procedure is continued until all 11 sensors are calibrated.

a calibration graph over multiple hops. An example is illustrated in Figure 5.4. Note that instead of forming a calibration path, where each sensor is calibrated by one parent sensor we adopt a graph structure in this chapter. That is, each sensor can be calibrated by more than one parent sensor, which all can exhibit different measurement uncertainties.

Heuristic filtering. Given a single uncalibrated sensor, which has been in rendezvous with one or multiple freshly calibrated sensors, we can collect a dataset $(M, \hat{m}, U_E(\hat{m}), U_{EA}(\hat{m}), U_A(\hat{m}))$ with uncalibrated sensor measurements M, virtual references $\hat{m}$ and the corresponding uncertainties $U_E(\hat{m}), U_{EA}(\hat{m})$ and $U_A(\hat{m})$ for each sample. Because the calibrated measurements may not be identical with the ground-truth measurements, we need to make sure that we remove potentially inaccurate samples before we perform the calibration for the uncalibrated sensor. In order to do this, we apply a simple heuristic to filter samples with general high uncertainty. In particular, we treat a calibrated sample $\hat{m}_i$ as a confident sample if the following rule applies:

$$(U_E(\hat{m}_i) \leq p_E) \wedge (U_A(\hat{m}_i) \leq p_A) \wedge (U_{EA}(\hat{m}_i) \leq p_{EA}), \tag{5.5}$$

where $p_{\{E,A,EA\}}$ are thresholds. We set these thresholds by percentiles over all the available values for each uncertainty metric of the dataset, e.g., p_E is set to the 90-th percentile of all $U_E(\hat{m})$ in the dataset. We will investigate the effect of this threshold in Section 5.7. As we have seen in Section 5.5.1, any situation with a high uncertainty metric is highlighting a potentially inaccurate calibrated measurement. The heuristic given in (5.5) is therefore removing any measurement that exhibits at least one of the three uncertainties with a too large value according to the defined threshold.

5.6 Simulations

In this section, we conduct simulations to validate the effectiveness of our method to estimate the uncertainties.

5.6.1 Setup

We evaluate our scheme using artificial data samples $(x, y(x))$, which are generated as follows:

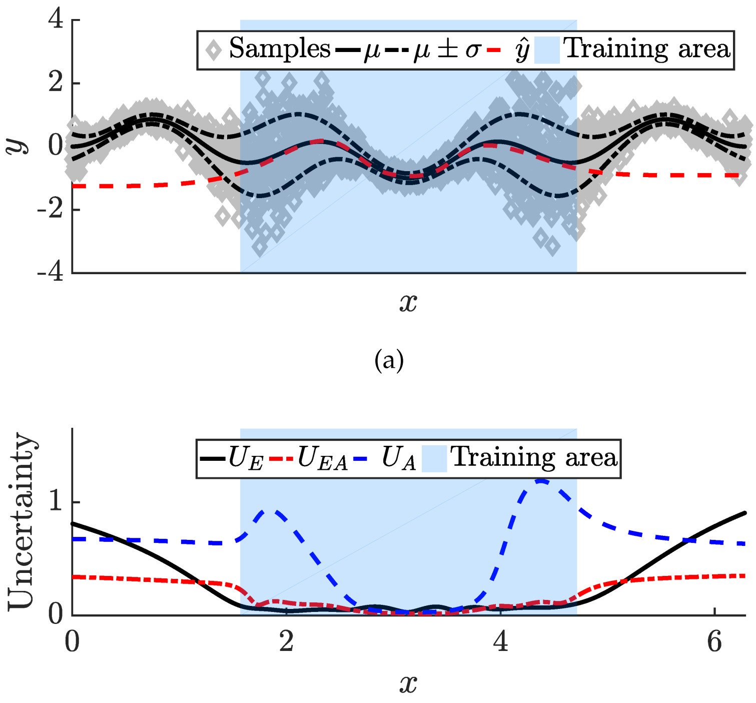
$$\begin{aligned} x &\sim U(0, 2\pi) \\ y(x) &\sim \mu(x) + N(0, \sigma(x)) \\ \mu(x) &= \sin\left(5 \cdot \frac{x}{2}\right) \cdot \sin\left(3 \cdot \frac{x}{2}\right) \\ \sigma(x) &= \frac{3}{20} + \frac{1}{4} \cdot \sin(4 \cdot x) + 2 \cdot \cos\left(\frac{6}{5} \cdot x\right) \end{aligned}$$

The dataset is sampled 2000 times and the samples, where $x \in \left[\frac{\pi}{2}, \frac{3 \cdot \pi}{2}\right]$, are used to train our models in phase 1 of the scheme. Because the underlying function $\mu(x)$ is non-linear, we use an ensemble of non-linear neural networks to approximate the function. The neural network uses the same optimized structure as the one to predict the aleatoric uncertainty, i.e., one shared hidden structure and p outputs. Specifically, we apply for each of the two networks one shared hidden layer with 16 neurons with *tanh*-activation functions and $p = q = 20$ ensemble outputs.

5.6.2 Result

Figure 5.5 shows the simulation results. First of all, in Figure 5.5a we observe that within the training area, the approximated function $\hat{y}(x)$ fits the true function $\mu(x)$. Outside the training area the approximation is not fitting at all. This result is not surprising because the model has to extrapolate. Accordingly, the epistemic uncertainties U_E and U_{EA} (see Figure 5.5b) capture this behaviour. Within the training area the uncertainty is small and grows outside the training area. Further, we also observe a difference between the two uncertainties U_E and U_{EA}. As mentioned in Section 5.4, both capture the same behaviour, but use, however, different underlying models in the respective ensembles and, consequently, output different epistemic uncertainty values.

Secondly, the samples used for training follow an input-dependent normal distribution. Especially at the borders of the training area the variation is significantly higher than in the centre. The aleatoric uncertainty U_A is highlighting this effect, see Figure 5.5b. It follows in fact the variance of the noise distribution $\sigma(x)$. This result is expected, because the optimization function of the aleatoric estimator is set to exactly estimate this variance, see (5.3).



(a)

(b)

Figure 5.5: Performance on artificial data. Figure 5.5a shows that our model is able to perform typical regression tasks. In Figure 5.5b we see the typical behaviour of the epistemic uncertainty, which starts to grow outside the training area, and that we are able to estimate the noise variance in form of the aleatoric uncertainty.

Overall, we can conclude that our scheme is able to *i)* perform typical regression tasks while also estimating *ii)* epistemic (U_E and U_{EA}) and *iii)* aleatoric (U_A) uncertainty.

5.7 Evaluations

In this section, we apply our scheme and evaluate its performance on real-world sensor data.

5.7.1 Setup

Dataset. The dataset consists of measurements from 11 different low-cost metal oxide ozone sensor prototypes, which are also used for evaluation in Section 3.5.2. Each sensor is also paired with a temperature sensor, which is used to tackle the gas sensors dependencies on environmental conditions. Similar to the setup in Section 3.5.2 we use measurements from two sensing layers per device coupled with temperature measurements. The sensors are placed next to the same governmental ozone sensor as the one described in Section 5.3 and shown in Figure 1.4 that serves as reference. We placed all sensors in the same ventilated box and collected with a sampling interval of 10 min approximately 4000 samples of each sensor during the month of October, 2014.

Test setup. We artificially build a multi-hop calibration setup where a subset of the sensors is calibrated by the reference sensor, which in return are used to calibrate another subset of sensors, and so on until all sensors are calibrated. We use five different calibration graph structures, for example the one in Figure 5.4, with varying number of parent nodes, i.e., references, (1 to 4) and number of hops (1 to 5). Each pair of sensor nodes collects 200 samples, i.e., we assume they are 200 times in rendezvous. To simulate a setup with diverse sensors, in particular diverse uncertainties, we randomly choose different measurement distributions, for instance see Figure 5.1a, for each sensor by varying their overall range of collected measurements during rendezvous. Each sensor covers between $[25, 60]\%$ of the total ozone range during the measurement period. The effective measurement ranges are randomly sampled in every experiment. The performance of each calibrated sensor is evaluated on 200 separate testing samples and evaluated by using the normalized $RMSE_\sigma$, see Section 2.4.3: (2.8). Overall, we perform 200 experiments with different graph setups, arrangement of the sensors in the graph and measurement distributions in each experiment.

Data filtering. In order to show the impact of our data filtering, we investigate the effect of different filter thresholds $p_{\{E,A,EA\}}$, see (5.5). As described in Section 5.5.2, the thresholds are defined by the percentile values of each individual uncertainty of a calibration dataset. We use 6 different levels at $\{100, 90, 80, 70, 60, 50\}\%$, where 100% leads to no measurements filtered. For simplicity we use the same percentage-level for all three uncertainties.

Calibration models. We apply our scheme on three different calibration models, our linear SCAN model presented Chapter 3, non-linear regression trees [LDC18] and non-linear neural networks [LDC18, CLL+14], similar to the one used in Chapter 4. Our linear SCAN regression model is specifically developed to reduce error accumulation in multi-hop calibration, see also Section 3.4. Regression trees [LDC18] and neural networks [LDC18, CLL+14] have mainly been used in one-hop calibration, i.e., calibration to perfect references, and allow the modelling of more complex calibration functions like in Chapter 4. Although these methods might suffer from additional error accumulation over multiple hops, their complex modelling capabilities can be helpful to calibrate graphs with small overall number of hops. Similar to Section 5.6, the neural network that learns the aleatoric uncertainty uses one hidden layer with

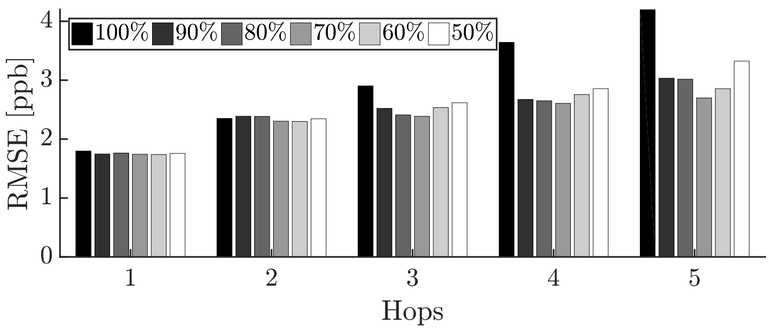


Figure 5.6: Calibration error of SCAN over multiple hops at different threshold percentile values. The higher the number of hops, the more important it is to reduce the impact of potentially inaccurate measurements.

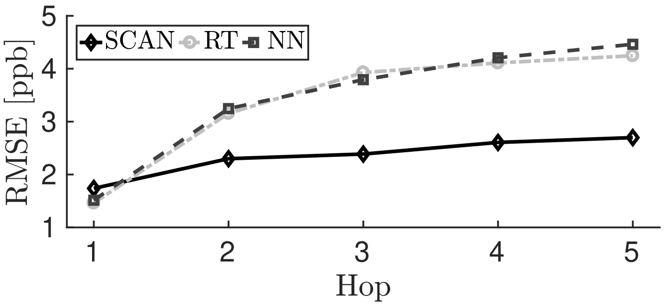
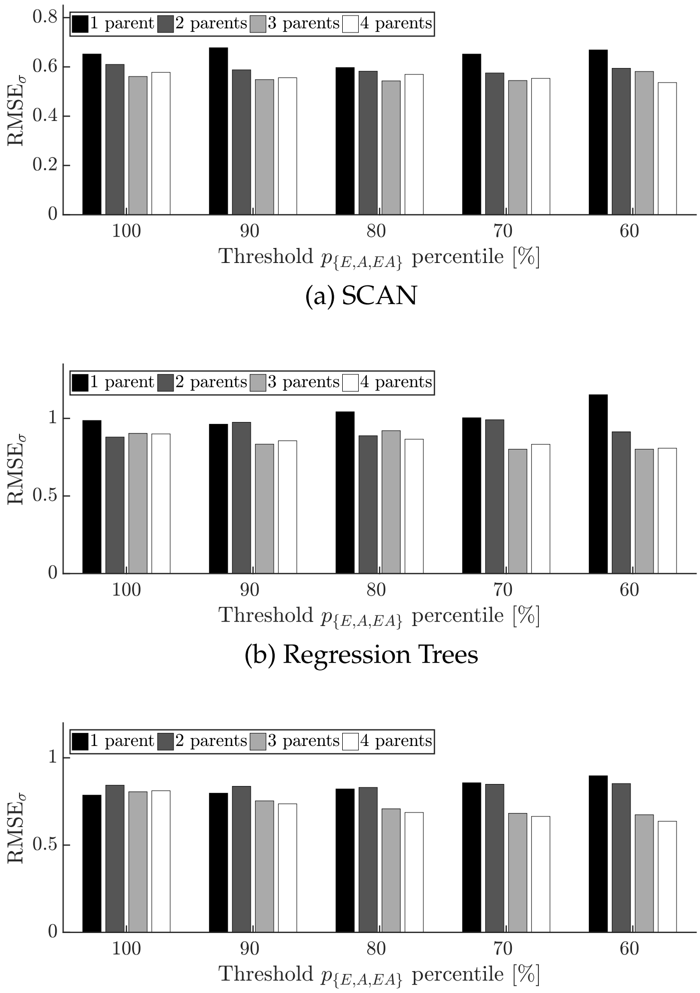


Figure 5.7: Lowest calibration error over multiple hops of the three methods SCAN, Regression Trees (RT) and neural networks (NN) over all threshold values. SCAN clearly outperforms the two other methods with an up to 69% lower error.

16 neurons with tanh-activation functions and 20 ensemble members. The same network structure is also applied when we use neural networks to perform the calibration in phase 1.

5.7.2 Results

Error over multiple hops. Figure 5.6 shows the average calibration error of the sensors when calibrated at different hops in the graph. The different bars correspond to different percentile levels of the data filtering. We can observe that data filtering is able to reduce the calibration error. Especially at hops 3 to 5, using only samples with high confidence performs notably better than using all samples for calibration. While the relative improvement in terms of calibration error compared to no data filtering at hops 1 and 2 is at most 5%, it is in average 11% at hop 3 and 25% at hop 4 and 5. We also observe that the best performing percentile level is 70% at hops 3 and larger. Using lower percentiles results in a loss of accuracy, because too many measurements have been removed. In future work, it will be important to enhance our scheme with capabilities to find the optimal thresholds. Note that the two other



(a) SCAN

(b) Regression Trees

(c) Neural Networks

Figure 5.8: Calibration error at different threshold levels for different number of parent nodes. Filtering samples with high uncertainty becomes more important with more parents for all three methods.

calibration models, regression trees and neural networks, do not perform equally well compared to SCAN. This can be observed in Figure 5.7, where we show the lowest calibration error of all threshold values at each individual hop for the three methods. SCAN achieves an up to 69% lower error than regression trees and 63% than neural networks. The data filtering is decreasing the error in average only by 3% for regression trees and neural networks. The error accumulation over multiple hops has in fact a big impact on the overall accuracy and begins to dominate at later hops. This is mainly because the methods are not developed for multi-hop calibration with large number of hops. However, they can be powerful in setups with small hop numbers, which is presented in the following.

Error versus number of parents. Figure 5.8 shows the average calibration error of our sensors at different filtering thresholds with different number of hops. We use a graph structure where these sensors are all 2 hops away from the reference, meaning they are calibrated by 1 to 4 parents that have been calibrated by the reference. For all three calibration models, SCAN in Figure 5.8a, regression trees in Figure 5.8b and neural networks in Figure 5.8c we can observe a similar trend. The more parents are used to calibrate a sensor, the more effective is our uncertainty based data filtering. This result is expected due to the fact that more parents with different uncertainties at different concentrations can potentially induce high levels of noise in the calibration dataset. By only keeping the most confident samples from the different parents we are able to generate an improved dataset that reduces the calibration error. In fact, we are able to achieve the lowest calibration error for all three calibration models by using 4 parents for calibration in combination with our data filtering.

5.8 Summary

Uncertainty in sensor calibration is an omnipresent phenomenon. In this chapter, we present a scheme to estimate two major uncertainties in typical regression tasks, epistemic and aleatoric uncertainty. The former is related to the ignorance of a calibration model and grows if we do not present the model with sufficient data samples or use a model that is not powerful enough to find the underlying calibration function. The latter describes the calibration error as a function of the calibrated measurements. Our scheme works on any calibration model and exploits bootstrapped ensembles to generate estimates of the two uncertainties. We apply our approach to improve the calibration of low-cost ozone sensors in a multi-hop setup from Chapter 3, an important practice to maintain high data quality over time. By estimating uncertainties of our sensor measurements and using this gained information in an uncertainty-based data filtering, we are able to reduce the calibration error in the multi-hop setup by up to 25% compared to existing approaches.

6

Conclusion and Outlook

Air pollution monitoring has gained increasing interest from various research institutions, governmental authorities and the private sector in the last few years. With the growing availability of various low-cost air pollution sensors there is a strong trend to extend the current monitoring efforts, which are conducted by fixed and expensive installations, with distributed and affordable sensor deployments. These large-scale sensor networks are able to provide air pollution data with high spatio-temporal resolution, which can be further used for applications like fine-grained concentration modelling, city planning or investigations of health effects on citizens.

Unfortunately, common low-cost air pollution sensors suffer from multiple limitations. Different aspects, both due to technology principles and external influences like environmental conditions, can affect a sensor's performance. As a result, the collected data is often affected by substantial inaccuracies when compared to high-end devices.

The aim of this thesis is to counteract the limiting factors and improve the overall data quality of low-cost air quality sensors by performing dedicated sensor calibration. We develop different strategies to transform raw sensor measurements into a calibrated form that describes a concentration of interest as accurate as possible. We focus on various deployment scenarios and provide tailored calibration models to uncover and remove important error sources in low-cost sensor signals. By performing appropriate sensor calibration both before and during a deployment we are able to gather air pollution information with the accuracy fit for future applications.

In the remainder of this chapter we present the main contributions of this thesis and layout possible future research directions.

6.1 Contributions

Pre-Deployment Testing (Chapter 2). Information about cross-sensitivities or dependencies on environmental effects of low-cost sensors is often only sparsely described in datasheets, yet both can have serious effects on the data quality. In order to uncover and counteract these effects we propose an in-field pre-deployment testing approach. Our approach allows to *(i)* judge a sensor on its feasibility for air pollution monitoring in a given environment *(ii)* uncover major cross-sensitivities and dependency on environmental conditions and *(iii)* propose an optimized measurement setup in the form of a sensor-array for stable and accurate measurements.

Multi-hop Sensor Array Calibration (Chapter 3). Low-cost sensors tend to drift over time and, thus, need to be frequently re-calibrated to maintain consistent data quality during their deployment. We design SCAN, a novel collaborative post-deployment algorithm for mobile deployments. SCAN is based on multi-hop calibration and can be applied to sensor arrays to compensate for cross-sensitivities and dependencies on environmental conditions. To the best of our knowledge, we are the first to formulate SCAN as constrained least-squares regression, which can be applied to calibrate sensor arrays during their deployment while being free of regression dilution and, thus, minimizing error accumulation over multiple hops.

Sensor Arrays On Wearables (Chapter 4). We design a wearable platform for personal air pollution monitoring using low-cost metal oxide sensors. Through an extensive measurement study we uncover substantial interference from human gas emissions, e.g., from natural skin oils or respiration, on the air pollution sensors. We counteract this interference by utilizing non-linear neural network calibration. In order to reduce training efforts we build a tailored neural network with shared hidden layers and apply a semi-supervised learning technique to allow post-deployment model updates with little user involvement. Our final design is able to collect accurate air pollution measurements even under the presence of human interference.

Uncertainty in Calibration (Chapter 5). Finally, we investigate two uncertainty types, epistemic and aleatoric, which are common in sensor calibration models. We first develop a scheme to estimate these uncertainties for each calibrated measurement of a model by utilizing bootstrapped ensembles. In a second part, we integrate the uncertainty metrics into multi-hop calibration and design a heuristic data filtering to remove potentially faulty measurements at each hop and, consequently, improve network wide calibration accuracy.

6.2 Possible Future Directions

Automated Sensor Degradation Detection. The data quality degradation of sensors over time is a severe problem that can prevent successful long-term data collection. In this thesis, we tackle it by frequently re-calibrating sensors during their deployment. However, it might not always be possible to perform frequent re-calibration, e.g., the sensor is not in rendezvous with any other sensors or there is in general a lack of accurate and reliable references. In this case, an automated technique, which estimates a sensor's degradation and evaluates its current expected accuracy, can pinpoint measurements with low quality for post-processing purposes and inform about the need for a, potentially manual, re-calibration. Possible directions include the appliance of statistical methods that detect anomalies by comparing measurement distributions over time or the appliance of spatio-temporal prediction models that capture the expected signal range of sensors in the deployment.

Optimal Calibration Path Selection. In this thesis, we applied multi-hop calibration to relatively small sensor networks with at most 11 sensors. With the increasing availability and popularity of low-cost sensors for air pollution monitoring, large-scale deployments with hundreds of distributed sensors pose additional calibration challenges. Especially, how multi-hop calibration scales for large-scale deployments is an open question. We conducted some initial tests using a public dataset of GPS traces from 537 taxis in San Francisco, USA [PSDG09] and discovered that each taxi is in average with 180 other taxis in rendezvous over the course of 30 days. If each taxi acts as a mobile air pollution sensor array, an interesting challenge is how to optimally choose calibration pairs and consequently calibration graphs to achieve the best calibration performance over the whole network. Potential solutions involve the application of uncertainty metrics presented in Chapter 5 in combination with graph theory concepts such as shortest path algorithms.

A Universal Network Calibration Method. In Section 1.5 we present a review of three different network calibration approaches, i.e., *blind* (Section 1.5.1), *collaborative* (Section 1.5.2) and *transfer* (Section 1.5.3) calibration. All these concepts and the specific methods within have their different strengths and weaknesses, as summarized in Section 1.5.4. The existing low-cost air pollution sensor network deployments all have different characteristics, e.g., mobility of the sensor nodes, and calibration requirements, e.g., non-linear calibration models. Unfortunately, a universal and deployment-independent network calibration approach is currently missing. One possibility is a combination of the three approaches by developing a scheme that exploits individual aspects. Offering different calibration opportunities to all the participating nodes is an important task, especially in deployments with heterogeneous sensors nodes, e.g., mobile and static, like in crowdsensing applications.

Exploiting Air Pollution Maps for Sensor Calibration. Creating air pollution maps of urban areas with high spatio-temporal resolution is a thriving research area. These maps offer estimations of current and past pollutant concentrations based on, for instance, sophisticated chemical process models or data-driven approaches. A possible future direction is the utilization of air pollution maps for low-cost sensor calibration. However, air pollution maps are often not able to provide measurements with equal accuracy over the whole area and time-horizon of interest. Therefore, a future challenge is the detection of locations and times where one can trust the air pollution map to be accurate and, hence, the measurements can be used to calibrate low-cost sensors. This could be achieved by applying a data-driven air pollution model coupled with uncertainty metrics, like the neural network structure we present in Chapter 5.

Self-Calibrating Sensor Nodes. Sensor calibration performed in laboratory setups, where the sensor under test is exposed to fixed target pollutant concentrations, is a popular approach that allows quick recordings of the sensor's response. In the last few years different low-cost gas concentration generators[1] have become available on the market. These devices are for instance based on hydrogen cells or ozone-generating ionization modules, have small form factors and allow controlled gas generation with little effort. By integrating these gas generators into air pollution sensor nodes it is possible to reconstruct a laboratory setup and, thus, to perform re-calibration or drift-detection in a controlled manner during deployment. Important aspects that need to be addressed in a potential future work are the accuracy, precision and reliability of the gas generation by these low-cost devices.

[1] E.g. *Varta Hydrogen Generator Cells* or *Murata Electronics Ionissimo Ionizer Modules*

Bibliography

[Aer16] Aeroqual. SM50 ozone sensor circuit (datasheet). https://www.aeroqual.com/wp-content/uploads/SM50-User-Guide-V2.1.pdf, 2016. Accessed: 2019-05-01.

[AHM+09] P. Aoki, R. J. Honicky, A. Mainwaring, C. Myers, E. Paulos, S. Subramanian, and A. Woodruff. A vehicle for research: Using street sweepers to explore the landscape of environmental community action. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (SIGCHI)*, pages 375 – 384, 2009.

[Alp13a] Alphasense. Alphasense air, sensors for air quality networks (webpage). http://www.alphasense.com/index.php/air/, 2013. Accessed: 2019-05-01.

[Alp13b] Alphasense. NO2-B43F 4-Electrode nitrogen dioxide sensor (datasheet). http://www.alphasense.com/WEB1213/wp-content/uploads/2018/12/NO2B43F.pdf, 2013. Accessed: 2019-05-01.

[Alp15] Alphasense. CO-B4 4-Electrode carbon monoxide sensor (datasheet). http://www.alphasense.com/WEB1213/wp-content/uploads/2015/04/COB41.pdf, 2015. Accessed: 2019-05-01.

[AMM15] A. Arfire, A. Marjovi, and A. Martinoli. Model-based rendezvous calibration of mobile sensor networks for monitoring air quality. In *2015 IEEE SENSORS*, pages 1–4, Nov 2015.

[AMM16a] A. Arfire, A. Marjovi, and A. Martinoli. Enhancing measurement quality through active sampling in mobile air quality monitoring sensor networks. In *IEEE International Conference on Advanced Intelligent Mechatronics (AIM)*, pages 1022–1027, 2016.

[AMM16b] A. Arfire, A. Marjovi, and A. Martinoli. Mitigating slow dynamics of low-cost chemical sensors for mobile air quality monitoring sensor networks. In *Proceedings of the 2016 International Conference on Embedded Wireless Systems and Networks (EWSN)*, pages 159 – 167. Junction Publishing, 2016.

[Amp17] Amphenol Advanced Sensors. Telair t6713 series co_2 module. https://www.amphenol-sensors.com/en/telaire/co2/525-co2-sensor-modules/3399-t6713, 2017. Accessed: 2019-05-01.

[Amp19] Amphenol Advanced Sensors. Telaire 6703 series CO2 module (datasheet). https://www.amphenol-sensors.com/en/telaire/co2/525-co2-sensor-modules/3400-t6703, 2019. Accessed: 2019-05-01.

[ams17] ams AG. CCS811 VOC sensor (datasheet). https://ams.com/ccs811, 2017. Accessed: 2019-05-01.

[ANSY15] E. Austin, I. Novosselov, E. Seto, and M. G. Yost. Laboratory evaluation of the shinyei ppd42ns low-cost particulate matter sensor. *PLOS ONE*, 10(9):1–17, 09 2015.

[ARG06] C. C. Austin, B. Roberge, and N. Goyer. Cross-sensitivities of electrochemical detectors used to monitor worker exposures to airborne contaminants: False positive responses in the absence of target analytes. *Journal of Environmental Monitoring*, 8(1):161 – 166, 2006.

[BBR+17] Z. A. Barakeh, P. Breuil, N. Redon, C. Pijolat, N. Locoge, and J.-P. Viricelle. Development of a normalized multi-sensors system for low cost on-line atmospheric pollution detection. *Sensors and Actuators B: Chemical*, 241:1235 – 1243, 2017.

[BEMRB13] M. Budde, R. El Masri, T. Riedel, and M. Beigl. Enabling low-cost particulate matter measurement for participatory sensing scenarios. In *Proceedings of the 12th international conference on mobile and ubiquitous multimedia (MUM)*, pages 19:1–19:10. ACM, 2013.

[BGvdS+13] M. Bruins, J. W. Gerritsen, W. W. van de Sande, A. van Belkum, and A. Bos. Enabling a transferable calibration model for metal-oxide type electronic noses. *Sensors and Actuators B: Chemical*, 188:1187–1195, 2013.

[BKB15] M. Budde, M. Köpke, and M. Beigl. Robust in-situ data reconstruction from poisson noise for low-cost, mobile, non-expert environmental sensing. In *Proceedings of the 2015 ACM International Symposium on Wearable Computers (ISWC)*, pages 179 – 182. ACM, 2015.

[BKW07] 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.

[BL11] R. M. Balabin and E. I. Lomakina. Support vector machine regression (SVR/LS-SVM)-an alternative to neural networks (ANN) for analytical chemistry? comparison of nonlinear methods on near infrared (NIR) spectroscopy data. *Analyst*, 136:1703–1712, 2011.

[BN07] L. Balzano and R. Nowak. Blind calibration of sensor networks. In *Proceedings of the 6th International Conference on Information Processing in Sensor Networks (IPSN)*, pages 79–88. IEEE/ACM, 2007.

[BR14] S. Banerjee and A. Roy. *Linear Algebra and Matrix Analysis for Statistics*. Taylor & Francis Group, 2014.

[Bre96] L. Breiman. Bagging predictors. *Machine learning*, 24(2):123 – 140, 1996.

[Bro03] R. Bro. Multivariate calibration what is in chemometrics for the analytical chemist? *Analytica Chimica Acta*, 500:185 – 194, 2003.

[BS17] R. Baron and J. 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.

[CBKL10] S. Chen, J. S. Brantley, T. Kim, and J. Lach. Characterizing and minimizing synchronization and calibration errors in inertial body sensor networks. In *Proceedings of the Fifth International Conference on Body Area Networks (BodyNets)*, pages 138–144. ACM, 2010.

[CDS+17] N. Castell, F. R. Dauge, P. Schneider, M. Vogt, U. Lerner, B. Fishbain, D. Broday, and A. Bartonova. Can commercial low-cost sensor platforms contribute to air quality monitoring and exposure estimates? *Environment International*, 99(Supplement C):293 – 302, 2017.

[CLL+14] Y. Cheng, X. Li, Z. Li, S. Jiang, Y. Li, J. Jia, and X. Jiang. Aircloud: A cloud-based air-quality monitoring system for everyone. In *Proceedings of the 12th ACM Conference on Embedded Network Sensor Systems (SenSys)*, pages 251–265. ACM, 2014.

[CMC+01] M. C. Carotta, G. Martinelli, L. Crema, C. Malagù, M. Merli, G. Ghiotti, and E. Traversa. Nanostructured thick-film gas sensors for atmospheric pollutant monitoring: quantitative analysis on field tests. *Sensors and Actuators B: Chemical*, 76(1):336 – 342, 2001.

[CWL+17] 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.

[DAK+09] P. Dutta, P. M. Aoki, N. Kumar, A. Mainwaring, C. Myers, W. Willett, and A. 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)*, pages 349 – 350. ACM, 2009.

[Das91] B. V. Dasarathy. *Nearest neighbor (NN) norms: NN pattern classification techniques*. IEEE Computer Society Press, 1991.

[DAS+05] A. Dobre, S. Arnold, R. Smalley, J. Boddy, J. Barlow, A. Tomlin, and S. Belcher. Flow field measurements in the proximity of an urban intersection in london, uk. *Atmospheric Environment*, 39(26):4647–4657, 2005.

[Dat14] Data Canvas. Data Canvas: Sense Your City — building independent, open and participative sensing networks to understand our changing cities, 2014.

[DE92] Y. Danon and M. Embrechts. Least squares fitting using artificial neural networks. *Intelligent Engineering Systems through Artificial Neural Networks*, 2, 1992.

[DES+18] S. De Vito, E. Esposito, M. Salvato, O. Popoola, F. Formisano, R. Jones, and G. D. Francia. Calibrating chemical multisensory devices for real world applications: An in-depth comparison of quantitative machine learning approaches. *Sensors and Actuators B: Chemical*, 255:1191 – 1210, 2018.

[DKC+15] P. J. Dacunto, N. E. Klepeis, K.-C. Cheng, V. Acevedo-Bolton, R.-T. Jiang, J. L. Repace, W. R. Ott, and L. M. Hildemann. Determining PM2.5 calibration curves for a low-cost particle monitor: common indoor residential aerosols. *Environmental Science: Processes & Impacts*, 17:1959–1966, 2015.

[DKJ+14] S. Deshmukh, K. Kamde, A. Jana, S. Korde, R. Bandyopadhyay, R. Sankar, N. Bhattacharyya, and R. Pandey. Calibration transfer between electronic nose systems for rapid in situ measurement of pulp and paper industry emissions. *Analytica Chimica Acta*, 841:58 – 67, 2014.

[DMP+08] S. De Vito, E. Massera, M. Piga, L. Martinotto, and G. D. 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.

[Dov] Dovelet. TP-401A indoor air quality gas sensor (datasheet). `https://seeeddoc.github.io/Grove-Air_Quality_Sensor/res/TP-401A_Indoor_Air_quality_gas_sensor.pdf`. Accessed: 2019-05-01.

[DPDR15] C. Dorffer, M. Puigt, G. Delmaire, and G. Roussel. Blind calibration of mobile sensors using informed nonnegative matrix factorization. In *Latent Variable Analysis and Signal Separation*, pages 497–505. Springer International Publishing, 2015.

[DPDR16a] C. Dorffer, M. Puigt, G. Delmaire, and G. Roussel. Blind mobile sensor calibration using an informed nonnegative matrix factorization with a relaxed rendezvous model. In *Procedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2941–2945. IEEE, 2016.

[DPDR16b] C. Dorffer, M. Puigt, G. Delmaire, and G. Roussel. Nonlinear mobile sensor calibration using informed semi-nonnegative matrix factorization with a vandermonde factor. In *Proceedings of Sensor Array and Multichannel Signal Processing Workshop (SAM)*, pages 1–5. IEEE, 2016.

[DPMF09] S. De Vito, M. Piga, L. Martinotto, and G. D. 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.

[DY97] N. Draper and Y. Yang. Generalization of the geometric mean functional relationship. *Computational Statistics and Data Analysis*, 23(3):355 – 372, 1997.

[EDS$^+$16] E. Esposito, S. De Vito, M. Salvato, V. Bright, R. 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.

[EDS$^+$17] E. Esposito, S. De Vito, M. Salvato, G. Fattoruso, and G. Di Francia. *Computational Intelligence for Smart Air Quality Monitors Calibration*, pages 443–454. Springer International Publishing, 2017.

[EDS$^+$18] E. Esposito, S. De Vito, M. Salvato, G. Fattoruso, V. Bright, R. L. Jones, and O. Popoola. Stochastic comparison of machine learning approaches to calibration of mobile air quality monitors. In *Sensors*, pages 294–302. Springer International Publishing, 2018.

[EK12] W. Eugster and G. W. Kling. Performance of a low-cost methane sensor for ambient concentration measurements in preliminary studies. *Atmospheric Measurement Techniques*, 5(8):1925–1934, 2012.

[EPC12] EveryAware Project Consortium. Report on: sensor selection, calibration and testing; EveryAware platform; smartphone applications. `http://www.everyaware.eu/`, 2012. Accessed: 2019-05-01.

[ESV$^+$18] E. Esposito, M. Salvato, S. D. Vito, G. Fattoruso, N. Castell, K. Karatzas, and G. D. Francia. *Assessing the Relocation Robustness of on Field Calibrations for Air Quality Monitoring Devices*, pages 303–312. Springer International Publishing, 2018.

[ET94] B. Efron and R. J. Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.

[Eur17] European Research Area Network. Convergence: Frictionless energy efficient convergent wearables for healthcare and lifestyle applications. `https://www.flagera.eu/wp-content/uploads/2016/02/FLAG-ERA_JTC2016_Project_flyer_Convergence_v0.3.pdf`, 2017. Accessed: 2019-05-01.

[FB17a] X. Fang and I. Bate. Issues of using wireless sensor network to monitor urban air quality. In *Proceedings of the First ACM International Workshop on the Engineering of Reliable, Robust, and Secure Embedded Wireless Sensing Systems (FAILSAFE)*, pages 32–39. ACM, 2017.

[FB17b] X. Fang and I. Bate. Using multi-parameters for calibration of low-cost sensors in urban environment. In *Proceedings of the 2017 International Conference on Embedded Wireless Systems and Networks (EWSN)*, pages 1–11. Junction Publishing, 2017.

[FCAB10] G. F. Fine, L. M. Cavanagh, A. Afonja, and R. Binions. Metal oxide semi-conductor gas sensors in environmental monitoring. *Sensors*, 10(6):5469 – 5502, 2010.

[FFGG+16] J. Fonollosa, L. Fernández, A. Gutiérrez-Gálvez, R. Huerta, and S. Marco. Calibration transfer and drift counteraction in chemical sensor arrays using direct standardization. *Sensors and Actuators B: Chemical*, 236:1044 – 1053, 2016.

[FMT+99] J. Farringdon, A. J. Moore, N. Tilbury, J. Church, and P. D. Biemond. Wearable sensor badge and sensor jacket for context awareness. In *Digest of Papers. Third International Symposium on Wearable Computers*, 1999.

[FP99] J. D. Fenske and S. E. Paulson. Human breath emissions of vocs. *Journal of the Air & Waste Management Association*, 49(5):594–598, 1999.

[FPPJ99] B. J. Finlayson-Pitts and J. N. Pitts Jr. *Chemistry of the upper and lower atmosphere: theory, experiments, and applications*. Elsevier, 1999.

[FRD17] K. Fu, W. Ren, and W. Dong. Multihop calibration for mobile sensing: k-hop calibratability and reference sensor deployment. In *Proceedings of Conference on Computer Communications (INFOCOM)*, pages 1–9. IEEE, 2017.

[FT00] C. Frost and S. G. Thompson. Correcting for regression dilution bias: comparison of methods for a single predictor variable. *Journal of the Royal Statistical Society Series A*, 163(2):173–189, 2000.

[GD98] M. Gardner and S. Dorling. Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmospheric Environment*, 32(14):2627 – 2636, 1998.

[GG16] Y. Gal and Z. Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning (ICML)*, pages 1050–1059, 2016.

[GL80] G. H. Golub and C. F. V. Loan. An analysis of the total least squares problem. *SIAM Journal on Numerical Analysis*, 17(6):883–893, 1980.

[Goo17] Google. Tensorflow mobile. `https://www.tensorflow.org/`, 2017. Accessed: 2019-05-01.

[GS01] H. A. Gowadia and G. S. Settles. The natural sampling of airborne trace signals from explosives concealed upon the human body. *Journal of Forensic Science*, 46(6):1324 – 1331, 2001.

[GWL$^+$08] M. Gallagher, C. Wysocki, J. Leyden, A. Spielman, X. Sun, and G. Preti. Analyses of volatile organic compounds from human skin. *British Journal of Dermatology*, 159(4):780 – 791, 2008.

[GYI16] J. Gong, X.-D. Yang, and P. Irani. Wristwhirl: One-handed continuous smartwatch input using wrist gestures. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology (UIST)*, pages 861 – 872. ACM, 2016.

[HHU$^+$10] S. Herberger, M. Herold, H. Ulmer, A. Burdack-Freitag, and F. Mayer. Detection of human effluents by a MOS gas sensor in correlation to VOC quantification by GC/MS. *Building and Environment*, 45(11):2430 – 2439, 2010.

[HIVF$^+$18] D. H. Hagan, G. Isaacman-VanWertz, J. P. Franklin, L. M. M. Wallace, B. D. Kocar, C. L. Heald, and J. H. Kroll. Calibration and assessment of electrochemical air quality sensors by co-location with regulatory-grade instruments. *Atmospheric Measurement Techniques*, 11(1):315–328, 2018.

[How06] J. Howe. The rise of crowdsourcing. *Wired Magazine*, 14(6):1 – 4, 2006.

[HPSS14] D. M. Holstius, A. Pillarisetti, K. R. Smith, and E. Seto. Field calibrations of a low-cost aerosol sensor at a regulatory monitoring site in california. *Atmospheric Measurement Techniques*, 7(4):1121–1131, 2014.

[HSST12] D. Hasenfratz, O. Saukh, S. Sturzenegger, and L. Thiele. Participatory air pollution monitoring using smartphones. In *2nd International Workshop on Mobile Sensing*, pages 1–5. ACM, 2012.

[HST12] D. Hasenfratz, O. Saukh, and L. Thiele. On-the-fly calibration of low-cost gas sensors. In *Proceedings of the 9th European Conference on Wireless Sensor Networks (EWSN)*, pages 228–244. Springer-Verlag, 2012.

[HSW$^+$14] 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 *Proceedings of IEEE International Conference on Pervasive Computing and Communications (PerCom)*, pages 69–77, 2014.

[HSW$^+$15] D. Hasenfratz, O. Saukh, C. Walser, C. Hueglin, M. Fierz, T. Arn, J. Beutel, and L. Thiele. Deriving high-resolution urban air pollution maps using mobile sensor nodes. *Pervasive and Mobile Computing*, 16:268 – 285, 2015.

[Jan92] J. Janata. Chemical sensors. *Analytical Chemistry*, 64(12):196–219, 1992.

[JHW+16] 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.

[JLT+11] Y. Jiang, K. Li, L. Tian, R. Piedrahita, X. Yun, O. Mansata, Q. Lv, R. P. Dick, M. Hannigan, and L. Shang. Maqs: a personalized mobile sensing system for indoor air quality monitoring. In *Proceedings of the 13th International Conference on Ubiquitous Computing (UbiComp)*, pages 271 – 280. ACM, 2011.

[JLT+18] R. Jayaratne, X. Liu, P. Thai, M. Dunbabin, and L. Morawska. The influence of humidity on the performance of a low-cost air particle mass sensor and the effect of atmospheric fog. *Atmospheric Measurement Techniques*, 11(8):4883 – 4890, 2018.

[Jon99] A. P. Jones. Indoor air quality and health. *Atmospheric Environment*, 33(28):4535 – 4564, 1999.

[JSBT+15] M. Jovašević-Stojanović, A. Bartonova, D. Topalović, I. Lazović, B. Pokrić, and Z. Ristovski. On the use of small and cheaper sensors and devices for indicative citizen-based monitoring of respirable particulate matter. *Environmental Pollution*, 206:696 – 704, 2015.

[Kah73] H. D. Kahn. Note on the distribution of air pollutants. *Journal of the Air Pollution Control Association*, 23(11):973–973, 1973.

[KBBS07] G. Korotcenkov, I. Blinov, V. Brinzari, and J. Stetter. Effect of air humidity on gas response of SnO2 thin film ozone sensors. *Sensors and Actuators B: Chemical*, 122(2):519 – 526, 2007.

[KBP06] M. Kamionka, P. Breuil, and C. Pijolat. Calibration of a multivariate gas sensing device for atmospheric pollution measurement. *Sensors and Actuators B: Chemical*, 118(1):323 – 327, 2006. Eurosensors XIX.

[KCS14] J.-Y. Kim, C.-H. Chu, and S.-M. Shin. Issaq: An integrated sensing systems for real-time indoor air quality monitoring. *IEEE Sensors Journal*, 14(12):4230 – 4244, 2014.

[KD09] A. D. Kiureghian and O. Ditlevsen. Aleatory or epistemic? does it matter? *Structural Safety*, 31(2):105 – 112, 2009. Risk Acceptance and Risk Communication.

[KESN+18] F. Kizel, Y. Etzion, R. Shafran-Nathan, I. Levy, B. Fishbain, A. Bartonova, and D. M. Broday. Node-to-node field calibration of wireless distributed air pollution sensor network. *Environmental Pollution*, 233:900 – 909, 2018.

[KG17] A. Kendall and Y. Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, pages 5574 – 5584, 2017.

[KPG10] S. Kim, E. Paulos, and M. D. Gross. Wearair: Expressive t-shirts for air quality sensing. In *Proceedings of the fourth international conference on Tangible, embedded, and embodied interaction (TEI)*, pages 295 – 296. ACM, 2010.

[KSB+16] P. Kumar, A. N. Skouloudis, M. Bell, M. Viana, M. C. Carotta, G. Biskos, and L. Morawska. Real-time sensors for indoor air monitoring and challenges ahead in deploying them to urban buildings. *Science of The Total Environment*, 560-561:150 – 159, 2016.

[KSL+18] J. Kim, A. A. Shusterman, K. J. Lieschke, C. Newman, and R. C. Cohen. The berkeley atmospheric CO2 observation network: field calibration and evaluation of low-cost air quality sensors. *Atmospheric Measurement Techniques*, 11(4):1937–1946, 2018.

[LCL+12] X. Liu, S. Cheng, H. Liu, S. Hu, D. Zhang, and H. Ning. A survey on gas sensing technology. *Sensors*, 12(7):9635–9665, 2012.

[LDC18] Y. Lin, W. Dong, and Y. Chen. Calibrating low-cost sensors by a two-phase learning approach for urban air quality measurement. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)*, 2(1):18:1 – 18:18, March 2018.

[LFS+12] J. J. Li, B. Faltings, O. Saukh, D. Hasenfratz, and J. Beutel. Sensing the air we breathe - the OpenSense Zurich dataset. In *AAAI Conference on Artificial Intelligence*, 2012.

[Lip89] M. Lippmann. Health effects of ozone: a critical review. *Journal of Air & Waste Management Association*, 39(5):672 – 695, 1989.

[LLE+16] A. C. Lewis, J. D. Lee, P. M. Edwards, M. D. Shaw, M. J. Evans, S. J. Moller, K. R. Smith, J. W. Buckley, M. Ellis, S. R. Gillot, and A. White. Evaluating the performance of low cost chemical sensors for air pollution research. *Faraday Discuss.*, 189:85–103, 2016.

[LMM14] S. Laurent, G. Michel, and A. Manuel. Laboratory and in-situ validation of ozone microsensors, α-sense, model B4 O3 sensors. Technical report, Publications Office of the European Union, 2014.

[LPB17] B. Lakshminarayanan, A. Pritzel, and C. Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, pages 6402 – 6413, 2017.

[LZW+17] S. Liu, Z. Zheng, F. Wu, S. Tang, and G. Chen. Context-aware data quality estimation in mobile crowdsensing. In *Proceedings of IEEE Conference on Computer Communications (INFOCOM)*. IEEE, 2017.

[MBS$^+$16] J.-F. Markert, M. Budde, G. Schindler, M. Klug, and M. Beigl. Private rendezvous-based calibration of low-cost sensors for participatory environmental sensing. In *Proceedings of the Second International Conference on IoT in Urban Space (Urb-IoT)*, pages 82–85. ACM, 2016.

[MBS$^+$18] J.-F. Markert, M. Budde, G. Schindler, M. Klug, and M. Beigl. Privacy-preserving collaborative blind macro-calibration of environmental sensors in participatory sensing. *EAI Endorsed Transactions on Internet of Things*, 18(10), 1 2018.

[MLCOS08] E. Miluzzo, N. D. Lane, A. T. Campbell, and R. Olfati-Saber. *CaliBree: A Self-calibration System for Mobile Sensor Networks*, pages 314–331. Springer Berlin Heidelberg, 2008.

[MLE$^+$15] S. Moltchanov, I. Levy, Y. Etzion, U. Lerner, D. M. Broday, and B. Fishbain. On the feasibility of measuring urban air pollution by wireless distributed sensor networks. *Science of The Total Environment*, 502:537 – 547, 2015.

[MMH17] M. Mueller, J. Meyer, and C. Hueglin. Design of an ozone and nitrogen dioxide sensor unit and its long-term operation within a sensor network in the city of zurich. *Atmospheric Measurement Techniques*, 10(10):3783–3799, 2017.

[Mon01] C. Monn. Exposure assessment of air pollutants: a review on spatial heterogeneity and indoor/outdoor/personal exposure to suspended particulate matter, nitrogen dioxide and ozone. *Atmospheric Environment*, 35(1):1–32, 2001.

[Mor07] I. Morsi. A microcontroller based on multi sensors data fusion and artificial intelligent technique for gas identification. In *Proceedings of 33rd Annual Conference of the IEEE Industrial Electronics Society (IECON)*, 2007.

[Mor08] I. Morsi. Electronic noses for monitoring environmental pollution and building regression model. In *IECON*, 2008.

[MPH15] 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.

[MPS$^+$13] M. Mead, O. Popoola, G. Stewart, P. Landshoff, M. Calleja, M. Hayes, J. Baldovi, M. McLeod, T. Hodgson, J. Dicks, A. Lewis, J. Cohen, R. Baron, J. Saffell, and R. Jones. The use of electrochemical sensors for monitoring urban air quality in low-cost, high-density networks. *Atmospheric Environment*, 70:186 – 203, 2013.

[MSVA99] M. A. Martin, J. Santos, H. Vásquez, and J. Agapito. Study of the interferences of NO2 and CO in solid state commercial sensors. *Sensors and Actuators B: Chemical*, 58(1–3):469 – 473, 1999.

[MZK+17] C. R. Martin, N. Zeng, A. Karion, R. R. Dickerson, X. Ren, B. N. Turpie, and K. J. Weber. Evaluation and environmental correction of ambient CO2 measurements from a low-cost NDIR sensor. *Atmospheric Measurement Techniques*, 10(7):2383–2395, 2017.

[NKSdD15] M. Nakayoshi, M. Kanda, R. Shi, and R. de Dear. Outdoor thermal physiology along human pathways: a study using a wearable measurement system. *International Journal of Biometeorology*, 59(5):503 – 515, 2015.

[NVL15] E. Niforatos, A. Vourvopoulos, and M. Langheinrich. Weather with you: Evaluating report reliability in weather crowdsourcing. In *Proceedings of the 14th International Conference on Mobile and Ubiquitous Multimedia (MUM)*, pages 152 – 162. ACM, 2015.

[NVL17] E. Niforatos, A. Vourvopoulos, and M. Langheinrich. Understanding the potential of human - machine crowdsourcing for weather data. *International Journal of Human-Computer Studies*, 102:54 – 68, 2017.

[NVZ+12] N. Nikzad, N. Verma, C. Ziftci, E. Bales, N. Quick, P. Zappi, K. Patrick, S. Dasgupta, I. Krueger, T. v. Rosing, and W. 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)*, pages 11:1 – 11:8. ACM, 2012.

[NW95] D. A. Nix and A. S. Weigend. Learning local error bars for nonlinear regression. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, pages 489 – 496, 1995.

[Nyf01] U. Nyffeler. Das Nationale Beobachtungsnetz für Luftfremdstoffe. In *BUWAL*, 2001.

[OB15] D. Oletic and V. Bilas. Design of sensor node for air quality crowdsensing. In *Proceedings of IEEE Sensors Applications Symposium (SAS)*, pages 1 – 5. IEEE, 2015.

[OBPVR16] I. Osband, C. Blundell, A. Pritzel, and B. Van Roy. Deep exploration via bootstrapped dqn. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, pages 4026 – 4034, 2016.

[OKR+00] L. Oglesby, N. Künzli, M. Röösli, C. Braun-Fahrländer, P. Mathys, W. Stern, M. Jantunen, and A. Kousa. Validity of ambient levels of fine particles as surrogate for personal exposure to outdoor air pollution. results of the european EXPOLIS-EAS study (swiss center Basel). *Journal of the Air & Waste Management Association*, 50(7):1251 – 1261, 2000.

[ORL+13] A. Overeem, J. C. Robinson, H. Leijnse, G.-J. Steeneveld, B. K. Horn, and R. Uijlenhoet. Crowdsourcing urban air temperatures from smartphone battery temperatures. *Geophysical Research Letters*, 40(15):4081 – 4085, 2013.

[Par13] W. S. Parker. Ensemble modeling, uncertainty and robust predictions. *Wiley Interdisciplinary Reviews: Climate Change*, 4(3):213 – 223, 2013.

[PEA08] PEAN 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.

[PM16] T. Pieri and M. P. Michaelides. Air pollution monitoring in lemesos using a wireless sensor network. In *2016 18th Mediterranean Electrotechnical Conference (MELECON)*, pages 1–6, April 2016.

[PPS+99] C. Pijolat, C. Pupier, M. Sauvan, G. Tournier, and R. Lalauze. Gas detection for automotive pollution control. *Sensors and Actuators B: Chemical*, 59(2–3):195 – 202, 1999.

[PR97] W. D. Penny and S. J. Roberts. Neural network predictions with error bars. *Dept. of Electrical and Electronic Engineering*, 1997.

[PSDG09] M. Piorkowski, N. Sarafijanovic-Djukic, and M. Grossglauser. CRAWDAD dataset epfl/mobility (v. 2009-02-24). `https://crawdad.org/epfl/mobility/20090224`, February 2009. Accessed: 2019-05-01.

[PSL+17] X. Pang, M. D. Shaw, A. C. Lewis, L. J. Carpenter, and T. Batchellier. Electrochemical ozone sensors: A miniaturised alternative for ozone measurements in laboratory experiments and air-quality monitoring. *Sensors and Actuators B: Chemical*, 240(Supplement C):829 – 837, 2017.

[PSMJ16] O. A. Popoola, G. B. Stewart, M. I. Mead, and R. L. Jones. Development of a baseline-temperature correction methodology for electrochemical sensors and its implications for long-term stability. *Atmospheric Environment*, 147:330 – 343, 2016.

[PXM+14] 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.

[RCK+10] R. K. Rana, C. T. Chou, S. S. Kanhere, N. Bulusu, and W. Hu. Ear-phone: An end-to-end participatory urban noise mapping system. In *Proceedings of the 9th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*, pages 105 – 116. ACM, 2010.

[RHB18] D. Rüffer, F. Hoehne, and J. Bühler. New digital metal-oxide (mox) sensor platform. *Sensors*, 18(4):1052, 2018.

[RKP$^+$17] A. C. Rai, P. Kumar, F. Pilla, A. N. Skouloudis, S. D. Sabatino, C. Ratti, A. Yasar, and D. Rickerby. End-user perspective of low-cost sensors for outdoor air pollution monitoring. *Science of The Total Environment*, 607-608:691 – 705, 2017.

[RKSM14] V. Radu, P. Katsikouli, R. Sarkar, and M. K. Marina. A semi-supervised learning approach for robust indoor-outdoor detection with smartphones. In *Proceedings of the 12th ACM Conference on Embedded Network Sensor Systems (SenSys)*, pages 280 – 294. ACM, 2014.

[RLC14] A. C. Rai, C.-H. Lin, and Q. Chen. Numerical modeling of volatile organic compound emissions from ozone reactions with human-worn clothing in an aircraft cabin. *HVAC&R Research*, 20(8):922–931, 2014.

[RP04] A. Ribas and J. Peñuelas. Temporal patterns of surface ozone levels in different habitats of the north western mediterranean basin. *Atmospheric Environment*, 38(7):985 – 992, 2004.

[RSC97] C. A. Redlich, J. Sparer, and M. R. Cullen. Sick-building syndrome. *The Lancet*, 349(9057):1013 – 1016, 1997.

[RTSH08a] C. R. Rao, H. Toutenburg, Shalabh, and C. Heumann. *The Multiple Linear Regression Model and Its Extensions*, pages 33 – 141. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

[RTSH08b] C. R. Rao, H. Toutenburg, Shalabh, and C. Heumann. *The Simple Linear Regression Model*, pages 7 – 31. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

[SAZP08] J. H. Sohn, M. Atzeni, L. Zeller, and G. Pioggia. Characterisation of humidity dependence of a metal oxide semiconductor sensor array using partial least squares. *Sensors and Actuators B: Chemical*, 131(1):230 – 235, 2008.

[Sch66] P. H. Schönemann. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1 – 10, 1966.

[SD17] G. Schmitt and D. Donath. ESUM: analysing trade-offs between the energy and social performance of urban morphologies. `http://esum.arch.ethz.ch/about`, 2017. Accessed: 2019-05-01.

[Sen16] Sensirion AG. SHTC1 humidity and temperature sensor IC (datasheet). `https://www.sensirion.com/fileadmin/user_upload/customers/sensirion/Dokumente/0_Datasheets/Humidity/Sensirion_Humidity_Sensors_SHTC1_Datasheet.pdf`, 2016.

[SFM99] O. Seppänen, W. Fisk, and M. Mendell. Association of ventilation rates and CO2 concentrations with health andother responses in commercial and institutional buildings. *Indoor Air*, 9(4):226 – 252, 1999.

[SG$^+$11] A. Shrivastava, V. B. Gupta, et al. Methods for the determination of limit of detection and limit of quantitation of the analytical methods. *Chronicles of young scientists*, 2(1):21, 2011.

[SGG$^+$13] L. Sánchez, V. Gutiérrez, J. A. Galache, P. Sotres, J. R. Santana, J. Casanueva, and L. Muñoz. Smartsantander: Experimentation and service provision in the smart city. In *Proceedings of 16th International Symposium on Wireless Personal Multimedia Communications (WPMC)*, pages 1–6. IEEE, 2013.

[SGK$^+$17] L. Spinelle, M. Gerboles, G. Kok, S. Persijn, and T. Sauerwald. Review of portable and low-cost sensors for the ambient air monitoring of benzene and other volatile organic compounds. *Sensors*, 17(7), 2017.

[SGV$^+$14] L. Spinelle, M. Gerboles, M. Villani, M. Aleixandre, and F. Bonavitacola. Calibration of a cluster of low-cost sensors for the measurement of air pollution in ambient air. In *Proceedings of IEEE SENSORS*, pages 21–24, Nov 2014.

[SGV$^+$15] L. Spinelle, M. Gerboles, M. G. Villani, M. Aleixandre, and F. 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(Supplement C):249 – 257, 2015.

[SGV$^+$17] L. Spinelle, M. Gerboles, M. G. Villani, M. Aleixandre, and F. 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(Supplement C):706 – 715, 2017.

[SGX08] SGX Sensortech. Mics-oz-47 ozone sensing head with smart transmitter pcb (datasheet). `https://gitlab.ethz.ch/tec/public/opensense/wikis/files/micsoz47.pdf`, 2008. Accessed: 2019-05-01.

[SGX13] SGX Sensortech. SGX Sensortech (webpage). `https://www.sgxsensortech.com/`, 2013. Accessed: 2019-05-01.

[Shi10] Shinyei. PPD42NS Particle Sensor (datasheet). `https://github.com/SeeedDocument/Grove_Dust_Sensor/raw/master/resource/Grove_-_Dust_sensor.pdf`, 2010. Accessed: 2019-05-01.

[SHT15] O. Saukh, D. Hasenfratz, and L. 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)*, pages 274–285. IEEE,ACM, 2015.

[SHWT14] O. Saukh, D. Hasenfratz, C. Walser, and L. Thiele. *On Rendezvous in Mobile Sensing Networks*, pages 29–42. Springer International Publishing, 2014.

[Sil17] Silicon Labs. Thunderboard sense kit (datasheet). `https://www.silabs.com/documents/public/user-guides/ug250-tb001-user-guide.pdf`, 2017. Accessed: 2019-05-01.

[SITN17] F. Sailhan, V. Issarny, and O. Tavares Nascimento. Opportunistic multiparty calibration for robust participatory sensing. In *Proceedings of IEEE 14th International Conference on Mobile Ad Hoc and Sensor Systems (MASS)*, Orlando, United States, 2017. IEEE.

[STP08] A. Szpakowski, C. Tyszkiewicz, and T. Pustelny. Multivariate analysis in gas sensing applications. *Acta Physica Polonica A*, 114:239 – 242, 2008.

[Str04] D. Stranneby. *Digital signal processing and applications*. Elsevier, 2004.

[SVC$^+$10] E. W. Steyerberg, A. J. Vickers, N. R. Cook, T. Gerds, M. Gonen, N. Obuchowski, M. J. Pencina, and M. W. Kattan. Assessing the performance of prediction models: a framework for some traditional and novel measures. *Epidemiology (Cambridge, Mass.)*, 21(1):128, 2010.

[SWLL91] H. Sundgren, F. Winquist, I. Lukkari, and I. Lundstrom. Artificial neural networks and gas sensor arrays: quantification of individual components in a gas mixture. *Measurement Science and Technology*, 2(5):464, 1991.

[SWN17] L. Sun, D. Westerdahl, and Z. Ning. Development and evaluation of a novel and cost-effective approach for low-cost NO2 sensor drift correction. *Sensors*, 17(8), 2017.

[SZW$^+$04] Y. Sun, G. Zhuang, Y. Wang, L. Han, J. Guo, M. Dan, W. Zhang, Z. Wang, and Z. Hao. The air-borne particulate pollution in beijing -- concentration, composition, distribution and sources. *Atmospheric Environment*, 38(35):5991 – 6004, 2004.

[TC09] A. Tiwary and J. Colls. *Air pollution: measurement, modelling and mitigation*. Taylor & Francis, 2009.

[TDMP16] R. Tian, C. Dierk, C. Myers, and E. Paulos. Mypart: Personal, portable, accurate, airborne particle counting. In *Proceedings of the 2016 Conference on Human Factors in Computing Systems (CHI)*, pages 1338–1348. ACM, 2016.

[THMA03] A. A. Tomchenko, G. P. Harmer, B. T. Marquis, and J. W. Allen. Semiconducting metal oxide sensor array for the selective detection of combustion gases. *Sensors and Actuators B: Chemical*, 93(1–3):126 – 134, 2003.

[Tho16] J. E. Thompson. Crowd-sourced air quality studies: A review of the literature & portable sensors. *Trends in Environmental Analytical Chemistry*, 11:23 – 34, 2016.

[TLL95] I. V. Tetko, D. J. Livingstone, and A. I. Luik. Neural network studies. 1. comparison of overfitting and overtraining. *Journal of chemical information and computer sciences*, 35(5):826 – 833, 1995.

[Tof02] C. Tofallis. Model fitting for multiple variables by minimising the geometric mean deviation. In *Total Least Squares and Errors-In-Variables Modeling: Algorithms, Analysis And Applications*, 2002.

[TPP12] P. Thunis, A. Pederzoli, and D. Pernigotti. Performance criteria to evaluate air quality modeling applications. *Atmospheric Environment*, 59(Supplement C):476 – 482, 2012.

[TSE80] K. Tadesse, D. Smith, and M. A. Eastwod. Breath hydrogen (h2) and methane (ch4) excretion patterns in normal man and in clinical practice. *Quarterly Journal of Experimental Physiology and Cognate Medical Sciences*, 65(2):85–97, 1980.

[TYIM05] W. Tsujita, A. Yoshino, H. Ishida, and T. Moriizumi. Gas sensor network for air-pollution monitoring. *Sensors and Actuators B: Chemical*, 110(2):304 – 311, 2005.

[Val14] D. Vallero. *Fundamentals of air pollution*. Academic press, 2014.

[VDR$^+$10] D. L. Vaughn, T. S. Dye, P. T. Roberts, A. E. Ray, and J. L. DeWinter. Characterization of low-cost NO2 sensors. Technical report, Sonoma Technology, Inc., 2010.

[VG08] S. Vaihinger and W. Goepel. *Multi-Component Analysis in Chemical Sensing*, pages 191 – 237. World Scientific, 2008.

[Wes16] C. J. Weschler. Roles of the human occupant in indoor chemistry. *Indoor air*, 26(1):6 – 24, 2016.

[WHO16] World-Health-Organization. Ambient air pollution: A global assessment of exposure and burden of disease. `https://www.who.int/phe/publications/air-pollution-global-assessment/en/`, 2016. Accessed: 2019-05-01.

[Wic18] Wicked Device LLC. Air Quality Egg: community-led sensing network. `http://airqualityegg.com`, 2018. Accessed: 2019-05-01.

[WLJ$^+$15] Y. Wang, J. Li, H. Jing, Q. Zhang, J. Jiang, and P. Biswas. Laboratory evaluation and calibration of three low-cost particle sensors for particulate matter measurement. *Aerosol Science and Technology*, 49(11):1063–1077, 2015.

[Woo41] E. B. Woolley. The method of minimized areas as a basis for correlation analysis. *Econometrica*, 9(1):38 – 62, 1941.

[WRH91] A. S. Weigend, D. E. Rumelhart, and B. A. Huberman. Generalization by weight-elimination with application to forecasting. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, pages 875 – 882, 1991.

[WW10] A. Wisthaler and C. J. Weschler. Reactions of ozone with human skin lipids: sources of carbonyls, dicarbonyls, and hydroxycarbonyls in indoor air. *Proceedings of the National Academy of Sciences*, 107(15):6568 – 6575, 2010.

[WWS+00] P. Wargocki, D. P. Wyon, J. Sundell, G. Clausen, and P. O. Fanger. The effects of outdoor air supply rate in an office on perceived air quality, sick building syndrome (SBS) symptoms and productivity. *Indoor Air*, 10(4):222 – 236, 2000.

[WYL+16] Y. Wang, A. Yang, Z. Li, X. Chen, P. Wang, and H. Yang. Blind drift calibration of sensor networks using sparse bayesian learning. *IEEE Sensors Journal*, 16(16):6249–6260, 2016.

[WYZ+10] C. Wang, L. Yin, L. Zhang, D. Xiang, and R. Gao. Metal oxide gas sensors: sensitivity and influencing factors. *Sensors*, 10(3):2088–2106, 2010.

[XBP+12] Y. Xiang, L. Bai, R. Piedrahita, R. P. Dick, Q. Lv, M. Hannigan, and L. Shang. Collaborative calibration and sensor placement for mobile sensor networks. In *Proceedings of ACM/IEEE 11th International Conference on Information Processing in Sensor Networks (IPSN)*, pages 73–84. ACM/IEEE, 2012.

[XWXJ08] G. Xing, T. Wang, Z. Xie, and W. Jia. Rendezvous planning in wireless sensor networks with mobile elements. *IEEE Transactions on Mobile Computing*, 7(12):1430–1443, 2008.

[YGTL14] H. Ye, T. Gu, X. Tao, and J. Lu. Sbc: Scalable smartphone barometer calibration through crowdsourcing. In *Proceedings of the 11th International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services (MOBIQUITOUS)*, pages 60 – 69, 2014.

[YLM+15] W. Y. Yi, K. M. Lo, T. Mak, K. S. Leung, Y. Leung, and M. L. Meng. A survey of wireless sensor network ased air pollution monitoring systems. *Sensors*, 15(12):31392–31427, 2015.

[YZ15] K. Yan and D. Zhang. Improving the transfer ability of prediction models for electronic noses. *Sensors and Actuators B: Chemical*, 220:115 – 124, 2015.

[YZ16] K. Yan and D. Zhang. Calibration transfer and drift compensation of e-noses via coupled task learning. *Sensors and Actuators B: Chemical*, 225:288 – 297, 2016.

[YZS+18] S. Yao, Y. Zhao, H. Shao, A. Zhang, C. Zhang, S. Li, and T. Abdelzaher. Rdeepsense: Reliable deep mobile computing models with uncertainty estimations. *Proc. of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)*, 1(4):173:1 – 173:26, 2018.

[ZG09] X. Zhu and A. B. Goldberg. Introduction to semi-supervised learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 3(1):1 – 130, 2009.

[ZGY17] D. Zhang, D. Guo, and K. Yan. *A Transfer Learning Approach for Correcting Instrumental Variation and Time-Varying Drift*, pages 137 – 156. Springer Singapore, Singapore, 2017.

[ZL07] Z.-H. Zhou and M. Li. Semisupervised regression with cotraining-style algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 19(11):1479 – 1493, 2007.

[ZLH13] Y. Zheng, F. Liu, and H.-P. 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)*, pages 1436 – 1444. ACM, 2013.

[ZLYX15] Y. Zhuang, F. Lin, E.-H. Yoo, and W. Xu. AirSense: A portable context-sensing device for personal air quality monitoring. In *Proceedings of the 2015 Workshop on Pervasive Wireless Healthcare (MobileHealth)*, pages 17 – 22. ACM, 2015.

[ZPK+18] 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.

[ZTK+11] L. Zhang, F. Tian, C. Kadri, B. Xiao, H. Li, L. Pan, and H. Zhou. On-line sensor calibration transfer among electronic nose instruments for monitoring volatile organic chemicals in indoor air quality. *Sensors and Actuators B: Chemical*, 160(1):899 – 909, 2011.

[ZW09] H. Zhang and J. Wang. Evaluation of peach quality attribute using an electronic nose. In *Sensors and Materials, Vol. 21*, 2009.

[ZZL+12] P. Zhou, Y. Zheng, Z. Li, M. Li, and G. Shen. Iodetector: A generic service for indoor outdoor detection. In *Proceedings of the 10th ACM Conference on Embedded Network Sensor Systems (SenSys)*, pages 113 – 126. ACM, 2012.

List of Publications

The following list includes publications that form the basis of this thesis. The corresponding chapters are indicated in parentheses.

B. Maag, Z. Zhou, L. Thiele. **A Survey on Sensor Calibration in Air Pollution Monitoring Deployments.** *In IEEE Internet of Things Journal (IoTJ), 2018, Vol 5., No. 6.* IEEE, 2018. (Chapter 1)

B. Maag, O. Saukh, D. Hasenfratz, L. Thiele. **Pre-Deployment Testing, Augmentation and Calibration of Cross-Sensitive Sensors .** *In Proceedings of the 13th European Conference on Wireless Sensor Networks (EWSN).* Graz, Austria, 2016. (Chapter 2)

B. Maag, Z. Zhou, O. Saukh, L. Thiele. **SCAN: Multi-Hop Calibration for Mobile Sensor Arrays.** *In Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, Vol.1, No.2 (IMWUT).* Maui, HI, USA. 2017. (Chapter 3)

B. Maag, Z. Zhou, L. Thiele. **W-Air: Enabling Personal Air Pollution Monitoring on Wearables .** *In Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, Vol.2, No.1 (IMWUT).* Singapore, 2017. (Chapter 4)

B. Maag, Z. Zhou, L. Thiele. **Enhancing Multi-Hop Sensor Calibration with Uncertainty Estimates** *In Proceedings of the 16th IEEE International Conference on Ubiquitous Intelligence and Computing (UIC).* Leicester, UK, 2019. ***Best innovation paper.*** (Chapter 5)

The following list includes publications that were written during the PhD studies, yet are not part of this thesis.

O. Saukh, B. Maag. **Quantle: Fair and Honest Presentation Coach in Your Pocket.** *In Proceedings of the IEEE/ACM International Conference on Information Processing in Sensor Networks (IPSN).* Montreal, Canada. 2019.

B. Maag, Z. Zhou, O. Saukh, L. Thiele. **BARTON: Low Power Tongue Movement Sensing with In-ear Barometers.** *In Proceedings of the International Conference on Parallel and Distributed Systems (ICPADS).* Shenzhen, China. 2017. ***Best paper.***

R. Lim R, B. Maag, L. Thiele. **Time-of-Flight Aware Time Synchronization for Wireless Embedded Systems.** *In Proceedings of the 13th European Conference on Wireless Sensor Networks (EWSN).* Graz, Austria. 2016.

R. Lim, B. Maag, B. Dissler, J. Beutel, L. Thiele. **A Testbed for Fine-Grained Tracing of Time Sensitive Behavior in Wireless Sensor Networks.** *In Proceedings of the 40th IEEE Conference on Local Computer Networks, Workshops (SenseApp).* Clearwater, FL, USA. 2015. ***Best paper.***