Improving data quality for low-cost environmental sensors

Xinwei Fang

Doctor of Philosophy

University of York
Computer Science
April 2018

Abstract

Using low-cost sensors to monitor the urban environment has become increasingly popular, as they can provide better data resolution than current practices. However, these low-cost sensors often produce poorer data quality, and so the data may not be utilised directly without processing.

This thesis presents a two-phase solution for improving the data quality of low-cost environmental sensors. The solution consists of a novel method for anomaly detection and removal, and a process of sensor calibration. In the first phase, an anomaly model is utilised to identify the anomalies, which is constructed using a Bayesian-based approach. New contextual information is used to build the anomaly model, that is to the best of our knowledge the first time it has been used for such purpose. The result shows that this solution is more practical and robust than the existing approaches. In the second phase, a systematic comparison of the state-of-the-art calibration approaches is performed. The comparison aims to understand the difference between the methods, and the result shows a regression based method could provide a more predicable result and require much less computational resources. As a result, a regression based method is used for calibrating sensors in this work. In contrast to the existing approaches, the proposed method for calibration is able to systematically and automatically select the calibration parameters. The parameter selection ensures the best set of parameters are used in the model, which makes the calibration process less sensitive to different environmental conditions.

The overall evaluations are performed using real datasets. The results show the data quality in terms of general accuracy against the reference instruments can be significantly improved, especially for sensors at roadside.

Contents

List of Figures

List of Tables

Acknowledgement

First, I would like to thank my parents and my wife, Yuan Zhuang, for their unconditional support and encouragement. Then, I would like to thank Dr. Iain Bate for his supervision and guidance, and Dr. Paul Mitchell for his comments. Finally, I would like to thank my friends for their help and company during my entire Ph.D.

This work was funded by the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no. 608014 (CAPACITIE)

Declaration

I declare that this thesis is a presentation of original work and I am the sole author. This work has not previously been presented for an award at this, or any other, University. All sources are acknowledged as References. Parts of this thesis have been published in or submitted to:

- X. Fang, I. Bate. *Using multi-parameters for calibration of low-cost sensors in urban environment.* **In** 2017 International Conference on Embedded Wireless Systems and Networks (EWSN).

  This paper presents a novel calibration method for low-cost sensors. The materials used in the paper are covered in Chapter 5.

- X. Fang, I. Bate. *Issues of using wireless sensor network to monitor urban air quality.* **In** 2017 International Workshop on the Engineering of Reliable, Robust, and Secure Embedded Wireless Sensing Systems (FAILSAFE).

  This paper discusses the issues that we encountered during the process of the sensor deployments and the data collection, which is covered in Chapter 3.

- X. Fang, I. Bate. *Improving data quality for the low-cost environmental sensors.* **Submitted to** ACM Transactions on Sensor Networks (TOSN) [**Under Review**].

  This paper proposes a two-phase solution to improve the data quality of low-cost sensors. The method consists of a novel method for anomaly detection and removal, and a process of sensor calibration. The materials used in the paper are covered in Chapter 5 and 6.

- X. Fang, I. Bate, D. Griffin. *Are Neural Networks Really the Holy Grail? - A Comparison of Multivariate Calibration for Low-cost Environmental Sensors*. **Submitted to** 2018 ACM Conference on Embedded Networked Sensor Systems (SenSys) [**Under Review**].

  This paper performs a systematic comparison of the state-of-the-art calibration approaches. The materials used in this paper are mainly from Chapter 4. In the paper, on top of my work, David tested a wider range of model parameters for the ANN-based method, and performed statistical tests for the determined results.

Chapter 1

Introduction

Pollution in urban environments has become the largest environmental cause of disease and premature death in the world today [44]. The BBC reported that in 2015, one in six premature deaths globally was related to pollution, two-thirds of which were linked to air pollution [74]. As a result, many studies have been carried out to understand pollution in cities.

1.1 Pollutants in an Urban Environment

Most cities in the world have serious air-quality issues [51,74]. According to WHO [87], the major pollutants in cities are Nitrogen Dioxide (NO_2), ground Ozone (O_3), and the particulate matters (PM_{10}) and ($PM_{2.5}$). PM_{10} and $PM_{2.5}$ stand for small-sized particles, smaller than 10 or 2.5 micrometers in diameter respectively.

According to [43, 51], pollutants in cities are mainly generated by human activities, such as traffic pollution and industry pollution, which could cause various adverse health effects on exposure. The health effects are predominantly respiratory and cardiovascular diseases, which could result in an increasing number of premature deaths. For example, $PM_{2.5}$ is of the greatest health concern to the general public, as they can easily pass through the nose and throat and accumulate deep inside the lungs.

Since pollution is related to human activities, the pollution levels at present are higher in industrialised cities and have clear daily patterns [87]. It is worth pointing out that the increase in NO_2 concentration often comes with a decrease of O_3. This relationship between NO_2 and O_3 is mainly due to the process of chemical reactions and the reaction ratio is related to many factors such as temperature and sunlight.

EU air quality standards clearly state limits on the concentration of a number of pollutants present in the air [25]. Exceeding these limits is likely to cause serious health effects and governments face fines if they fail to meet the annual limits. The limits for NO_2, O_3 and PM_{10} and $PM_{2.5}$ are summarised below.

- $PM_{2.5}$: 25 μg/m^3 annual mean (exceedances are not allowed)
- PM_{10}: 50 μg/m^3 annual mean (exceedances are not allowed)
- O_3: 120 μg/m^3 daily 8-hour mean (exceedances are allowed 25 times over 3 years)
- NO_2: 40 μg/m^3 annual mean (exceedances are not allowed)

It is noted that concentrations of chemicals in ambient air are typically measured in units of the mass of chemical per volume of air. Hence, micrograms per cubic meter (μg/m^3) is a unit that is often used in this context. However, the concentrations of chemicals may also be expressed as parts per million (ppm) or parts per billion (ppb) in other contexts. The relationship between ppb and (μg/m^3) follows a conversion equation as follows:

$$Concentration(\mu g/\mathrm{m}^3) = 0.0409 \times concentration(ppb) \times molecularweight \tag{1.1}$$

1.2 Current Practice of Monitoring and its Issues

The European Union and each government has developed an extensive body of legislation to support the mitigation of pollution in cities. The legislation not only names the pollutants that need to be monitored but also requires them to be monitored at an appropriate spatial and temporal resolution [15, 17, 18, 26]. However, it has been widely reported that the spatial resolution of the current monitoring approach is gradually becoming insufficient [10, 23, 37].

Currently, air quality monitoring in the UK uses: 1) passive samplers, and 2) high-quality electronic sensing stations [17]. These approaches are used for regulatory monitoring purposes and have been widely applied in countries across Europe [26]. Passive samplers such as diffusion tubes are cost-effective devices that enable a large number of devices to be deployed. For example, diffusion tubes are deployed at more than 230 locations in York, UK [12, 16]. However, since the measurements from these devices can only be obtained after manual collection and laboratory analysis, the temporal resolution of the data is significantly limited. For example, diffusion tubes have been used to provide quarterly averaged NO_2 data in York.

By contrast, high-quality electronic sensing stations provide a much higher temporal resolution as they sample the environment automatically and frequently store the measurements in digital format. However, due to the high market price of individual sensors, as well as the maintenance costs involved, it can be financially impractical to construct a dense network using those sensors. According to Table 1.1, the minimum cost of using a reference instrument for a year is more than £100k. As a result, there are only 134 reference instruments in the Automatic Urban and Rural Network (AURN) across the UK [15], only two of which are in the York region [12]. Thus, the spatial resolution of the data is considerably limited using this approach.

The assumption that pollutant concentrations measured by sensors are representative of the entire urban environment is a common practice for pollution assessment [84]. Thus, using data with a limited resolution would have an adverse impact on the accuracy of the assessment [37, 62]. Therefore, a new monitoring approach that is able to provide an improved spatial and temporal resolution would be significantly important [10, 23, 37].

Table 1.1: The estimated cost for using a reference instrument for a year (recommended by DEFRA) [13]

The Task	Estimated Cost
Six-month to one year monitoring survey contracted "all-inclusive" to specialist consultancy.	£10k - £25k
Purchase and installation of single gas-analyser in existing building with power and phone line already available.	£10k – £15k
Purchase and installation of a particulate monitor in an existing building with power and phone line already available.	£10k – £25k
Purchase and installation of multi-pollutant site including PM10 in purpose-built enclosure. Power and phone to be connected, calibration gases to be purchased, data collection software to be purchased.	£50k – £80k
Annual "all-inclusive" service and maintenance costs.	£3-8k per site
Annual data management and QA/QC costs.	£5-10k per site
Annual staff costs for site visits.	£5-10k per site
Annual cost of electricity/phone.	£2-3k per site
Web site commissioning costs.	£3-10k
Annual software and web site maintenance fees.	£1-2k
Annual filter weighing costs for gravimetric PM10 monitoring.	£3-10k per year
Total estimated costs, per site, per year	**£102-198k**

1.3 Low-cost Sensors

It has been widely reported that the use of low-cost sensors can improve spatial and temporal resolution significantly [32, 43, 52]. Therefore, using low-cost sensors is an option if they are sufficiently accurate.

Low-cost sensor units are defined as electronic sensing units that cost several orders less than existing reference instruments. With recent advances in electronics, one or more laboratory functions can be integrated on a single electronic circuit (e.g., Metal Oxide Sensor (MOS) and electrochemical sensor), which makes the cost of sensors considerably lower and they are more compact and easy to use [85]. More importantly, the costs incurred during sensor deployment and maintenance can also be significantly reduced as the use of low-cost sensors does not require infrastructure for their deployment or entail frequent manual handling for maintenance. It is noted that sensors are defined as individual devices that measure physical phenomenon (e.g. NO_2 sensor and O_3 sensor); whereas a sensor unit (e.g. ELM unit, AQ mesh unit) is a system that integrates one or more sensors. We further differentiate sensors and sensor units from reference instruments, and consider monitoring stations that are used for regulatory purposes or that fulfil regulatory standards as reference instruments. However, even though the use of low-cost sensors has many advantages over existing practices, they have not yet been used for regulatory monitoring purposes due to widely reported data quality issues [10, 49, 76, 77].

1.4 The Quality of Data

Data in this work is defined as the measurements of environmental parameters from sensors. The environmental parameters include, but are not limited to, temperature, humidity, nitrogen dioxide (NO_2) and ground ozone (O_3). According with this definition, variation in data pattern (e.g. spikes, variations) is related to: 1) the actual physical phenomenon (caused by environments); and 2) sensing issues (cause by sensors, including communication problems). Figure 1.1 illustrates how variation of the data would be associated with both factors.

The quality of data in this context is often defined by end-users as whether it is sufficient for their purposes. Since requirements from end-users can vary depending on the application, and most users are only interested in the actual physical phenomenon, data quality in this work is considered to be the general accuracy of the data with respect to the ground truth of the environment. Therefore, as illustrated in Figure 1.1, we believe that data quality can be maximised if the data patterns caused by the factor of sensing issues are identified and minimised.

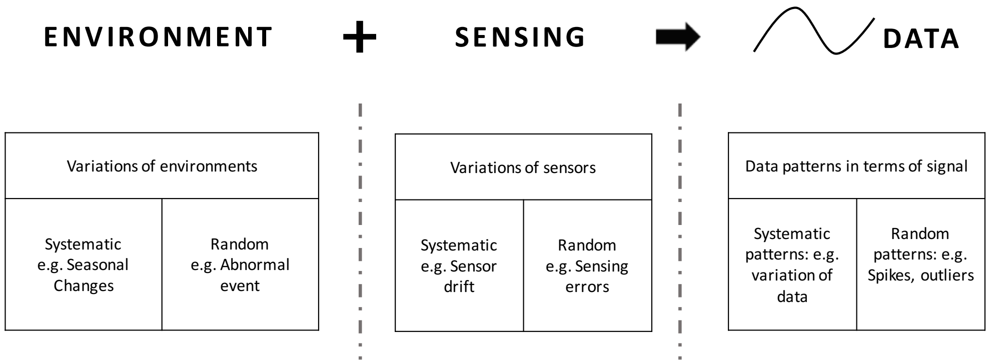

ENVIRONMENT
+
SENSING
DATA
Variations of environments
Systematic e.g. Seasonal Changes
Random e.g. Abnormal event
Variations of sensors
Systematic e.g. Sensor drift
Random e.g. Sensing errors
Data patterns in terms of signal
Systematic patterns: e.g. variation of data
Random patterns: e.g. Spikes, outliers

Figure 1.1: The structure of sensed data

However, identifying the causes of data patterns can be difficult. An example of a sensed data series is given in Figure 1.2, which is O_3 data obtained by a low-cost sensor in a city centre. In the figure, various data patterns can be observed, such as spikes and variations. In many studies, spikes would be considered as anomalies, if the environmental parameters are expected to vary smoothly [6]; but spikes could also be introduced by real but unusual events, such as spikes caused by a bus idling near a sensor. It is noted that anomalies in this work are differentiated from outliers. Anomalies are abnormal measurements caused by sensing issues, whereas outliers are genuine extreme data values. In addition, variation in the sensed data would be intuitively considered to be a daily variation of the environment; however any inconsistent responses of the sensor would also introduce a variation in the data. Since it is difficult to differentiate the causes of data patterns, the ground truth of the environment is important for this work. We assume that the data

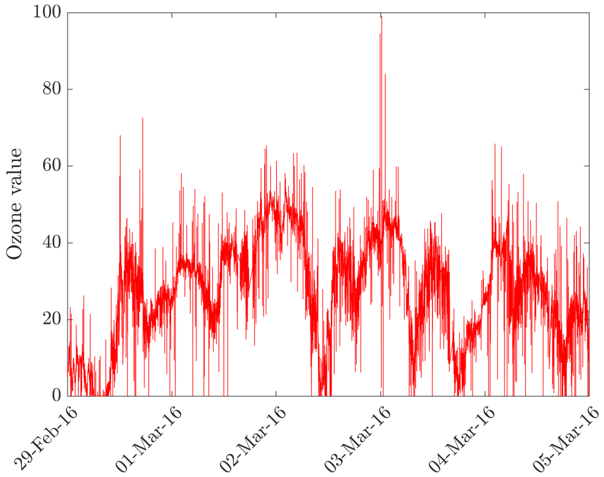


Figure 1.2: Ozone data obtained from a low-cost sensor

pattern is caused by sensing issues if it is inconsistent with the ground truth of the environment.

1.4.1 Ground Truth in the Environment

As reported in [75], it can be challenging to obtain the ground truth of an uncontrolled environment. As a result, assuming the data from reference instruments as the ground truth of the environment is a common practice [28, 49, 67, 76].

A reference instrument often contains multiple analysers, and each analyser only measures the target pollutant. For example, the Tapered Element Oscillating Microbalance (TEOM) analyser is used for monitoring particulate matter and UV absorption is used for monitoring O_3. The exact model and brand of the analysers are not specified in the purchasing guide distributed by DEFRA. However, the analysers used must meet the requirements as described in [13]. A service of the instruments is performed every 6 months by Ricardo, which calibrates all analysers in the instrument using the approved quality assurance and quality control (QA/QC) procedures [1, 14]. The calibration process includes leak tests, analyser reconfiguration and a linearity test, which is based on the standard calibration procedure described in [66].

Since no instrumentation could provide an absolute ground truth in the field, any errors that occurred in the reference measurements would make this work more difficult. The data quality objectives for ambient air quality assessment in [26] allow for uncertainty for $PM_{2.5}$ and PM_{10} of 25% and uncertainty for O_3 and NO_2 of 15%. The uncertainty of the assessment is evaluated in accordance with the standard method (ISO 5725:1994), and it is considered as the maximum deviation of the measured concentration over the period of consideration, without taking into account the timing of the events. Therefore, we expect the data quality of a target pollutant to be no worse than the stated accuracy in [26]. According to [1], the accuracy of the data from reference instruments is not routinely calculated, and the 'best possible' uncertainty for the reference instruments is estimated as ±15% for the measurements of NO_2 and O_3; and less than ±10% for the measurements of particulate matter ($PM_{2.5}$ and PM_{10}) at the annual averaged concentration. Since the reference instruments are used as part of regulatory monitoring and fulfil the EU requirements, we assume that the data quality is sufficient for end-users [49]. In the rest of this thesis, the reference measurements, i.e., the data from reference instruments, is considered to be the ground truth.

1.5 Problem Formulation and Research Questions

In comparison to the reference instruments, the data from low-cost sensor often encounter: 1) lower data accuracy, 2) a higher percentage of outliers, and 3) unexpected data patterns (i.e. constant values) [7,11,32, 47,81]. Admittedly, the calibration of sensors and detection of anomalies are able to alleviate reported data issues as demonstrated in [71, 80, 94]. However, according to [28, 49, 77], the existing methods would not sufficiently compensate for the issues of low-cost sensors, especially when they are in a polluted urban environment. Therefore, an investigation was performed to determine a process capable of enhancing the data quality of low-cost sensors, particularly in a more polluted environment.

1.5.1 Calibration of Sensors

The literature has demonstrated that sensor calibration is able to improve data quality [6, 28, 49]. Sensor calibration is used to determine a transferable model, which minimises the systematic differences between the signal of an uncalibrated sensor and the reference.

Since the response of a low-cost sensor may be related to certain environmental factors, the state-of-the-art method uses multivariate calibration. Unlike the univariate calibration process which only uses parameters of interest to construct the model, the multivariate calibration also uses supporting parameters such as [49, 75, 76]. The intuition of this is, if the response of NO_2 is related to temperature, a more accurate calibration of NO_2 can be determined if the calibration function includes temperature and subtracts its effect. This allows for a more accurate calibration model to be derived, as demonstrated in the literature [19, 20, 23, 24].

1.5.1.1 Selection of the Method

It has been widely reported that data from low-cost sensors are not able to provide sufficient information without proper calibration. According to the literature, regression and artificial neural network (ANN) are two of most widely used approaches for the calibration of low-cost sensors. However, the lack of work on effective comparison of calibration approaches makes it difficult to determine the most appropriate calibration solution. This leads to the first research question:

> *Research Question 1: Which is the appropriate calibration method (Regression or ANN) considering the needs of our application?*

1.5.1.2 How to Use Supporting Parameters

It is known that the response of a sensor can be affected by a lot of influences, which implies that calibration may require different supporting parameters, depending on the actual influence [60, 69]. As a low-cost sensor often monitors multiple parameters, a large number of parameters are available for multivariate calibration. The problem is when an appropriate parameter is available but not used in the calibration, the calibration error may remain large. Whereas, if an inappropriate parameter is used, the result would be negatively affected [30,53,78]. This leads to our second research question:

> *Research Question 2: How can we ensure calibration results by properly using supporting parameters?*

1.5.2 Detection of Anomalies

The detection and removal of anomalies is another well-known process for improving data quality. It is known that some of the anomalies may be associated with a systematic cause. Since we do not have access to the hardware during this study, the root causes of anomalies remains unknown. Therefore, we assume that anomalies in the data may not be compensated by a calibration process, and thus can only be removed.

In this work, anomalies are referred to as abnormal sensor readings (i.e. the sudden change of pollution concentration) caused by sensor issues that are uncorrelated to the underlying physical phenomena. However, as discussed in Section 1.4, anomalies are hard to differentiate from genuine data when the actual physical phenomena is unknown (e.g., a bus idling near the sensor). Intuitively, using data from reference instruments would reveal what actually happened, for example, how it is used in the calibration. However, the data from low-cost sensors often has a much higher temporal resolution (20 seconds) than reference instruments (hourly), which implies that a real short-term increase in value at higher temporal resolution may not be noticeable when looking at hourly data from reference instruments. Furthermore, it is noted that using averaged data would still be inappropriate, as the data after the aggregation would still be inconsistent with the data from reference instruments, as illustrated in [28]. Thus, the data from reference instruments may not be used as the ground truth for the detection of anomalies.

Admittedly, the consistency between data from low-cost sensors and reference instruments would be significantly improved after calibration. However, considering that the calibration process would affect the properties of anomalies (i.e. changes their magnitude), the detection of anomalies after calibration is inappropriate. Most importantly, anomalies in the data could also affect the calibration result. Therefore, we believe that the detection of anomalies before calibration is important. The above issues lead to the third research question:

> *Research Question 3: How can we accurately detect and remove anomalies to further improve data quality?*

1.6 Statement of Hypothesis

Based on the motivations and research questions highlighted in the previous section, the hypothesis of this thesis is formalised as follows:

> *Both regression and ANN-based methods are able to improve data quality for low-cost sensors. However, the regression-based method is more suitable for our application due to lower computational cost, reduced sensitivity to the model parameters used and the need for less training data. The data quality can be enhanced by a calibration process that properly uses the supporting parameters and data quality can be further improved by applying an accurate removal of anomalies before calibration.*

1.7 Organisation of the Thesis

To present this research, the rest of the thesis is summarised in this section.

- Chapter 2: This chapter describes the research background. The use of sensors and their deployments are presented. The data obtained from the deployments is then used to illustrate the properties of the data in terms of variability.

- Chapter 3: This chapter is a review of state-of-the-art work in this area. It covers the calibration of sensors and the detection of anomalies. The limitations of the current methods are also discussed in the review. At the end, we discuss the contributions of the thesis with respect to the limitation of the methods.

- Chapter 4: This is the first technical chapter. This chapter presents a systematic comparison of the calibration techniques. It focuses on determining the difference between two of the most used calibration methods, regression and artificial neural networks. In addition, this chapter also uses multiple sets of training and testing data to determine the sensitivity of each method to these data.

- Chapter 5: This chapter presents a modified regression-based method. In contrast to the existing method, the new method is able to maximise the dependency between input variables and automatically use the appropriate supporting parameters. The evaluation of the method is carried out using data obtained from different environments and the results are compared with the state-of-the-art method used in Chapter 4.

- Chapter 6: This chapter demonstrates the method for the detection of anomalies. New relevant contextual information, cross-sensitive parameters, is used to help identify anomalies. A Bayesian-based method is used to learn the information and construct the anomaly model. The evaluation is performed on both synthetic and real datasets, and the results are also compared with the state-of-the-art method.

- Chapter 7: This chapter concludes the work with a discussion of the main contributions and potential future work.

Chapter 2

Literature Review

This chapter presents a literature review which covers a wide range of studies related to this work. The main purpose of this chapter is to provide a background of the state-of-the-art research and to understand its strengths and limitations with respect to the problems formulated in Section 1.5.

To begin with, we discuss the the trade-offs between the on-line and the off-line process in Section 2.1. After that, the state-of-the-art methods for the calibration of sensors and the detection of anomalies are reviewed in Section 2.2 and 2.3 respectively. Finally, the limitations of the current methods and a set of important points are summarised in Section 2.4.

2.1 On-line and Off-line Process

For sensor related applications, data processing can be generally classified as an on-line or off-line process. An on-line process means that the data is processed on the sensor unit before being transmitted to a server; whereas an off-line process is performed in another place, e.g. a computer, by fetching the data from the server.

On-line processes ensure the data is processed in real time, and reduce communication overheads if anomalies are removed before the transmission. Reducing communication overheads can be extremely beneficial for sensors running on batteries, as the cost of transmission can be several orders higher than the processing [4]. However, the on-line process would permanently change the data. This can be a disadvantage if the process is inadequate, as the process would lead to important information being incorrectly permanently removed from the data. In addition, an on-line process would hinder using information from external sources, which makes it difficult to identify anomalies as reported in [57, 79].

By contrast, off-line processes only work on a copy of the data, and are able to use information from external sources. However, they require all sensor data to be transmitted, which increases the communication costs dramatically. Therefore, using an off-line process is not always the better option.

Existing studies show that the selection of an on-line or off-line process is closely related to the application requirement. Thus, it is important to balance the trade-off between the processes according to the application. For example, the trade-off between the communication costs and data integrity.

2.2 Calibration of Sensors

The calibration of sensors has been extensively studied for many years. In this section, firstly we present a review of the calibration of a single sensor unit. Then, we review state-of-the-art methods for the calibration of multiple sensor units (sensor networks). Finally, we summarise the review of sensor calibration.

2.2.1 Calibration of a Single Sensor Unit

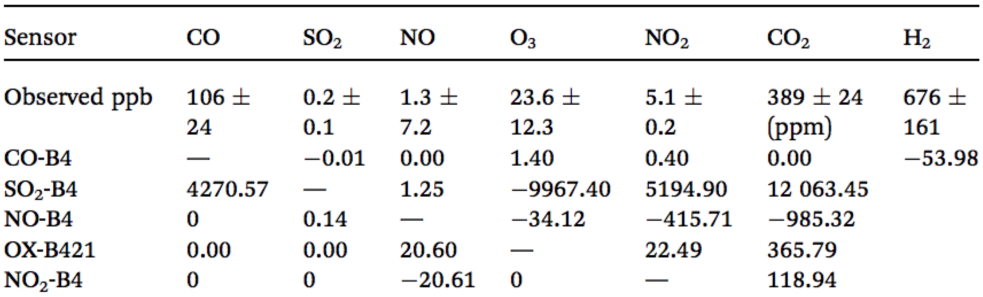
It is known that data from low-cost sensors are widely reported to be insufficient and may not be used without proper processing. Hence, many studies have been conducted trying to identify the potential causes.

An obvious question is what factors or variables would affect the response of low-cost sensors. In order to answer that question, Lewis et al. [46] performed a detailed laboratory-based analytical study on different electrochemical sensors, including O_3 and NO_2 sensors. The main purpose of their work is to determine whether co-pollutants would affect the response of the sensor.

In their first experiment, a controlled concentration of a particular gas was injected into clear air. They report that no abnormal sensors responses were observed. In the second experiment, the sensors were tested in ambient air. The air was mixed with a controlled percentage of different gases including O_3, NO, NO_2, SO_2, CO, H_2 and CO_2. The authors conclude that interference from co-pollutants in the response of sensors can be significant, and the degree of the interference depended on the ratio of co-pollutants. The results of the interference are shown in Figure 2.1. Even though they did not further investigate how different percentages of co-pollutants and various mixtures of the air would affect the response of the sensors, their findings are still important as they explain why low-cost sensors often behave unexpectedly in a real environment.

Finally, the sensors were exposed in the field (a real environment) and their response was evaluated against a reference instrument. The sensor data were linearly calibrated, and the results show the O_3 sensor has a good correlation ($R^2 = 0.9$) with the reference as shown in Figure 2.2; whereas the NO_2 sensor has poor statistical agreement ($R^2 = 0.25$) with the reference as shown in Figure 2.3. Lewis et al. believe that the NO_2 sensor was not measuring the target compound exclusively due to interference from co-pollutants. Moreover, the NO_2 sensor generally reported a significantly higher concentration than the reference, which considerably exceeded air quality standards (200 μg/m^3 1-hour mean). Their results suggest that NO_2 sensors would be more difficult to compensate and require a more comprehensive evaluation than O_3 sensors.

Similarly, Castell et al. [10] deployed 24 identical units of low-cost sensors in the field to evaluate how the data quality of the sensors compared to the reference instruments. Instead of testing sensors in just one location, as in [46], the sensors were deployed at different locations for 3 months. Thus, the spatial variation of the data could be obtained in their work. The sensors monitored multiple parameters including NO_2 and O_3. The sensors were firstly tested in a laboratory before the deploy-



Sensor	CO	SO_2	NO	O_3	NO_2	CO_2	H_2
Observed ppb	106 ± 24	0.2 ± 0.1	1.3 ± 7.2	23.6 ± 12.3	5.1 ± 0.2	389 ± 24 (ppm)	676 ± 161
CO-B4	—	−0.01	0.00	1.40	0.40	0.00	−53.98
SO_2-B4	4270.57	—	1.25	−9967.40	5194.90	12 063.45	
NO-B4	0	0.14	—	−34.12	−415.71	−985.32	
OX-B421	0.00	0.00	20.60	—	22.49	365.79	
NO_2-B4	0	0	−20.61	0	—	118.94	

Figure 2.1: First row: observed mean ambient pollution mixing ratio and one sigma range over 18 days. Subsequent rows show the impact of the signal induced by a co-pollutant expressed as a percentage of the mean ambient mixing ratio of the measurand (quoted from [46])

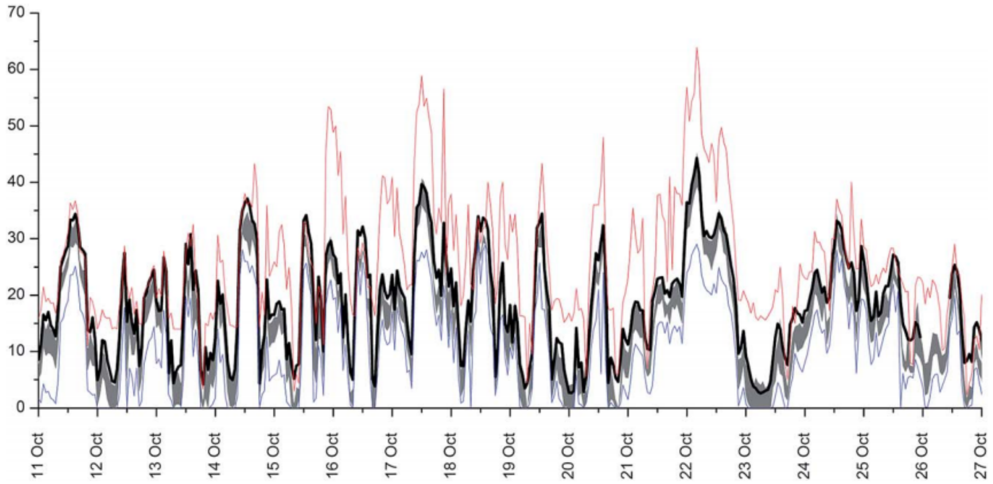


Figure 2.2: A time-series comparison of reference photometric O_3 instrument (black line), highest O_3 sensor (red line), and lowest O_3 sensor (blue line). Grey shaded area shows those sensors lying in the 25th to 75th percentile range. (quoted from [46])

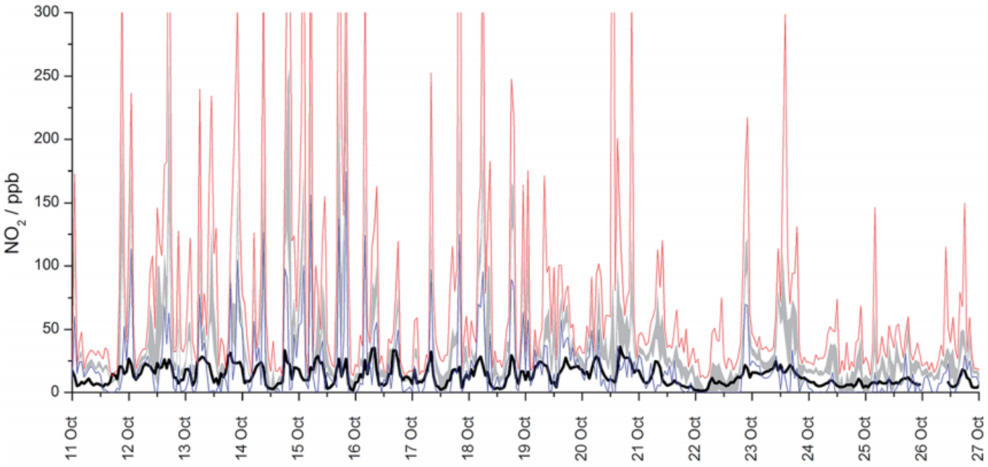


Figure 2.3: A time-series comparison of reference chemiluminescence NO_2 instrument (black line), highest NO_2 sensor (red line), and lowest NO_2 sensor (blue line). Grey shaded area shows those NO_2 sensors lying in the 25th to 75th percentile range. (quoted from [46])

ment, the test results show that the sensors had a linear response when only the target gas was injected into the clean air; but the response of the sensors became hard to predict when the air was mixed with multiple gases.

The results are in-line with findings reported in [46], which suggests that low-cost sensors are generally sensitive to co-pollutants. The data quality was determined and compared against the reference instruments. Their results show that the performance of sensors varies both spatially and temporally, and it also varies from sensor to sensor. The results also show that the variation in sensor performance is related to environmental variables, such as meteorological conditions and different air composition. Therefore, the authors conclude that all sensors may need to be evaluated individually and evaluation in the environment of operation is necessary.

Mueller et al. [54] further investigate the performance of low-cost sensors (i.e. NO_2 and O_3) in a real environment. In contrast to the study in [10, 46] where the environmental conditions of deployment were relatively consistent, the experiment in this work covered a wide range of environmental conditions, including urban roadside and urban background. The units were initially operated in the different environmental conditions for 3 months, for performance analysis and initial calibration. They were then relocated to different locations. The authors report that multivariate regression provides a good calibration result for NO_2 in a harsh environment, as shown in Figure 2.4. However, the result of calibration would drop significantly after relocation. This is largely related to the change in humidity between different locations, as reported in the paper.

Mueller et al. further identify using multivariate calibration as important for compensating for the data quality of low-cost sensors, especially for NO_2 sensors. However, they do not specify what supporting parameters are important to use in the model, but they suggest that the use of the parameters would be sensitive to different environmental conditions. Their results indicate that the calibration function may need to be adjusted ech time the environmental conditions change, which implies that the calibration of low-cost sensors in an urban environment may need to be frequently applied.

Studies in [10,46,54] shows consistent findings in the response of low-cost sensors, which are:

- The response of low-cost sensors is sensitive to their co-pollutants. Thus, co-pollutants should be considered in the calibration process.
- The calibration of an NO_2 sensor is more difficult than an O_3 sensor.
- The response of low-cost sensors varies in different locations. Thus, the assumption that calibration can hold when sensors are in different locations is not appropriate.
- Calibration may not hold over time due to changes in environmental conditions.

Since the response of low-cost sensors is sensitive to environmental conditions (e.g. co-pollutants, meteorological conditions), univariate calibrations which neglect those effects may be insufficient to compensate the data of low-cost sensors, especially for NO_2 sensors, as identified

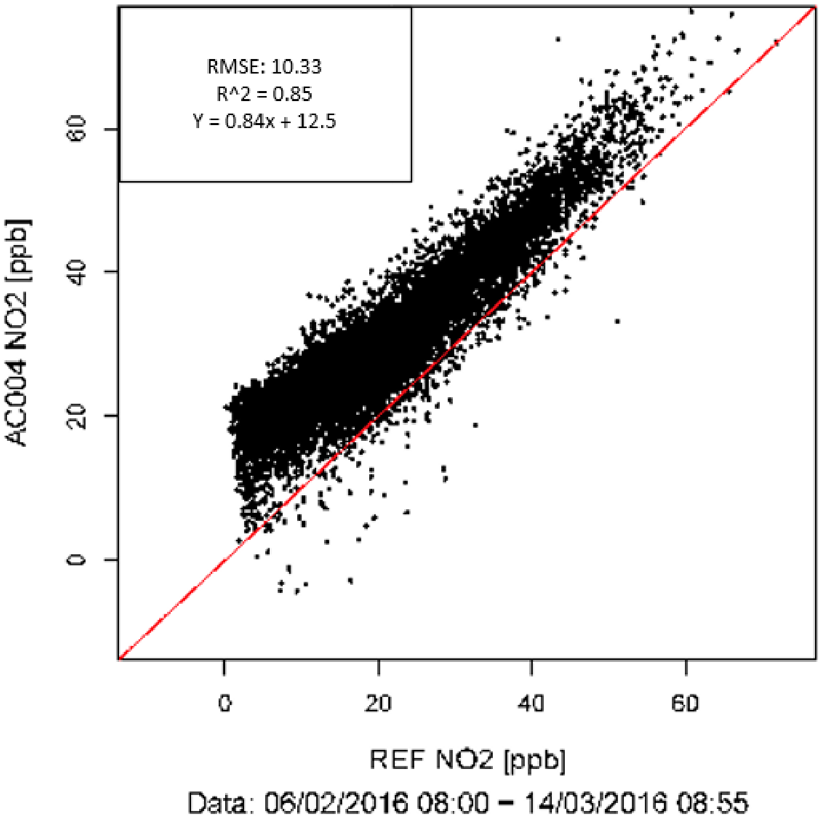


Figure 2.4: The result of calibration with respect to reference (quoted from [54])

in [23]. As a result, multivariate calibrations that use certain supporting parameters have become increasingly popular, and they are currently the most efficient and effective approach for the calibration of low-cost sensors.

In contrast to univariate calibration, multivariate calibration uses not only the parameter of interest (e.g. the target gas) but also other supporting parameters or useful parameters, such as co-pollutants or temperature and humidity. It is clear that the response of sensors could be strongly related to those effects. Hence, by including certain parameters, a calibration model would be expected to extract information from those parameters and subtract the influence from them. This is one reason why using multivariate calibration is believed to provide a better sensor calibration. According to [20], a multivariate calibration is generalised as the determination of an approximation Ψ in:

$$Ref_t = \Psi(Parameter(1)_t, Parameter(2)_t, ..., Parameter(n)_t) \qquad (2.1)$$

where Ref_t is the pollutant concentration measured by a reference instrument at time t, and $parameter(1)_t$ to $parameter(n)_t$ are the uncalibrated parameter of interest and other supporting parameters measured at the time t.

Devito et al. [20] performed an multivariate calibration of NO_2 using an ANN-based method. Their sensors monitored a list of parameters that were all used as supporting parameters in the calibration, which included CO, NO_x, O_3, temperature (T), and relative humidity (H). In their evaluation, the various combinations of the supporting parameters were tested and the calibration results were determined by the Mean Absolute Error (MAE) between the model output and the reference. It is noted that in their results, using an increasing number of supporting parameters in the calibration did not always lead to a better calibration result. This indicates that the optimal calibration would not be simply using all available parameters, which implies that selection of the supporting parameters would be needed.

A dynamic calibration of low-cost sensors has been proposed based on an ANN-based method in [23]. Dynamic calibration means the datasets used for training and testing were obtained at an ultra high temporal resolution, e.g. at the minute level. This reflects the rapid changes in concentrations in a real environment. According to the authors, this is the first dynamic calibration that has been performed in a real environment as the evaluation was limited by 1) the high temporal resolution reference data and 2) the inconsistent environmental conditions (both spatially and temporally) that results in the tested sensor and the reference not sampling the same phenomena. It is noted that unlike in a regression-based method, which requires the raw data to be averaged to a desired temporal resolution in advance, ANN-based methods use data with higher temporal resolution directly in the calibration. The evaluation shows using the raw data directly in the ANN calibration would consistently obtain a better result than the one using the averaged data. However, the authors did not specify how the network would deal with the inconsistent number of inputs caused by data gaps.

Spinelle et al. [77] compared a number of calibration methods for calibrating NO_2 and O_3 in the field. The methods include univariate linear regression, multivariate linear regression (MLR) and ANN. The results were evaluated according to root mean squared error (RMSE) only, which are illustrated in Figure 2.5. Their results show O_3 sensors are relatively easy to calibrate, as they can achieve a high correlation with the reference using just univariate linear regression. By contrast, the NO_2 sensors are more difficult to calibrate and they would require a more complex calibration, such as MLR and ANN. The results confirm the findings in [10, 46, 54]. Furthermore, their evaluation also indicates that using an ANN would obtain a better calibration result than using a MLR. However, it is noted that their calibration was carried out in a suburban environment, where the environmental conditions could be considerably different from a typical urban environment. Thus, an ANN-based method may not always be the best choice for the calibration of sensors.

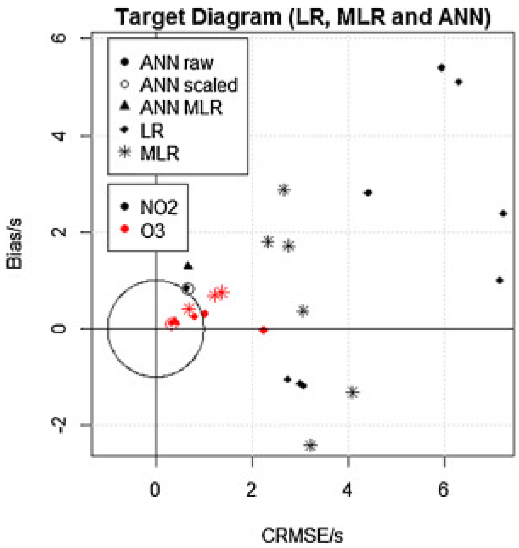


Figure 2.5: The result of calibration from multiple methods; centred root mean square error (CRMSE)(quoted from [77])

Maag et al. [49] also performed a multivariate calibration for low-cost sensors in the field. Their results suggest that the cross-sensitive parameter is the most important parameter for the calibration of NO_2. Cross-sensitivity is defined as sensitivity to one substance which renders the sensors sensitive to other substances. It is known that NO_2 is cross-sensitive to O_3; hence O_3 is also referred to as the cross-sensitive parameter of NO_2. This means that the readings from an NO_2 sensor would be dependent on the concentration of O_3 in the mixed air. Given an intuitive example, we assume an NO_2 sensor has a response to O_3 at a rate of 50% (this value can be both positive and negative, and would vary in different conditions) due to cross-sensitivity. Then, if the NO_2 sensor is exposed to 200ppm of O_3 only, the NO_2 sensor will report 50% of 200ppm. However, if the NO_2 sensor is exposed to 100ppm NO_2 and 200ppm O_3, the NO_2 sensor would provide readings of 100ppm + 50% 200ppm.

For this experiment, the monitored parameters are identical to the ones in [20]. However, it is noted that the supporting parameters used for constructing the final calibration model are different. This suggests that the use of the supporting parameter is not only dependent on the availability of the parameters, but also related to other factors, e.g. current environmental conditions. Their results emphasise the importance of selecting the supporting parameters and shows the significance of using cross-sensitive parameters in sensor calibration.

The existing studies demonstrate that multivariate calibrations are able to better calibrate low-cost sensors. The lessons learnt from those studies are:

- The existing studies often ignore the incompleteness or inconsistency of the data in their process.
- A larger number of supporting parameters used in the calibration does not always lead to a better calibration result.
- Temperature and humidity, as well as cross-sensitive parameters, are often used in multivariate calibration, and they have been reported to be useful in many applications.

- The use of supporting parameters is sensitive to environmental conditions. This shows the importance of selecting supporting parameters, especially for sensors in different locations.

The review suggests that using multivariate calibration is important to calibrate low-cost sensors, and the ANN-based method and the regression-based method are the most applied approaches. However, it is difficult to select an appropriate calibration method for an application, as the difference between the methods has not yet been comprehensively studied. Intuitively, a regression-based method can be easily applied and interpreted [49], but it may not be suitable for calibrations that have a complex relationship between input variables and the output [77]. By contrast, an ANN-based method would be able to solve such a problem with more complicated training and a higher computational cost. However, to the best of our knowledge, little work has demonstrated a systematic comparison of the approaches, which hinders understanding the difference between the methods.

One prominent existing comparison, [77], is limited to comparing the calibration results in terms of calibration accuracy, which is often represented as the averaged error between the model predictions and the reference, e.g. root-mean-squared error (RMSE) or mean-absolute error (MAE). Since two identical averaged errors may represent different error distributions, using the calibration accuracy as the only metric for the comparison would not be sufficient and would not give a deep understanding of their differences.

To solve that issue, Esposito et al. [24] and Devito et al. [19] provide a more detailed comparison for multivariate calibration approaches. In their work, the approaches are cross-compared not only for calibration accuracy (determined by the mean absolute error), but also for the capability to deal with different training scenarios. In [24], calibration results were compared by varying a different number of training and testing samples. However, it is noted that the variation of the training and testing samples was divided by a cut-off value. Hence, results are impacted by changes in both the training and testing samples, which makes it difficult to determine which changes are responsible for the variation in the output.

Devito et al. [19] analysed how the calibration accuracy was affected by using different model parameters. For example, they compare calibration accuracy by varying certain model parameters in the ANN network. In that case, they would have to assume that the model parameters of the ANN are independent or partially dependent. However, according to our analysis, this assumption is not valid as demonstrated in Section 4. Therefore, we believe that the existing comparisons are less effective and not able to find the best calibration approach for the needs.

2.2.2 Calibration of Sensor Networks

Calibration of sensor networks are different from the calibration of a single sensor unit. A number of surveys categorise the methods for the calibration of sensor networks into micro-calibration and macro-calibration, e.g., in [28, 82, 86]. Macro-calibration calibrates a network by using the consistency of the nearby environment and maximises the similarity of measurements from neighbouring sensors [6, 9, 29, 48]. Thus, no reference sensor would be required which significantly reduces the cost of calibration.

Balzano et al. [6] propose a blind calibration, which requires neither controlled environmental conditions nor high-fidelity reference sensors. The method in theory is able to automatically calibrate a group of sensors (i.e., sensor networks) in the field. The approach assumes the target gas would vary smoothly in the field (spatially), and the signal of the target gas is bandlimited (i.e. the signal can be sampled by a limited number of sensors). By having that assumption, based on the Nyquist theorem, the sensor network can over-sample the target gas and reconstruct its spatial distribution if the deployment of the sensors (spatial distance) is at least two times higher than the spatial variation of the signal. The reconstructed signal would be used as the ‘reference’ for the calibration, and each sensor then adjusts the gain and offset to minimise the difference to the ‘reference’. However, the method would require that every sensor in the network can be compensated by an univariate linear model (i.e., the model has only two unknowns, gain and off-set). It is clear that such the assumptions are often invalid in real practice as univariate calibration is often hard when compensating low-cost sensors, as reviewed in Section 2.2.1

Lipor et al. [48] improved the work in [6] by using total least squares estimation. The advantage of which is reducing the errors which occurred during the estimation of the 'reference'. The simulated result compares four methods, i.e. Least Square (LS), Partially-blind Least Square (PB-LS), Singular Value Decomposition (SVD) and Partially-blind Total Least Square (PB-TLS), in Figure 2.6. The figure indicates that the proposed method (PB-TLS) in theory outperforms the method (LS) in [6]. In the figure, the dashed line indicates the error of sensors without calibration, the subspace error is the error that occurs during the estimating of the 'reference'. However, even though this method is considerably more robust than the method in [6], the method still relies on the same assumptions as in [6], which are often invalid in practice.

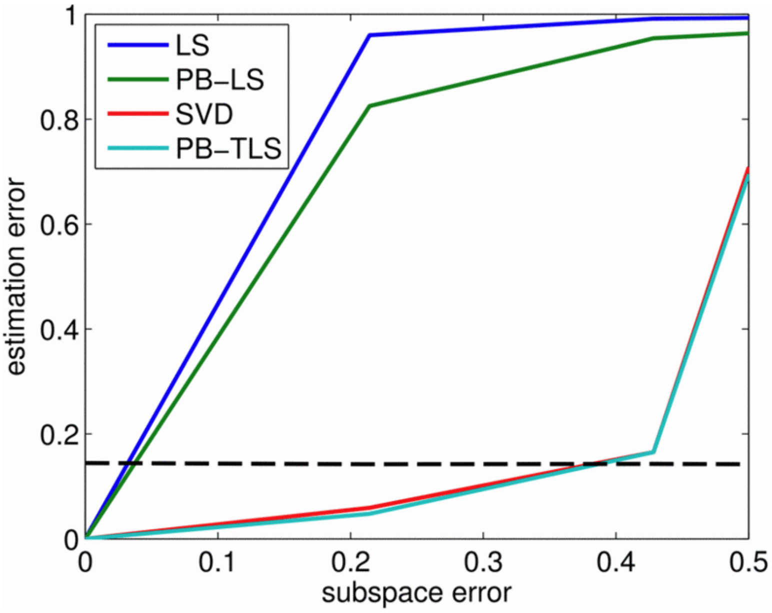


Figure 2.6: Error in gain estimation as a function of subspace error (quoted from [48])

In [9], Bychkovskiy et al. present a two-phase method for calibrating a group of sensors. In the first phase, all sensors are required to be co-located in a place to obtain a relative calibration, which determines the relationship between pairs of co-located sensors. In the second phase, the method optimises the determined calibration by maximising the consistency of the sensor measurements in the environment of the deployment. An obvious issue with this method is if the environmental conditions are different between the two phases of deployment, the calibration function obtained in the first phase may not provide a good calibration result in the second phase. According to Section 2.2.1, it may not be appropriate to assume the environmental conditions between the two phases of deployment are consistent. Thus, we believe that this method may not be suitable for applications in an urban environment.

The literature suggests that in theory macro-calibration could calibrate a large sensor network at a relatively low cost, as it requires neither the references nor on-site manual handling. However, such calibration often demands significant assumptions, and some of the assumptions are inappropriate for applications in an urban environment, e.g. assuming environmental conditions are spatially consistent. In addition, since macro-calibration calibrates a sensor using the outputs from a non-reference sensor or model, the determined calibration is only relative, and could be significantly different from the true value. Due to the identified limitations, macro-calibration is not ideal for real applications, especially in the urban environment.

Micro-calibration, on the other hand, is different from macro-calibration as it uses reference instruments or freshly calibrated low-cost sensors as a reference. Hence, the obtained calibration is more reliable and it does not require any assumptions about the behaviour of the sensors or the conditions of the environment [59, 67]. However, it can be impractical to co-locate a reference next to every low-cost sensor in the network due to cost and practical issues. Therefore, using freshly calibrated mobile sensors as a reference is commonly used in practice [35, 88, 89].

Saukh et al. [68] propose the idea of using *rendezvous* to calibrate sensor networks. Rendezvous is considered as the vicinity when a reference (freshly calibrated low-cost sensors or reference instruments) and an uncalibrated low-cost sensor are close in space and time. As a result, a rendezvous can be used to propagate the calibration function as sensors at the rendezvous are considered to monitor the same physical phenomena. In this context, the calibration is classified as a single-hop calibration if a low-cost sensor is calibrated directly from a reference instrument; whereas a multi-hop calibration means that a sensor is calibrated by the freshly calibrated low-cost sensors and rendezvous (e.g. propagated via multiple hops from the reference).

Hasenfratz et al. [35] present a multi-hop calibration to calibrate O_3 sensors using mobile sensors. The work uses simulations to emulate a scenario where O_3 sensors were placed on mobile platforms, such as a bus or tram. The constraints were 1) the platforms would pass a reference sensor every 40 minutes for rendezvous; and 2) the calibration would be propagated by a number of simulated low-cost sensors. Their simulated results suggests the accuracy of the calibration would decrease with an increasing number of hops. It indicates that the error propagation could be an issue for such a method. The other criticisms of this study are: 1) the results are based on the simulation, and therefore may not reflect a real case scenario; 2) the method may not be applicable to parameters that require a multivariate calibration, such as NO_2.

Maag et al. [50] offer a constrained least-square method for a multi-hop calibration, denoted as sensor array network calibration (SCAN). Their method was evaluated using both artificial data and real data. The result illustrated in Figure 2.7 shows a significant improvement in reducing error propagation over a number of hops in comparison to multiple least regression (MLR) and geometric mean regression (GMR). Thus, this method would alleviate the error propagation issue encountered in [35]. However, their work is only demonstrated for calibration parameters that can be compensated by a univariate calibration.

Other studies on the multi-hop calibration have taken a different research focus. For example, the authors in [31] consider multi-hop calibration as an optimisation problem to determine the best travel route for mobile platforms and in [89] to determine the optimal rendezvous to improve the calibration accuracy.

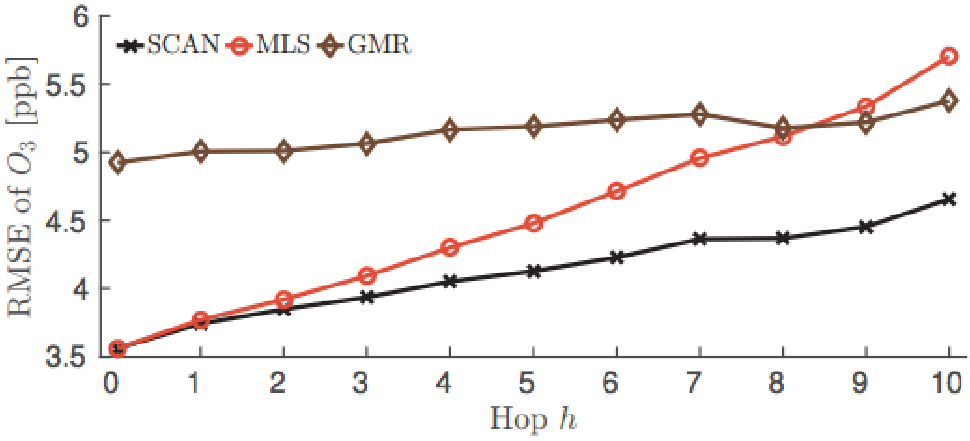


Figure 2.7: The calibration error in term of overall RMSE over a number of hops (quoted from [50])

The review illustrates that current studies on the calibration of a sensor network are mainly based on simulations and may not calibrate a parameter that would require multivariate calibration, such as NO_2. It is understood that real deployments and datasets are important for studying the calibration of sensor networks. However, as reported in [27], to deploy sensors in desired locations and to obtain the required data can be difficult in practice, which is the main barrier for such a study.

According to the review, the issues for the calibration of sensor networks are summarised as:

- A lot of studies rely on simulations due to the lack of real sensed data.

- The propagation of calibration errors is still an open challenge for the calibration of sensor networks.

- A lot of studies require assumptions which often do not hold in practice.

2.3 Detection of Anomalies

The detection of anomalies has been an active research area in many domains including fraud detection, image processing and sensor networks [39, 94, 95]. This review only focuses on anomaly detection for sensors and sensor networks.

In many existing studies, outliers and anomalies are used interchangeably. However, in this study, we differentiate them. We consider outliers to be the extreme values only. Thus, outliers are not necessarily anomalies, as extreme values can also be real measurements (e.g. a bus idling next to a sensor). By contrast, anomalies are defined as abnormal data caused by sensing issues. More specifically, we consider anomalies as a sudden change in the signal which is uncorrelated to the underlying physical phenomena. Thus, methods that only detect outliers would not be sufficient for this study and an ideal method is required to further separate anomalies from outliers.

A straightforward solution to detect anomalies is to determine a profile of the normal data or the anomalies, and use the profile to differentiate one from another. Such a profile is referred to as an anomaly model in this thesis. However, as discussed in Section 1.4, the underlying physical phenomena is not always available, which makes an accurate anomaly model difficult to obtain. Common practice is to use extra information to approximate physical phenomena (e.g. using information from other sensors). However, the existing solutions may not be directly applicable in this work. In this section, we review a set of methods that have been widely used for determining anomaly models. According to [64,94], these methods can be broadly classified as 1) statistical-based methods, 2) nearest-neighbour-based methods, 3) cluster-based methods, and 4) classification-based methods.

2.3.1 Statistical-based Methods

Statistical-based methods are the simplest method of the four types of method identified in [94]. They differentiate anomalies by building a statistical distribution that represents normal data. These methods can be further divided into parametric and non-parametric approaches. For the parametric approaches, the distribution parameters can be easily obtained if the data distribution to be used is known; whereas, determining the distribution parameters for non-parametric approaches can be difficult.

Palpanas et al. [57] propose a non-parametric on-line method which uses a kernel density estimator to determine the profile of the normal data. The method in theory can detect outliers in streaming data. However, the method only identifies outliers and is not able to further separate anomalies. More importantly, the proposed method was only discussed theoretically and not evaluated experimentally.

Subramaniam et al. [79] extended the work in [57] by performing an experiment on a synthetic dataset, and the authors also introduced a sliding window to up-date the profile of normality. The result shows that the method can obtain an accurate detection result in streamed data. However, since the method is also performed on-line, there is only limited information that can be used (as discussed in Section 2.1). Hence, it would not solve the issue encountered in [57], i.e. the method only detects outliers not anomalies.

Sheng et al. [73] use an off-line method to detect abnormal values in sensed data. The key idea of their work is instead of transmitting all sensed data to the sink for analysis, which would be extremely costly, they use a histogram to extract information from the data and only transmit the histogram back to the sink for the analysis. The simulation results show that the method is able to detect outliers, with communication costs being dramatically saved. Since the method is an off-line process, it enables the use of information from other sensors. However, the abstracted histogram contains too little information to help further separate anomalies from outliers.

Zhang et al. [92] use off-line spatial and temporal correlations respectively to differentiate anomalies from outliers. Temporal correlation has three steps. Firstly, the difference between any two consecutive data points, $x(s,t)$ and $x(s,t-1)$, is calculated, resulting in a new time series $\{x^{'}(s,t) = x(s,t) - x(s,t-1)\}$. It is noted that x is the dataset, s and t

indicate the parameter s taken at time t. The new time series indicates the change in the temporal property. It is noted that the authors ignore the natural variation of the data, e.g. daily and seasonal patterns, by taking data for a short period of time (i.e. data is only few hours long). In the second step, an auto-regressive moving average model is fitted, which is defined as $AR(p)$. $AR()$ stands for the model and p is the number of historical observations used in the model. As a result, the model $AR(p)$ suggests that the current observation is only correlated with the previous p observations. It is noted that a larger p would not always lead to a better result due to the variation of data. Hence, the current observation can be modelled using the previous observation according to Equation 2.2

$$\hat{x'}(s,t) = \sum_{i=1}^{p} \alpha_i x^{'}(s,t-i) + \epsilon_t \tag{2.2}$$

where $\alpha_i = \alpha_i : i = 1, 2, ..., p$ are model parameters and ϵ_t is white noise. In the third step, the anomalies are identified by comparing the current observation with the predicted one. If the difference is above a threshold value, the current observation will be considered to be an anomaly. However, such a method would require a complete time series, meaning no data gaps or missing values in the data are allowed. Otherwise, the result would be significantly compromised. The use of spatial correlation is a similar idea to using temporal correlation. The key difference being the predicted value $\hat{x'}(s,t)$ is derived from the spatial domain (i.e. neighbouring sensors) instead of the time domain. However, using the spatial correlation would require the neighbouring sensors reporting consistent measurements. The detection results using temporal correlation and spatial correlation are illustrated in Figure 2.8 and 2.9 respectively. It is noted that TOD and SOD in the figures stand for temporal outliers detected and spatial outliers detected. The authors conclude that using spatial correlations would result in communication overheads but lead to a better result on the detection accuracy. By contrast, the temporal correlations do not require data from outside but the results would be much less accurate.

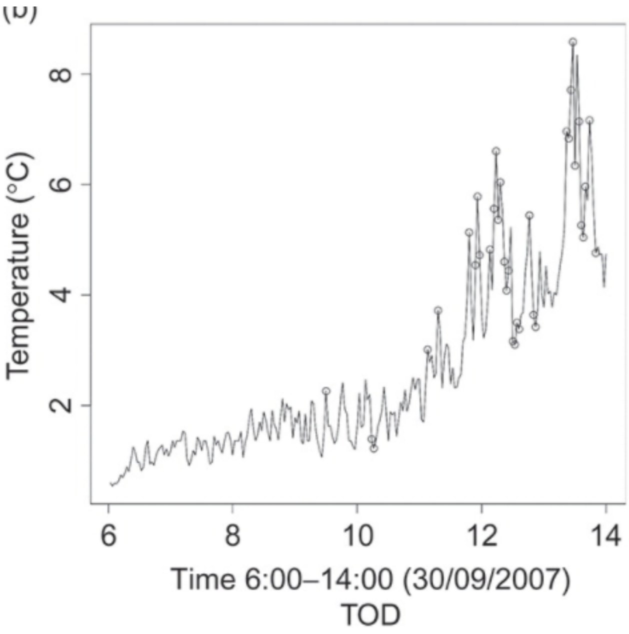


Figure 2.8: The result of anomaly detection using temporal correlation, the circle indicates the detected anomalies (quoted from [92])

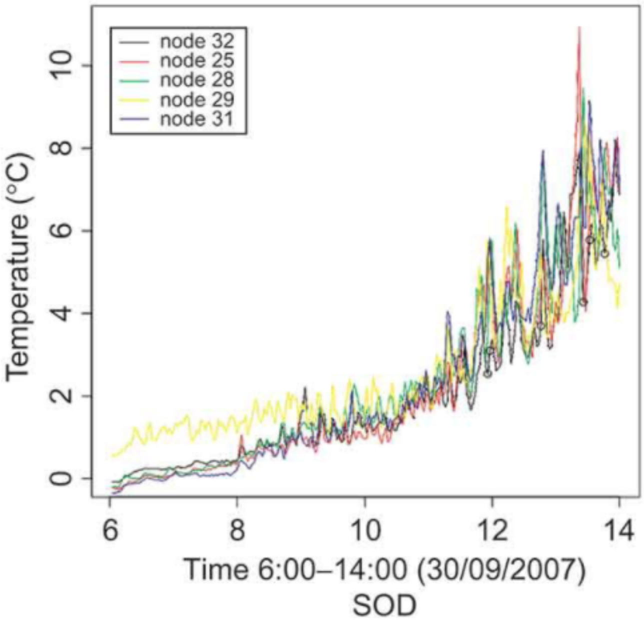


Figure 2.9: The result of anomaly detection using spatial correlation, the circle indicates the detected anomalies (quoted from [92])

In this section, we reviewed statistical-based methods. We summarise the limitations of such methods as follows:

- A lot of methods can only detect outliers, especially in the on-line processes.
- It would require extra contextual information, i.e. spatial and temporal correlations to further separate anomalies from outliers.
- The threshold value is often application dependent and determining a proper threshold value remains an open challenge.

2.3.2 Nearest-Neighbour-based Methods

The nearest-neighbour-based method assumes that normal data patterns would be found in a dense neighbourhood and abnormal data are far from this. In contrast to statistical-based method that the profile of normality is often determined by fitting a data distribution, the nearest neighbour based method require the data to be intensively processed to determine similarity measures. The similarity measure indicates the degree of a data point being normal or abnormal, e.g. a data point would be considered as an anomaly if its Euclidean distance to a dense neighbourhood in a certain feature space is below a given threshold.

Zhuang et al. [96] proposed a method for in-network (on-line) outlier cleaning for data collection. In their method, the data is transformed in the time-frequency domain. Then, the similarity measure of a data point is determined based on Dynamic Time Warping (DTW) distance within that domain. Abnormality is identified if the similarity measure is above a pre-defined threshold. An obvious drawback of this method is the result would be highly dependent on a pre-defined threshold. However, the authors did not analyse the relationship between the increasing threshold and the result. Thus, the threshold value may not be obvious to define, as the trade-off is not clear. Furthermore, it is noted that since this method is processed on-line, this method would face the same issues that are discussed in Section 2.1.

Other studies also use the nearest-neighbour based method to detect abnormal values [8,90,91]. However, their main focus is not on detection accuracy but on the balancing of trade-offs, such as the trade-off in using different network topologies (hierarchy vs flat) or the energy consumed in transmission and computation.

In summary, the limitations of the methods are:

- It is a pointwise process, thus, it can be computationally expensive.
- Similar to statistical-based methods, it can be difficult to separate anomalies from outliers if no extra contextual information is available.
- The selection of a proper threshold value is important in such a method.

2.3.3 Cluster-based Methods

The cluster-based method groups data with similar patterns or characteristics into clusters and identifies abnormal values according to their similarity measure, e.g. the Euclidean distance between a data point and a cluster, or between clusters. In contrast to a pointwise process such as the nearest-neighbour-based method, the cluster-based method can also label small clusters as abnormal. As a result, it would require less computational resources for a larger dataset than the nearest-neighbour-based method. However, the following limitations still exist:

- Even though its computational cost is less than the nearest-neighbour-based method, it is still computationally expensive.
- The detection result is highly dependent on the choice of cluster, which make the method extremely sensitive to the data and the model parameters.

2.3.4 Classification-based Methods

The classification-based method firstly determines a classifier (i.e. a model of the anomaly) using a training dataset and uses the determined classifier to classify normal and abnormal data into a different dataset. The classification-based method can be further divided into supervised learning and unsupervised learning, with the main difference being that the supervised learning requires a specified output in the training phrase, e.g. labelling data as anomalies and non-anomalies. It is clear that labelling anomalies in a real dataset can be difficult. Thus, supervised learning, e.g. using an artificial neural network (ANN), is rarely used in anomaly detection in WSN according to [64].

Elnahrawy et al. [22] present context-aware sensors by using a Naive Bayesian network to detect anomalies and predict missing values. The assumption is that sensors in the network would provide data that is both spatially and temporally correlated. A Naive Bayesian network is then employed to learn such correlations by calculating the joint probabilities of the current reading between 1) the current readings from neighbouring sensors (spatial correlation), and 2) its previous readings (temporal correlation). As a result, the current reading can be predicted with a confidence level by its previous readings and the current readings from its neighbouring sensors. If the confidence level of the predication is above a threshold, then the predicted reading will be used as reference to identify anomalies or to fill the data gap if missing values occur. However, it is noted that if the assumption is invalid (i.e. data are not spatial and temporal consistent), the method would not obtain a sufficient result.

Janakiram et al. [40] propose a method that not only uses spatial and temporal correlations, like [22], but also explores the dependencies from the observation of sensor attributes. A Bayesian Belief Network is used to model the temperature value. Apart from the spatial and temporal correlations, attributes like relative humidity, barometric pressure, light intensity and mote voltage are also used. Their evaluation shows that detection results benefit from including the attribute dependencies, but the improvement is not significant. Therefore, the authors conclude that the selected attribute may not be used for detection of anomalies solely, and the method would not be applicable if the spatial and temporal correlation is weak.

According to [94], apart from Bayesian-based methods, support vec-

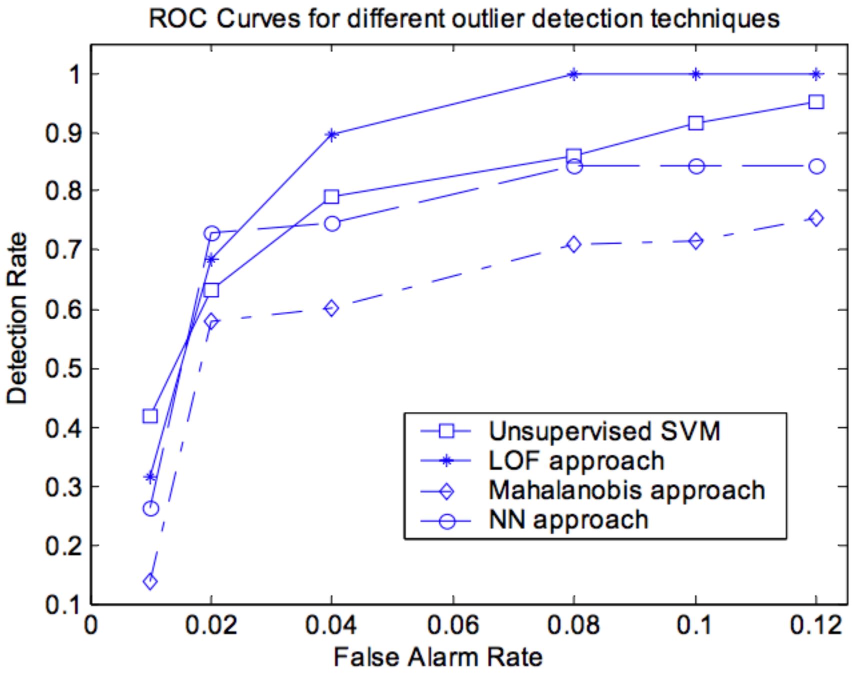


Figure 2.10: ROC curves showing the performance of anomaly detection algorithms (quoted from [45])

tor machine (SVM) based methods are also a widely used unsupervised method for detection of abnormal values in WSN. The idea of using an SVM is to separate data belonging to different classes using a hyperplane in a higher dimensional feature space. However, finding an optimal hyperplane is often reported to be difficult e.g. [63,70,93], as the hyperplane would be sensitive to the dataset, kernel functions and the use of model parameters.

Lazarevic et al. [45] present a receiver operating characteristic (ROC) curve to illustrate the trade-off between the detection rate and false alarms. The ROC for a number of different methods is shown in Figure 2.10.

According to the review, the issues in using a classification-based method can be summarised as:

- For supervised learning, such as an ANN, labelling the training dataset is required, which is often impractical.

- For SVM-based methods, to determine an optimal model parameter can be difficult as the model will be sensitive to the dataset and the selection of kernel functions, and this can be very costly in terms of computation.

- Using appropriate contextual information can be important.

2.4 Summary of Literature Review

The review of calibration techniques indicates that the current calibration of sensor networks is still an open challenge. Macro-calibration relies on significant assumptions (e.g. assuming the environment is spatially dependent), which may not be applicable in an urban environment. In comparison, micro-calibration is more practical. However, micro-calibration would not suit a long calibration path or large calibration errors. It is noted that many studies on sensor network calibration are based on simulations. Thus, it is important to deploy sensors to collect real datasets for such studies.

Existing studies on sensor calibration show that the NO_2 sensor is more difficult to calibrate than the O_3 sensor, especially in urban environments. It is clear that using multivariate calibrations can significantly alleviate this issue. However, the review suggests that the use of supporting parameters could be dependent on many factors, which implies that the selection of supporting parameters is important to ensure the calibration results work in a different environment.

Furthermore, according to the review, the regression-based method and the ANN-based method are two of the most widely used approaches for multivariate calibration. However, the lack of studies with an effective comparison of the calibration methods hinders the selection of the calibration, which may lead to an inappropriate calibration being used.

According to the review, we list a few important findings for sensor calibration:

- The difference between the widely used calibration methods is not clear, which hinders the most appropriate method being used.

- Calibration of the NO_2 sensor is currently problematic and difficult, especially in an urban environment.
- Calibration of sensors needs to be carried out under their working conditions.
- Calibration needs to be frequently applied as long as environmental condition changes. This implies that calibration needs to be a lightweight process in terms of computational cost and training complexity.
- Selecting the supporting parameters from the available parameters is important to ensure the calibration results work in different environments.

The review of anomaly detection shows that there are many methods and techniques available for the detection of anomalies. Each method has certain advantages and disadvantages, and needs to balance different trade-offs (e.g. on-line and off-line processes). For example, some techniques prefer to remove outliers on-line as it reduces the communication overhead and further saves battery-power. However, those techniques would not be important for sensors running on mains power as saving power is not the major concern. Therefore, the use of methods for anomaly detection needs to be tailored to the purpose.

Since time series data are often considered as one dimensional data, many existing methods may struggle to differentiate anomalies from outliers, as not enough information is provided. Fortunately, the method that uses appropriate contextual information shows a great advantage as it would not only improve the detection result, but also be able to further differentiate anomalies. As identified in [71, 94], this is due to the fact that the correct measurements are often contextually related, while anomalies are stochastically unrelated. Hence, we believe that using appropriate contextual information is important for the detection of anomalies in our application.

It is known that widely used contextual information is summarised as spatial dependency, temporal dependency and attribute dependency according to [71]. Our review shows that the detection result can be significantly compromised if the spatial and the temporal dependencies are not sufficient [5]. Considering the spatial and temporal information is often inconsistent in our context, as shown in Section 3, we believe that new contextual information will be essential for the detection of anomalies in our application.

According to the review, we list a number of findings for the methods of anomaly detection:

- Evaluation in a real dataset can be difficult as the reference of the anomalies is hard to obtain.
- Using appropriate context information is important (i.e. spatial and temporal information) as it could help to differentiate anomalies from outliers.
- The threshold value is often application dependent. However, knowing the effect of changing the threshold would be helpful.

Chapter 3

Research Background

The main purpose of this chapter is to show the characteristics of real data, and to illustrate the issues that need to be addressed in the thesis. It presents the background of the research, and justifies the thesis contributions with respect to the limitations of the current studies.

To begin with, ELM units which are used as low-cost sensing units in this work are introduced in Section 3.1. Following that, three deployments performed during this research are discussed in Section 3.2. Then, the characteristics of the data in terms of the variation caused by the types of sensors and environments are illustrated in Section 3.3. In Section 3.4, the issues that need to be addressed are summarised. Finally, we justify the thesis contributions in Section 3.5.

3.1 ELM Units

ELM units, a product from Perkin Elmer, are used as the low-cost sensors in this work [58]. An ELM can measure multiple parameters including nitrogen dioxide (NO_2), ozone (O_3), nitrogen oxide (NO), temperature (T), humidity (H), volatile organic compound (VOC), dust and noise. The parameter of dust stands for particulate matter, which combines PM_{10} and $PM_{2.5}$; The parameter of noise represents the amplitude of sound in decibels. It is noted that the sensors used in each unit are off-the-shelf sensors. Thus, the monitored parameters can be tailored according to application requirements.

An ELM unit is about the size of a shoe box. It is designed to have a life time of about 18 months, as some of the sensors provide data via chemicals that degrade. By default, the data is uploaded to a server using GSM. However, when the GSM server is not available, data is temporally stored (within the limits of available resources) in an on-board data logger and uploaded again when GSM communication recovers. The temporal resolution of data for all parameters is 20 seconds.

ELM units are powered by main supplies rather than battery, which addresses the power limitations that many low-cost sensors have. As a result, an off-line process is more appropriate than an on-line process for this application. However, the locations of the deployments are then bound by the mains supply, which means facing many practical issues when deploying them. For example, sensors may not be deployed in a desired place if the power supply is not available [27].

It is noted that, like most end-users, we do not have direct access to the sensors' hardware or software during and after the deployments. The deployments were performed and managed by engineers from the Department of Electronics, University of York. The data is obtained through the service provider (i.e. Perkin Elmer) and downloaded from their server directly via an API [58].

3.2 Deployments

Three deployments were carried out in York, UK, during this study. The first deployment was in 2015. For this deployment, the aim was to understand the performance of ELM units in an uncontrolled environment to compare it with a datasheet describing how it behaved in a laboratory. We wanted to know how accurate and consistent the sensors could be with a simple calibration, e.g. using a univariate calibration. Hence, 20 ELM units were co-located with a reference instrument for more than two months in the Wolfson Atmospheric Chemistry Laboratories (WACL) as illustrated in Figure 3.1. The parameters monitored in all units are identical, NO_2, O_3, T, H, VOC, *dust*, and *noise*. The reference instrument was maintained by WACL in accordance with regulatory

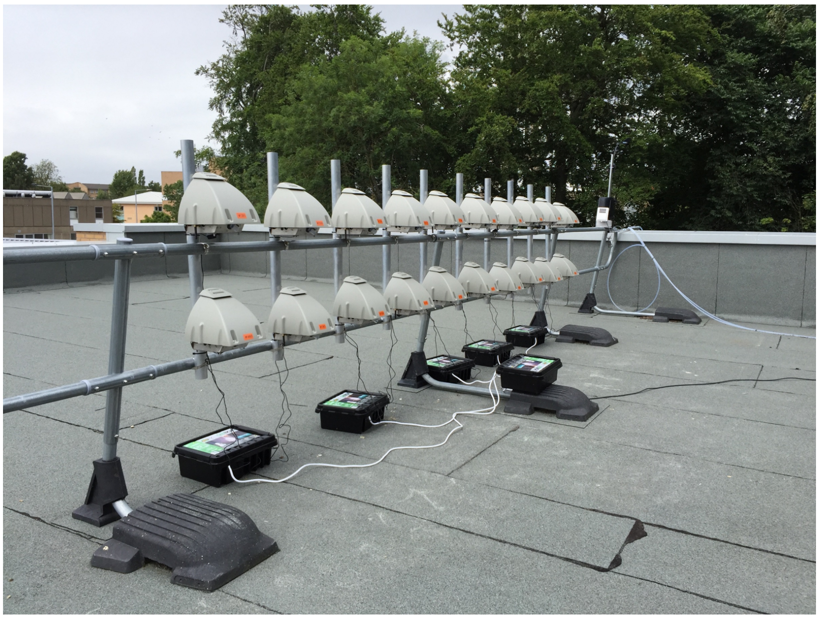


Figure 3.1: Deployment on WACL

requirements [1, 14], which provides NO_2 data with a temporal resolution of an hour (not publicly available). The WACL is on the Heslington West site of the University of York, which is outside the city centre and surrounded by green infrastructure, such as trees. The environmental conditions at the WACL are similar to an *urban background* or *suburban* condition, as defined by DEFRA in [2], which is referred to as *mild* in this thesis. The pollution concentration in a mild environment is expected to vary insignificantly over time and space and its annual averaged concentration is expected to be far below the annual limitation (e.g. 40 μg/m^3 for NO_2). The data from this deployment is about 2 month's worth.

Sensors may have non-unique responses in different environmental conditions as identified in [10, 46]. Hence, the aim of the second deployment was to understand how ELM units would perform in a typical urban environment and to determine how the response of the sensors would differ from those in the mild environment. This deployment was located on Fishergate, which is in the centre of York next to a busy junction. This environment is classified as *traffic* by DEFRA according to [2] and it is referred to as *harsh* in this work. In contrast to the mild environment,

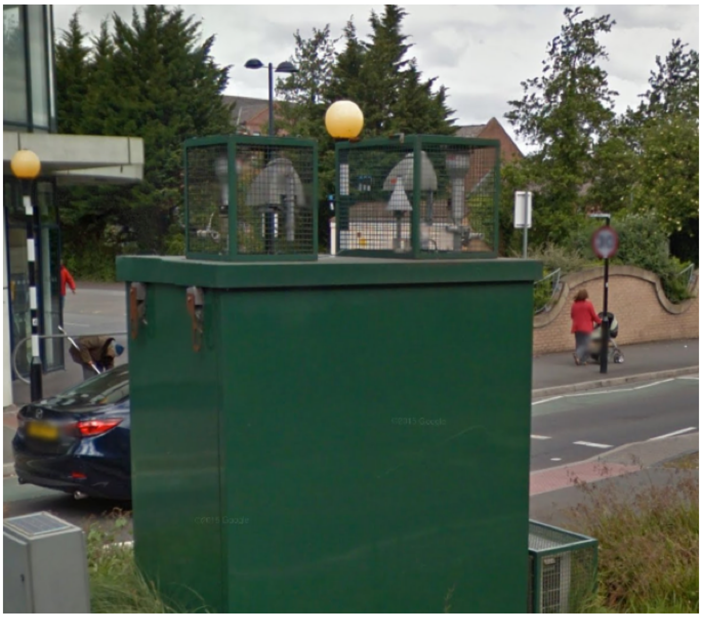


Figure 3.2: Deployment in the Fishergate

the pollution concentration in the harsh environment is expected to vary significantly over time and space and its annual averaged concentration is expected to be around the annual limitation (e.g. 40 μg/m^3 for NO_2). Two ELM units were co-located with a reference instrument at Fishergate for more than 8 months in early 2016 as illustrated in Figure 3.2. The measurements of the ELMs are NO_2, NO, O_3, T, H. The reference instrument (EU Site ID: GB0919A) was managed by the City of York Council and it is a part of Automatic Rural and Urban Networks (ARUN). The reference data can be easily accessed from the on-line portal with the temporal resolution of an hour [17]. It is noted that one of the units stopped transmitting data shortly after the deployment and has not been recovered since, the root cause for that is unknown. This deployment collected about 6 months worth of data.

The third deployment was designed for studying the calibration of sensor networks. The original plan had two phases. The first one was to co-locate all 20 units of ELM at WACL to pre-determine a calibration function, for which the set-up was identical to the first deployment. The second phase was to remove 18 units and deploy them in groups of 3 in a linear fashion heading on to Heslington East alongside Lakeside Way, using the CCTV infrastructure as mounting/power points. With two units remaining on WACL to tie in the data with that of the reference, the remaining 6 groups of 3 sensors were equally spaced at a distance of about 100 meters. Having 3 units at each mounting point was to gain statistical confidence if one of the units malfunctioned. The obtained data would have been useful to study the calibration of sensor networks and answer the question of how the different distance between neighbouring sensors would affect the propagation of the calibration.

The first phase of the deployment started in the middle of 2016. However, shortly after deployment, an increasing number of sensors stopped working. The exact root cause was not clear after the on-site visit, but the engineers suspected that was partially due to the hardware failure of certain sensors, and the dust and bugs accumulated within the units. A large number of sensors stopped working again after the affected sensors were replaced and cleaned. At the end of the three month co-location period, only 9 units of ELM remained working. This experiment suggests that even though a physical inspection may allow the units to be corrected (temporally), the units could fail at any time during the deployment if the root causes were not correctly compensated for. Therefore, we decided not to continue the phase two deployment as it would be very costly in terms of the labour costs and managements.

It is understood that for all the deployments, it is important to ensure that the co-located sensors (e.g. reference and uncalibrated ELM units) are sufficiently close. The sensors need to be in the same micro-environment and to monitor the same phenomena. Otherwise, the data from the reference instrument may not represent the ground truth of the low-cost sensors. We are fully aware that sufficient distance between the low-cost sensor and reference instrument would be sensitive to environmental conditions [65]. According to the literature [59,67], we considered a sufficient distance for co-location as tens of meters in a mild environment and a meter in a harsh environment. These constraints were applied in all of our deployments. As a result, this thesis is focused on the calibration of a single sensor unit.

3.3 Characteristics of Environmental Data

As discussed in Section 1.4, the characteristics of environmental data are associated with the environment and the sensors. In this section, we firstly use the data from our real deployment to illustrate how the data would vary with respect to those two factors. Then, we illustrate the issues of data gaps and the dependency between the monitored parameters.

3.3.1 Data Variation Caused by Environment

The variation of urban environments can be illustrated by comparing reference instruments in two locations as shown in Figure 3.4. The figure shows week-long NO_2 data obtained by reference instruments in Fishergate (harsh) and at the WACL (mild) respectively. The dashed line indicates the 40 μg/m^3 annual limitation of the NO_2. The distance between two locations is about a mile as shown in Figure 3.3. Data gaps can be observed in the dataset from Fishergate (in the black circle). Since we do not have information on how this particular dataset is processed, it is not clear what the actual root causes are. However, according to the data quality control procedures in [1, 14], the data gap is likely caused by the manual removal of the suspected reading.

The temporal variation of environmental parameters can be observed in many different levels, such as daily, seasonal and annual. In Figure 3.4, a clear daily pattern can be observed from the instrument located in Fishergate, where the concentration of NO_2 in the day is consistently higher than the night. In contrast, the daily variation is not very clear in the dataset collected from WACL. It suggests that the pollution concentration may not be spatially consistent.

Furthermore, given the prior knowledge that NO_2 in cities is mainly contributed by vehicle emissions, and considering the volume of traffic is higher during the day than at night, and higher in the harsh environment than the mild, we believe that the temporal variation of the environment is related to certain environmental factors. Therefore, the change of environmental factors would affect the temporal consistency, e.g. abnormally

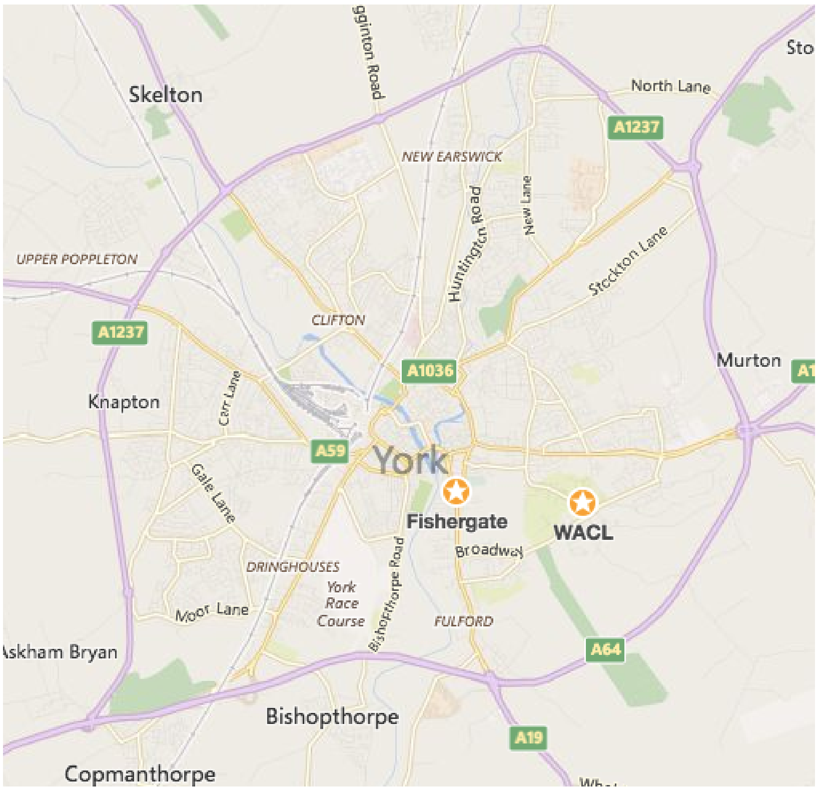


Figure 3.3: The locations of the reference instruments (The map and the pinpoint service are provided by the Bing Maps, Microsoft, 2018)

heavy traffic at night time would have an impact on the daily variation. As a result, the environment is spatially and temporally inconsistent.

3.3.2 Data Variations Caused by Low-costs Sensors

As discussed in Section 1.4, the data from low-cost sensors can be affected by the environment and the sensor simultaneously. Thus, knowing the actual environment is important to determine how a low-cost sensor would impact the data. In the following, we illustrate the variation of low-cost sensors by comparing them to the co-located references.

Figures 3.5 and 3.6 illustrate the NO_2 data obtained from low-cost sensors and the references in the mild and harsh environments. Comparing the data from the low-cost sensors to their references in Figure 3.5, we can observe that the data from the low-cost sensors has a consider-

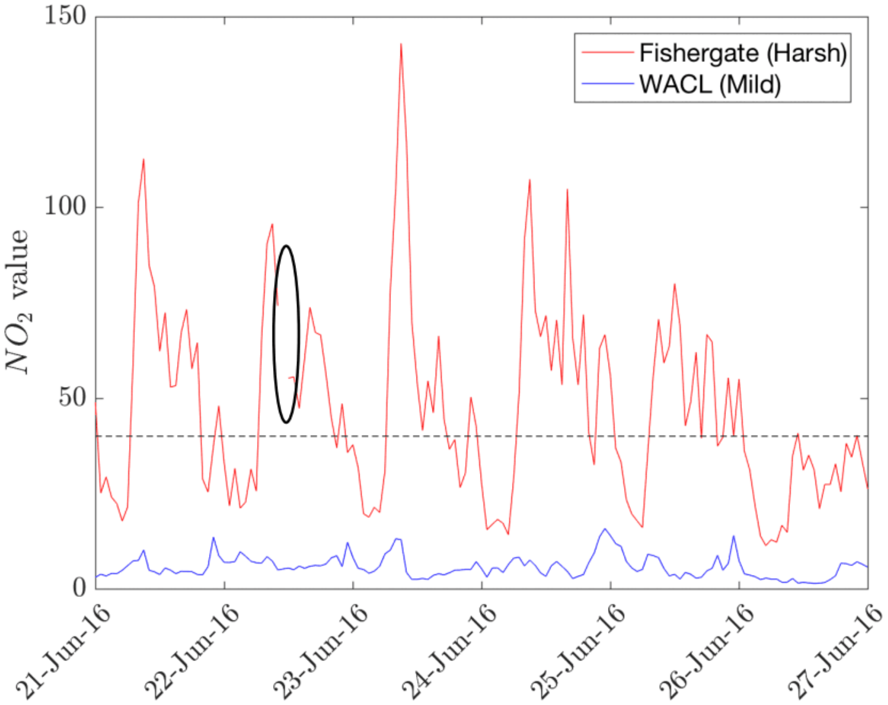


Figure 3.4: NO_2 data obtained from reference instruments at mild and harsh environment respectively

ably higher percentage and magnitude of outliers. It is understood that the outliers are not necessarily anomalies. However, as the variation and the magnitude of the outliers observed in low-cost sensors are so significant in comparison with the reference, we believe that the outliers in the data are dominated by anomalies. It suggests that low-cost sensors would introduce anomalies into the data and that the anomalies are more significant in the harsh environment.

Figure 3.6 is rescaled from Figure 3.5 where the data of the outliers are excluded. It is noted that in Figure 3.6 the data pattern for the ELM unit in the city is significantly different from the reference, in which more than 50% of data from the low-cost sensors are zero values. The zero values are more problematic as they cannot be rescaled during the calibration. It suggests that the harsh environment has a greater influence on the low-cost sensors, which implies that a sensor in a harsh environment could be more difficult to compensate.

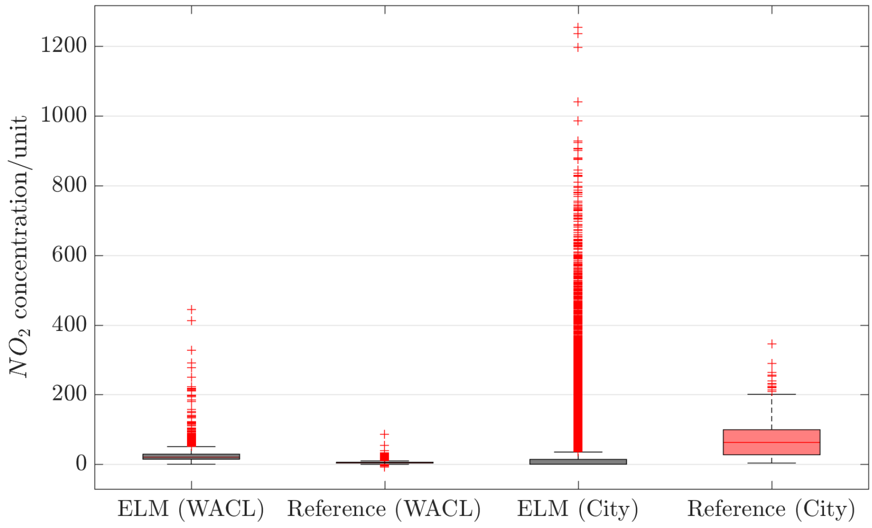


Figure 3.5: NO_2 comparison (Raw data)

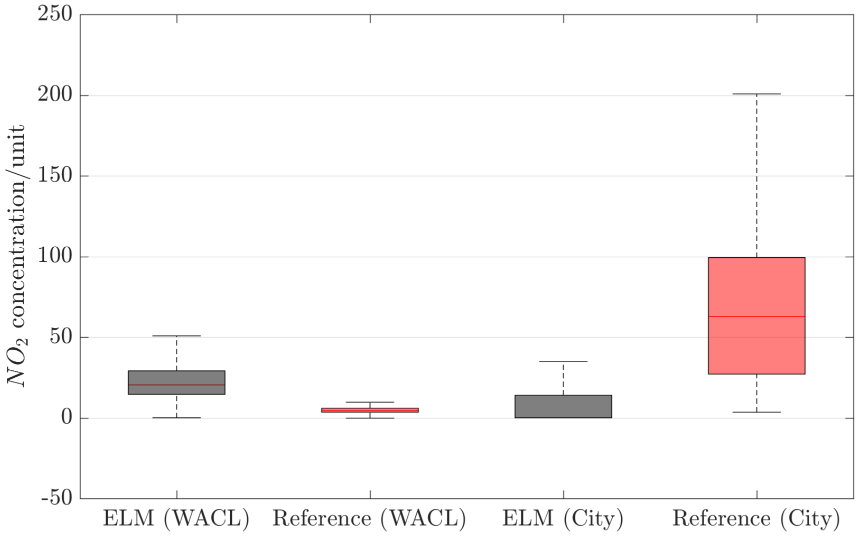


Figure 3.6: NO_2 comparison (With outliers excluded)

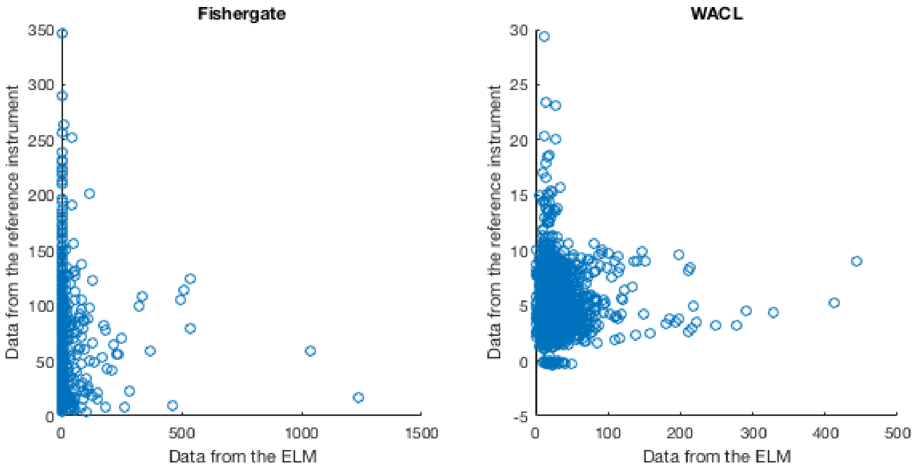


Figure 3.7: Scatter plots between ELM data and reference data at two locations

Figure 3.7 shows the scatter plot between the data from the ELM sensors and reference instruments at two locations. Since the data pattern between the mild and harsh environments in the figure are significantly different, we believe that the calibration being determined in one environment is not necessarily applicable to sensors in another environment. Hence, it is important to calibrate sensors each time when the surrounding environment changes.

3.3.3 Data Gaps

Data gaps may frequently occur, and these may have a significant influence on certain processes, e.g. data aggregation. Figure 3.8 shows the completeness of week long data received from 9 ELM units deployed at the WACL. The colour is associated with the percentage of the data received in an hour. In the figure, we can see that there is only a small percentage of time that all data was successfully received by the server. For the rest of the time, data gaps frequently occur. It is clear that the data gaps were also present in the data from the reference instruments as shown in Figure 3.4. However, according to the figure, the data gaps observed in the low-cost sensors are more frequent and significant. As a result, the temporal consistency of the data would be significantly affected, especially at a higher temporal resolution.

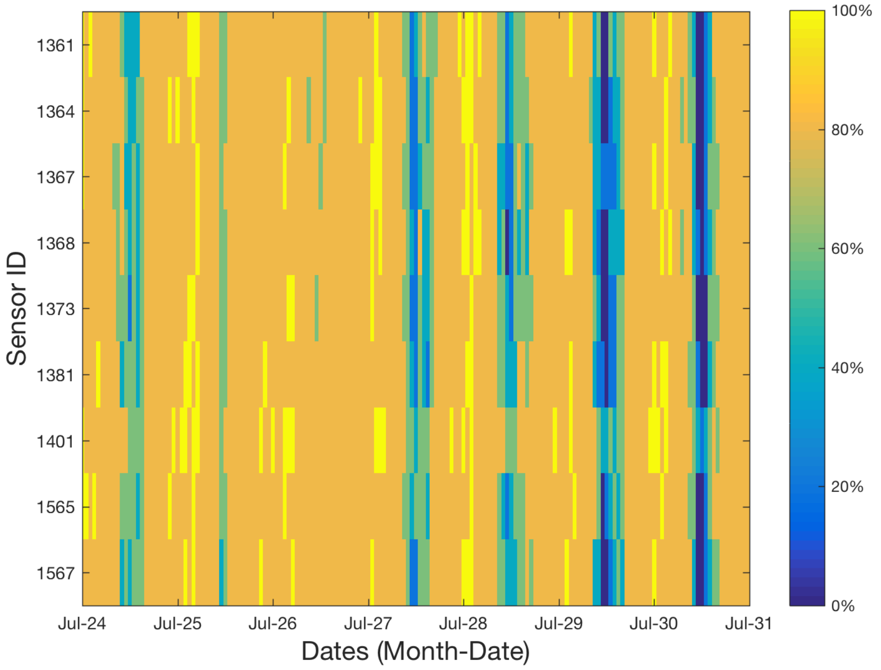


Figure 3.8: Data completeness

3.3.4 Dependency of the Parameters

It is known that if one or more parameters are severely linear dependent, the calibration model, especially constructed by regression, may be negatively affected due to the multicollinearity. Hence, we calculate the cross-correlation between all monitored parameters in both mild and harsh environments and illustrate their linear dependency.

Figures 3.9 and 3.10 show the correlation coefficient using Pearson's R (Pearson correlation coefficient) for all pairs of parameters in the mild and harsh environments. It is understood that multicollinearity or collinearity is not dependent on a definite threshold value. However, O'Brien [56] suggests using the variance inflation factor (VIF) as an index to determine the significance of the collinearity. The VIF is calculated based on Equation 3.1.

$$VIF = \frac{1}{1 - R^2} \tag{3.1}$$

O'Brien states that if the VIF is above 10, collinearity is likely to be an issue for the process. This implies that if the correlation coefficient (R) between any pair of the parameters is over 0.94 in the figures, the issue of collinearity may need to be considered. Fortunately, the maximum correlation coefficient from both figures is 0.70, which is below 0.94. Hence, the result indicates that collinearity may not significantly affect our process.

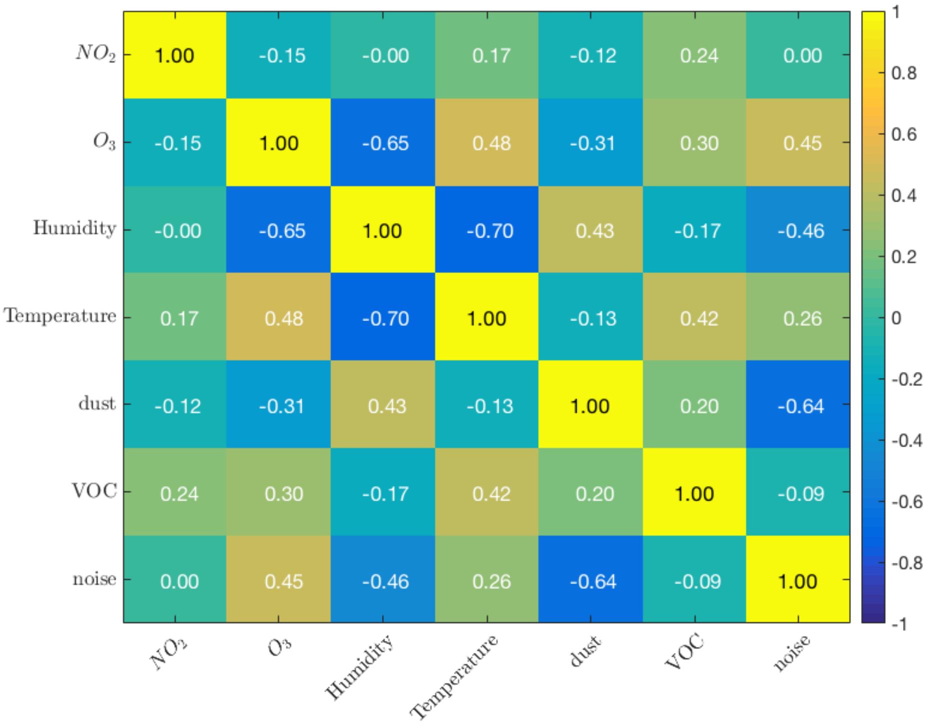


Figure 3.9: Cross-correlation from a sensor unit at WACL (mild)

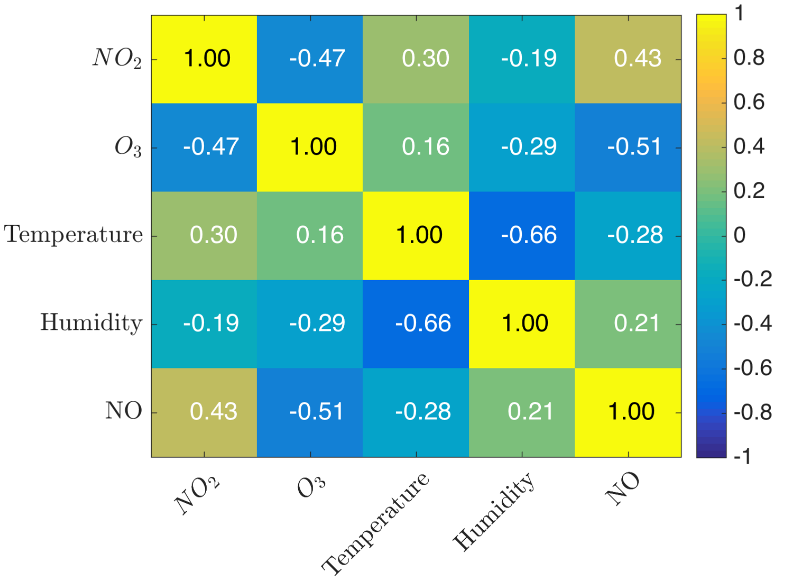


Figure 3.10: Cross-correlation from a sensor unit at Fishergate (harsh)

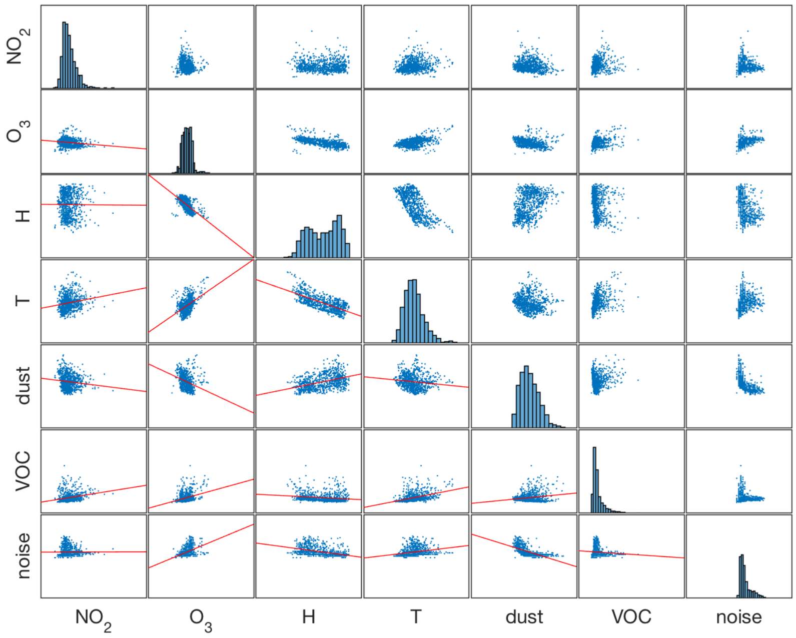


Figure 3.11: Cross-correlation from a sensor unit at WACL (mild)

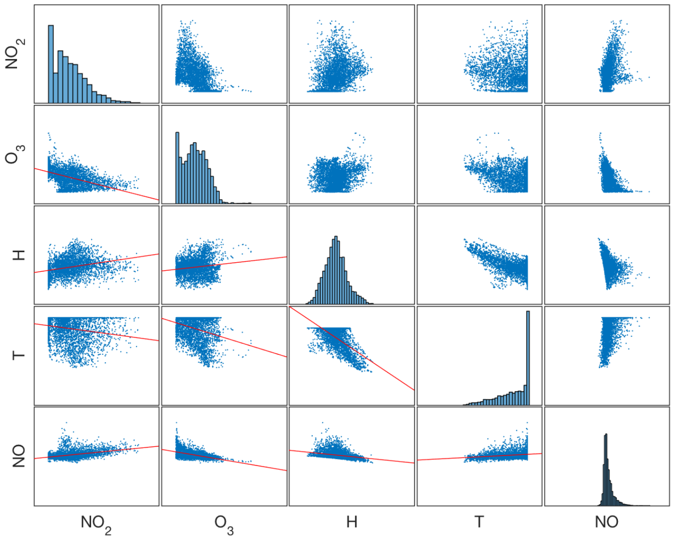


Figure 3.12: Cross-correlation from a sensor unit at Fishergate (harsh)

We also generate scatter plots across the parameters in both locations in Figure 3.11 and 3.12. From the figures, the temperature and the humidity show strong negative correlations in both locations. In the mild environment, humidity has a strong negative correlation to the O_3. However, the correlation gets weak in the harsh environment. The results show that the correlation for the same pair of parameters in different locations can be inconsistent. It suggests that dependencies between parameters can be non-unique, which implies that the use of supporting parameters may vary for the calibration in different locations.

3.4 Issues in Improving Data Quality

In this chapter, we introduced the ELM units and discussed three of our deployments. From this, we noticed that the location of deployment can be bounded by many practical constraints and the failure of sensors can frequently occur after deployment [27]. Since we do not have physical access to the sensors, the root causes of abnormalities in the data are hard to identify and compensate for. Hence, we believe having easy access to physical sensors is important for the deployment of low-cost sensors. Furthermore, as we failed to deploy the sensors to form a sensing network, this thesis is focused on a single sensor unit.

The data from our real deployments are used to illustrate the characteristics of the environmental data. Using data from the reference instruments, we determined that the data obtained from the urban environment is neither spatially nor temporally consistent. As a result, anomalies in the data may be hard to identify, as reviewed in Section 2.3. Furthermore, comparing the data from the low-cost sensors to their references in different environmental conditions, we concluded that low-cost sensors would be more sensitive to the harsh environment than the mild environment. This implies that compensating for the data issues associated with the sensors in the harsh environment is more difficult.

Finally, we summarise a list of issues that need to addressed with respect to our data and the limitations of the current methods:

- Data is not spatially and temporally consistent. Hence, such information cannot be used to determine an anomaly model.

- The difference between ANN-based and regression-based calibration is not clear, which hinders the most effective method being used.

- A selection of supporting parameters is important for the calibration, as the use of supporting parameters is non-unique.

3.5 Thesis Contributions

The main contribution of this thesis is a two-phase solution to improve the data quality of low-cost NO_2 sensors in an urban environment. The solution consists of the novel detection and removal of anomalies with a comprehensive calibration process, in which anomalies are removed before calibration. With this solution, the data from low-cost sensors is able to achieve significantly enhanced accuracy than before in a harsh environment. Under the main contribution, a list of other contributions is also given, which addresses the issues identified in Section 3.4.

- Chapter 4 presents a systematic comparison of state-of-the-art calibration techniques, which focus on determining the difference between two of the most used calibration methods, i.e. ANN and regression-based approaches. To the best of our knowledge, this is the most effective comparison for comparing the calibration method. The result is able to support the selection of a calibration method.

- Chapter 5 proposes a calibration method that systematically and automatically uses supporting parameters for multivariate calibration, which ensures the optimal set of supporting parameters are used according to local conditions.

- Chapter 6 presents a method for the detection of anomalies, which uses new contextual information (i.e. cross-sensitive parameter) to detect and remove anomalies. The results show that anomalies can be better differentiated from outliers when using the new contextual information. To the best of our knowledge, this is the first research to use this information for the detection of anomalies in air quality sensors.

Chapter 4

The Comparison of Calibration Methods

This chapter aims to answer the first research question, which is quoted below:

> *Research Question 1: Which is the appropriate calibration method (Regression or ANN) considering the needs of our application?*

Since the calibration process may need to frequently be applied with a change in the environmental conditions, a light-weight process in terms of complexity and computational cost is preferred for sensor calibration in urban environments. In addition, considering the life-time of low-cost sensors is bound by the degradation of the sensors, the dataset collected from the existing deployments is often small (e.g. less than a year's worth of data in our application). Therefore, this application would need a light-weight calibration process that works better on a relatively small dataset.

The review in Section 2.2 indicates that multivariate calibration is the best practice for the calibration of low-cost sensors. Regression-based and ANN-based methods are two of the most used approaches for such a purpose. Intuitively, a regression-based method can be easily applied and interpreted, but it may not suit calibrations that have a complex relationship between the inputs and output. By contrast, an ANN-based method is able to solve the problem with a more complicated training process. However, to the best of our knowledge, the difference between these approaches in dealing with different training and testing scenarios has not been thoroughly discussed in the literature.

This chapter presents a systematic comparison of state-of-the-art calibration methods, i.e. regression-based methods and ANN-based methods. Instead of comparing only calibration accuracy, this work uses multiple training and testing datasets to determine the sensitivity of the methods to these datasets and to understand their differences.

In the rest of this chapter, we firstly explain how the calibration models of both methods can be constructed and illustrate which model parameters are needed for constructing them in Section 4.1. Then, we demonstrate the determination of the model parameters for both methods using a dataset obtained from one of our deployments in Section 4.2. Following that, both calibration methods are cross-compared using different training and testing data in Section 4.3. The research validity is discussed in Section 4.4. Finally, we summarise the findings and answer the research question in Section 4.5.

4.1 Sensor Calibrations

In this section, we illustrate how an ANN-based method and a regression-based method can be used for the calibration of sensors, and discuss which model parameters are important for both methods.

4.1.1 Calibration Using an ANN-based Method

An ANN operates in a similar way to a biological neural network in animal brains, which propagates the information via neuron connections. In an ANN, tasks are performed using the knowledge learnt from the training process. More specifically, the training process is used to determine the weights of neurons and their propagation path.

Assuming that calibrating X_1 requires X_2 and X_3 as supporting parameters, a graphical structure of an ANN is illustrated in Figure 4.1.

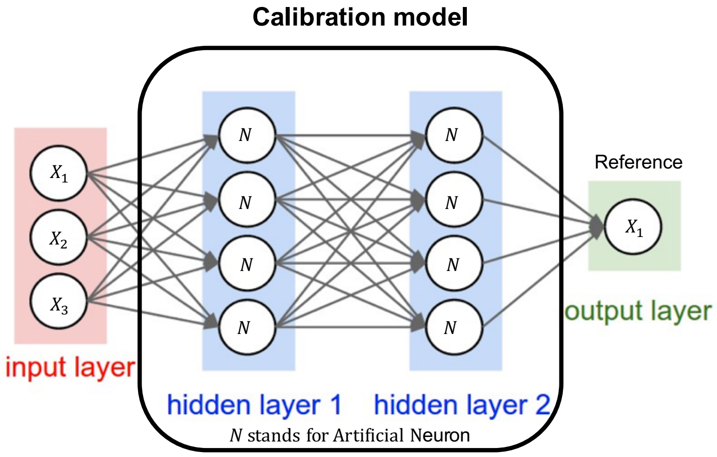
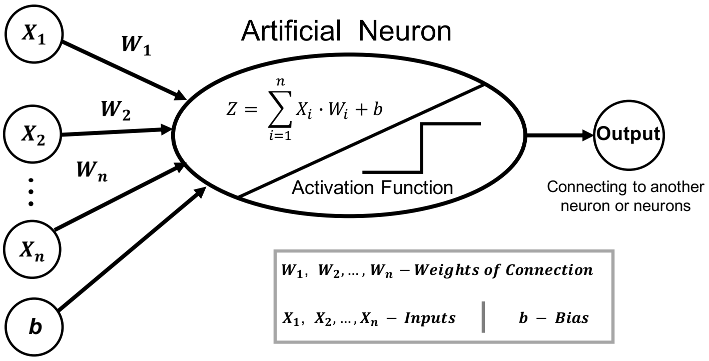


Figure 4.1: An ANN structure for the calibration of X_1

The training process is used to determine the calibration model as shown in the figure. The determined model can then provide an approximation of X_1 (calibrated) by the given corresponding inputs, X_1 (uncalibrated), X_2 and X_3. In theory, an ANN can be programmed without any task-specific rules (e.g. without knowing the relationship between inputs and outputs). Thus, the calibration can be performed without any prior knowledge.

An artificial neuron is an important part of constructing an ANN. The graphical structure of a neuron is shown in Figure 4.2, which works in the same way as the neurons in Figure 4.1. A neuron can have multiple inputs, which can be either inputs from a network or from the output of another neuron. However, a neuron often has just one output, but the output can connect to multiple neurons, as shown in Figure 4.1.

A neuron calculates the weighted sum of inputs (Z) and passes Z to an activation function as shown in Figure 4.2. The outcome of this controls the neuron output. For example, if the outcome is above a certain threshold, the neuron is on (On indicates that the information will be propagated); otherwise, the neuron is off (Off indicates that the information will not be propagated). Therefore, an activation function is important for an ANN.



X_1
W_1
Artificial Neuron
$Z = \sum_{i=1}^{n} X_i \cdot W_i + b$
W_2
X_2
W_n
Activation Function
Output
Connecting to another neuron or neurons
X_n
b
$W_1, W_2, \ldots, W_n$ – Weights of Connection
$X_1, X_2, \ldots, X_n$ – Inputs
b – Bias

Figure 4.2: Inside a Neuron

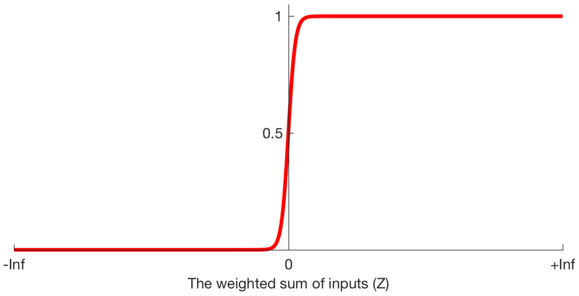


Figure 4.3: Sigmoid Function

4.1.1.1 Activation Function

According to the literature, e.g. [3,55], the main purpose of an activation function is to transfer the weighted sum of inputs, Z. It is clear that the variation of Z is unbounded before the transformation, as it can vary from $-Inf$ to $+Inf$, which would hinder the optimisation process in determining the weights. However, after the transformation, e.g. using the sigmoid function, which is defined as Equation 4.1, $S(Z)$ is normalised and shown in Figure 4.3.

$$S(Z) = \frac{1}{1 + e^{-Z}} \tag{4.1}$$

According to the literature, using an activation function is essential for an ANN, and it has three main purposes:

- The weighted sum of inputs becomes bounded, which avoids unstable convergence during optimisation.

- It helps to decide whether the neuron is firing or not. Firing means the inputs from this neuron will be propagated to another neuron, otherwise not. For example, the neuron will fire if the $S(Z)$ is larger than 0.5 in the example in Figure 4.3.

- A non-linear activation function enables the ANN to explore a complex non-linear relationship, especially with multiple neurons and layers.

It is noted that there are many activation functions available. We list a number of activation functions that have been widely used in Figure 4.4.

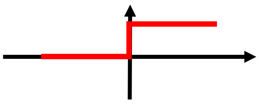
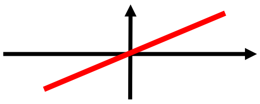
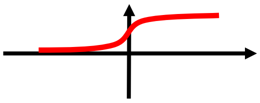
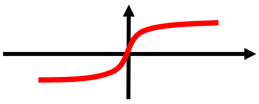
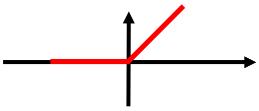
Activation function	Equation	Plot
Binary step	$\emptyset(\mathbf{z}) = \begin{cases} 0 & z < 0 \\ 1 & z \geq 0 \end{cases}$	
Identity	$\emptyset(\mathbf{z}) = z$	
Sigmoid	$\emptyset(\mathbf{z}) = \dfrac{1}{1 + e^{-z}}$	
TanH	$\emptyset(\mathbf{z}) = \dfrac{2}{1 + e^{-2z}} - 1$	
Rectified linear unit (ReLU)	$\emptyset(\mathbf{z}) = \begin{cases} 0 & z < 0 \\ z & z \geq 0 \end{cases}$	

Figure 4.4: Summary of activation functions

4.1.1.2 Structure of a Network

The structure of the network is also important for constructing an ANN. A network structure specifies the type of neurons, the number of neurons and layers and how each neuron is connected. The design of a network structure often requires expert knowledge and its improvement relies on trial by error. Therefore, having an optimal structure for a specific dataset can be difficult. It is noted that the deep neural network has become popular in recent years. This is an ANN with a large number of layers which potentially enables the modelling of complex data. However, the computational cost required for the training would increase dramatically. Furthermore, considering the small size of our data, it is unlikely to obtain a stable deep neural network. Thus, this type of network is not considered in this thesis. We illustrate a few widely used network structure in Figure 4.5.

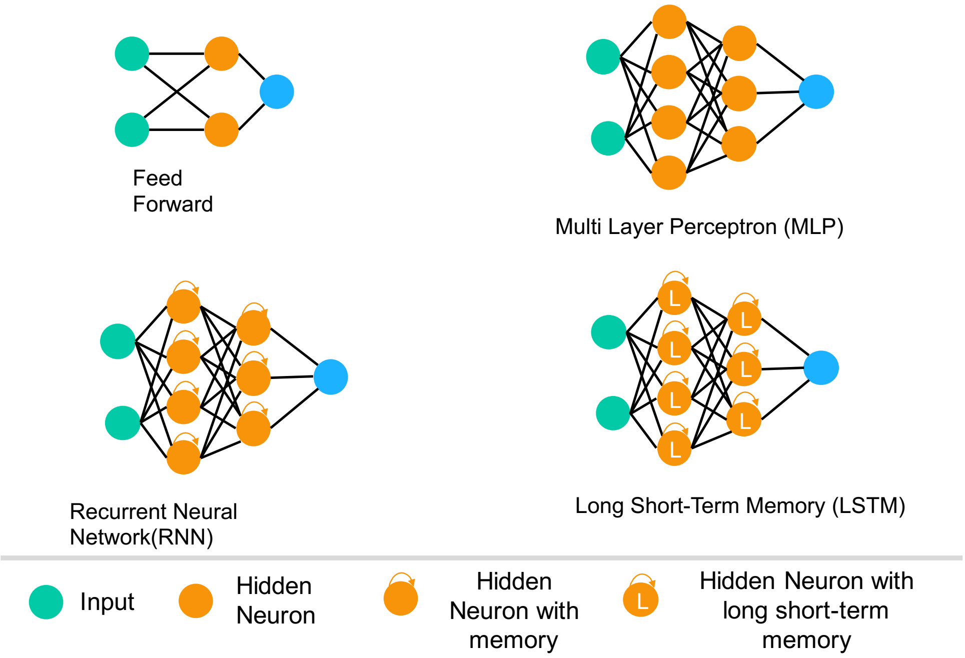

Feed Forward
Multi Layer Perceptron (MLP)
Recurrent Neural Network(RNN)
Long Short-Term Memory (LSTM)
L
Input
Hidden Neuron
Hidden Neuron with memory
Hidden Neuron with long short-term memory

Figure 4.5: A list of network structures

4.1.1.3 Other Parameters

Apart from the parameters that have been discussed above, a number of epochs and batch sizes, the loss function and the optimisation method are also important for an ANN-based method.

The loss function in an ANN is a function that we want to maximise or minimise during training. Considering the purpose of calibration is to minimise errors between the model output and the reference, the loss function is often used to describe calibration errors. Therefore, implementing the loss function has a number of available options, such as mean absolute error, mean squared error and mean squared percentage error.

Optimisation is a method to determine the weights of all connections that minimise (or maximise) the loss function. In practice, there are also many optimisation methods available, such as, gradient descent and Adam optimiser.

In most cases, a loss function can have multiple modes. Hence, an optimisation method may only find local minima/maxima rather than the global minima/maxima. It is known that the initial points of an optimisation method often start randomly, which could result in different model outputs. In order to minimise variation in the model output, using multiple epochs is desired. One epoch indicates the entire dataset being passed into the training. A number of epochs indicates the number of the times that the entire dataset has gone through training. However, even though a higher number of epochs would minimise the variation of model output, it would not completely solve the problem. Furthermore, it is clear that using an extremely large number of epochs would increase the computational time dramatically. Therefore, the use of epochs needs to balance the trade-off between variations in the model output and computational time.

The batch size indicates the number of samples used in an iteration, and both the batch size and iteration are associated with the total number of samples for a training dataset. For example, there are 100 samples in a training dataset. If we select 5 as the batch and 3 as the epoch, the number of iterations in every epoch is $100/5 = 20$, and for the entire training, it is $20 \times 3 = 60$.

4.1.1.4 Summary of Model Parameters

In this section, we explained how an ANN-based method can be used for multivariate calibration. From the discussion, we identified a number of model parameters that are important for an ANN-based method, which are summarised in Table 4.1.

Table 4.1: A number of model parameters that need to be determined for an ANN

Model parameters	Examples
Activation function	Sigmoid, ReLU,...
Type of neuron	Dense, LSTM,...
Number of Neurons	1 to $+\infty$
Number of layers	1 to $+\infty$
Batch size	1 to the total number of training sample
Epoch	1 to $+\infty$
Loss function	Mean square error, Mean absolute error, ...
Optimisation method	Gradient descent, Adam,...

4.1.2 Calibration Using a Regression-based Method

In this section, we illustrate how multivariate calibration can be performed using a regression-based method and which parameters are important for this model. In contrast to the ANN-based method, a regression-based method needs to pre-determine the relationship between input variables.

Again, using the calibration of X_1 as an example, which requires X_2 and X_3 as supporting parameters. Assuming a linear relationship between the inputs, a multivariate regression using the corresponding coefficients β can be constructed based on Equation 4.2.

$$Y(i) = \beta_0 + \beta_1 \cdot X_1(i) + \beta_2 \cdot X_2(i) + ... + \beta_n \cdot X_n(i) + \varepsilon(i) \tag{4.2}$$

In Equation 4.2, the ε stands for error term and the i indicates that the measurements are taken from the same time frame, Y is the reference of X_1; n presents the number of parameters used in the model. The calibration model is then to determine the coefficient β using Equation 4.3.

$$\mathcal{E} = minimise \sum_{i=1}^{N} \varepsilon_i^2 \tag{4.3}$$

Note that the example in Equation 4.2 uses a linear combination of first order terms to describe the relationship between the input variables and the output (i.e. linear). If a more complex non-linear relationship is important, the relationship needs to be pre-defined before training (e.g. include non-linear terms or apply a non-linear transformation). Therefore, we consider the relationship between input variables and output as the only model parameter for a regression-based method.

4.2 Determining the Model Parameters

In this section, we demonstrate the determination of the model parameters and discuss the practical issues encountered during the process. Firstly, we present the data and programming environment used for this experiment. Then, we determine the model parameters for both methods.

4.2.1 Data and Programming Environment

Since our sensors are expected to work in the harsh environment, and the calibration of sensors in the environment of operation is important, the experiment in this chapter uses the data obtained from ELM unit at Fishergate (harsh).

The data was pre-processed by aggregating the raw data into an hourly basis and excluding data gaps. The process is based on Algorithm 1 in Section 5.1.1. The dataset after pre-processing contained around 4,000 samples with a temporal resolution of an hour, and the available parameters are NO_2, O_3, NO, T and H, where the T and H present temperature and relative humidity respectively.

The regression-based method was programmed in Matlab, and the ANN-based method was programmed in Python using Keras library [41] and $TensorFlow$ [83]. Both programs were running on a Mac Book Pro laptop with 2.7 GHz Intel Core-i5 (no dedicated GPU).

Since it was not clear how the size of the training and testing datasets would affect the calibration, the dataset was divided sequentially into two equally sized partitions. The first 2,000 samples were used as training and the rest of the samples were used as testing. This practice also maximised the temporal order of the data. Furthermore, since calibrating NO_2 is often reported to be problematic and requires multivariate calibration, as discussed in Section 2.2, the calibration of NO_2 is used as an example.

4.2.2 Model Parameters for A regression-based Method

It is clear that a pre-determined relationship between the inputs and output is important for using a regression-based method. This experiment is to analyse if using a non-linear relationship would improve the model prediction. For this experiment, the non-linear relationship is considered as using higher order terms in the model. The first 2,000 samples of the dataset are used for training and another 2,000 samples are used for testing.

For the experiment, the first model uses a linear combination of first order terms, which is identical to Equation 4.2, and expressed as $f(NO_2\ , O_3\ , NO\ , T\ , H)$. The following models are constructed by gradually including a second order term into the existing model, as well as their interactions [36]. We express the second model as $f(NO_2\ , O_3\ , NO\ , T\ , H\ , NO_2^2)$ and the last model as $f(NO_2\ , O_3\ , NO\ , T\ , H\ , NO_2^2\ , O_3^2\ , NO^2\ , T^2\ , H^2)$. The experiment tests all the possible combinations, i.e. $\binom{0}{5} + \binom{1}{5} + \binom{2}{5} + \binom{3}{5} + \binom{4}{5} + \binom{5}{5}$, which means in total 32 models were used.

Figure 4.6 shows the results of Root-Mean-Squared-Error (RMSE) and correlation coefficient (R) between the predictions and the reference, and the time cost for the training in seconds (Time). In the figure, X-axis (1) indicates the linear model; whereas X-axis (2) to (32) indicates one or more higher order terms being used in the model. The result suggests that the linear model is the best model in comparison to the models using higher order terms. Furthermore, since the time spent in the training is less than a second, we believe that the regression-based method is a lightweight process.

The boxplot in Figure 4.8 shows the error distribution between the

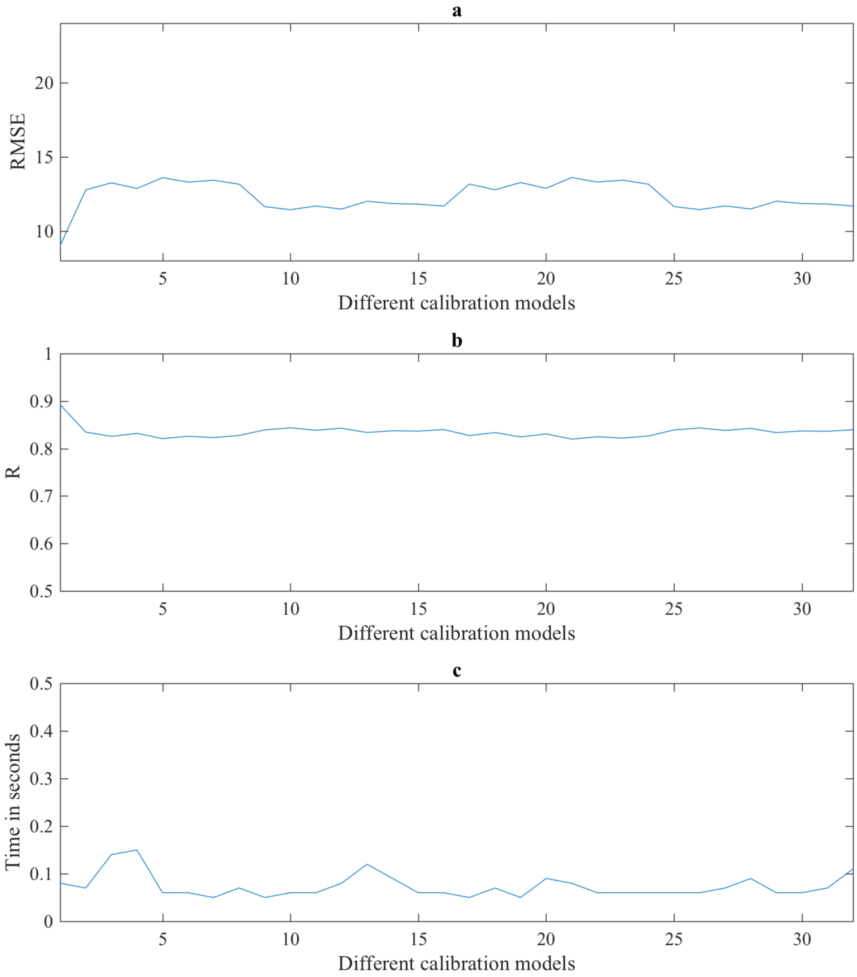


Figure 4.6: The results of RMSE, R and time for different model settings where (1) is the linear model, (2) to (32) are the non-linear models

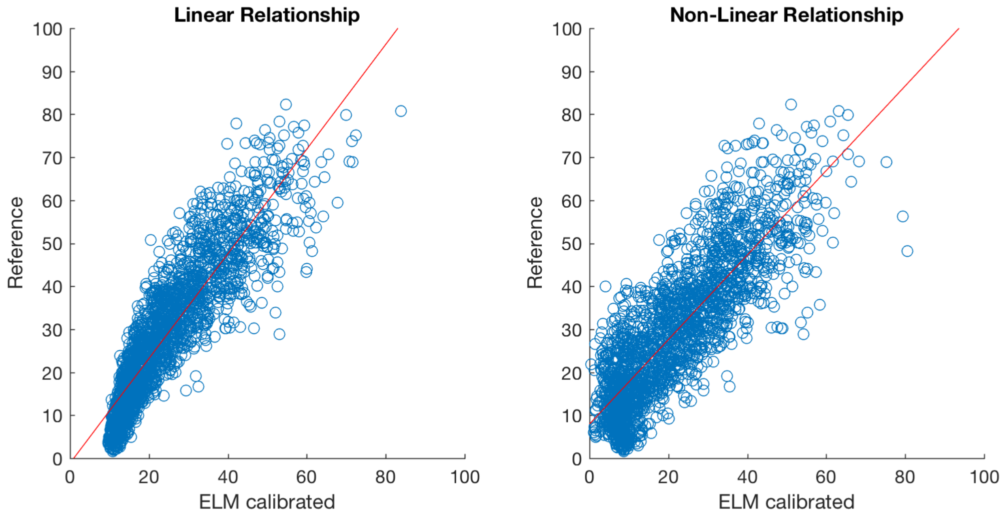


Figure 4.7: The scatter plots between the results of using the linear relationship and a non-linear relationship

model predictions and the reference. The error is defined as the difference between the model output ($y^{'}$) and the reference (Y), given by Equation 4.4. It is noted that i indicates the number of samples.

$$error(i) = Y(i) - y^{'}(i) \tag{4.4}$$

It is clear that the result in Figure 4.8 is in-line with Figure 4.6, as the result with a lower RMSE and a higher R value corresponds to better error distributions. The results from both figures suggest that using a non-linear relationship in the regression-based method does not necessarily improve the calibration result.

We plot a scatter plot to compare the results of using the linear relationship and a non-linear relationship, which is shown in Figure 4.7. The plot for the non-linear relationship uses $f(NO_2\ , O_3\ , NO\ , T\ , H\ , NO_2^2\ , O_3^2\ , NO^2\ , T^2\ , H^2)$, and it shows a wider spread of points in comparison to the one using the linear relationship. The result indicates that the calibration model is unlikely to benefit from a complex relationship without a proper reason. Therefore, a linear relationship is chosen for constructing the model for the regression-based method as discussed in Equation 4.2.

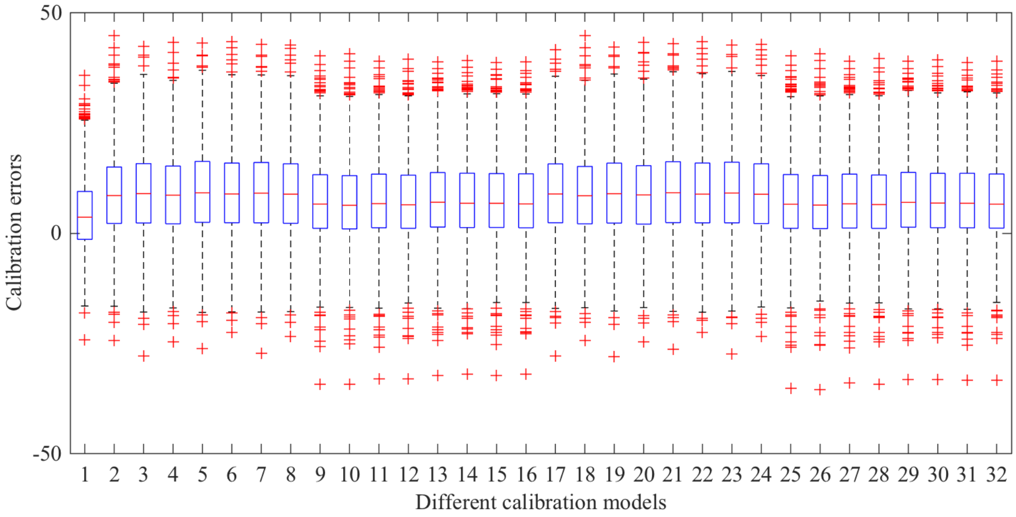


Figure 4.8: The absolute error distribution for different model settings where (1) is the linear model, (2) to (32) are the non-linear models

4.2.3 Model Parameters for an ANN-based Method

In the following, we discuss how the model parameters of the ANN-based method were determined and selected for this experiment. It is noted that the determination of optimal model parameters is still an open challenge in the artificial intelligence (AI) community, and trial by error is currently the best practice for this purpose. It is understood that trying all possible combinations of different parameter settings would not be practically feasible. Therefore, the variation of parameters is tested in a certain range and for certain parameters only, for which the decision is made based on existing work.

Activation function It is clear that there are many activation functions to select from. In this work, we use a sigmoid function as the activation function because 1) the sigmoid function is often used in an ANN for sensor calibration [19, 24], and 2) it is able to uncover the hidden relationship (e.g complex non-linear) between inputs and output [33]. It is clear that each neuron can have a different activation function. However, since it is practically difficult to test this, the same activation function is applied to all neurons.

Structure of the network It is noted that determining the structure of the network would be highly reliant on expert knowledge, as the structure of a network can be sensitive to the use of data and the modification of the network is mainly based on trial by errors. It is known that the current concentration level of an environmental parameter is often associated with previous concentrations. Thus, in order to maximise this property, the LSTM is used as the neuron type, as it is able to use the information from the previous training [38, 55].

Number of neurons and layers We vary the number of neurons in each layer as [5 20 35] and the number of layers in [1 2 3 4 5]. The same number of neurons are used in each layer.

Batch size and epoch We test the number of batch size in [1 6 11 16 21 26] and epoch in [1 6 11 16 21 26], which is 1 to 26 with an incremental of 5.

Loss function Considering the quality of a calibration is often evaluated using mean square error, this is used as the objective function in the experiment.

Optimisation method Gradient descent is used in this work for optimisation, as it is one of the mostly used methods for this purpose [21, 61].

To sum up, four parameters need to be selected in this experiment, number of layers, number of neurons, epoch and batch size.

Since the experiment has a large number of trials, it is not practical to evaluate the error distribution for all results. Thus, we use RMSE and R between the predictions and the reference, and the time spent in training to approximate the calibration result. It is understood that two identical RMSE or R may represent different error distributions. However, according to the results in Section 4.2.2, we believe a result with a low RMSE and a high R value is sufficient to represent a good calibration result. The calibration result, in terms of RMSE, R and the time, from the use of different model parameters are presented in Figures 4.9 to 4.13 (at the end of this chapter). For this experiment, the first 2,000 samples of the dataset are used for training and another 2,000 samples are used for testing in each trial. In other words, the data used for training and testing are identical across all trials.

Figures 4.9 to 4.13 differentiate the number of layers used in the model. Hence, the effect of the layer can be derived by comparing the plots across figures. Within each figure, the number of neurons used in the model is distinguished by different rows. The first, second and third column of the figures are RMSE, R and the time respectively. In each plot, the value is dependent on different epochs and batch sizes.

From the figures, the RMSE and R do not have a predictable response for different model parameters. For example, the use of parameters to obtain the best RMSE value does not always secure the highest correlation coefficient. Furthermore, the RMSE and R do not have consistent trends across different plots, which suggests they may be sensitive to all model parameters. Time increases when larger epochs are used; but, using a bigger batch size would compensate for this. It is noted that a larger batch size would lead to a significant degradation in the quality of the model according to [42]. Hence, it may not be appropriate to simply increase the batch size to reduce computational cost. Moreover, the figures show that the time spent in training an ANN-model is significantly higher than training a regression-based model as fitting one model could take up to almost 1,000 seconds.

Since the results did not show a distinctive mode, selecting optimal model parameters is difficult. We decided to use a 20 neuron and 3 layer network with 26 epochs and 26 batch size for the model parameters in the following experiment, as it provides a relatively good result in terms of RMSE and R, and the balance of time spent on training the model. Admittedly, the parameters used may not be the most optimal ones and potentially result in the calibration result being less accurate. However, it reveals one drawback of using an ANN-based method, that obtaining the optimal parameter settings is often difficult.

4.3 Comparison of Calibration Methods

Once the model parameters for both methods have been determined, the models are evaluated in three ways. Firstly, the variability of model generation is evaluated. Then, the model outputs are cross-compared by varying the training dataset and testing dataset. In contrast to the existing research in [19], the training and testing datasets are altered individually in this experiment; hence, the result can help to understand how different training and testing would affect the result of the calibration respectively. The data used in this section is identical to the previous section which has been discussed in Section 4.2.1.

4.3.1 Variability of Model Generation

With the determined parameters, we train and test both models using the same data for multiple iterations to determine the variability of the result during the model generation process. For this experiment, the dataset from Fishergate is firstly averaged using Algorithm 1 in Section 5.1.1. The training and testing datasets are then divided sequentially in the same way as the previous experiment, for which the first 2,000 samples of the dataset are used as training and the rest of the samples are used as testing. Both models are trained using the same model settings and the data over multiple iterations (1,000 iterations). The testing result in terms of RMSE, R and time spent for training are shown in Figure 4.14 to 4.15.

Figure 4.14 shows the results from the regression-based method. It is clear that a regression-based method would provide a consistent result as long as the model settings and the use of the data are identical. Hence, the RMSE and R show no variation over the 1,000 iterations. It is noticed that there is a variation on the time spent on training, but the variation is extremely small as it is below 0.15 seconds.

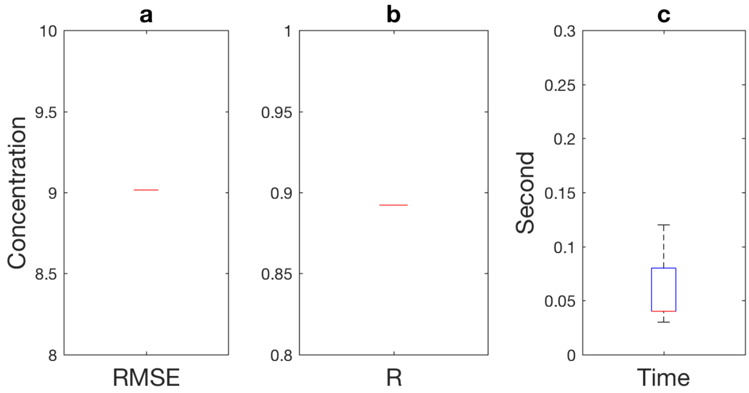


Figure 4.14: The variation of objectives over 1000 repetitions using the regression-based method

Figure 4.15 shows the results from the ANN-based method. In comparison to Figure 4.14, all three objectives show more significant variations. The result indicates that an ANN-based method would have a large variation in the model generation process, which would result in a large uncertainty in the calibration result.

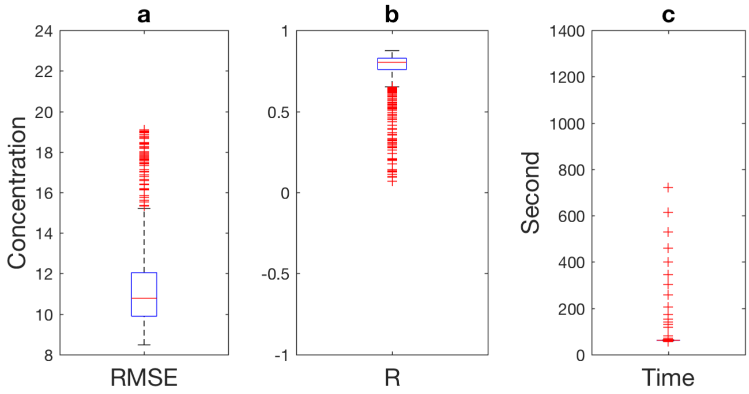


Figure 4.15: The variation of objectives over 1000 repetitions using ANN

In Figure 4.15-c, a few extreme outliers can be observed. They are suspicious, as the time spent on training the same model is expected to have much less variation. To further investigate that, the time spent on each repetition was plotted, as shown in Figure 4.16. It can be observed that the magnitude of spikes gradually increases with the number of iterations. In order to rule out experiment error, we performed the same experiment again, the result is similar in terms of the patterns, with the main difference being the spikes occurred in different iterations and different magnitudes. We further correlated the time to R and RMSE respectively to determine if the spikes caused any abnormality in those objectives. The correlations are shown in Figure 4.17.

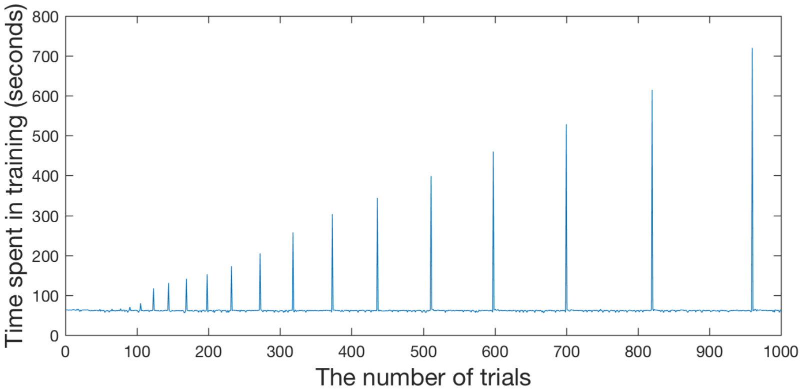


Figure 4.16: The time variation over 1000 repetitions

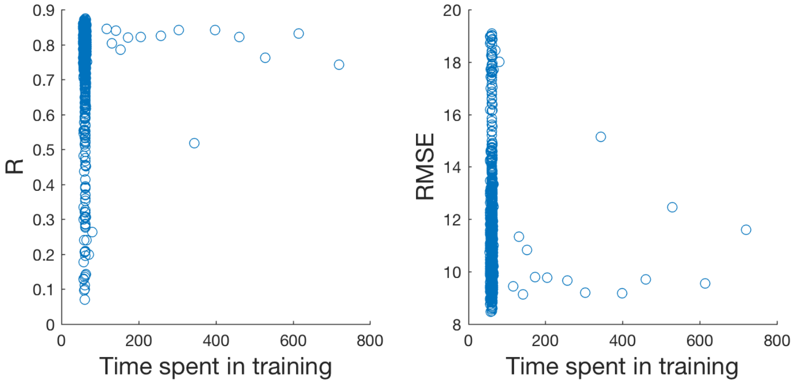


Figure 4.17: The correlation of (time vs R) and (time vs RMSE)

The figure shows that the iterations with high time cost do not have an observable impact on RMSE and R values. However, we noticed that during the experiment, memory consumption increases with the number of iterations. Therefore, we suspect the time spikes may be related to the Keras library or the garbage collection of the system. Since it does not affect the RMSE and R values significantly, and the root cause may be out of the scope of this work, we did not investigate it further. However, it suggests that applying an ANN using the existing library may introduce unknown errors into the process.

4.3.2 Comparing the Models under Different Scenarios

This section compares the model outputs that use different training and testing datasets in terms of data size. The results indicate the sensitivity of the methods for the different calibration scenarios. The experiment is firstly performed by varying the training dataset, followed by varying the testing dataset.

4.3.2.1 Varying the Training Dataset

This experiment is designed to understand how the increasing size of the training dataset would affect the calibration results for both methods. For this experiment, the dataset from Fishergate is firstly averaged using Algorithm 1 in Section 5.1.1. The processed dataset is then sequentially divided into 10 equally sized partitions with each partition having 10 percent of the data. The calibration model is determined by using the training dataset with different data partitions; and the result of the calibration is evaluated in the same testing dataset. All available parameters are considered in the calibration model, which are NO_2, O_3, NO, T and H. The classification and the use of the training and testing dataset are illustrated in Figure 4.18.

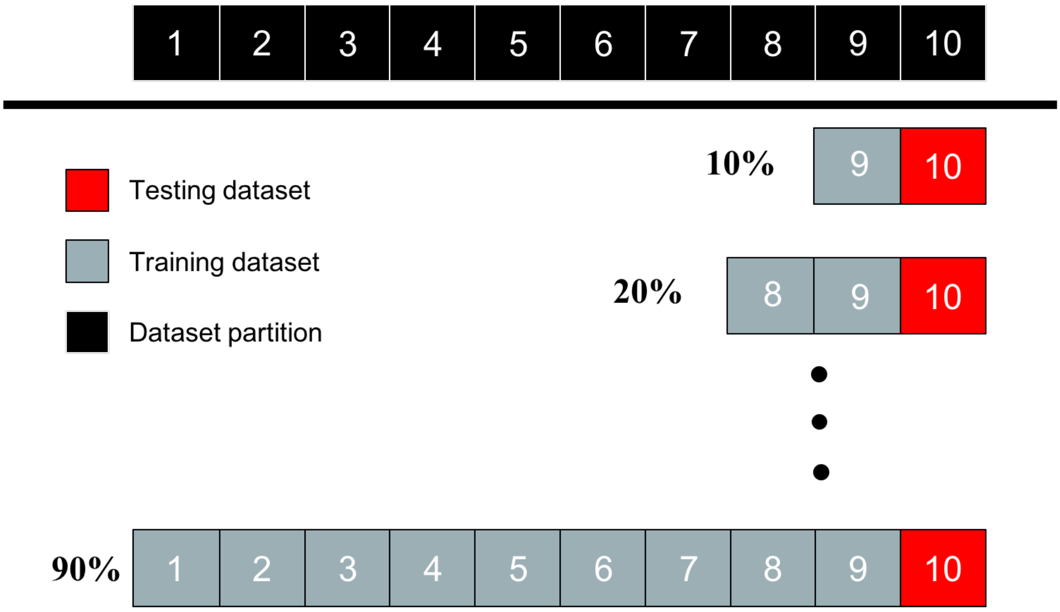
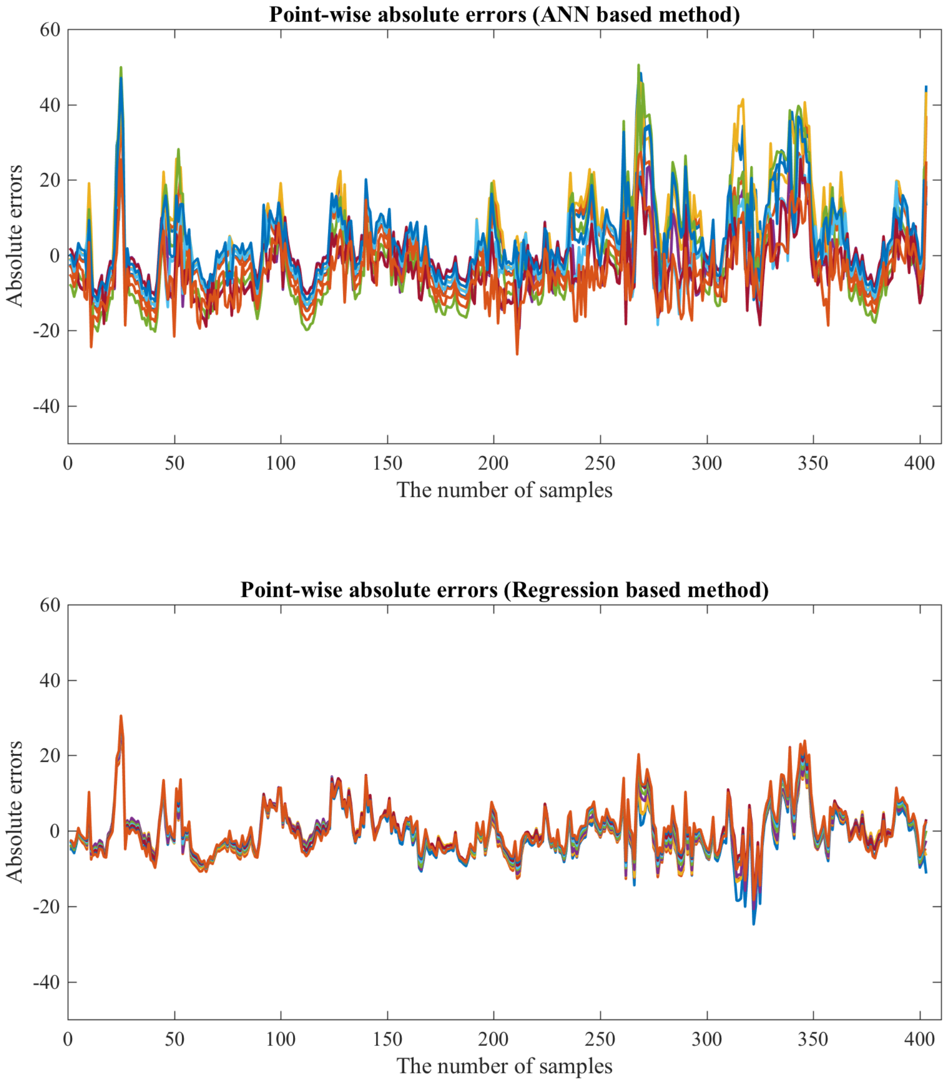


Figure 4.18: Varying the training datasets

Figure 4.18 shows how the data is divided into ten equal partitions, numbered from (1) to (10). It is noted that the partitions (1) to (10) follow the temporal order. For the testing dataset, the last partition (10) is used, and for the training dataset, different combinations of the partitions are applied. As illustrated in Figure 4.18, the training dataset steadily increased from 10 percent of the data to 90 percent of the data with each step being 10 percent. In order to preserve the temporal dependencies of the data, the first experiment uses Partition (9) (to preserve the dependencies with Partition (10) in the training). More data is added to the later experiments by going backwards from Partition (9) e.g. the second experiment uses Partitions (8) and (9). We label the different training datasets as 10% to 90% to simplify the labelling in the later plots.

The results in terms of the errors from the models using different training datasets are illustrated in Figure 4.19. From the figure, we can observe that errors from using the ANN-based method are less consistent than the regression-based method when the training datasets are increased. This suggests that the ANN-based method may be more sensitive to the training datasets. To further investigate, we present the errors using boxplots to show their distributions in Figure 4.20.


Point-wise absolute errors (ANN based method)

Point-wise absolute errors (Regression based method)

Figure 4.19: The absolute errors of the result when using different training datasets

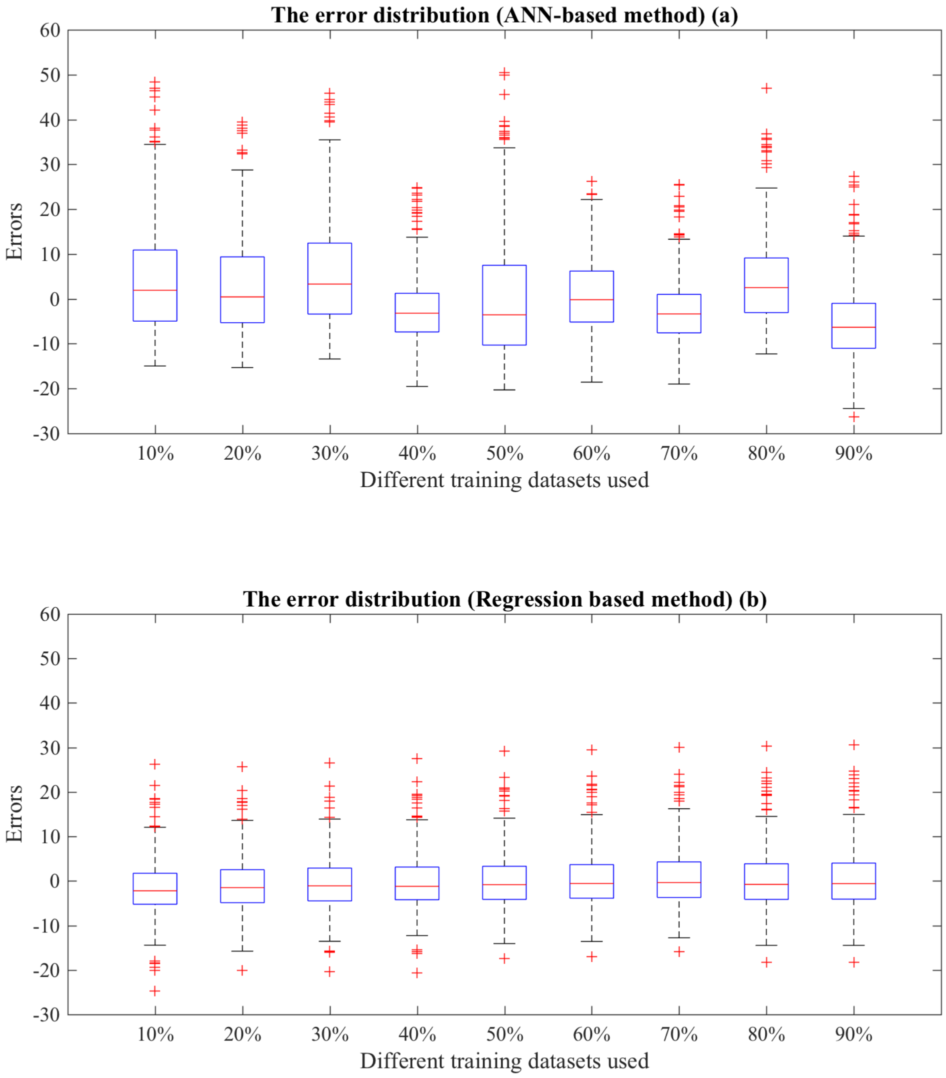


Figure 4.20: The boxplot of errors when using different training datasets

Comparing Figures 4.20 (a) and (b), the figures show that the variation of errors from using the regression-based method is generally less than the ANN-based method, and the errors are more consistent across different training datasets. The scatter plots of the results are presented in Figure 4.21, which show the results when using 10%, 50% and 90% of the data for the training. From the figure, we observe that the plots from the ANN-based methods shows discrete pattern and not consistent over the different training datasets. By contract, the results from the regression-based methods are more consistent over the different training scenarios. The result suggests the importance of looking at the correlation between reference and calibrated value when evaluate the calibration result.

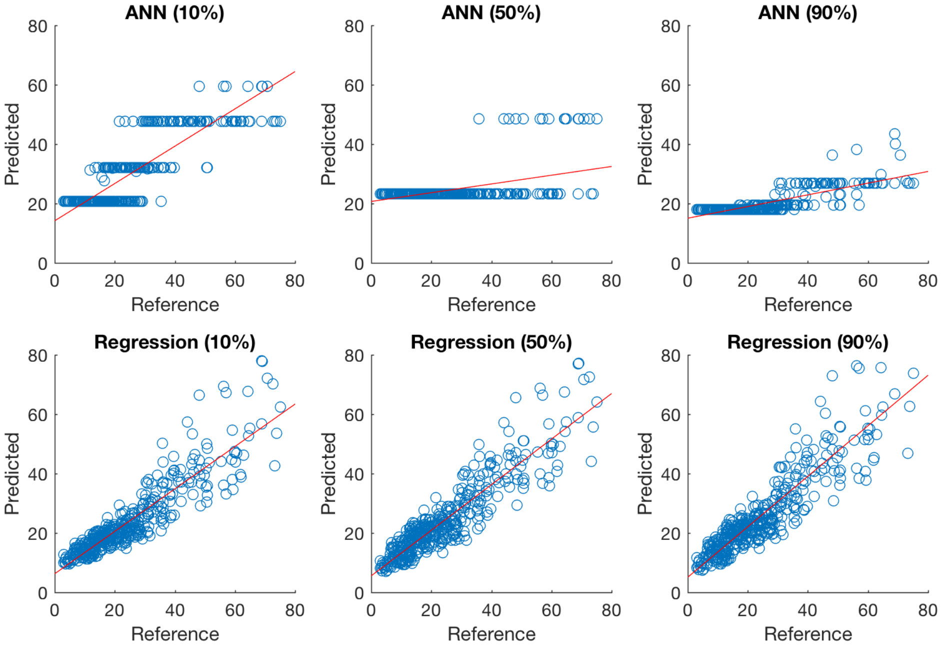


Figure 4.21: The scatter plots of the result when using different training datasets

We further present and plot the mean and the standard deviation of the errors, which are shown in Table 4.2 and Figure 4.22. In the figure, the line indicates the changes of the error mean and the bar stands for the standard deviation. Ideally, errors closer to zero and a smaller standard deviation indicate a better result.

Table 4.2: Mean and standard deviation of the errors by varying the training datasets

datasets	1	2	3	4	5	6	7	8	9
ANN-based method									
Mean	-5.09	3.81	-2.56	0.44	0.14	-2.30	5.78	2.54	4.57
STD.	8.34	9.92	7.42	7.92	13.86	7.72	12.05	10.58	12.59
Regression-based method									
Mean	0.71	0.56	0.85	0.56	0.11	-0.23	-0.45	-0.81	-1.44
STD.	6.75	6.68	6.52	6.49	6.40	6.35	6.26	6.27	6.37

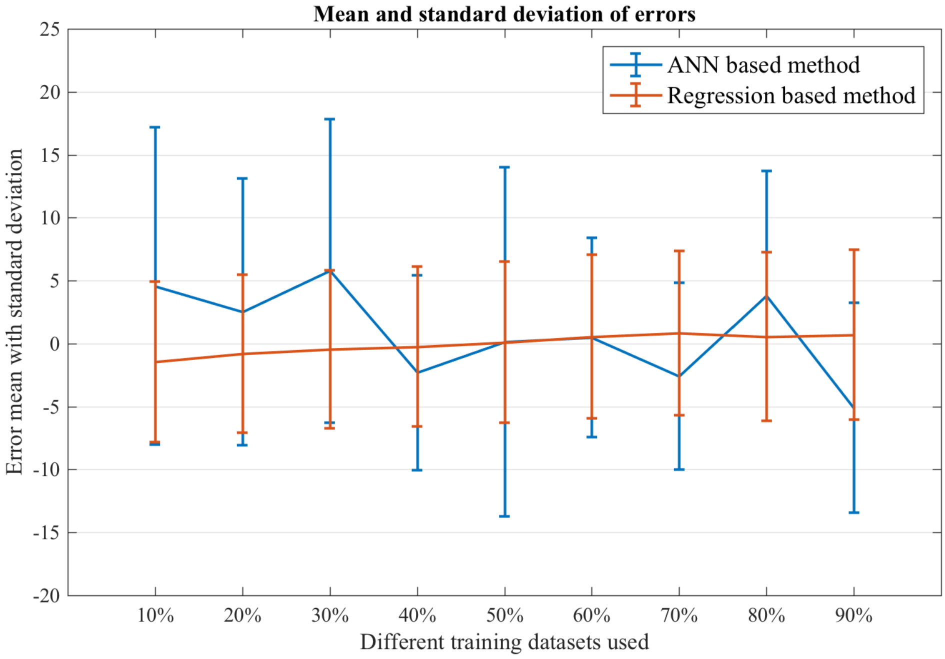
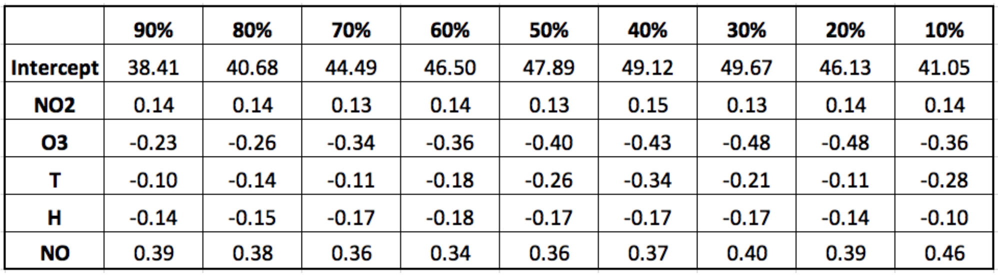


Figure 4.22: Mean and standard deviation of errors for different methods and training datasets

In Figure 4.22, we can observe that the regression-based method over predicts when the training dataset is relatively small (smaller than 50 percent of the dataset), and under predicts when the training dataset is relatively large (larger than 50 percent of the dataset). It suggests that too large or too small a training dataset is not ideal for regression-based calibration.

We further investigated the change in the coefficients from the regression-based method to understand how the variation of the training dataset would affect the calibration model. The results are shown in Figure 4.23.



	90%	80%	70%	60%	50%	40%	30%	20%	10%
Intercept	38.41	40.68	44.49	46.50	47.89	49.12	49.67	46.13	41.05
NO2	0.14	0.14	0.13	0.14	0.13	0.15	0.13	0.14	0.14
O3	-0.23	-0.26	-0.34	-0.36	-0.40	-0.43	-0.48	-0.48	-0.36
T	-0.10	-0.14	-0.11	-0.18	-0.26	-0.34	-0.21	-0.11	-0.28
H	-0.14	-0.15	-0.17	-0.18	-0.17	-0.17	-0.17	-0.14	-0.10
NO	0.39	0.38	0.36	0.34	0.36	0.37	0.40	0.39	0.46

Figure 4.23: The change of regression coefficients with different training datasets

The figure shows that the coefficients of the regression did not change significantly in most of the parameters when different datasets were used for the training. However, we can observe a larger variation of the coefficient in parameters O_3 and T. The overall result shows a model with good stability. It is noted that this result would be data dependent, so it may not represent a general trend.

It is understood that the result for the ANN-based method will also be affected by the model generation process, which makes it difficult to determine which change in result is related to the varying of the dataset. We have demonstrated the model generation process would have a large impact on the error mean as the RMSE value would vary significantly. However, we identify that the model generation process would have much less influence on the standard deviation of the error, which suggests that it can be used to reflect the actual change in the results. To prove this, we trained an ANN 20 times, for which the process is identical to Section 4.3.1. The mean and the standard deviation of the error from the 20 iterations were obtained and are summarised in Table 4.3.

In the table, the coefficient of variance (CV) was used to indicate the significance of the variance. CV is a measure of relative variability and it is the ratio of the standard deviation (σ) to the mean (μ), which is calculated using Equation 4.5. A smaller CV indicates a smaller variation

Table 4.3: Summarised mean and standard deviation of the model error

from the 20 means			from the 20 STD.		
$STD._{mean}$	$Mean_{mean}$	CV_{mean}	$STD._{std}$	$Mean_{std}$	CV_{std}
1.7407	-1.526	-114.06	1.231	11.172	11.02

in the result.

$$CV = \frac{\sigma}{\mu} * 100 \tag{4.5}$$

Since CV_{std} is considerably smaller than CV_{mean}, it suggests the variation of the standard deviation of the errors would be much less affected by the model generation process. As a result, we use the standard deviation of the error to approximate and compare the response of both models, which is presented in Figure 4.24.

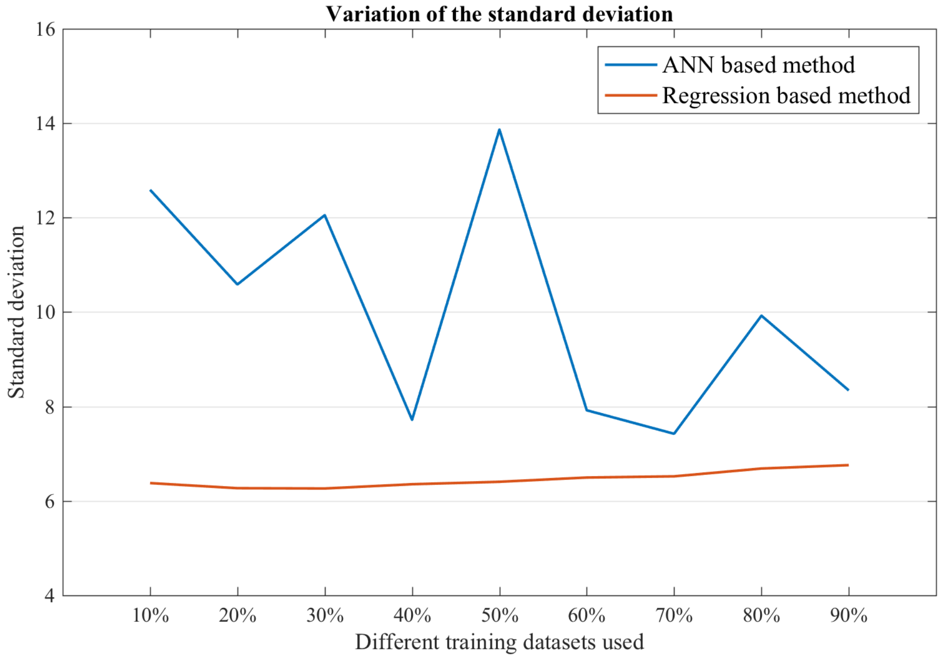


Figure 4.24: The variation of the standard deviation for different methods and training datasets

From Figure 4.24, we can observe that the variation of the standard deviation for the regression-based method is smaller than for the ANN-based method. This suggests that the regression-based method is better than the ANN-based method, as the errors have less variation. However, even though the variation in errors between the two methods is considerably different, the methods show the opposite trend when more training datasets are used. The error for the ANN-based method shows a trend of declining with more historical data used for training, whereas the error for the regression-based method gradually increases. This suggests that an ANN-based method would potentially benefit from a larger training dataset, where a regression-based method is more suited for smaller datasets. However, it is noted that such improvements may be unobservable and insignificant with respect to the variation of the ANN model.

4.3.2.2 Varying the Testing Dataset

In this section, the experiment is designed to understand how the calibration result is affected by increasing the size of the testing dataset. For this experiment, the dataset from Fishergate is firstly averaged using Algorithm 1 in Section 5.1.1. The processed dataset is then sequentially divided into 10 equally sized partitions with each partition having 10 percent of the data, for which process is identical to the one in Section 4.3.2.1. In contrast to the previous experiment, the calibration model is determined by using the same training dataset, and the testing is performed on different combinations of the datasets. The use of the training and testing dataset is explained in Figure 4.25. All available parameters in the datasets are considered in the calibration model, which are NO_2, O_3, NO, T and H. It is noted that since the training dataset is identical (Partition (1)), the same ANN model is applied in this experiment. Thus, the results are not affected by the model generation process, and the variation in results would be directly associated with the use of different datasets.

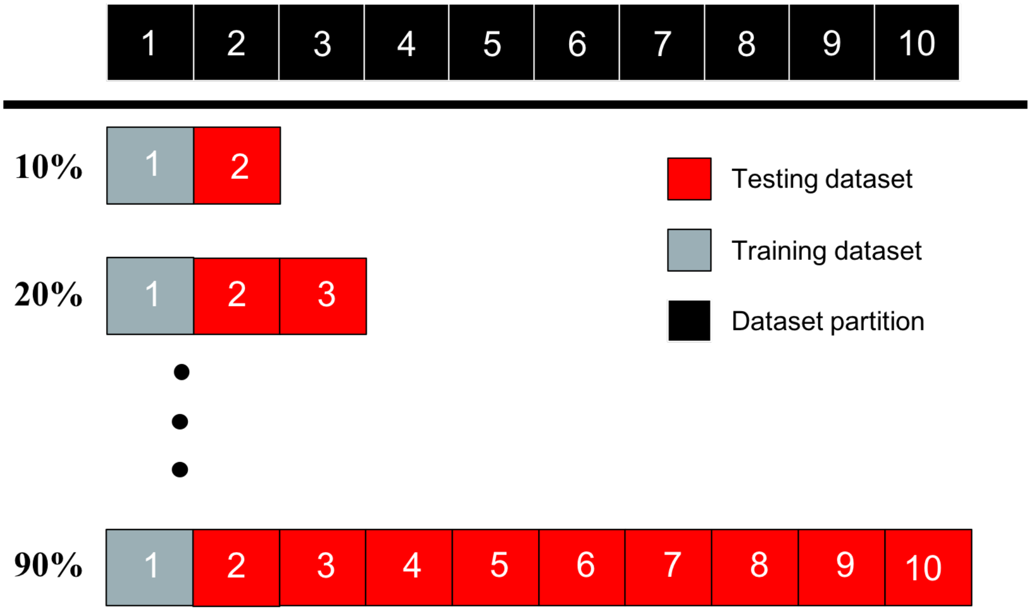


Figure 4.25: Training models by varying the testing datasets

The calibration results are illustrated in Figure 4.26, where the boxplots represent the error distribution and the x-axis indicates that the testing dataset increases from 10 percent to 90 percent of the dataset according to Figure 4.25. From the figure, the error distributions between the two methods do not have an observable difference, which suggests that varying the testing dataset would have a similar impact on both methods.

The scatter plots showing the result of varying the testing datasets are presented in Figure 4.27. Even though the error patterns of both methods are similar. as shown in Figure 4.26, the correlations between the reference and their calibrated results are significantly different. The correlations for the ANN-based method are much worse than the regression-based method. The results suggest that it is important to look at the scatter plot when comparing the calibration results as the errors can be misleading.

Using only partition (2) as the testing dataset shows the best calibration errors compared to the others, which suggests that the calibration function obtains a better result if the testing dataset and the training dataset are close in time and have a similar data size. To further in-

Table 4.4: Mean and standard deviation of the errors by varying the testing datasets

datasets	1	2	3	4	5	6	7	8	9
ANN-based method									
Mean	-0.54	1.51	4.22	6.08	7.52	6.74	7.00	7.00	6.55
STD.	8.64	10.28	11.52	12.62	12.97	12.93	13.44	13.53	13.57
Regression-based method									
Mean	-0.73	1.35	3.17	4.39	5.83	5.83	6.54	7.04	7.19
STD.	7.46	8.63	9.45	10.15	10.49	10.18	10.41	10.41	10.26

vestigate, we plot the mean and standard deviation of the errors in Figure 4.28, for which the data is also presented in Table 4.4.

Figure 4.28 shows the mean value of the error gradually increases with more testing data used. It suggests that both calibrations would degrade over time with a similar trend. However, it is noted that the ratio of the decrease is higher at the beginning and gets lower towards the end. The results imply that the degradation of the calibration in an urban environment may not be linear. However, due to the availability of the data (no other dataset available), we did not investigate it further. We finally plot the standard deviations of the errors in Figure 4.29, the result is in line with Figure 4.28.

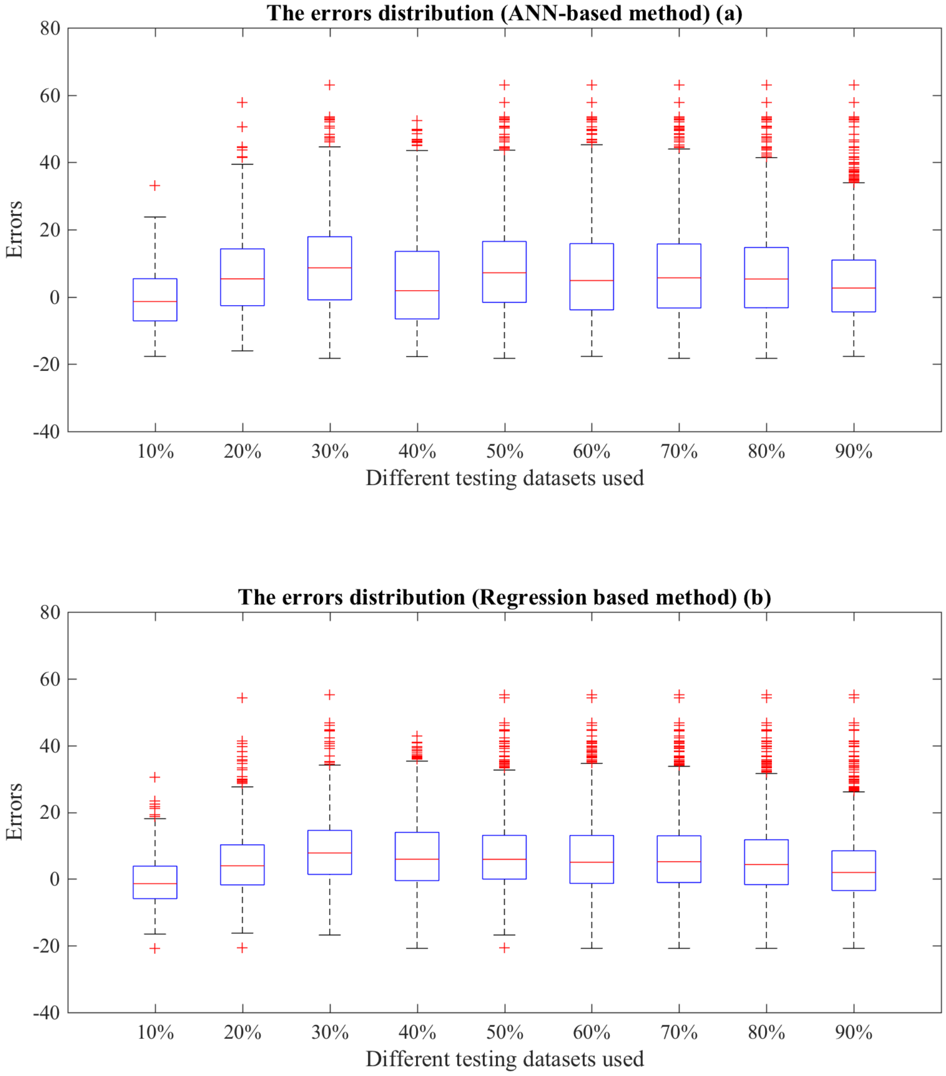


Figure 4.26: The boxplot of errors when using the different testing datasets

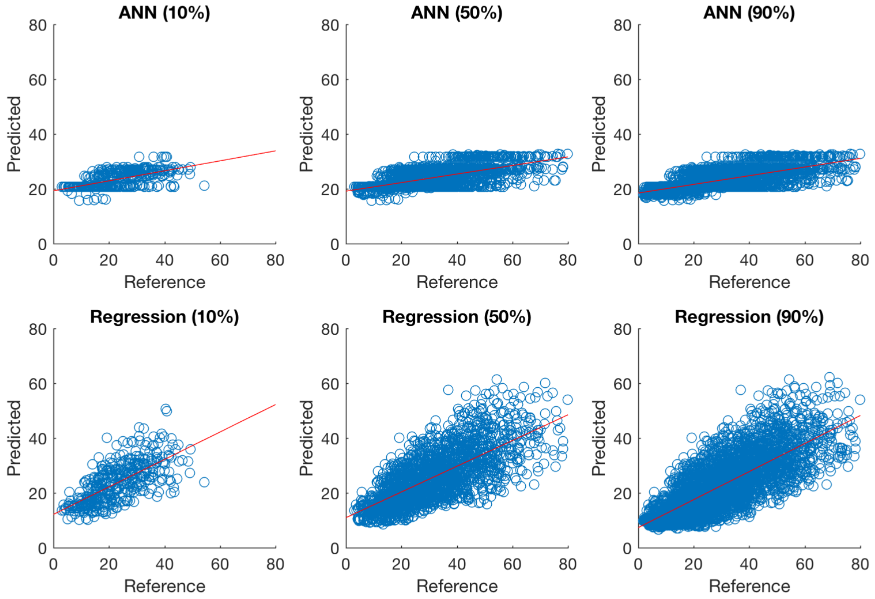


Figure 4.27: The scatter plots of the result when using different testing datasets

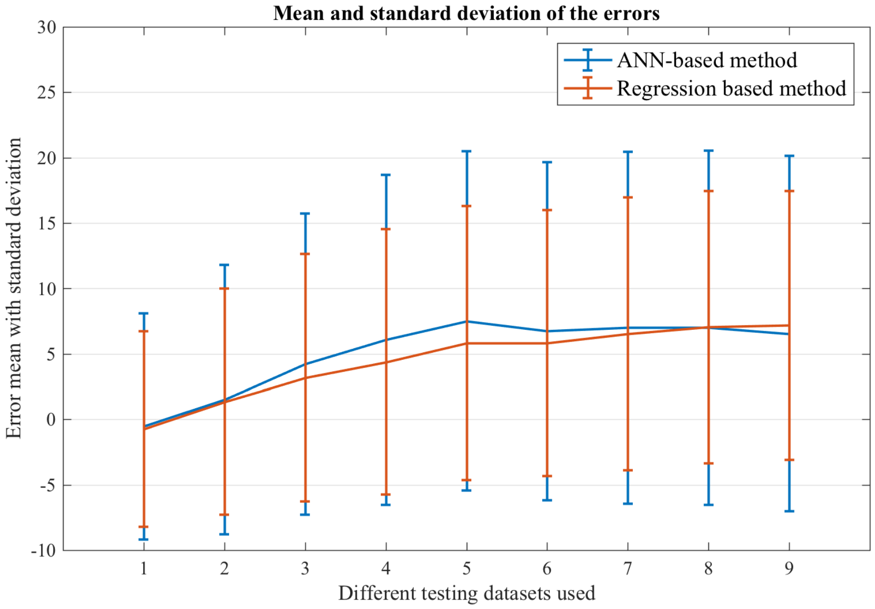


Figure 4.28: Mean and standard deviation of the errors for the different methods and testing datasets

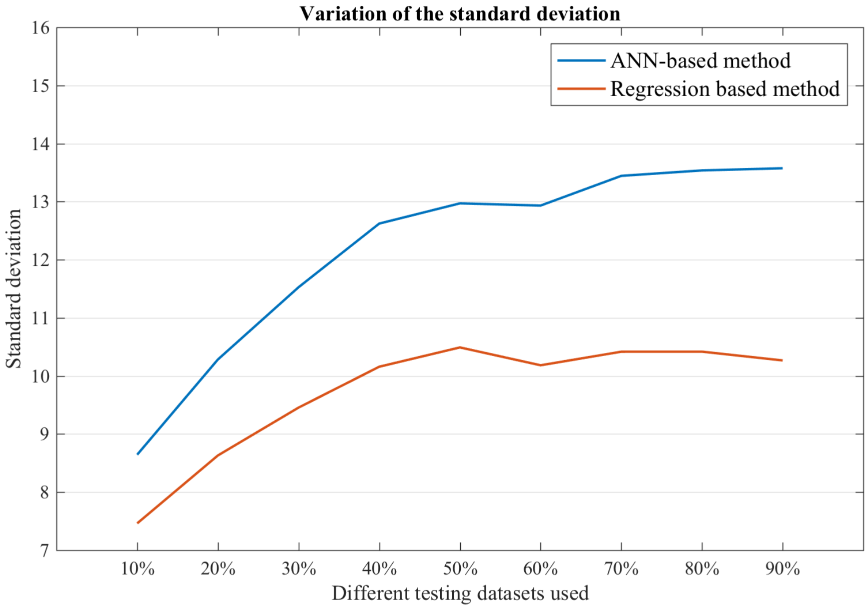


Figure 4.29: The variation of the standard deviation for the different methods and training datasets

4.4 The Limitation of Validity

In this experiment, the ANN-based method was constructed using Python with a well-built library, which hindered some of the programming features and logic. As a result, the causes of abnormal results are difficult to identify and correct, as discussed in Section 4.3.1.

Determination of the optimal model parameters is often constrained by practical limitations. It is understood that the model parameters are selected from a large parameters space, and trial by error to determine the optimal model output is the best practice for such a purpose. Since it is not feasible to test all combinations of the parameters, the selected model parameters may not always be globally optimal. As a result, the calibration results would be compromised accordingly.

It is clear that the model generation process will introduce variations in model output discussed in Section 4.3.1. We determined that the standard deviation of the error is much less affected than the error mean in the process, and used the standard deviation to approximate the calibration result. However, such an evaluation may not be appropriate for applications that are interested in the error mean.

Finally, the evaluations in Section 4.3.2.1 and 4.3.2.2 were only used with one dataset. Hence, statistical confidence in the results may not be obtained.

4.5 Summary

This section provides a systematic comparison of state-of-the-art calibration methods. There are a regression-based method and an ANN-based method. The comparison demonstrates the practicality of using both methods in terms of constructing calibration models and testing variations in the model generation processes. We further used multiple training and testing data to determine the sensitivity of each method to these data. The results show that the ANN-based method is sensitive to the use of model parameters and random variations in the model generation process, which could lead to a large variation in the calibration results. By contrast, the regression-based method provides a more predictable result and requires much fewer computational resources.

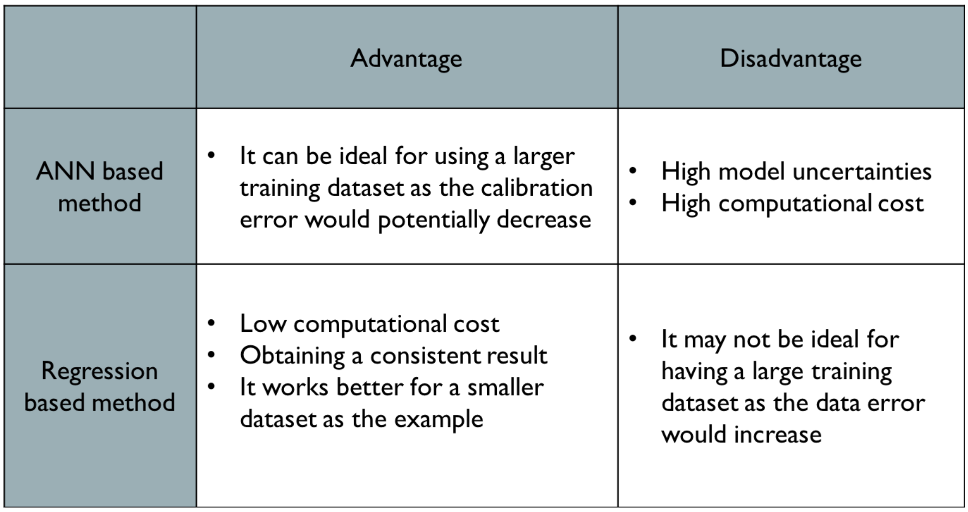
The evaluation performed by varying training datasets suggests that the ANN-based method would benefit from using a larger training dataset, whereas the regression-based method is more suited for a relatively small training dataset. By varying the testing datasets, the calibration results for both methods gradually decreased as more testing dataset were used. The results suggest that both calibrations would degrade over time and their degradation would be similar. Our experiment also indicates that the degradation of calibration in an urban environment may not be linear, but would require more evidence to confirm it. More importantly, our analysis shows the importance of looking at the scatter plots when comparing calibration results, as calibration errors can be misleading.

Finally, we summarise the advantage and disadvantage of both methods determined in our experiment in Figure 4.30, and answer the research question which is quoted below:

> *Research Question 1: Which is the appropriate calibration*

method (Regression or ANN) considering the needs of our application?

Considering the needs of our application are a light-weight process that can work on a relatively small dataset, we believe that a regression-based method would be more appropriate in this work. However, it is noted that the existing regression-based methods have a lot of limitations, as discussed in Section 2. Hence, they may not be directly applicable to our application.



	Advantage	Disadvantage
ANN based method	• It can be ideal for using a larger training dataset as the calibration error would potentially decrease	• High model uncertainties • High computational cost
Regression based method	• Low computational cost • Obtaining a consistent result • It works better for a smaller dataset as the example	• It may not be ideal for having a large training dataset as the data error would increase

Figure 4.30: The advantages and disadvantages of the two methods

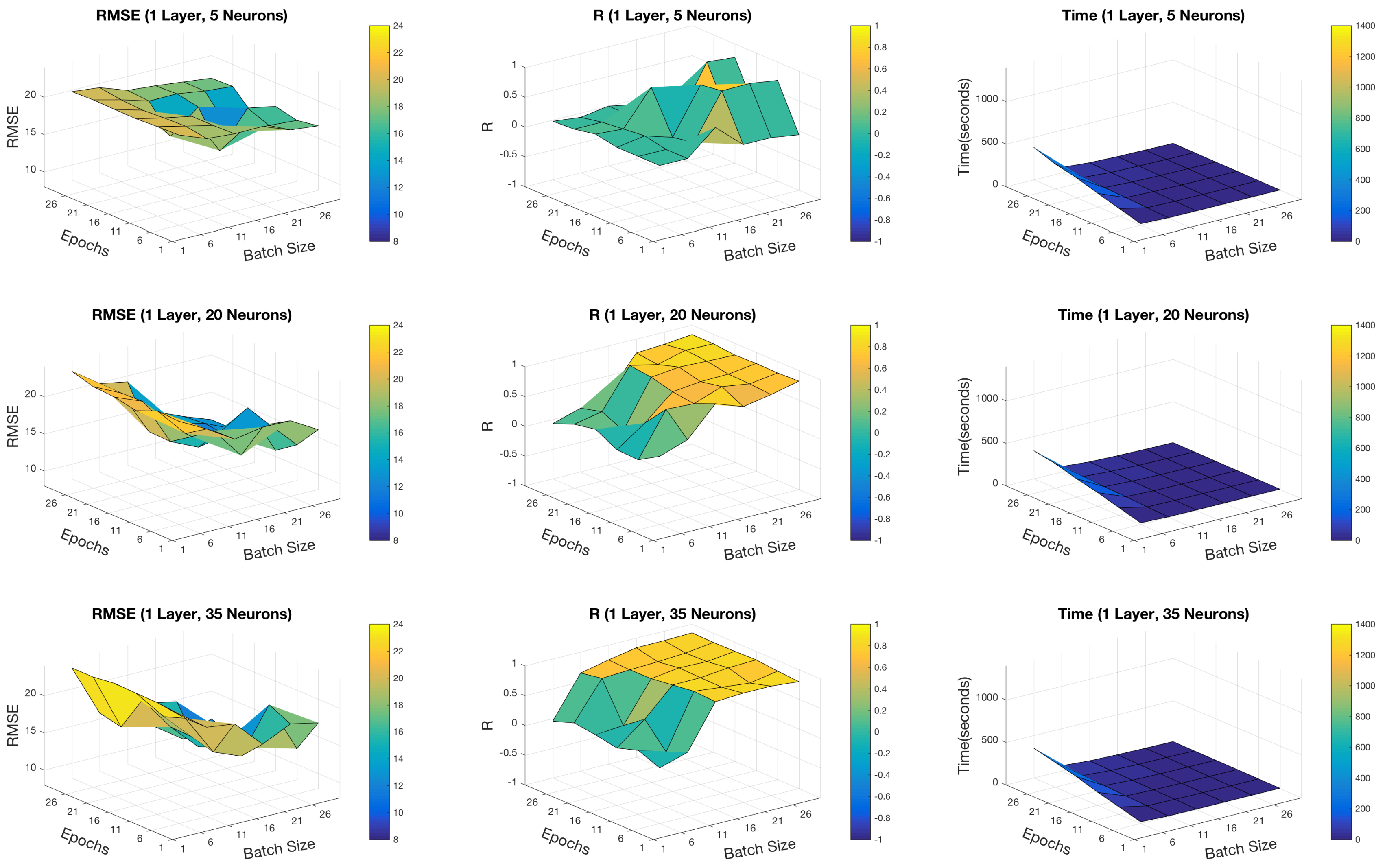


Figure 4.9: Results from 1 Layer ANN

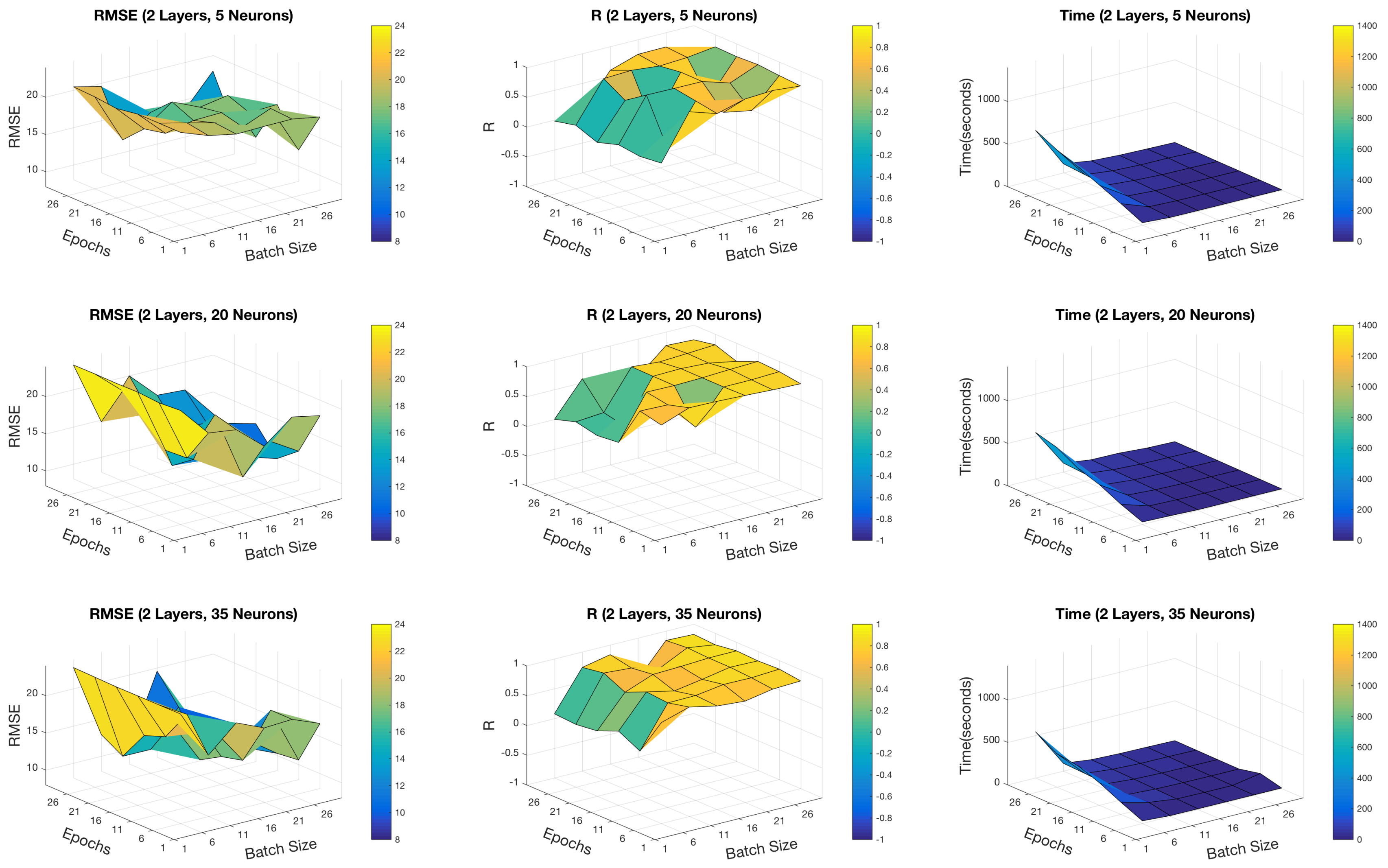


Figure 4.10: Results from 2 Layer ANN

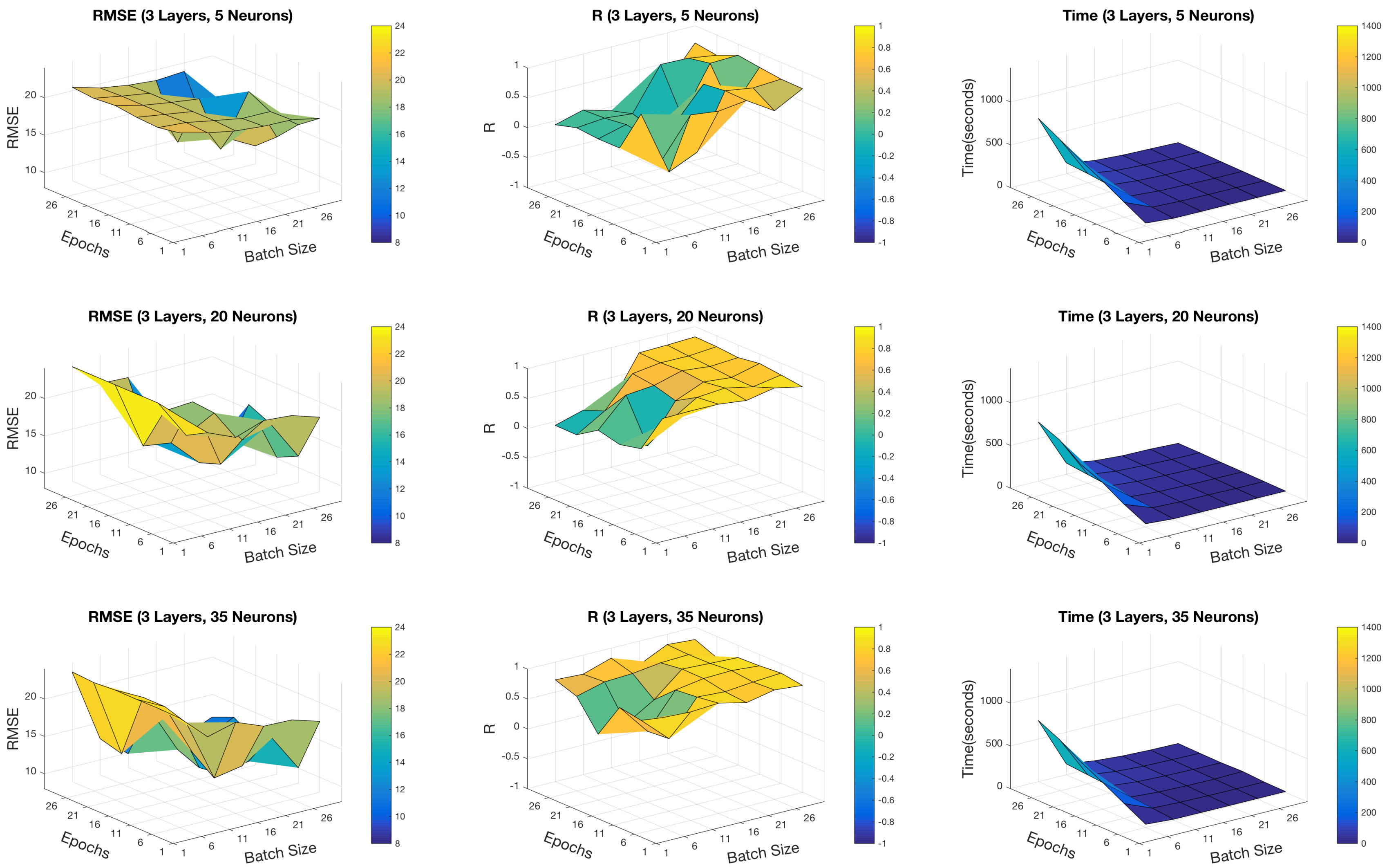


Figure 4.11: Results from 3 Layer ANN

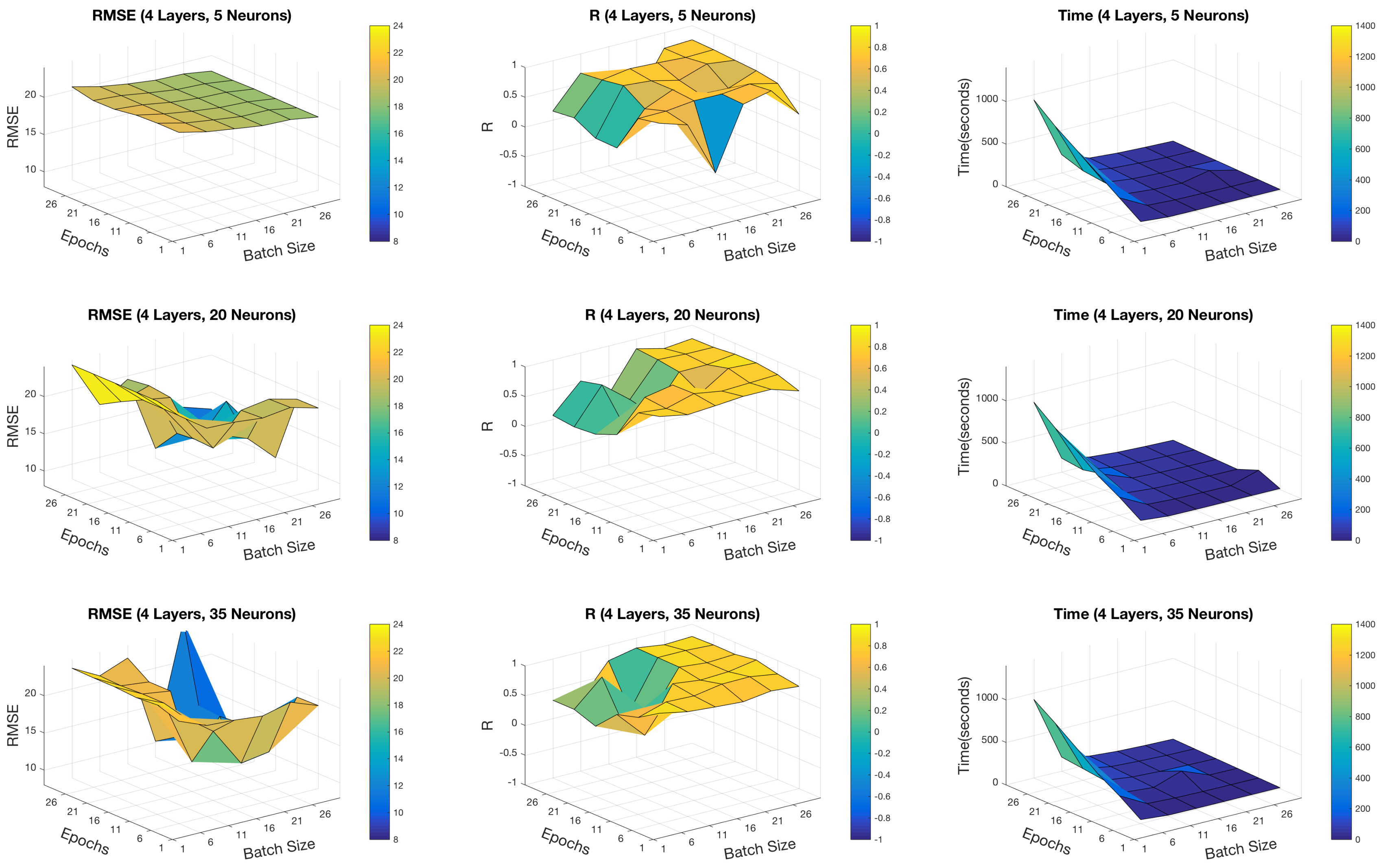


Figure 4.12: Results from 4 Layer ANN

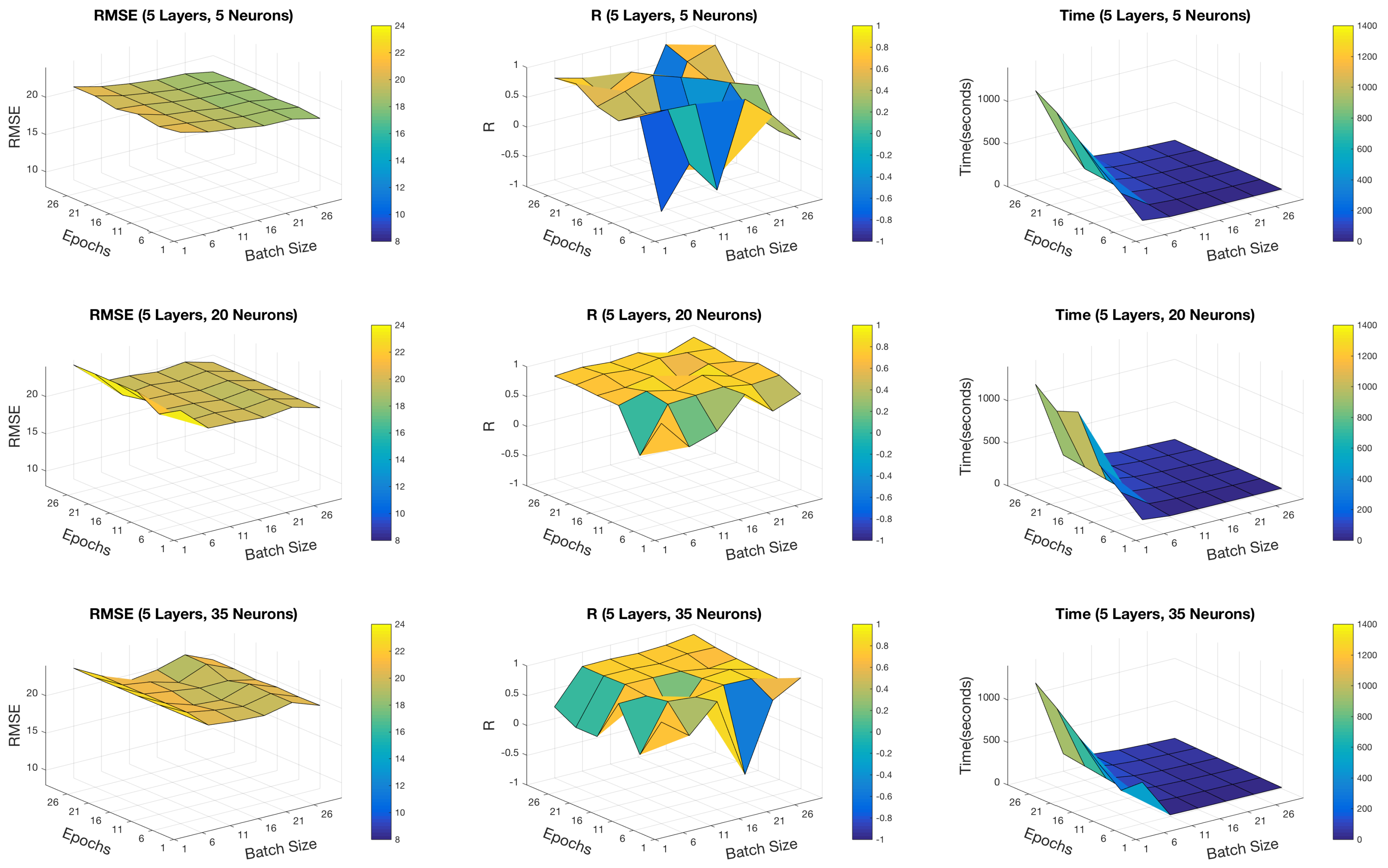


Figure 4.13: Results from 5 Layer ANN

Chapter 5

Regression based Method for the Calibration of Sensors

This chapter aims to answer the second research question, which is quoted below:

> *Research Question 2: How can we ensure calibration results by properly using supporting parameters?*

In Chapter 3, we demonstrate that environmental interference can have a significant impact on the response of the sensors. As a result, many studies reviewed in Chapter 2 use supporting parameters in the calibration as it is believed to improve calibration results by subtracting those interferences. However, according to the review, the use of supporting parameters is not only dependent on the availability of the parameters, but also related to many other factors, e.g. the current environmental conditions. Since failing to use appropriate parameters may result in calibration errors remaining large, and using an inappropriate parameter would bias the calibration results [30, 53, 78], the selection of supporting parameters is important for the calibration process.

As demonstrated in Chapter 4, a regression-based method is more appropriate for this research. However, the current calibration methods will not automatically select the supporting parameters. As a result, the existing calibration cannot be directly applied to sensors situated in different environmental conditions as the use of supporting parameters could be different. Performing a manual selection of the supporting parameters before each calibration is able to solve such an issue [20, 49]. However, the manual selection of supporting parameters is not practical and desirable because 1) there could be a large number of parameters to select from and 2) the calibration process may frequently be applied. Hence, a calibration method that can automatically select the optimal supporting parameters from the available dataset is important.

This section proposes a novel regression-based calibration method. In contrast to the existing method, the proposed method is able to automatically select the optimal supporting parameters from an available dataset. Hence, the method is believed to be less sensitive to a change of environmental conditions.

In the rest of this chapter, the method of calibration is firstly discussed in Section 5.1. Then, the evaluation is carried out in the Section 5.2 using datasets from both mild and harsh environments, and the results are cross-compared with the state-of-the-art method which has been described in the previous section. Following that, we discuss the limitations of validity in Section 5.3 and summarised the findings and answer the research questions in Section 5.4.

5.1 Calibration Method

The proposed calibration method has three main steps. The first step is to pre-process the data making the data suitable for the process. Then, a two-way interaction term is introduced to the model, which is believed to maximise the relationship between inputs. Finally, stepwise regression is introduced to construct the calibration model, which can statistically use supporting parameters from the available dataset.

5.1.1 Data Pre-processing

It is known that the temporal resolution from most regulatory sites is an hour, whereas most low-cost sensors provide data at a much higher temporal resolution, e.g. 20 seconds for ELM units. Hence, for the regression-based method, it is important to aggregate the ELM data (20 seconds) into the same resolution as the reference (hourly).

In this work, the data from the low-cost sensors is averaged by the time-stamp. A wide range of hourly windows were tested to average the data. The correlations to the reference from using different hourly windows shows no significant difference. Thus, it suggests that the use of different moving windows is not significantly important unless it is known how the reference data is produced. For this work, the hourly data is averaged based on the window from the current whole hour until the next whole hour. For example the hourly averaged value for 12:00:00 is obtained from the samples between (>=) 12:00:00 and (<) 13:00:00.

For the techniques of data aggregation, arithmetic mean and median are commonly used. The arithmetic mean is the sum of the received values divided by the number of received values. However, by definition, the arithmetic mean is sensitive to the sample size, which implies that the mean will have a different confidence if the sample size is different. Considering the number of samples in an hour window is likely to be significantly different due to data gaps, using the arithmetic mean can considerably affect the confidence level of the averaged value. Moreover, the mean is also sensitive to extreme values. For example, the mean value could be largely influenced by anomalies with an extreme value. This can be a particular problem for data that contains high values and biased anomalies (e.g. non zero means). However, it does not imply that using a median value is always a better option. As the median value is a single value, it will not be representative for other samples. If the spikes are caused by real events, taking the median value would ignore important information. Moreover, if the percentage of anomalies is more than 50% of an averaged sample, the median value is likely to be biased.

For this research, as shown in Section 3.3, the number of sample received in an hour can be very inconsistent and the anomalies are unlikely to be more than 50% of the hourly samples as reported from the existing literature. Hence, we believe that the median is more appropriate for the aggregation in our application. The process of data pre-processing is illustrated in detail in Algorithm 1

ALGORITHM 1: Pseudo code for data pre-processing

```
Data:
1) Dataset from low-cost sensors, $D_{m \times n}$. It has the size of $m \times n$.
(n indicates the number of columns; $m$ indicates the number of rows)
2) The first column is a time array, which stores the time when the sample was taken. The rest of the column stores the measurements taken at the corresponding time. The number of rows indicates the number of samples
3) Reference, $Ref_{r \times 2}$, The first column is a time array, $t_i \subset T$. $T(:,1)$, which stores a consistent time-stamp with the date on an hourly basis (Date.Month.Year 00:00:00,Date.Month.Year 01:00:00,Date.Month.Year 02:00:00 ...). The second column stores the reference value for the parameter of interest. (Hourly reference which may contain NAN).
5) for $i$ = 0 to m-1 do
    for $j$ = 2 to $n$ do
        $temp_D$ = D(find($t_i \leqslant D(:,1) < t_{i+1}$),j) (Determine all values that measured within that hour)
        if the number of samples in $temp_D < 5$ then
            T(i,j) = NAN (Not a Number)
        else
            T(i,j) = nan-median($temp_D$) (The process ignores the NAN)
        end
    end
end
Result: Hourly averaged data for low-cost sensors (contains NAN)
6) Join the $Ref$ with T according to the time-stamp and remove all NAN instances in the dataset.
Result: The dataset that the first column stores the time and the second column stores the reference data. The rest of the columns are the averaged data from low-cost sensors.
```

It is noted that if there is not enough data to be averaged (the number of samples within a window is less than 5) or if data gaps occurred in the reference as shown in Figure 3.4, the relevant data from the corresponding sensor will be removed, for consistency.

5.1.2 Two-way Interaction

Once the data has been aggregated, it is ready to process. For example, according to Section 4.1.2, a calibration model using just one supporting parameter constructed by the current method is presented in Equation 5.1.

$$Y(i) = \beta_0 + \beta_1 \cdot X_1(i) + \beta_2 \cdot X_2(i) + \varepsilon(i) \tag{5.1}$$

Assuming X_1 is the parameter of interest, and X_2 is the supporting parameter. The equation indicates that the variation of X_1 and X_2 is independent, as every one unit increment of Y is constantly associated with β_1 units of X_1 and β_2 units of X_2. This suggests that the current method would not consider the potential dependency between the inputs.

To solve that issue, an interaction term, which is a multiplication of any two variables is used in our method. The interaction term is also known as a moderation term [36]. Adding an interaction term onto the Equation 5.1, the result is shown in Equation 5.2:

$$Y(i) = \beta'_0 + \beta'_1 \cdot X_1(i) + \beta'_2 \cdot X_2(i) + \beta'_3 \cdot (X_1(i) \cdot X_2(i)) + \varepsilon(i) \tag{5.2}$$

which can be re-written as Equation 5.3:

$$Y = \beta'_0 + (\beta'_1 + \beta'_3 \cdot X_2(i)) \cdot X_1(i) + \beta'_2 \cdot X_2(i) + \varepsilon(i) \tag{5.3}$$

In this case, the variable X_1 is now associated with the variable X_2 as the variation of the X_2 would impact the coefficient of the X_1. Hence, we believe that the calibration result can benefit from using the interaction terms, as dependencies between inputs are now considered in the model.

It is noted that the interaction terms are also considered to be supporting parameters in the following process. As a result, the number of available parameters would increase dramatically, which emphasises the importance of using an automatic process to make the parameter selection.

5.1.3 Stepwise Regression

In this section, we adopt stepwise regression to systematically select the useful parameters and calibrate the sensor.

Stepwise regression is similar to multivariate regression with the key difference being that it performs a systematic selection of inputs and only uses parameters that make a positive contribution to the calibration.

The method starts with fitting a model using just one input. At each step, the sum of squared error (SSE) and the p-value are calculated to test models with and without a new term. If the term is not currently in the model, the null hypothesis is that the added new term would have a zero coefficient in the model. If there is sufficient evidence to reject the null hypothesis, the term with the lowest p-value is added to the model. Conversely, if a term is currently in the model, the null hypothesis is that the term has a zero coefficient. If the null hypothesis fails to be rejected, the term with the highest p-value is removed from the model. The method proceeds as follows:

1. Construct the initial model using just one term.

2. Terms not in the current model have p-values less than a threshold ($p < 0.05$), add the one with the lowest p-value and repeat this step; otherwise, go to step 3.

3. Terms in the model have p-values less than a threshold ($p > 0.05$), remove the one with the highest p-value and go to step 2; otherwise, end.

Since the method terminates when no single step improves the model, a different sequence of steps would not lead to a better result. Thus, the sequence of adding the parameters is unlikely to affect the result.

The overall method for the proposed sensor calibration is illustrated in Algorithm 2.

ALGORITHM 2: Pseudo code for the method

```
Data: Data from Algorithm 1 as $Data_{m \times n}$, Reference:
      $Y = Data(:,2)$  Uncalibrated data trace: $Data(:,3)$  Other
      monitored parameters : $Data(:,4:end)$
for $i$ = $3$ to $n$-$1$ do
    for $j$ = $i$+$1$ to $n$ do
        $terms_{i,j} = x_{j_{m \times 1}} \; x_{i_{m \times 1}}$
    end
end
Result: Obtaining two-way interaction terms for all parameters,
        $terms_{m \times t}$, t is for number of interaction terms
$X_{m \times (t+n-2)} = [Data(:,3:end) \; terms(:,:)]$;
Result: Combining interaction terms and measured parameters as
        independent variables, X
while improvement can be determined do
    (1)Constructing the initial model using just one term.
    (2)Terms not in the current model have p-values less than a
    threshold ($p < 0.05$), add the one with the smallest p-value
    and repeat this step; otherwise, go to step 3.
    (3) Terms in the model have p-values less than a threshold ($p
    > 0.05$), remove the one with the largest p-value and go to step
    2; otherwise, end.
end
Result: Using stepwise regression to determine the use of
        variables and the calibration function
```

5.2 Calibration Evaluation

In this section, the proposed method is evaluated in both the mild and harsh environments and the result is compared to the state-of-the-art method discussed in Section 4.1.2.

The evaluation performed in the mild environment focuses on the quantitative analysis of the method, which is to determine how the performance of the proposed method compares to the existing method across multiple sensor units. The evaluation in the mild environment first compares the results from using datasets that contain different set of parameters. Then, it investigates how the results of the calibration would be affected if the characteristics of the testing dataset are different from the training dataset. Those two evaluations use two months' worth of data from eleven ELM units deployed on WACL. The monitored parameters are NO_2, O_3, H for relative humidity, T for temperature, *dust* for particulate matter PM_{10} and $PM_{2.5}$, VOC, and *noise* for the magnitude of sound in decibels.

By contrast, the evaluation in the harsh environment is to demonstrate that the proposed method can further improve the calibration error in a typical urban environment. Since sensors at the Fishergate monitor less parameters than those in WACL, the evaluation in the harsh environment only uses parameters: NO_2, O_3, NO, T and H.

5.2.1 Varying the Available Parameters in the Dataset

This experiment tests how different available parameters affect the calibration results. Two-months' worth of data from eleven ELMs at WACL were used. The experiment gradually adds one parameter into the dataset to simulate sensor units having different sensors on-board. The dataset from WACL was aggregated into hourly basis data using Algorithm 1. The training dataset for this experiment is based on indices that were randomly selected from 50% of the data, and the rest of the data was used for testing. This process is to avoid the influence from the testing dataset which has different characteristics from the training dataset. It is noted that the same indices are used for selecting the training and testing datasets in the experiment.

Figure 5.1 shows a comparison of the methods from eleven sensors. The Y-axis of Figures (a) (b) and (c) represents the RMSE value, the standard deviation of the error and mean error. The X-axes of the figures indicate the parameters that are available in the dataset. The plus (+) sign indicates the current parameter is added into the previous dataset. For example, in the first dataset only NO_2 is included, and in the second dataset O_3 is added into the NO_2. In the third dataset humidity is then added into the NO_2 and O_3, and so on. The boxplot represents the variation in the results from eleven sensors, and the colour differentiates the methods. The MLS method uses all parameters in each dataset and the calibration model is constructed according to Equation 4.2. By contrast, the calibration model for the proposed method is then constructed according to the steps discussed in Section 5.1. It is noted that the sequence of adding the parameter may affect the result of MLS, but it is unlikely to influence the proposed method as discussed Section 5.1.3. Since the experiment only compares the results between the methods within a given dataset, the sequence of adding a parameter would not affect the conclusion.

In Figure 5.1-(a), both methods show an improved calibration result with a larger number of parameters in the dataset; but the proposed method shows a better result than the existing method. In Figure 5.1-(b), the results from the variation of the errors is in line with the result of the calibration accuracy (RMSE). The error mean depicted in Figure 5.1-(c) shows no significant difference as the boxplots for both methods have a similar variation.

From the figure, we can confirm that calibration benefits from using supporting parameters as the results show significant improvement when multiple parameters are used in the calibration. The proposed method shows better results than the existing method in general, especially when the number of parameters in the dataset is relatively large. This could be due to the removal of inappropriate variables and the use of two-way interaction terms as justified in Section 5.1.

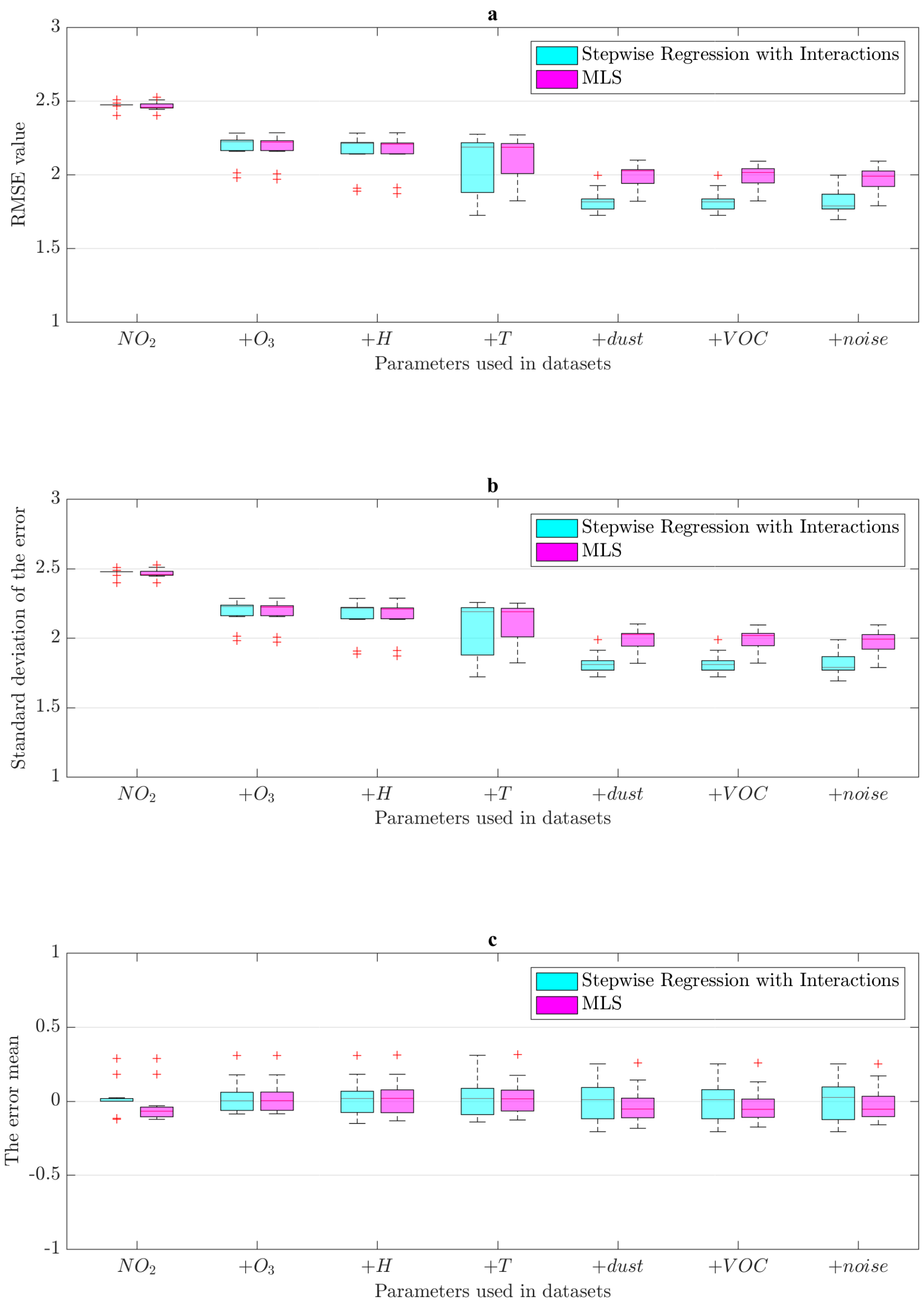


Figure 5.1: Calibration result between the two methods over the different datasets

5.2.2 Varying the Data Characteristics

In the previous experiment, the training and testing datasets are constructed by random sampling to avoid any potential influence from data patterns. In this section, we investigate how the characteristics of the testing dataset being different from the training dataset would affect the result of calibration. For this experiment, the same two-months' worth of data used in Section 5.2.1 was used. The data from one randomly selected ELM was used. The training dataset was also determined based on randomly selected indices with a size of 50% of the data, and the rest of the indices were used for testing. It is noted that the indices for training and testing were only generated once, and the same calibration model is used for all the testing.

To add different characteristics into the data, we artificially manipulated the pattern of the testing dataset. It is clear that the characteristics of the data can be different in many ways. In this experiment, we focus on 1) the constant value, 2) offset and 3) higher standard deviation.

Table 5.1: The modification of data characteristics

Constant value	STD.	= 0
	Mean	Not changed
Offset mean	STD.	Not changed
	Mean	2*mean
Higher standard deviation	STD.	2*STD.
	Mean	Not changed

The modification of the testing dataset was performed as shown in Table 5.1. The changes in the mean and standard deviation were with respect to the original testing data. Since different parameters may contribute to the calibration result differently, the modification was tested on all parameters. It is noted that for each experiment, only one parameter was modified.

Figures 5.2, 5.3 and 5.5 and Tables 5.2, 5.3 and 5.5 show the calibration results when the testing dataset of one parameter is modified according to different rules. The figures and tables differentiate the different modifications which are offset, constant value and higher standard deviation. The boxplots in each figure represent the calibration result in terms of the error distribution when the modification was taken in a particular parameter. The colour of the boxplot indicates the different methods used for the calibration.

The tables summarise the RMSE, standard deviation and mean value of the errors from both methods. The labels indicate which parameter has been modified for the calibration. It is noted that the label *original* stands for the calibration using the data without any modification. This has been used as the benchmark for the evaluation.

Table 5.2: The calibration results when using the testing dataset with constant value

Constant value									
		Original	NO_2	O_3	H	T	Dust	VOC	Noise
Proposed Method	RMSE	1.72	1.80	2.21	2.26	2.12	2.25	2.25	2.35
	STD.	1.72	1.80	2.20	2.18	2.04	2.17	2.17	2.27
	Mean	-0.03	-0.08	0.17	-0.60	-0.60	-0.60	-0.60	-0.60
MLS Method	RMSE	1.84	1.91	2.14	2.13	2.05	2.18	2.18	2.27
	STD.	1.83	1.91	2.14	2.13	2.05	2.17	2.17	2.27
	Mean	-0.15	-0.15	-0.15	-0.15	-0.15	-0.15	-0.15	-0.15

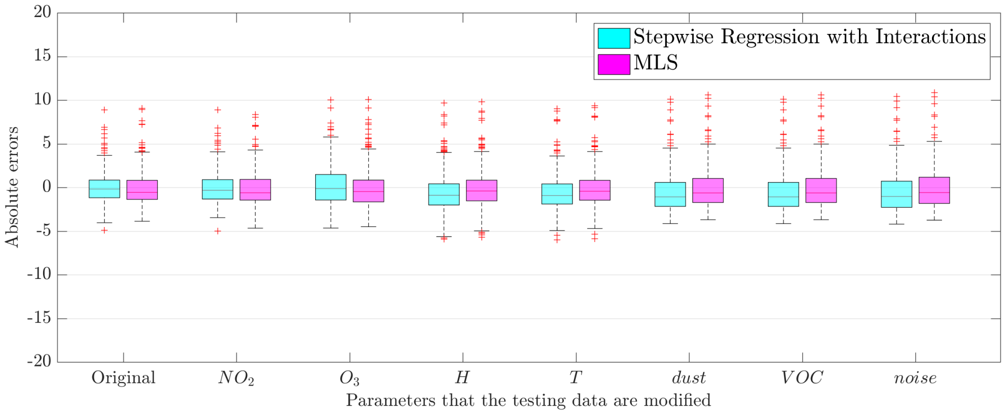


Figure 5.2: The calibration errors when using the testing dataset with constant value

Figure 5.2 and Table 5.2 present the results when the testing data becomes constant. Figure 5.2 shows no observable difference in terms of absolute errors, which suggests that the constant value would have a small impact on both calibration methods. The table indicates that the constant value in the testing data only slightly influences the RMSE and the standard deviation, and it has even less impact on the error mean, especially for MLS method.

Table 5.3: The calibration results when using the testing dataset with offset mean

Off-set mean									
		Original	NO_2	O_3	H	T	Dust	VOC	Noise
Proposed Method	RMSE	1.72	2.41	5.50	8.95	27.43	26.52	26.50	60.76
	STD.	1.72	1.89	5.08	8.96	9.68	9.95	9.95	16.39
	Mean	-0.03	-1.49	2.13	-0.12	-25.67	-24.59	-24.59	-28.51
MLS Method	RMSE	1.84	2.28	3.80	8.00	10.53	9.89	9.88	16.55
	STD.	1.83	1.83	1.83	1.83	1.83	1.83	1.83	1.83
	Mean	-0.15	-1.35	3.33	7.78	10.37	9.82	9.71	16.44

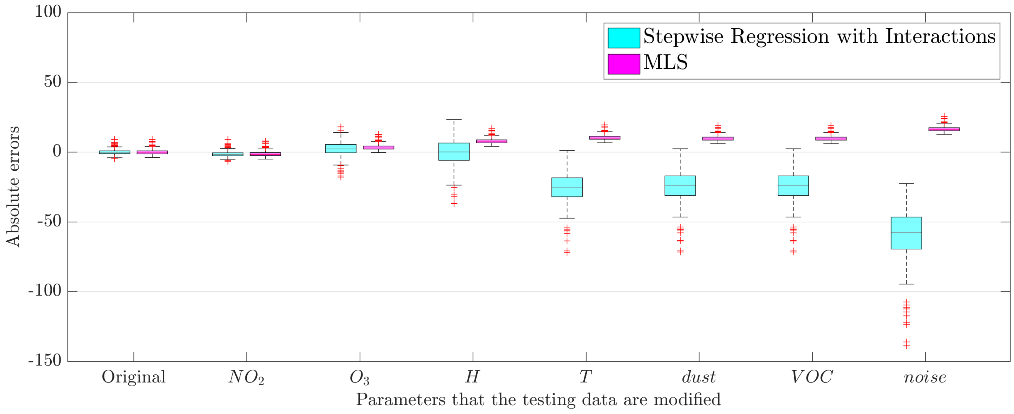


Figure 5.3: The calibration errors when using the testing dataset with offset mean

Figure 5.3 and Table 5.3 illustrates the result when the mean value of the testing data is doubled. The figure suggests the change in mean value would have a considerable impact on the calibration. We observe that for certain parameters the influence are more significant than others. Furthermore, the figure suggests that it can have a higher influence on the proposed method than the MLS method, which may be related to the use of the interaction terms. The table further confirms the impact on the RMSE, standard deviation and mean are significantly weaker compared to the result using the unmodified data. This implies that a recalibration may be needed if the testing dataset has a different mean value from the training. We further present a scatter plot for results that the parameter noise being modified in Figure 5.4. The results confirm that a recalibration is needed.

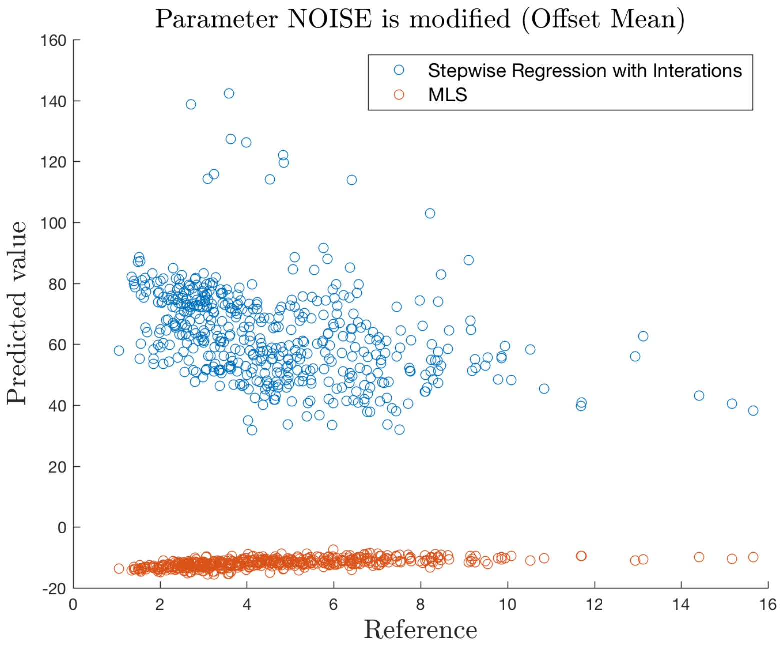


Figure 5.4: The scatter plots when the testing dataset of the parameter *noise* is changed with the offset mean

Table 5.4: The calibration results when using the testing dataset with a higher standard deviation

Higher variation										
		Original	NO_2	O_3	H	T	Dust	VOC	Noise
Proposed Method	RMSE	1.72	1.82	2.46	2.90	3.36	4.15	4.15	4.08
	STD.	1.72	1.82	2.46	2.66	3.11	3.67	3.67	3.73
	Mean	-0.03	0.01	-0.14	1.16	1.28	1.94	1.94	1.66
MLS Method	RMSE	1.84	1.89	2.10	2.16	2.12	2.22	2.22	2.32
	STD.	1.83	1.88	2.10	2.15	2.12	2.22	2.22	2.32
	Mean	-0.15	-0.15	-0.15	-0.15	-0.15	-0.15	-0.15	-0.15

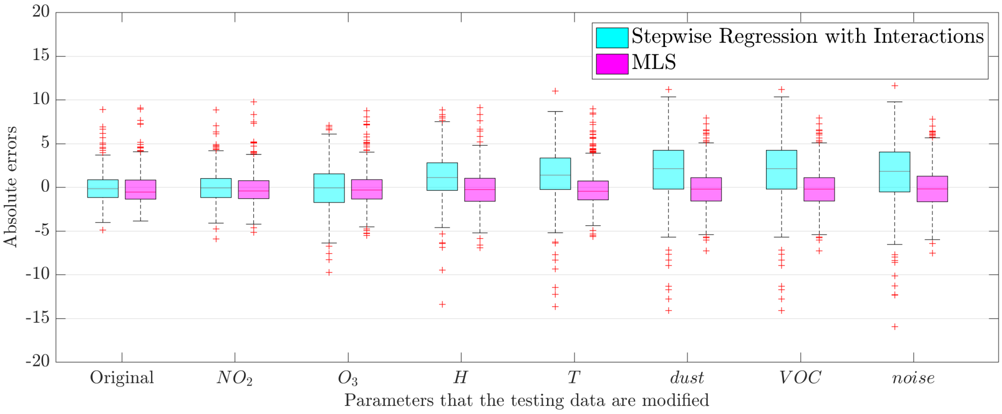


Figure 5.5: The calibration errors when using the testing dataset with a higher standard deviation

The results from using the testing data with a higher variation are presented in Figure 5.5 and Table 5.4. The figure suggests that the MLS method obtains a better error profile than the proposed method as the variation of the error is generally smaller. The table shows the result from the MLS method is relatively consistent in comparison to the proposed method, whereas the result of the proposed method varies more significantly, and the errors are further magnified if the modification is made in certain parameters. We also present a scatter plot for the results of the parameter noise being modified, in Figure 5.6, which can be cross-compared with Figure 5.4. The comparison shows that the influence from the variation in testing data is much smaller than the change in the mean value.

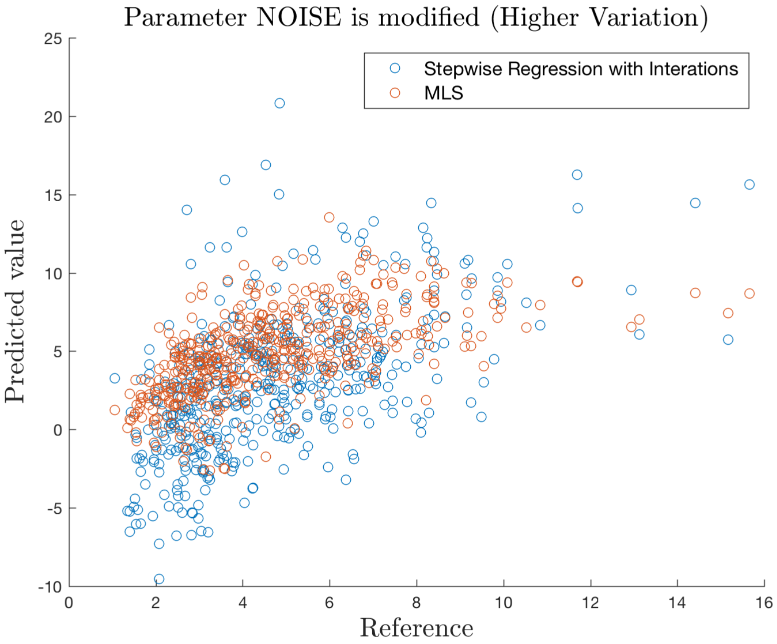


Figure 5.6: The scatter plots when the testing dataset of the parameter *noise* is changed with the higher variation

The evaluation in the mild environment suggests the proposed method can potentially obtain a better result than the MLS method in a mild environment, especially when the number of available parameters in the dataset is relatively large. However, the proposed method would be more sensitive to the difference between training and testing datasets, particularly the difference in the mean value. Therefore, if the mean values are considerably different between the training and testing datasets, recalibration may need to be considered as its influence on the calibration result is significant.

We have demonstrated that the proposed method is able to further enhance the calibration result in the mild environment. Next, we will evaluate the method in the harsh environment.

5.2.3 Evaluation in a Harsh Environment

For this evaluation, six months' worth of data from the sensor at Fishergate was firstly pre-processed by aggregating the raw data into hourly based data and excluding any data gaps. The process is based on Algorithm 1 in Section 5.1.1. After pre-processing, the dataset contains around 4,000 samples with a temporal resolution of an hour, and the available parameters are NO_2, O_3, NO, T and H, where the T and H present temperature and relative humidity respectively. In Chapter 4, we determined that using a slightly larger training dataset than the testing dataset would get a better calibration result. Thus, the dataset is sequentially and evenly divided into three partitions. The first two partitions are used for training the calibration model, and the last partition is used for testing calibration results. The calibration results are represented in Figure 5.7.

Figure 5.7 shows a series of scatter plot of an ELM sensor against a reference. Figure 5.7-a shows the raw data. The raw data is ELM data averaged into hourly data using the median without calibration (data obtained after Algorithm 1). From the figure, we can see the range of ELM data varies from 0 to 200 as emphasised by the red, which is much greater in comparison to the reference. Furthermore, a significant number of zero readings can be observed in the ELM data, which would often be considered as anomalies in existing work [72].

In Figure 5.7-b, an univariate calibration was applied to the ELM data and the result is compared against the reference. The function obtained in Figure 5.7-a is used as the calibration function of the univariate method. From the figure, apart from the range of ELM data being rescaled, improvement in the data is barely noticeable, leading to a strange data pattern (zero values) and a low correlation between the calibrated data and the reference. This shows and confirms that the univariate calibration is insufficient for the calibration of low-cost sensors, especially for the NO_2 sensor in a harsh environment. The finding is also in line with [23, 46, 77].

Figure 5.7-c shows a calibration being done by an MLS method used

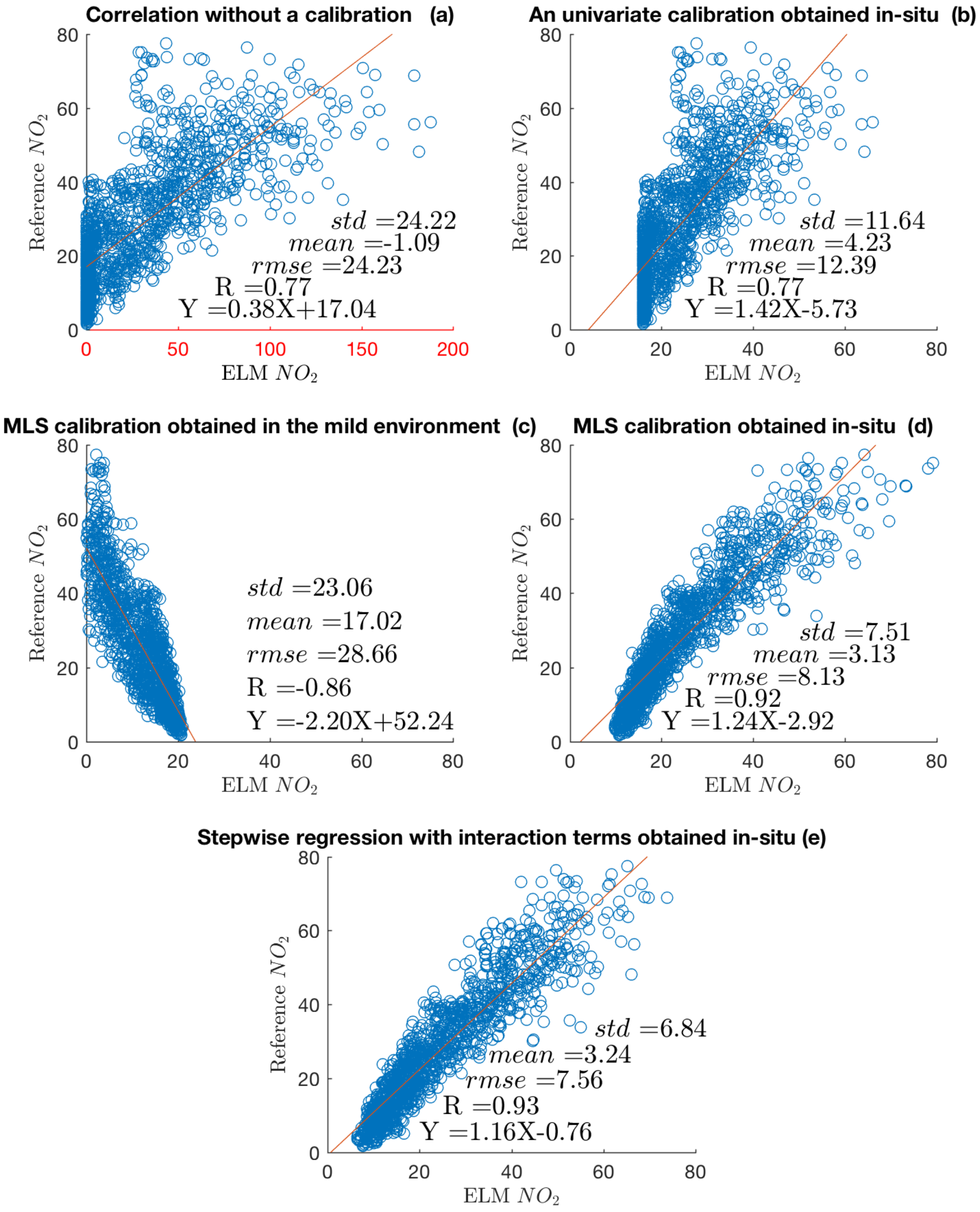


Figure 5.7: Calibration result in harsh environment

in [49]. It is noted that the calibration function in this figure was determined in the mild environment, at WACL, and applied to the sensor located in the harsh environment without any modification. We can observe a negative correlation between the calibrated data and the reference in the figure, with the worst linearity between them (i.e. the slope and offset are the worst). Furthermore, the RMSE and error mean are even worse than the data without calibration. The result confirms that a calibration determined in one place is not necessarily applicable to another place due to the different environmental conditions.

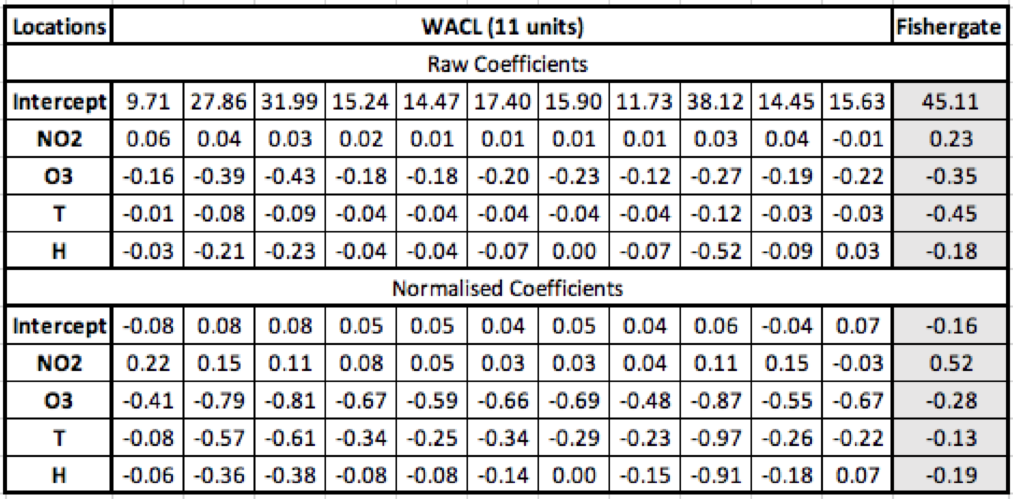
We then applied the MLS method that the calibration model was determined by in Fishergate. Figure 5.7-d shows that the performance is significantly improved compared to Figure 5.7-c. It shows the importance of calibrating sensors in the real working environment. Furthermore, compared to the univariate calibration in Figure 5.7-b, the large number of constant values are compensated for by using the supporting parameters. The correlation between the calibrated data and the reference has also improved greatly from 0.77 to 0.92, and the relationship between the calibrated data and the reference is getting closer to linear. Most importantly, the RMSE, standard deviation and mean all improved. The result confirms the importance of using multiple parameters in the calibration and suggests that including certain parameters can indeed help to compensate for constant values in the data and reduce the calibration errors.

Finally, the proposed method was applied to the ELM data. The result is shown in Figure 5.7-e. Comparing this with the result in Figure 5.7-d, the result in Figure 5.7-e shows further improved correlation and better linearity between the calibrated data and the reference, as well as RMSE, standard deviation and mean. The result suggests that the proposed method can also obtain a better calibration result in the harsh environment.

5.2.4 Generalisability of Sensor Calibration

This evaluation tests the generalisability of sensor calibration across different sensor units and environmental conditions by comparing the changes in the regression coefficients. We first analyse the variation in the regression coefficients for the 11 sensor units in the mild environment (WACL) to determine the variation in coefficients across the sensor units; then we compare the coefficients determined in the mild environment (WACL) to those determined in the harsh environment (Fishergate) to further understand the variation in coefficients across different environments.

As discussed in Section 3.2, the parameters monitored in the mild and the harsh environments are different. In order to cross-compare the regression coefficients in both environments, the calibration of NO_2 in this section is constructed using the parameters they have in common, which are NO_2, O_3, T and H. The training regimes in the mild and harsh environments are identical to the ones used in Section 5.2.1 and Section 5.2.3 respectively. The regression coefficients for sensors in the mild and the harsh environments are shown in Figure 5.8.



Locations	WACL (11 units)											Fishergate
Raw Coefficients												
Intercept	9.71	27.86	31.99	15.24	14.47	17.40	15.90	11.73	38.12	14.45	15.63	45.11
NO2	0.06	0.04	0.03	0.02	0.01	0.01	0.01	0.01	0.03	0.04	-0.01	0.23
O3	-0.16	-0.39	-0.43	-0.18	-0.18	-0.20	-0.23	-0.12	-0.27	-0.19	-0.22	-0.35
T	-0.01	-0.08	-0.09	-0.04	-0.04	-0.04	-0.04	-0.04	-0.12	-0.03	-0.03	-0.45
H	-0.03	-0.21	-0.23	-0.04	-0.04	-0.07	0.00	-0.07	-0.52	-0.09	0.03	-0.18
Normalised Coefficients												
Intercept	-0.08	0.08	0.08	0.05	0.05	0.04	0.05	0.04	0.06	-0.04	0.07	-0.16
NO2	0.22	0.15	0.11	0.08	0.05	0.03	0.03	0.04	0.11	0.15	-0.03	0.52
O3	-0.41	-0.79	-0.81	-0.67	-0.59	-0.66	-0.69	-0.48	-0.87	-0.55	-0.67	-0.28
T	-0.08	-0.57	-0.61	-0.34	-0.25	-0.34	-0.29	-0.23	-0.97	-0.26	-0.22	-0.13
H	-0.06	-0.36	-0.38	-0.08	-0.08	-0.14	0.00	-0.15	-0.91	-0.18	0.07	-0.19

Figure 5.8: The regression coefficients across sensor units and environments

Figure 5.8 presents not only the raw coefficients but also the nor-

Table 5.5: The mean and the standard deviation of the coefficients for the 11 sensor units at WACL

Raw Coefficients				
Intercept	NO_2	O_3	T	H
19.32±9.11	0.02±0.02	-0.23±0.1	-0.05±0.03	-0.12±0.16
Normalised Coefficients				
Intercept	NO_2	O_3	T	H
0.04±0.05	0.09±0.07	-0.65±0.14	-0.38±0.25	-0.21±0.27

malised coefficients. The normalised coefficients are also referred to as the standard coefficients, which are determined from the regression that converted all the variables using z-scores. Since the magnitude of the measurements in the mild and harsh environments are significantly different, as discussed in Section 3.3, using the normalised coefficients enables a better comparison, as the variations of variables are all referred to as their standard deviation.

In Figure 5.8, we can observe a large variation in the coefficients for all variables among the 11 sensor units in the WACL. Since the units at the WACL were co-located in the same environmental condition as discussed in Section 3.2, we consider the variation in the coefficients in these 11 sensor units is the result of using different sensor units. The results show that the coefficients of the calibration function are inconsistent for different sensor units in the mild environment, which suggests that calibration from one sensor unit could give significant errors if applied to another sensor unit.

We further calculate the mean and standard deviation of the coefficients for the 11 sensor units at the WACL in Table 5.5. We then compare the regression coefficients determined from the unit in the harsh environment (i.e. the grey zone in Figure 5.8) to Table 5.5 to further determine how the coefficients vary in different conditions. The training datasets for the mild and the harsh environments have a different number of samples, which could affect the variation in the coefficients. However, as presented in Figure 4.23 and then discussed in Chapter 4, the size of the training dataset does not have a significant impact on the variation in the coefficients, and thus the influence of the dataset size is not considered further in this analysis. The comparison of the coefficients between the mild (i.e. the coefficients in Table 5.5) and the harsh (i.e. the coefficients in the grey zone in Figure 5.8) shows that the calibration coefficients determined in the harsh environment are considerably different from the coefficients for the mild environment, especially for the intercept, NO_2 and O_3. The results also suggest that the calibration function may not easily generalise across different environments.

In summary, the results presented in this section show that the coefficients of the calibration functions are sensitive to the individual sensor unit and may be sensitive to environmental conditions. Therefore our recommendation is in-line with the conclusions in [10, 46] that individual sensor units are required to be calibrated in the location of the operation, although further data is needed to definitively conclude the importance of the location.

5.3 Limitations of Validity

The proposed method shows the ability to systematically select the supporting parameters. However, it is noted that the objective of the proposed method is to minimise the difference between the model output and the reference. Hence, the parameters used in the calibration model would indicate the importance of these parameters in general. Similarly, the coefficients of the parameters may not present the importance of the parameters, as the result is data dependent. Therefore, we did not further analyse what parameters are removed in each calibration and how the coefficients are changed. For the same reason, we believe that it would not be appropriate to feed the selected parameter into an ANN-based method.

Furthermore, due to the limitations of the dataset, the evaluation in the harsh environment was performed using the same dataset as the one in Chapter 4. As a result, certain decisions that are based on the conclusion from the previous experiment could be biased, such as using more data in the training dataset than the testing dataset. In addition, in Figure 5.7-c, the calibration function was determined from a different sensor, environment and time of the measurement. Hence, the result might also be influenced by those factors.

5.4 Summary

In this section, we have demonstrated how the supporting parameters can be better used. The evaluations show that the proposed method is able to reduce the calibration errors in both the mild and the harsh environment significantly more than the existing method. The evaluation also confirms that univariate calibration can be insufficient for calibrating low-cost sensors and suggests that some constant values in the uncalibrated data can be alleviated by the calibration process. The result indicates that the calibration result benefits from the use of the appropriate parameters and consideration of their interactions.

The evaluation in the mild environment suggests the proposed method would have a better performance when the number of available parameters in the dataset is relatively large. Furthermore, it is demonstrated that the calibration results can be considerably affected if the data pattern between the training and testing dataset are inconsistent, especially if the difference is in the mean value.

The evaluation in the harsh environment confirms that the calibration function applied in one location may not be directly applicable to another location, and the result illustrates that the proposed method is able to further reduce calibration errors.

With the evidence provided above, the second research question, which is quoted below, is answered:

> *Research Question 2: How can we ensure the calibration result by properly using supporting parameters?*

Chapter 6

The Detection of Anomalies

This chapter aims to answer the last research question, which is quoted below:

> *Research Question 3: How can we accurately detect and remove anomalies to further improve data quality?*

According to our review, the main difficulty in the detection of anomalies is to differentiate anomalies from outliers. The state-of-the-art research suggests using contextual information to identify anomalies, as the correct measurements are often contextually related, while anomalies are stochastically unrelated [94]. It is known that spatial and temporal dependencies are the most commonly used contextual information, and considerable research has demonstrated that they can sufficiently improve the detection results and are capable of separating anomalies from outliers. However, as illustrated in Section 3.3, spatial and temporal dependencies are not sufficient for anomaly detection in our data. Therefore, it is important to explore new contextual information for such a purpose.

It is understood that the response of a low-cost sensor would be significantly affected by its cross-sensitive parameters due to sensor properties [85]. As reviewed in Chapter 2, using cross-sensitive parameters is able to improve calibration results significantly, as it would provide complementary information for the parameter of interest. As a result, we believe that a certain dependency exists between the parameter of interest and its cross-sensitive parameters. This allows anomalies to be differentiated from abnormal events, i.e. a higher than normal value of NO_2 may be considered as an anomaly rather than an event, if its cross-sensitive parameter O_3 exhibits a significantly different trend.

In this chapter, we explore the cross-sensitive parameter as new contextual information to determine if the results of anomaly detection can benefit from it. A Bayesian-based method is firstly justified and introduced in Section 6.1 to determine the anomaly model. Then, the proposed method is evaluated in both the synthetic data and real data in Section 6.2. Finally, we discuss the research validity and conclude this chapter in Sections6.3 and 6.4.

6.1 Method of Anomaly Detection

According to the review, determining an anomaly model is important for the detection of anomalies. For this work, a learning-based method is ideal as the dependency between the parameters of interest and its cross-sensitive parameter needs to be determined. It is understood that there are many learning-based methods available, such as an ANN, SVM and Bayesian-based method. Considering the detection of anomalies is applied with calibration, a lightweight process is ideal as the process may be applied frequently. Thus, a Bayesian-based method is used in this work to learn the contextual information and determine the anomaly model.

6.1.1 Learning the Information

Learning the contextual information in a Bayesian-based method is used to determine the joint probability between two events, in our case, between the parameter of interest and its cross-sensitive parameter. We characterise the learning method as follows:

- The set of measurements, I for the parameter of interest; and C for the cross-sensitive parameter.

- An index of the measurements, i, where $i \in \mathrm{Z}^+$ and Z^+ stands for all positive integers.

- A number of classes (bins) in I and C as j and k, where j and $k \leq max(i)$.

- A joint probability, $P(\mathrm{I}, \mathrm{C})$

- A conditional probability distribution, $P(\mathrm{I}|\mathrm{C}_k)$

- A conditional probability, $P(\mathrm{I}_j|\mathrm{C}_k)$

In practice, due to the missing values, the set of measurements I and C may not have the same number of samples. Since the method determines the joint probability between two sets of measurements, $P(\mathrm{I}_i, \mathrm{C}_i)$, some simple processing is needed to unify the sample size, ensuring both sets have a measurement at the same time stamp. For example, if at a given time, only the I has a reading, then this reading needs to be removed for consistency.

Bayesian methods only deal with discrete data, so determining a proper bin size is important. A small bin size could result in the histogram having a non-distinct mode, which would make anomalies inseparable from the data; however, a large bin size could reduce the precision of the method, which would then increase the number of false positives. Since the precision of the sampled data is two significant digits, using the sampled data directly would result in the bin size becoming too small, especially when the number of samples in a dataset is relatively small. Hence, a new discretisation process is required for the measurement sets, I and C.

We determine the bin size using a two dimensional histogram, which is similar to the method used in [72]. In practice, domain knowledge can be important and trying a wide range of bin sizes is ideal for determining appropriate bin sizes as identified in [72]. The process of discretisation classifies the set I and C into j and k classes. We then determine the joint probability as $P(\mathrm{I}, \mathrm{C})$ which is also referred to as the joint probability table, according to the class number [34]. The determined joint probability, $P(\mathrm{I}, \mathrm{C})$, is considered as the model of anomalies as it will help to identify anomalies. It is noted that the process can also be applied to learning spatial and temporal information by determining the dependencies between the parameters of interest in different locations as $P(\mathrm{I}_{\mathrm{location1}}, \mathrm{I}_{\mathrm{location2}})$ and the dependency between the current measurement and its previous measurement as $P(\mathrm{I}_{\mathrm{t}}, \mathrm{I}_{\mathrm{t-1}})$.

6.1.2 Inferencing

Once the joint probability table, $P(\mathrm{I}, \mathrm{C})$, is determined, it can be used to make an inference and statistically identify anomalies. For any testing dataset, the probability distribution of I at a given value of C can be obtained according to the class number k, which is $P(\mathrm{I}|\mathrm{C}_k)$. The probability of the actual I_j, at a given value of C_k, can be determined as $P(\mathrm{I}_j|\mathrm{C}_k)$. If the probability is less than a threshold value, this measurement is considered as an anomaly and removed from the data. The threshold value again is sensitive to the use of the data and would require domain knowledge to determine. The selection of the threshold value will be discussed in Section 6.2. The method of the detection of anomalies is in Algorithm 3.

6.2 Evaluation

The evaluation was performed using both a synthetic dataset and a real dataset. The use of synthetic data is for evaluating the classification accuracy as it is often impossible to label the anomalies in real datasets. Furthermore, the evaluation of real data determines how the detection and removal of anomalies would enhance the calibration. For this evaluation, the NO_2 is used as the parameter of interest and the O_3 is used for the cross-sensitive parameter.

6.2.1 Synthetic Data

The synthetic data was constructed by injecting anomalies into a clean dataset. The base signals of the clean dataset are taken from a reference

ALGORITHM 3: Pseudo code for detection of anomalies

```
Data: Define:
$C_{m\times1}$ = the measurements of the cross-sensitive parameter with m samples
$I_{i\times1}$ = the measurements of the parameter of interest with i samples
$m$ and $i$ indicate the length of the measurement, ($i$<$m$)

Removing measurements in $C_{m\times1}$ for consistency according to time stamps for both training and testing datasets
Result: $C_{m\times1}$ becomes $C_{i\times1}$ which have same number of samples
        as $I_{i\times1}$

1) Classify $C_{i\times1}$ into $k$ classes and $I_{i\times1}$ into $j$ classes.
2) (max(C)-min(C)) / $k$ → step for each bin in $C$; (max(I)-min(I)) / $j$ → step for each bin in $I$.
3) Using *bin size* classifies $C_{i\times1}$ into $k$ bins and $I_{i\times1}$ into $j$ bins.
4) Assigning the bin number to the raw data → $I_{i\times2}$ and $C_{i\times2}$
Result: Discretisation. (Training and testing datasets)

Declare an empty table = $table_{k\times j}$, which have k rows and j columns
for (Training dataset) i = 1 to ALL do
    Add 1 on $table_{k\times j}$ according to index $C(i,2)$ and $I(i,2)$
end
Divide the frequency table by the total number of measurements, $i$
Result: Joint probability table, $P(I,C)$,

for (Testing dataset) i = 1 to ALL do
    1) Obtaining the conditional distribution I by using the index, $P(I|C(i,2))$

    If the index is out of the range, discard the data instance and label as anomaly

    if $P(I(i,2)|C(i,2))$ < a threshold then
        $I(i,1)$ is an anomaly
    else
        $I(i,1)$ is a correct measurement
    end
end
Result: Labelling all the measurement
```

sensor with a temporal resolution of a minute. The dataset contains four days of measurements of NO_2 and O_3. We manually removed any suspicious measurements and filled them using linear interpolation. The process is to maximise the consistency of temporal information, as we compare our method to the one that uses temporal dependencies. The clean dataset after this process is free from anomalies and is temporally consistent. The anomalies are then randomly injected into the clean signals.

The magnitude of anomalies in reality is often unknown. Considering that an extremely high magnitude of anomalies can be classified by a simple threshold value, and an extremely small magnitude of anomalies would not significantly affect the data process (e.g. calibration), the range of the magnitude for artificial anomalies is randomly chosen between 10% to 60% of the maximal values of the clean signal. We reckon that the anomalies in that range are problematic as they are difficult to detect and remove, and have an adverse impact on the data process.

In our previous work [28], the boxplot suggests that there are about 8% of outliers in the dataset. Considering outliers from low-cost sensors are likely to be dominated by anomalies, as justified in Section 3.3.2, 8% of anomalies were injected into the clean dataset. The constructed synthetic data for NO_2 is illustrated in Figure 6.1. In the figure, the base signal is in red and the injected anomalies are in blue.

6.2.2 Discretisation

Discretisation is an important step when using a Bayesian-based method as mentioned previously. Once the training data is available, the minimum and maximum values of the parameters need to be estimated to set the boundaries of the joint probability table. This process aims to avoid a scenario in which a real measurement in the testing dataset is greater than in the training dataset. Setting a large boundary would avoid this problem; however, setting a larger boundary takes much more computational resources and individual bins can be left with too few samples for sound statistical analysis. Hence, such a trade-off needs to be balanced

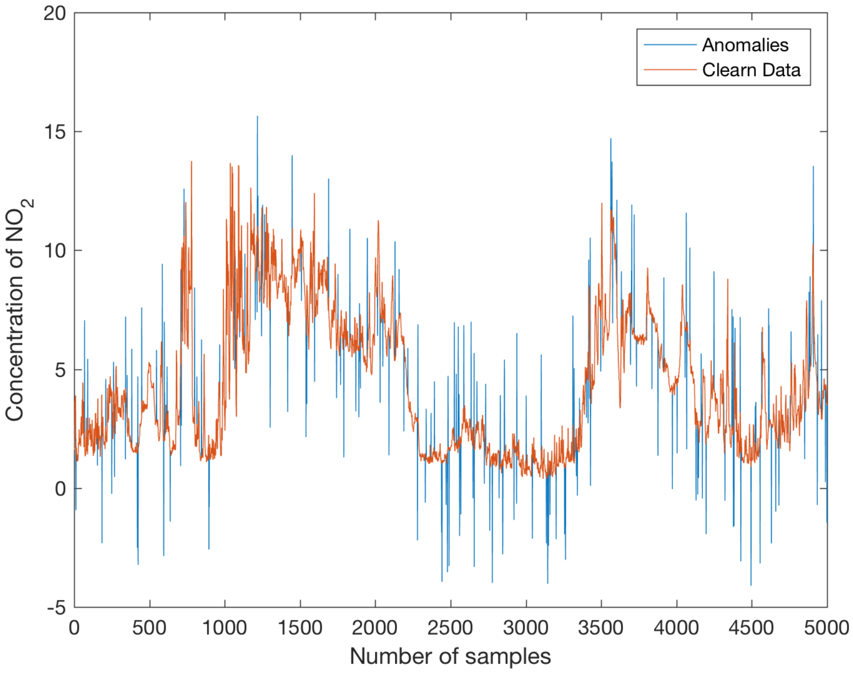


Figure 6.1: The synthetic dataset

according to priori knowledge, such as knowing the distribution of the data in a given time period. For the synthetic data, we know exactly the maximum and minimum values of the data. Hence, we directly use those as the boundaries of the table. For the real dataset, we set the boundary by adding 20% onto the minimal and maximal values of the training dataset. This is sufficient for the detection of anomalies over a relatively short period, i.e., the month's worth of data used in this evaluation. However, if the dataset varies more significantly, the boundary may need to be extended accordingly, as data instances that exceed the range will be discarded and considered as anomalies.

Once the boundary is determined, the data is classified into a number of bins according to the *bin size*. As the boundary of the data is fixed, the bin size and the number of the bins are complementary and represent the same thing. The selection of a bin size can be dependent on the type of data or a requirement from the user. Our data has 5,000 samples and varies from -4 to 18 as shown in Figure 6.1. We found 15 bins are sufficient for NO_2 as there is a joint probability between the selected parameters and each bin has a sufficient number of counts for the analysis. It is noted that the bin size needs to be adjusted if the number of samples or the boundary of the data are significantly different from the example. For instance, if the variation in data is more significant, the number of bins would need to increase accordingly.

6.2.3 Threshold Value

According to the literature review in Section 2, the selection of a threshold value is data dependent. Therefore, the determination of a threshold is often difficult and requires expert knowledge. In this section, we demonstrate how the results of anomaly detection are affected by an increase in the threshold.

The results of the anomaly detection are evaluated in terms of accuracy, precision and completeness, which have been widely used for such a purpose. Those metrics are defined as follows:

$$Accuracy = \frac{\text{(Number of True Positive + Number of True Negative)}}{\text{Number of Total}} \tag{6.1}$$

$$Precision = \frac{\text{Number of True Positive}}{\text{Number of Test Outcome Positive}} \tag{6.2}$$

$$Completeness = \frac{\text{Number of True Positive}}{\text{Number of Condition Positive}} \tag{6.3}$$

where a conditional positive in Equation 6.3 indicates the number of real positive cases in the data. All values of evaluation metrics are normalised in the range from 0 to 1.

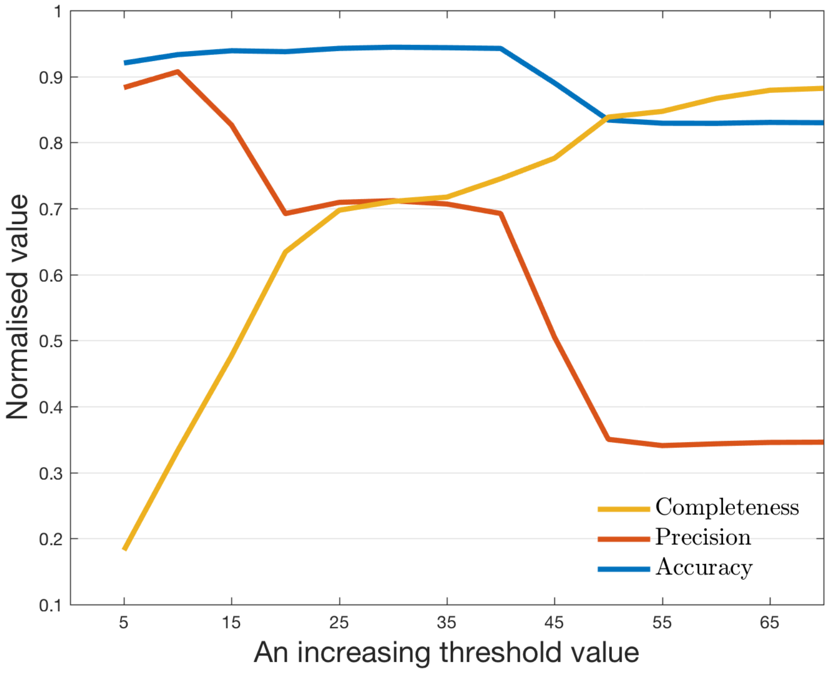


Figure 6.2: The detection results when using different threshold value

Figure 6.2 shows the detection result when the value of the threshold is gradually increased. The x-axis indicates the threshold value. We can see the accuracy of the detection is not significantly affected by increasing the threshold value. However, the precision and the completeness are extremely sensitive to the value of the threshold. We can see a clear trade-off between those two metrics. The result of the completeness starts around 0.2 at the beginning and gradually increases up to 0.87 at the end. However, the trend of the precision is opposite to the trend of the completeness. It starts at a higher precision and drops to around 0.35 in the end. From Figure 6.2, we can conclude that the determination of a threshold is not only sensitive to the use of data but also relies on the requirements of the end-users, as the trade-off between precision and completeness need to be balanced. In this study, precision is more important than completeness, as we do not want to remove too much correct data. Hence, according to Figure 6.2, our threshold value is determined as 15. However, it is noted that in practice, we do not have the labels for the anomalies. Hence, we cannot rely on Figure 6.2 to determine the threshold value. As a result, expert knowledge is required to determine an appropriate threshold value by balancing precision and completeness. It can be difficult to obtain a perfect balance, however, it is clear that a smaller threshold value often indicates higher precision but lower completeness, and a larger threshold value would lead to an opposite result.

6.2.4 Evaluation in Synthetic Data

For this evaluation, we compare the results from using the cross-sensitive parameter against those using the temporal information. The same learning process was performed for both contextual information. Figure 6.3 shows the results in term of detection accuracy, precision and completeness. We inject anomalies into the clean signal for 100 interations to minimise any potential bias caused by the injection. The process used the rule discussed in Section 6.2.1, and each boxplot indicates the result from the 100 tests. The results in Figure 6.3 suggest that using cross-sensitive parameters is able to produce a more reliable detection result, as the accuracy and the completeness are significantly better than the one using the temporal information, and the precision is also no worse than using the temporal information.

In the first experiment, anomalies are only injected in the NO_2. As anomalies can affect all parameters, in the following experiment, anomalies are injected in both NO_2 and O_3 data. The percentage and magnitude of the O_3 anomalies were determined from real data, which was done in the same way as in Section 6.2.1. Therefore, the injection of NO_2 remains the same as in the previous experiment. 10% of the samples in the clean O_3 data were randomly replaced with anomalies that have a magnitude randomly selected in the range 10% to 60%.

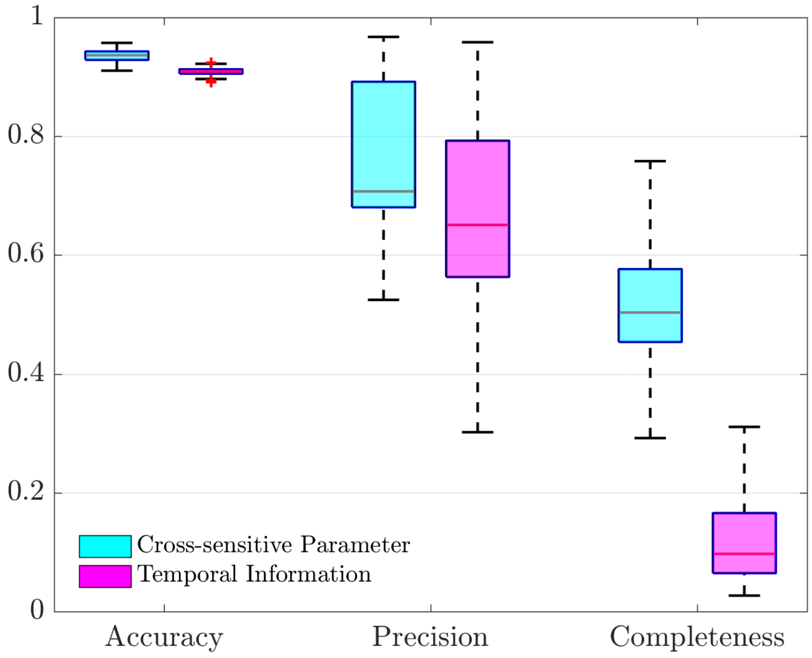


Figure 6.3: Anomalies are injected in NO_2

The result of 100 tests is shown in Figure 6.4. In the figure, the accuracy and the completeness of using the cross-sensitive parameter is still significantly better than using the temporal information after 10% of anomalies being added into the O_3 data. In terms of precision, the result using the cross-sensitive parameter has much less variation than the one using the temporal information. Hence, in general, using the cross-sensitive parameters still has a better performance than using the temporal information.

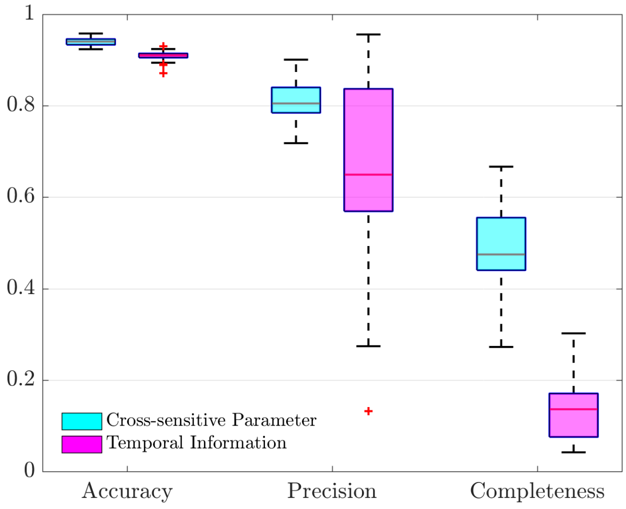


Figure 6.4: Anomalies are injected in NO_2 and O_3

To consider the variation in the results, we use the mean value to summarise the two experiments in Table 6.1. The results show that there is only a small influence on the result by using the cross-sensitive parameters. Accuracy and precision are not significantly affected, and completeness is reduced by 5%. By contrast, the anomalies in the O_3 do not affect the result of using the temporal information at all. In general, using the cross-sensitive parameters obtains a better result than using the temporal information. Therefore, the results suggest that using the cross-sensitive parameters is able to sufficiently detect anomalies.

Table 6.1: Mean value from two experiments

	Cross-sensitive Parameter			**Temporal Information**		
	Accuracy	**Precision**	**Completeness**	**Accuracy**	**Precision**	**Completeness**
Exp. 1	0.9398	0.8028	0.5113	0.9103	0.6842	0.1268
Exp. 2	0.9381	0.7977	0.4657	0.9101	0.6859	0.1243

6.2.5 Evaluation Using Real Data

The next stage of the evaluation uses real data obtained from the sensors at the WACL. As the sensed data is likely to contain anomalies, the objective of this evaluation is to determine how accurate the resulting signal (after anomaly detection and calibration) is compared to the reference signal. This objective reflects the overall goal of the research work.

We have justified why the detection of anomalies needs to be performed before calibration. However, it is not clear if the process needs to be applied before or after data aggregation. For this experiment, we assume that a better anomaly detection and removal would lead to a better calibration result.

In the first experiment, the process of anomaly removal is applied before data aggregation. Then, the data is aggregated and calibrated according to Algorithms 1 and 2 respectively. The training and testing datasets from WACL are randomly sampled to ensure the same number of training and testing datasets are used in the comparison. The results of the calibrations over a number of NO_2 sensors are shown in Figures 6.5 and 6.6

In the figures, *raw data* presents the data that is calibrated without removing anomalies. *Non-parametric* indicates the anomalies identified using a boxplot, which are defined as data points which are greater than $q3 + w \times (q3–q1)$ or less than $q1–w \times (q3–q1)$. The w is the maximum whisker length, and $q1$ and $q3$ are the 25th and 75th percentiles of the sample data respectively. *Temporal* and *cross-sensitive* represent the use of related contextual information. Figure 6.5 shows the calibration accuracy in terms of RMSE value. We observe that the results from using the cross-sensitive parameter are slightly better than using the other methods. However, in Figure 6.6, no significant difference is observed in the standard deviation of the error for the compared methods. We suspect that it is because the aggregation process hinders the impact of anomalies.

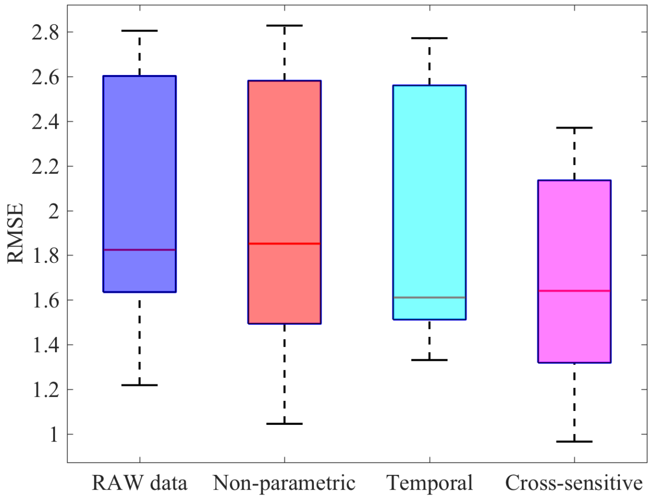


Figure 6.5: Result of calibration in terms of RMSE

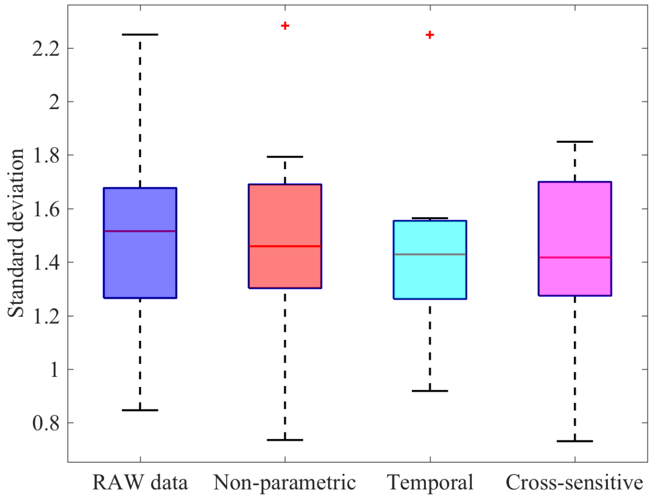


Figure 6.6: Result of calibration in terms of standard deviation of the errors

For the second experiment, the data is firstly aggregated according to Algorithm 1. The anomalies are then detected and removed from the aggregated datasets before calibration. The calibration results are illustrated in Figures 6.7 and 6.8. It is noted that the results in this experiment are not comparable to Figure 6.5 and 6.6. From the figures, the results suggest that using contextual information may further help the calibration result as both variations of errors and calibration accuracy are smaller. The results suggest that the calibration result would benefit from the anomalies being removed after data aggregation.

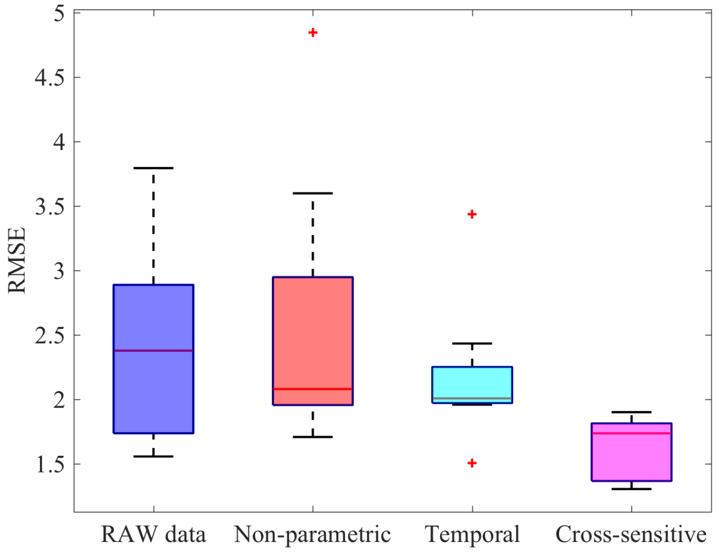


Figure 6.7: Result of Calibration in terms of RMSE

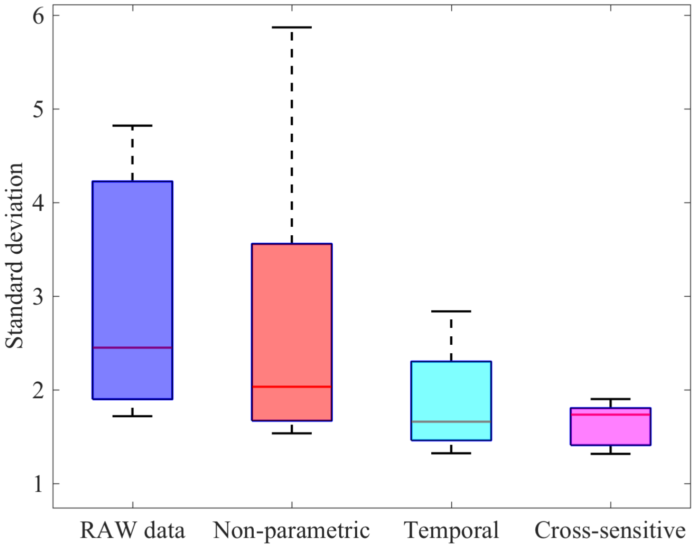


Figure 6.8: Result of Calibration in terms of standard deviation of the errors

In the third experiment, the process of anomaly detection is applied to the dataset collected from Fishergate, York. We firstly remove the anomalies in the NO_2 data after the aggregation process and then perform the calibration according to Section 5. The calibration result can be cross-compared to the result in Figure 5.7-d, which is shown in Figure 6.9. Compared to Figure 5.7-d where the sensor was calibrated without anomaly removal, Figure 6.9 shows an improved linearity as the slope is closer to one. Comparing other evaluation metrics, it suggests that the calibration result is further improved in terms of reduced RMSE, as well as the mean and standard deviation of the error.

The results indicate that the proposed anomaly detection in combination with the calibration is able to further reduce calibration errors and improves the data quality for low-cost sensors.

6.3 Limitation of Validity

It is understood that evaluation of anomaly detection is still an open challenge, as using synthetic data will never fully reflect the reality, and

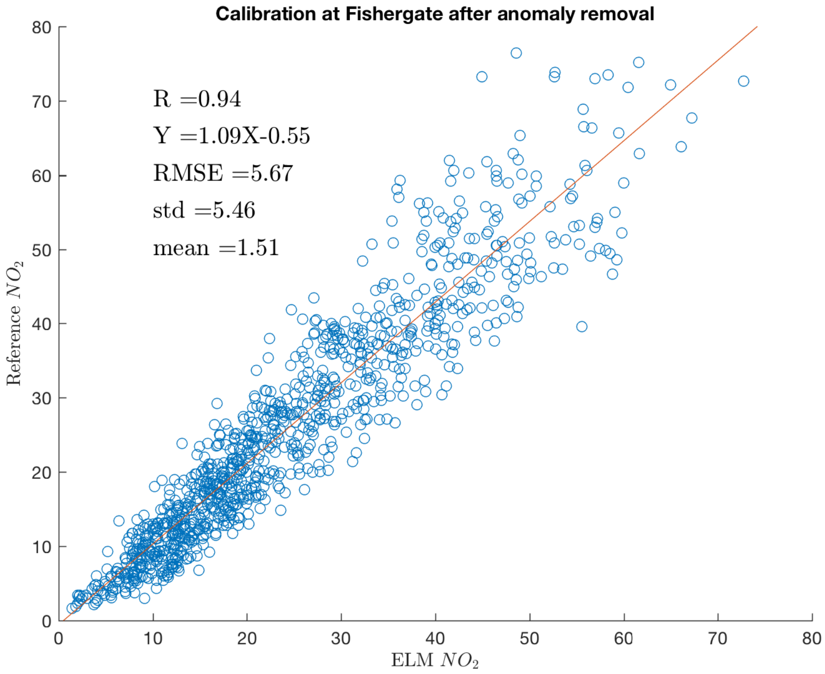


Figure 6.9: Result of Calibration in the Fishergate

labelling real datasets often introduces errors.

The proposed method uses the dependence between the parameter of interest and its cross-sensitive parameters (i.e. NO_2 and O_3). Our evaluation shows the determined dependency can be sufficient and accurately detect anomalies. However, since the actual dependence between NO_2 and O_3 would be sensitive to many factors, it is important to carry out further studies of changes of the dependence. Furthermore, the results of the proposed method would be sensitive to model parameters, such as the threshold value. Hence, the result is not cross-comparable for different applications.

The result of sensor calibration is associated with many factors, such as the use of sensors and number of training and testing datasets. Hence, the assumption that a better anomaly detection would lead to a better calibration result may not always hold. In addition, anomaly removal applied before and after data aggregation would remove different data, as the dependency of NO_2 and O_3 over 20s is expected to be different from an hour. Therefore, we believe that a deep understanding of the cause of anomalies and dependencies between cross-sensitive parameters is important for future work.

6.4 Summary

The detection of anomalies in the sensed data is difficult as anomalies do not have a distinctive pattern. Using contextual information is believed to help anomaly detection, as correct measurements are often contextually related, while anomalies are stochastically unrelated [94].

Since widely used contextual information (i.e. spatial and temporal dependencies) is often inaccessible, we propose to use cross-sensitive parameters as new contextual information. The evaluation using synthetic data shows the proposed method is able to improve detection accuracy. However, the trade-off between completeness and precision needs to be balanced, depending on the user requirements.

It is understood that the evaluation of anomaly detection is still an open challenge, as accurate labelling for anomalies can be difficult. Hence, for evaluation on a real dataset, the evaluation objective is to determine how accurate the resulting signal (after anomaly detection and calibration) is compared to the reference signal. We consider better anomaly detection would lead to a better calibration result in certain circumstances and this objective reflects the overall goal of the research work.

The anomalies were removed before and after data aggregation. The results show no observable differences in the calibration results for which anomalies are removed before data aggregation. In the results where anomalies are removed after data aggregation, using contextual information shows a significant improvement in the calibration results, and the results from using the cross-sensitive parameter is even better than the one using temporal dependency. It is understood that the calibration result is sensitive to many factors. Thus, our evaluation using the real dataset may only be indicative.

Finally, we apply our two-phase solution to the data obtained from

Fishergate. The results suggest using cross-sensitive parameters is able to accurately detect and remove anomalies, and the result of calibration can be further improved. The material presented in this chapter answers the last research question, which is quoted below:

> *Research Question 3: How can we accurately detect and remove anomalies to further improve data quality?*

Chapter 7

Conclusion

Using low-cost sensors to monitor the urban environment has become increasingly popular, as they provide better data resolution than current practices. However, low-cost sensors often produce poorer data quality and so the data may not be used directly without processing. This thesis presents a two-phase solution to improve the data quality of low-cost sensors. It consists of a method for the detection and removal of anomalies and a process of sensor calibration. The evaluation shows that the proposed solution is better than state-of-the-art methods and is able to improve data quality, especially for sensors in a harsh polluted environment.

In Chapter 2, a detailed review of the state-of-the-art research was given, which focused on the calibration of sensors and the detection of anomalies. The review shows that multivariate calibration is the best practice for calibrating low-cost sensors, as it includes environmental influences as supporting parameters and subtracts the related effects. A number of studies suggest that the use of supporting parameters would be dependent on the current environmental conditions, and calibration of sensors in a harsh environment can be more difficult than in a mild environment, especially for NO_2 sensors. Furthermore, the literature review also reveals that there is a lack of work on an effective comparison of calibration approaches, which makes it difficult to determine the most appropriate solution for the needs. Finally, the review indicates that in the existing methods for anomaly detection it is difficult to separate anomalies from outliers, especially when there is insufficient contextual information, e.g., the spatial and temporal data dependencies are weak.

In Chapter 3, we discussed the deployment of the sensors and illustrated the characteristics of the environmental data. The illustration of real data indicates that the data is neither spatially nor temporally consistent, and the response of the sensors is closely associated with influences such as cross-sensitive parameters or environmental variables (e.g. temperature). With the understanding of the data and the limitations of the state-of-the-art method, we have demonstrated the validity of the research questions. They are revisited and summarised below:

> *Research Question 1: Which is the appropriate calibration method (Regression or ANN) considering the needs of our application?*

The first research question was answered in Chapter 4. The needs of the calibration were derived from the context of the application in Chapters 2 and 3 as a lightweight process that works better on a relatively small training dataset. Chapter 4 presented a systematic comparison of state-of-the-art calibration methods, i.e. a regression-based method and an ANN-based method. The evaluation shows that the ANN-based method is sensitive to the use of model parameters and the random variation in the model generation process, which can lead to a large variation in the calibration result. By contrast, the regression-based method provides a more predictable result and has a better performance for a relatively small training dataset. Moreover, we demonstrated that the same error may be associated with different correlations between the reference and calibrated values. Thus, it is always important to look at both evaluation metrics. Using the comparison, we believe a regression-based method is more appropriate for our application.

> *Research Question 2: How do we ensure the calibration result by properly using supporting parameters?*

This research question was answered in Chapter 5. The review in Chapter 2 indicated that using supporting parameters was able to improve calibration, but the calibration result would be compromised if 1) an appropriate parameter is available but is not used; and 2) an inappropriate parameter is used in the calibration. Therefore, we introduced a novel regression-based calibration in Chapter 5. In contrast to the state-of-the-art method, the proposed method automatically selects the optimal supporting parameters from an available dataset. The method uses stepwise regression with interaction terms, which not only maximises the information from the supporting parameters, but also ensures the most appropriate parameters are used in the calibration. The evaluation was carried out in both mild and harsh environments, which shows the proposed method is significantly better than the state-of-the-art method in terms of calibration accuracy.

> *Research Question 3: How can we accurately detect and remove anomalies to further improve the data quality?*

The above research question was answered in Chapter 6. Chapter 2 shows that using contextual information is important to detect anomalies, and the most commonly used information is temporal and spatial dependencies. Since such information is often unavailable or insufficient in our data, the problem becomes finding new contextual information that is available in our dataset, and also providing a better detection result than existing practices. In Chapter 6, we used new contextual information, i.e. cross-sensitive parameters, to construct an anomaly model and identify anomalies. The anomaly model was constructed using a Bayesian-based approach and the evaluations were carried out on both synthetic and real data. The evaluation using synthetic data shows using the cross-sensitive parameter is able to obtain a better detection result in terms of accuracy, completeness and precision than using the temporal dependency. The evaluation using a real dataset suggests the proposed method is able to further improve data quality, as the calibration result is further enhanced after the anomalies have been removed by the proposed method.

Based on the above discussion, the evidence provided in this thesis fails to reject the thesis hypothesis as restated below:

Both regression and ANN-based methods are able to improve data quality for low-cost sensors. However, the regression-based method is more suitable for our application due to lower computational cost, reduced sensitivity to the model parameters used and the need for less training data. The data quality can be enhanced by a calibration process that properly uses the supporting parameters and data quality can be further improved by applying an accurate removal of anomalies before calibration.

7.1 Key Findings

In this section, we discuss the key findings that we learnt throughout the thesis.

Systematic Methods for Evaluating Sensor Calibration

In Chapters 2 and 3, we identified that using multivariate calibration is essential for improving the data quality of low-cost sensors. However, to the best of our knowledge, there is a lack of sufficient comparison between calibration methods, which not only hinders the difference between the methods, but also prevents an appropriate calibration method being used. We identified that it is important to have a systematic method for evaluating calibration methods, as using an inappropriate evaluation metric could create artefacts in the evaluation and bias the conclusion.

In Chapter 4, we compared calibration methods between an ANN-based method and a regression-based method and used various metrics to evaluate them. The evaluation in Section 4.3.1 shows that both methods could calibrate the sensor and produce a similar RMSE value. The evaluation in Section 4.3.2.2 (Figure 4.26) shows that both methods produce a similar error distributions of the calibrations. It is noted that the RMSE value tends to be the only metric chosen by most environmental scientists and regulators for the evaluation of calibrations. This is part of the reason why the existing comparisons of calibration methods are insufficient, as they only compare the calibration results using a single evaluation metric, e.g. RMSE, which would not show a clear difference between the methods.

In the subsequent comparison in Chapter 4, we demonstrated that the scatter plot is a good evaluation metric as the scatter plots in Figures 4.21 and 4.27 were able to show a clear difference between the methods. The figures show that the predicted values from the ANN-based method are always categorical, whereas the predicted values from the regression-based method were not. This indicates that using a scatter plot is important to evaluate the results of calibration. The work in Chapter 4 suggests that certain evaluation metrics could create an artefact that hinders the difference between calibrations. Therefore, we believe it is important to investigate a systematic evaluation method for the calibration of sensors. However, for the current evaluation, it is important to look at a wide range of evaluation metrics to avoid artefacts in the evaluation.

This study presents the difference between the two calibration methods. The results show that the predicted values from the ANN-based method are always categorical. However, it was difficult to determine the reason why the ANN-based method produced such a pattern. It could be associated with the use of the data, the combination of model parameters or the training regimes.

Understanding Generalisability of Sensor Calibration

Chapter 3 illustrates that the calibration of low-cost sensors is closely related to the current environmental conditions. Thus, understanding the generalisability of sensor calibration is important as it indicates whether a calibration function can be used across different sensor units or environments. A good generalisability means that a calibration can be reused directly on new sensor units or sensors in different environments. We have tested the generalisability of sensor calibration throughout the thesis. In Section 4.3.2.1, we demonstrated that the calibration function is relatively stable with different sizes of training datasets being used. Therefore, in practice, we would have confidence in determining a good calibration function when different sizes of training datasets are used. However, in Section 5.2.4, we have shown that the calibration function has a limited generalisability when different sensor units are in use or sensors are in different environmental conditions. This implies that we have to apply new calibrations each time the use of sensor units or the environmental conditions change. Due to the limitation of the dataset, the conclusion may not be definitive. For example, we only tested the variation of calibration coefficients for a group of sensor units in the mild environment. Thus, we would not know how the calibration of coefficients would vary across sensor units in the harsh environment. Furthermore, there was only one sensor unit available in the harsh environment, which makes it impossible to obtain statistical confidence for how the environments would impact on the generalisability of the calibration function in a particular location.

It is worth pointing out that this study only suggests the calibration coefficients vary depending on the sensor units used and the environments. It was not able to identify what the factors that cause the variations were and how to compensate for them. Therefore, more knowledge would be required to answer the following questions: what is the factor that causes a variation in calibration coefficients across sensor units? How does the factor affect the behaviours of the sensor units in the mild environment?

Issues for Anomaly Detection

Data in general reflects underlying physical phenomena but it is also affected by sensing issues. Since anomalies and outliers often appear in data with similar patterns, the detection of anomalies, especially separating anomalies from outliers, can be difficult. Chapter 3 shows that anomalies are unrelated to the underlying physical phenomena, thus anomalies are often identified with respected to their actual physical phenomena. However, in practice, it is often difficult to determine the actual physical phenomena due to the fact that reference instruments often monitor the environment with different temporal resolutions, as discussed in Section 1.5.2. It is noted that the properties of anomalies change with respect to the data processing. Thus, it would be inappropriate to detect anomalies after calibration. For example, the pattern of anomalies observed before and after calibration would be different.

Unknown underlying physical phenomena not only makes the detection of anomalies more difficult but also hinders studying anomaly detection in real datasets, as we would not easily get the reference of anomalies in a real dataset for evaluation. According to the literature review on anomaly detection, the state-of-the-art methods often use relevant information (i.e. contextual information) to estimate underlying physical phenomena. For example, if consistent data patterns were presented in neighbouring sensors (spatial information), these measurements are likely to be a reflection of real physical phenomena. In Chapter 3, we have shown that the existing contextual information is not applicable in this application. Thus, we propose the use of new contextual information (i.e. cross-sensitive parameters) to estimate the underlying physical phenomena and to detect anomalies. However, it is noted that the proposed contextual information would also be unavailable in certain scenarios, e.g. a sensor unit only monitors the parameters of interest. Therefore, the use of contextual information to estimate the underlying physical phenomena would be application dependent.

The anomalies were identified in this work without knowing what their root causes were. We believe that some of the anomalies are associated with systematic causes, e.g., dust accumulated in the sensor affects the measurements and compensating for those root causes would significantly reduce the anomalies. Finding out the causes of anomalies is a challenging but important task.

7.2 Future Work

An obvious future study is to deploy sensor units in a way that they are able to conclude findings for the generalisability of sensor calibration. For example, firstly, we need to deploy a group of sensor units in lab conditions, where the environmental conditions can be fully controlled to determine how exactly the calibration coefficients vary across sensor units. Then, in the second step, the same group of sensor units with an identical set-up in different environments (e.g., the relative distance between the neighbouring sensor units needs to be identical for the deployment in different environments) needs to be deployed in different environmental conditions to determined how they would vary in different environments. The above steps would ensure the use of sensors and the environment are the only changed variables that are responsible for any change in measurements in the experiment. We would also recommend that it is essential to use a large number of sensor units in each deployment as it could provide stronger statistical confidence for the analysis.

Furthermore, we consider that it is important to better understand what experimental methods, including statistical analysis, should be performed on individual sensors. A first step would be to perform selection methods to understand the principal factors that contribute to the errors in sensor data and under what conditions. A second step would be to look at different evaluation metrics to better understand the errors related to sensors. For example, using a large dataset and data from a large group of sensor units to understand the reliability and confidence of the measurements. This would give the required confidence level under defined operational conditions. This type of information could then be used by environmental scientists in their selection of appropriate sensors.

7.3 Open Issues and Challenges

This section discusses a list of open issues and challenges related to this thesis.

Understanding the Sensing Errors

In this work, we generalised all errors observed in the data as systematic errors or random errors. We assume that systematic errors can be compensated for by the process of calibration and random errors would be removed by the process of anomaly detection. However, data errors can have many different forms and causes. For example, errors that randomly occur in the data may be associated with a systematic cause (e.g. dust accumulated in the sensing unit). The challenge would be establishing a link between the errors and their causes. By having this link, the errors can be tolerated at their source.

The Ground Truth in Uncontrolled Environments

Currently, reference sensors are used to provide the ground truth in an uncontrolled environment. However, due to the cost of the sensors, it is often impossible to have a sufficient number of references in a network, especially for large-scale, high-density sensing applications. Furthermore, as the data from the reference instrument is assumed to be the ground truth, errors incurred in the reference instrument would hinder the understanding of the errors in low-cost sensors. Therefore, the challenge is how to obtain an accurate reference for sensing applications.

Understanding the Errors Propagation and Effects

After finishing the data quality check, the sensed data will be further analysed to provide information to support decision-making. However, it is still not clear how the errors incurred at the sensing stage would be propagated throughout the data analysis and affect the final decision. We believe that a better process can be designed if we understand which errors would have the most impact on the decision-making. Therefore, understanding error propagation and its effects would be important but challenging.

Long-term Performance

The nature of the environment as well as the performance of sensors will change over the time. As a result, the data will be affected differently with respect to the original effect. The challenge is how to design a process to cope with those changes to achieve long-term performance. This is important for future work as it could dramatically reduce the maintenance costs.

Bibliography

[1] AEA Technology. QA/QC Procedures for the UK Automatic Urban and Rural Air Quality Monitoring Network (AURN) , May 2009.

[2] AEA Technology. Automatic Urban & Rural Network: Assessment of Site Classifications , 2010.

[3] F. Agostinelli, M. Hoffman, P. Sadowski, and P. Baldi. Learning activation functions to improve deep neural networks. 2014.

[4] I. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci. Wireless sensor networks: a survey. *Computer networks*, 38(4):393–422, 2002.

[5] S. Araki, H. Shimadera, K. Yamamoto, and A. Kondo. Effect of spatial outliers on the regression modelling of air pollutant concentrations: A case study in Japan. *Atmospheric Environment*, 2017.

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

[7] 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.

[8] J. Branch, C. Giannella, B. Szymanski, R. Wolff, and H. Kargupta. In-network outlier detection in wireless sensor networks. *Knowledge and information systems*, 34(1):23–54, 2013.

[9] V. Bychkovskiy, S. Megerian, D. Estrin, and M. Potkonjak. A collaborative approach to in-place sensor calibration. In *Information Processing in Sensor Networks*, pages 301–316. Springer, 2003.

[10] N. Castell, F. 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*, 99:293–302, 2017.

[11] 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*, pages 251–265. ACM, 2014.

[12] City of York Council. Air quality monitoring, Mar. 2017.

[13] Department for Environment Food & Rural Affairs. A Guide for Local Authorities Purchasing Air Quality Monitoring Equipment , Mar. 2006.

[14] Department for Environment Food & Rural Affairs. Quality Assurance and Quality Control (QA/QC) Procedures for UK Air Quality Monitoring under 2008/50/EC and 2004/107/EC, May 2012.

[15] Department for Environment Food & Rural Affairs. Automatic urban and rural network (AURN), Mar. 2015.

[16] Department for Environment Food & Rural Affairs. Local air quality management (LAQM), 2015.

[17] Department for Environment Food & Rural Affairs. Monitoring networks, Mar. 2017.

[18] Department for Environment Food & Rural Affairs. New air quality plan published for consultation, May 2017.

[19] S. Devito, E. Esposito, M. Salvato, O. Popoola, F. Formisano, R. Jones, and G. Difrancia. Calibrating chemical multisensory devices for real world applications: An in-depth comparison of quan-

titative machine learning approaches. *Sensors and Actuators B: Chemical*, 255:1191–1210, 2018.

[20] S. Devito, M. Piga, L. Martinotto, and G. Difrancia. *co*, no_2 and no_x urban pollution monitoring with on-field calibrated electronic nose by automatic bayesian regularization. *Sensors and Actuators B: Chemical*, 143(1):182–191, 2009.

[21] J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159, 2011.

[22] E. Elnahrawy and B. Nath. Context-aware sensors. In *European Workshop on Wireless Sensor Networks*, pages 77–93. Springer, 2004.

[23] E. Esposito, S. Devito, 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.

[24] E. Esposito, S. Devito, M. Salvato, G. Fattoruso, and G. Difrancia. Computational intelligence for smart air quality monitors calibration. In *International Conference on Computational Science and Its Applications*, pages 443–454. Springer, 2017.

[25] European Commission. Air quality standards, 2017.

[26] European Union. Directive 2008/50/ec of the european parliament and of the council of 21 May 2008 on ambient air quality and cleaner air for europe. *Official Journal of the European Union*, 2008.

[27] 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)*. ACM, 2017.

[28] X. Fang and I. Bate. Using multi-parameters for calibration of low-cost sensors in urban environment. In *International Conference on Embedded Wireless Systems and Networks (EWSN)*, 2017.

[29] J. Feng, S. Megerian, and M. Potkonjak. Model-based calibration for sensor networks. In *Proceedings of IEEE Sensors*, volume 2, pages 737–742. IEEE, 2003.

[30] P. Filzmoser, M. Gschwandtner, and V. Todorov. Review of sparse methods in regression and classification with application to chemometrics. *Journal of Chemometrics*, 26(3-4):42–51, 2012.

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

[32] Y. Gao, W. Dong, K. Guo, X. Liu, Y. Chen, X. Liu, J. Bu, and C. Chen. Mosaic: A low-cost mobile sensing system for urban air quality monitoring. In *IEEE International Conference on Computer Communications*, pages 1–9. IEEE, 2016.

[33] J. Han and C. Moraga. The influence of the sigmoid function parameters on the speed of backpropagation learning. In *International Workshop on Artificial Neural Networks*, pages 195–201. Springer, 1995.

[34] W. Härdle, S. Klinke, and B. Rönz. *Introduction to Statistics: Using Interactive MM * Stat Elements*. Springer, 2015.

[35] D. Hasenfratz, O. Saukh, and L. Thiele. On-the-fly calibration of low-cost gas sensors. In *European Conference on Wireless Sensor Networks*, pages 228–244. Springer, 2012.

[36] A. Hayes. *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach.* Guilford Press, 2013.

[37] I. Heimann, V. Bright, M. McLeod, M. Mead, O. Popoola, G. Stewart, and R. Jones. Source attribution of air pollution by spatial scale

separation using high spatial density networks of low cost air quality sensors. *Atmospheric Environment*, 113:10–19, 2015.

[38] S. Hochreiter and J. Schmidhuber. Long short-term memory. 9:1735–80, 12 1997.

[39] V. Hodge and J. Austin. A survey of outlier detection methodologies. *Artificial intelligence review*, 22(2):85–126, 2004.

[40] D. Janakiram, V. Reddy, and A. Kumar. Outlier detection in wireless sensor networks using bayesian belief networks. In *First International Conference on Communication System Software and Middleware, 2006*, pages 1–6. IEEE, 2006.

[41] Keras. The Python Deep Learning library, 2017.

[42] N. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. Tang. On large-batch training for deep learning: Generalization gap and sharp minima. In *International Conference on Learning Representations*, 2016.

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

[44] P. Landrigan, R. Fuller, N. Acosta, O. Adeyi, R. Arnold, A. Baldé, R. Bertollini, S. Bose-O'Reilly, J. Boufford, and P. Breysse. The lancet commission on pollution and health. *The Lancet*, 2017.

[45] A. Lazarevic, L. Ertoz, V. Kumar, A. Ozgur, and J. Srivastava. A comparative study of anomaly detection schemes in network intrusion detection. In *Proceedings of the 2003 SIAM International Conference on Data Mining*, pages 25–36. SIAM, 2003.

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

[47] J. Li, B. Faltings, O. Saukh, D. Hasenfratz, and J. Beutel. Sensing the air we breathe-the opensense zurich dataset. In *Proceedings of the National Conference on Artificial Intelligence*, volume 1, pages 323–325, 2012.

[48] J. Lipor and L. Balzano. Robust blind calibration via total least squares. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4244–4248. IEEE, 2014.

[49] B. Maag, O. Saukh, D. Hasenfratz, and L. Thiele. Pre-deployment testing, augmentation and calibration of cross-sensitive sensors. pages 169–180. ACM, 2016.

[50] B. Maag, Z. Zhou, O. Saukh, and L. Thiele. Scan: Multi-hop calibration for mobile sensor arrays. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(2):19, 2017.

[51] H. Mayer. Air pollution in cities. *Atmospheric environment*, 33(24-25):4029–4037, 1999.

[52] M. Mead, O. Popoola, G. Stewart, P. Landshoff, M. Calleja, M. Hayes, J. Baldovi, M. McLeod, T. Hodgson, and J. Dicks. The use of electrochemical sensors for monitoring urban air quality in low-cost, high-density networks. *Atmospheric Environment*, 70:186–203, 2013.

[53] T. Mehmood, K. Liland, L. Snipen, and S. Sæbø. A review of variable selection methods in partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, 118:62–69, 2012.

[54] 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.

[55] M. A. Nielsen. *Neural Networks and Deep Learning*. Determination Press, 2015.

[56] R. Obrien. A caution regarding rules of thumb for variance inflation factors. *Quality & Quantity*, 41(5):673–690, 2007.

[57] T. Palpanas, D. Papadopoulos, V. Kalogeraki, and D. Gunopulos. Distributed deviation detection in sensor networks. *ACM SIGMOD Record*, 32(4):77–82, 2003.

[58] Perkin Elmer. Elm sensor, 2015.

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

[60] A. Ponzoni, C. Baratto, N. Cattabiani, M. Falasconi, V. Galstyan, E. Nunez-Carmona, F. Rigoni, V. Sberveglieri, G. Zambotti, and D. Zappa. Metal oxide gas sensors, a survey of selectivity issues addressed at the sensor lab, brescia (italy). *Sensors*, 17(4):714, 2017.

[61] N. Qian. On the momentum term in gradient descent learning algorithms. *Neural networks*, 12(1):145–151, 1999.

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

[63] S. Rajasegarar, C. Leckie, M. Palaniswami, and J. Bezdek. Quarter sphere based distributed anomaly detection in wireless sensor networks. In *IEEE International Conference on Communications, 2007.*, pages 3864–3869. IEEE, 2007.

[64] M. Rassam, A. Zainal, and M. Maarof. Advancements of data anomaly detection research in wireless sensor networks: A survey and open issues. *Sensors*, 13(8):10087–10122, 2013.

[65] B. Resch, M. Mittlboeck, F. Girardin, R. Britter, and C. Ratti. Live geography: Embedded sensing for standarised urban environ-

mental monitoring. *International Journal on Advances in Systems and Measurements*, 2009.

[66] RICARDO. Automatic Urban and Rural Network: Site Operator?s Manual, May 2017.

[67] 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*, pages 274–285. ACM, 2015.

[68] O. Saukh, D. Hasenfratz, C. Walser, and L. Thiele. On rendezvous in mobile sensing networks. In *Real-World Wireless Sensor Networks*, pages 29–42. Springer, 2014.

[69] A. Schütze, M. Leidinger, B. Schmitt, T. Sauerwald, M. Rieger, and C. Alépée. A novel low-cost pre-concentrator concept to boost sensitivity and selectivity of gas sensor systems. In *SENSORS*, pages 1–4. IEEE, 2015.

[70] N. Shahid, I. Naqvi, and S. Qaisar. Quarter-sphere svm: attribute and spatio-temporal correlations based outlier & event detection in wireless sensor networks. In *Wireless Communications and Networking Conference (WCNC)*, pages 2048–2053. IEEE, 2012.

[71] N. Shahid, I. Naqvi, and S. Qaisar. Characteristics and classification of outlier detection techniques for wireless sensor networks in harsh environments: a survey. *Artificial Intelligence Review*, 43(2):193–228, 2015.

[72] A. Sharma, L. Golubchik, and R. Govindan. Sensor faults: Detection methods and prevalence in real-world datasets. *ACM Transactions on Sensor Networks (TOSN)*, 6(3):23, 2010.

[73] B. Sheng, Q. Li, W. Mao, and W. Jin. Outlier detection in sensor networks. In *Proceedings of the 8th ACM international symposium on Mobile ad hoc networking and computing*, pages 219–228. ACM, 2007.

[74] K. Silver. Pollution linked to one in six deaths (BBC news), 2017.

[75] K. Smith, P. Edwards, M. Evans, J. Lee, M. Shaw, F. Squires, and A. Lewis. Clustering approaches to improve the performance of low cost air pollution sensors. *FARADAY DISCUSSIONS*, pages 1–15, 2017.

[76] 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 *SENSORS*, pages 21–24. IEEE, 2014.

[77] L. Spinelle, M. Gerboles, M. 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*, 215:249–257, 2015.

[78] A. Studenmund. *Using Econometrics: A Practical Guide*. Addison-Wesley Series in Economics. Pearson Addison Wesley, 2006.

[79] S. Subramaniam, T. Palpanas, D. Papadopoulos, V. Kalogeraki, and D. Gunopulos. Online outlier detection in sensor data using non-parametric models. In *Proceedings of the 32nd international conference on Very large data bases*, pages 187–198. VLDB Endowment, 2006.

[80] L. Sun, D. Westerdahl, and Z. Ning. Development and evaluation of a novel and cost-effective approach for low-cost NO_2 sensor drift correction. *Sensors*, 17(8):1916, 2017.

[81] B. Szulczyński and J. Gebicki. Currently commercially available chemical sensors employed for detection of volatile organic compounds in outdoor and indoor air. *Environments*, 4(1):21, 2017.

[82] R. Tan, G. Xing, Z. Yuan, X. Liu, and J. Yao. System-level calibration for data fusion in wireless sensor networks. *ACM Transactions on Sensor Networks (TOSN)*, 9(3):28, 2013.

[83] Tensorflow. An open-source software library for Machine Intelligence, 2017.

[84] S. Vardoulakis, B. Fisher, K. Pericleous, and N. Gonzalez-Flesca. Modelling air quality in street canyons: a review. *Atmospheric environment*, 37(2):155–182, 2003.

[85] C. Wang, B. Esse, and A. Lewis. *Low-cost multispecies air quality sensor*, volume 198, pages 105–116. WITPress, 6 2015.

[86] K. Whitehouse and D. Culler. Calibration as parameter estimation in sensor networks. In *Proceedings of the 1st ACM International Workshop on Wireless Sensor Networks and Applications*, pages 59–67. ACM, 2002.

[87] World Health Organization. *Air quality guidelines global update 2005: particulate matter, ozone, nitrogen dioxide, and sulphur dioxide*. World Health Organization, 2006.

[88] Y. Xiang, L. Bai, R. Piedrahita, R. Dick, Q. Lv, M. Hannigan, and L. Shang. Collaborative calibration and sensor placement for mobile sensor networks. In *Proceedings of the 11th International Conference on Information Processing in Sensor Networks*, pages 73–84. ACM, 2012.

[89] Y. Xiang, L. Bai, R. Piedrahita, R. Dick, Q. Lv, M. Hannigan, and L. Shang. Collaborative calibration and sensor placement for mobile sensor networks. In *Proceedings of the 11th International Conference on Information Processing in Sensor Networks*, pages 73–84. ACM, 2012.

[90] M. Xie, J. Hu, S. Han, and H. Chen. Scalable hypergrid k-nn-based online anomaly detection in wireless sensor networks. *IEEE Transactions on Parallel and Distributed Systems*, 24(8):1661–1670, 2013.

[91] K. Zhang, S. Shi, H. Gao, and J. Li. Unsupervised outlier detection in sensor networks using aggregation tree. *Advanced data mining and applications*, pages 158–169, 2007.

[92] Y. Zhang, N. Hamm, N. Meratnia, A. Stein, M. van de Voort, and P. Havinga. Statistics-based outlier detection for wireless sensor networks. *International journal of geographical information science*, 26(8):1373–1392, 2012.

[93] Y. Zhang, N. Meratnia, and P. Havinga. An online outlier detection technique for wireless sensor networks using unsupervised quarter-sphere support vector machine. In *International Conference on Intelligent Sensors, Sensor Networks and Information Processing, 2008.*, pages 151–156. IEEE, 2008.

[94] Y. Zhang, N. Meratnia, and P. Havinga. Outlier detection techniques for wireless sensor networks: A survey. *IEEE Communications Surveys & Tutorials*, 12(2):159–170, 2010.

[95] X. Zhou, C. Yang, and W. Yu. Moving object detection by detecting contiguous outliers in the low-rank representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(3):597–610, 2013.

[96] Y. Zhuang and L. Chen. In-network outlier cleaning for data collection in sensor networks. In *CleanDB*, 2006.