An Improved Sensor Calibration with Anomaly Detection and Removal

Xinwei Fang[a], Iain Bate[a]

[a] *Department of Computer Science, University of York, York, UK*

Abstract

Sensor calibration is a widely adopted process for improving data quality of low-cost sensors. However, such a process may not address data issues caused by anomalies. Anomalies are considered as data errors that are inconsistent to the actual physical phenomena. This paper presents an improved sensor calibration, which applies a process for detection and removal of anomalies before the sensor calibration process. A Bayesian-based method is used for anomaly detection that takes advantage of cross-sensitive parameters in a sensor array. The method utilises dependencies between cross-sensitive parameters, which allows underlying physical phenomena to be modelled and anomalies to be detected. The calibration approach is based on stepwise regression, which automatically and systematically selects suitable supporting parameters for a calibration function. The evaluation for anomaly detection shows that the results are better than the state-of-the-art methods, in terms of accuracy, precision and completeness. The overall evaluation confirms that data quality can be further enhanced when anomalies are removed before the calibration.

1. Introduction

Regulatory environmental monitoring is a key approach to understand urban environment and has been enforced by law <1>. Currently, high-quality monitoring instruments, which are also called reference instruments, are used for such purpose <2; 3>. Reference instruments provide data with good quality but they are extremely expensive. This limits the number of reference instruments that have been deployed, which further affects the spatial and temporal resolution of data <4; 5; 6>.

To overcome the existing limitations on data, larger numbers of low-cost sensor have been considered <7>. Individual low-cost sensors often trade data quality with usability, price, power consumption and size <8; 9; 10; 11>. For example, the data from an individual low-cost sensor may be significantly inconsistent with reference data. This issue further magnifies when sensors work in dynamic environments such as a city centre <12>. Compared to reference instruments, data from low-cost sensor could have: 1) lower data accuracy, 2) a high percentage of anomalies, and 3) unexpected data patterns (i.e., constant values) <13; 14; 15>. As a result, data from low-cost sensors should not be utilised without proper processes <16; 17>.

Sensor calibration is a widely adopted process for improving data quality. The state-of-the-art approaches use multiple variables to construct a calibration function. These approaches are referred to as the multivariate calibration in this paper <16; 12; 18; 19; 20>. Using the calibration of NO_2 as an example, a multivariate calibration not only uses the parameters of NO_2 but also uses other monitored parameters. The other parameters, such as temperature and humidity, are referred to as the supporting parameters in the multivariate calibration. The intuition of this is, if the response of NO_2 is related to temperature and humidity, a more accurate calibration of NO_2 can be derived when the calibration function includes these parameters and considers their effects <16; 21; 22; 23>.

Even though existing studies show that multivariate calibrations can significantly improve the data quality, their calibration errors are still large <12; 22>. Many of data errors can occur randomly without a systematic pattern nor with zero mean. Without knowing the actual root causes, such errors are difficult to compensate <24>. We assume that those data errors are responsible for the large calibration errors as they would not be compensated for by calibrations <25; 26>. In this work, we consider those data errors as anomalies and further assume that data quality can be improved if those anomalies are accurately detected and removed. Anomalies should not be confused with outliers as outliers are genuine but unusual measurements, e.g. a data spike caused by a bus idling next to a sensor, whereas anomalies are data spikes caused by a variety of reasons such as communication errors. Simple techniques for isolating data spikes such as using the difference between the data and a running mean are not utilising contextual information. These methods, therefore, may not be appropriate for anomaly detection in this application as they may not be able to differentiate anomalies from outliers (e.g. anomalies and outliers have a similar mean value).

To identify anomalies, the '***normal***' needs to be defined. This '***normal***' is denoted as the anomaly model in this work. An anomaly is identified if a data instance deviates from the anomaly model with a pre-defined threshold. Thus, the challenge of anomaly detection becomes how to determine a good anomaly model. The state-of-the-art method utilises contextual information to estimate anomaly models. This is due to the fact that anomalies tend to be stochastically unrelated to contextual information <27>. The spatial and the temporal de-

Email addresses: xinwei.fang@york.ac.uk (Xinwei Fang), iain.bate@york.ac.uk (Iain Bate)

pendencies are the mostly used contextual information for such purposes <28>. Using a spatial dependency as an example, anomalies are identified if the data from nearby sensors have a significantly and un-explainable different data pattern. However, as spatial and temporal dependencies are often weak in urban environments <29>, obtaining an accurate anomaly model is a challenging task.

This paper provides an improved sensor calibration, which applies a process for detection and removal of anomalies before the sensor calibration. The method of calibration is from <29> and based on our previous publication <12> where advantage is taken from using stepwise regression. This allows suitable supporting parameters to be automatically and systematically selected according to the context of use. Furthermore, a novel anomaly detection approach is proposed using a Bayesian-based method that takes advantage of cross-sensitive parameters in a sensor array. The method utilises dependencies between cross-sensitive parameters, which allows anomaly model to be determined and anomalies to be identified, i.e., a higher than normal value of X may be considered as an anomaly if it deviate from the estimated model by a threshold value.

The contributions of this particular paper over our previous works in <12; 30; 29> are listed below.

- Demonstrate that the cross-sensitivity of sensors can be used to derive an anomaly model (To the best of our knowledge this is the first work).
- Demonstrate that the derived anomaly model is able to obtain a better result than the one utilising temporal dependency.
- Demonstrate an improved calibration result when anomalies are removed.
- Application of the techniques to a second case study.

The rest of this paper is organised as follows. The novel method for anomaly detection is firstly proposed in Section 2, which is then followed by the method of calibration in Section 3. Evaluations are presented in Section 4. Finally, in Section 5, we summarise and discuss our findings.

2. Detection of anomalies

As discussed in Section 1, using a threshold value would only identify outliers, which is not ideal for anomaly detection in this work. Furthermore, as the data from the low-cost sensors and reference instruments are inconsistent, using co-located reference data to identify anomalies can be insufficient <12; 22>. Therefore, the challenge for anomaly detection is how to obtain a good anomaly model <28; 27>.

The state-of-the-art methods utilise contextual information to estimate anomaly model, such as spatial or temporal dependencies of data. For example, measurements from neighbouring sensors or adjacent time stamps are expected to be similar as they are sensing a similar environment. In that case, an anomaly is identified if the measurement exhibits a significantly different value (defined by a threshold) to it neighbouring sensors (spatial) or adjacent measurements (temporal). However, the confidence of an anomaly model would be significantly affected when the spatial or the temporal dependencies are weak. Considering the spatial and temporal dependency are generally weak in an urban environment <28; 29>, we propose to use new contextual information which is the dependency between cross-sensitive parameters to estimate anomaly model. To the best our knowledge, this is first time this feature being used for anomaly detection. In this section, we first explain what cross-sensitive parameters are, and then discuss how this dependency is utilised for anomaly detection.

ALGORITHM 1: The pseudo code for detection of anomalies

```
Data:
  $C_{n\times1}$: the measurements of the cross-sensitive
  parameter
  $I_{n\times1}$: the measurements of the parameter of interest
  ($n$ indicates the number of measurements)
Result: the measurements with anomalies labelled
/* Remove invalid data (marked as NaN).
   Assume the number of measurements
   changes to $m$ after this step.             */
for $i$ = 1 to $n$ do
    if $(C_i == NaN)$ or $(I_i == NaN)$ then
        remove $C_i$ and $I_i$;
    end
end
begin
    /* Produce the joint probability table,
       $P(\mathrm{I}, \mathrm{C})$                  */
    classify $C_{m\times1}$ into $k$ classes;
    classify $I_{m\times1}$ into $j$ classes;
    declare a frequency table: $T_{k\times j}$;
    for $i$ = 1 to $m$ do
        $x = C(i).class$;
        $y = I(i).class$;
        $T(x, y) \leftarrow T(x, y) + 1$;
    end
    $P \leftarrow T./m$;

    /* Labelling all the measurements          */
    for $i$ = 1 to $m$ do
        if $I_i$ is not in $P(I,C)$ then
            mark_as_anomaly(I($i$));
        end
        /* the conditional distribution I by
           using the index class of $C$        */
        $p = P(\mathrm{I} \mid \mathrm{C}(i).class)$;
        if $p < threshold$ then
            mark_as_anomaly(I($i$));
        end
    end
end
```

2.1. Cross-sensitive parameters

Cross-sensitivity is defined as the sensitivity to one substance which renders the sensors sensitive to other substances. It is known that low-cost sensors are often cross-sensitive to each other, such as NO_2 and O_3. For example, the readings from an NO_2 sensor would be dependent on the concentration of O_3 in the mixed air. In this case, O_3 is considered to be a cross-sensitive parameter of the NO_2 sensor. Assuming an NO_2 sensor has a response to O_3 at a rate of 50% due to the 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. 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. This implies that changes in one sensor would lead to an effect on another sensor. It is noted that the rate of cross-sensitivity can vary in different conditions, and thus the dependency does not have a fixed and a functional relationship <31>. In the next section, we utilise this property of low-cost sensors to construct anomaly model.

2.2. Learning the information

In this work, we employ a Bayesian-based method to learn the dependency between cross-sensitive parameters (e.g., NO_2 and O_3). A Bayesian-based method is chosen as it allows the likelihood of a particular value (i.e. the value sensed for the parameter of interest) to be predicted given a number of other values (i.e. its cross-sensitive parameters) occurring. If the likelihood is lower than a threshold then it can conclude an anomaly exists. 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, $P(\mathrm{I} \mid \mathrm{C}(i).class)$
- A conditional probability, $P(\mathrm{I}(i).class \mid \mathrm{C}(i).class)$

To determine the joint probability, it requires that the I and C have the same number of measurements. As data may contain invalid values (marked as NaN) <30>, a pre-process is used first to unify the number of measurements (i.e. if one set have a value of NaN for a particular time stamp then the corresponding value is removed from the other set).

Bayesian methods rely on the data being assigned to bins, and hence a bin size is important. A bin size that is too small could result in the histogram having an non-distinct mode, which makes anomalies inseparable from the data. A bin size that is too large could reduce the precision of the method leading to more false positives. Our data is with two significant digits. Using the data directly without further discretisation would result in the bin size becoming too small, especially when the number of samples in the dataset is relatively small (i.e., 4000 samples). Hence, a further discretisation is required for the measurement sets, I and C. We determine the bin size using a two dimensional histogram approach based on <32>. This involves a wide range of bin sizes being tested. The "best" bin size is not unique as long as each bin has a sufficient number of counts and the histogram has a distinct mode. The process of discretisation classifies the set I and C into j and k classes. Then, a joint probability table can be obtained <33>.

2.3. Inferencing and detection

Once the joint probability table is determined, it can be used as the anomaly model to make an inference and statistically identify anomalies. The probability distribution of I at a given value of $C(i)$ can be obtained according to the class number, which is $P(\mathrm{I} \mid \mathrm{C}(i).class)$. The probability of $\mathrm{I}(i).class$, at given value of $\mathrm{C}(i).class$, can be determined as $P(\mathrm{I}(i).class \mid \mathrm{C}(i).class)$. If this probability is less than a threshold value, this measurement is considered as an anomaly and removed from the data. As the threshold value is sensitive to the use of data, expert or domain knowledge is required to determine a 'good' threshold value. We will discuss some observations for selecting a 'good' threshold value in Section 4.3.2. The overall method for the detection of anomalies is in Algorithm 1.

3. Calibration

In this section, we introduce the calibration method. The method was used in <29> and is a refinement of the work <12>. The key difference is this method tests for whether both adding and removing of a variable would improve the calibration, whereas the method in work <12> only considered effects of adding variables. As this method terminates when no single step improves the model, the calibration result by design is unlikely sensitive to a different sequence of steps.

3.1. Regression for calibration

Regression methods have been widely used for the calibration of low-cost sensors, such as uni-variate calibrations <34> and multivariate calibrations <18>. Regression is often optimised by the ordinary least square method that minimises the difference between the "independent variables" and the "dependent variable" <35>. For the sensor calibration, the independent variables and the dependent variable can be considered as the parameter of interest with other supporting parameters and the reference respectively.

Assuming there are j number of independent variables X and one dependent variable Y. A multivariate linear regression model at any time instance i can be constructed as shown in Equation 1.

$$Y_i = \beta_0 + \beta_1 \cdot X_{i,1} + \beta_2 \cdot X_{i,2} + \cdots \beta_j \cdot X_{i,j} + \varepsilon_i \qquad (1)$$

The calibration function is to determine the coefficients, β, when the error term, ε, is minimised as shown in Equation 2.

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

3.2. Two-way interaction terms

According to Equation 1, the variables X in the calibration function are independent, which means the changing of one variable would not affect another one. However, this is not the case in real environment. For example, a higher temperature could result in a higher chemical reaction rate. This rate can then affect the actual concentration of a parameter and further reflect on the sensor readings. In this section, two-way interaction terms are utilised to uncover the relationships between parameters <36>.

An interaction term, which is also known as a moderation term, is a multiplication of any two variables. Assuming the calibration of a parameter of interest only needs one supporting parameter. A calibration function can be simplified presented in Equation 3.

$$Y \sim \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 \tag{3}$$

Adding a two-way interaction term into Equation 3 gives Equation 4:

$$Y \sim \beta_0' + \beta_1' \cdot X_1 + \beta_2' \cdot X_2 + \beta_3' \cdot (X_1 \cdot X_2) \tag{4}$$

which can be re-written as Equation 5:

$$Y \sim \beta_0' + (\beta_1' + \beta_3' \cdot X_2) \cdot X_1 + \beta_2' \cdot X_2 \tag{5}$$

According to the Equation 5, the variable X_2 is associated with the variable X_1 as the variation of the X_2 would impact the coefficient of the X_1. Hence, the variation of one variable can also affect the value of another. In this work, we include interaction of parameters in the calibration process to maximise the relationship among parameters.

3.3. Stepwise regression

The parameters available depends on the sensor units used; and the important parameters are dependent on both the particular sensor unit and where it is deployed. As a result, the selection of the parameters can be important to ensure the calibration result. Stepwise regression differs from multivariate regression by first performing a systematic selection of the parameters to be used in the regression. Only the parameters that have positive contribution to the calibration are chosen. This enables automatic and systematic selection of supporting parameters, which significantly reduces the human intervention in this process.

For all monitored parameters including their interaction terms, the method starts with fitting a model using just one term. At each step, p-value for an F-test of the change in the sum of squared error (SSE) when adding or removing one term are calculated. The p-value is used to make decisions whether to add to remove terms. If the term is not currently in the model, the

ALGORITHM 2: The pseudo code for calibration using stepwise regression

```
Data: /* V indexes the use of parameters in X */
    $X = \{x_1(n), x_2(n), \cdots, x_m(n)\} \in \mathcal{R}^{m\times n}$;
    $Y = \{y(n)\} \in \mathcal{R}^{1\times n}$;
    $V = \{V_1, V_2, \cdots, V_m\}$.
Result: Calibration Function: Y = f(V)
/* add intersection terms */
begin
    $k \leftarrow m + 1$;
    for $i = 1$ to $m$ do
        for $j = i + 1$ to $m$ do
            $x_k \leftarrow x_i \times x_j$;
            $X$.append($x_k$);
            $V$.append($V_k$);
            $k \leftarrow k + 1$;
        end
    end
end
/* perform stepwise regression */
begin
    $V_c = \{\varnothing\}$;
    do
        /* try to add terms */
        do
            for $V_i$ in $V$ do
                P_candidate = evaluate_p_value_with($V_i$)
                  & p<0.05;
            end
            /* $V_k$ is the candidate term */
            k ← find_the_lowest_p(P_candidate);
            $V_c$.append($V_k$);
            $V$.remove($V_k$);
        while P_candidate $\neq \varnothing$;
        /* try to remove terms */
        for $V_i$ in $V_c$ do
            P_candidate = evaluate_p_value_without($V_i$)
              & p>0.05;
        end
        k ← find_the_highest_p(P_candidate);
        $V_c$.remove($V_k$);
        $V$.append($V_k$);
    while $V_c$_has_changed();
    calculate_coefficients($Y$, $X$, $V_c$);
end
```

null hypothesis is that the added new term would have a zero coefficient in the model. If there is sufficient evidence ($p < 0.05$) 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 ($p > 0.05$), the term with the highest p-value is removed from the model. This method applies the following steps:

1. Construct the initial model using just one term.
2. For terms not in the current model with 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 with 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. Compared to the method used in <12> which only considered how results would improve by adding terms, this method also tests whether remove a term in the existing model would further improve the result. The revised approach can be more robust as the sequence of adding the parameters is unlikely to influence the result. The overall method for the proposed sensor calibration is illustrated in Algorithm 2.

4. Experiment and evaluation

In this section, we first introduce the real datasets used in the evaluation. Then, we discuss a data pre-processing that performed before the calibration, i.e., aggregating data with different temporal resolutions. Finally, we evaluate the methods in both artificial and real datasets.

4.1. Datasets

There were two real datasets used in the evaluation. One dataset was obtained from our own deployment by an ELM unit. The ELM unit is a product from Perkin Elmer <37>. Multiple parameters were monitored, which are nitrogen dioxide (NO_2), ozone (O_3), nitrogen oxide (NO), temperature (T) and humidity (H). The NO_2 and O_3 were measured by metal-oxide sensors and a dielectric film was used to measure temperature and humidity. The ELM unit was located at the city centre next to a busy junction and co-located with a reference instrument in early 2016 as illustrated in Figure 1. The reference instrument (EU Site ID: GB0919A) is jointly managed by the City of York Council and Automatic Rural and Urban Networks (ARUN). The temporal resolution is 20 seconds for the ELM data and 1 hour for the reference data.

The second dataset was obtained in Beijing in 2017 and used in <38>. The dataset contains NO_2, O_3, carbon monoxide (CO), T, H and volatile organic compounds (VOC). It is noted that parameters, NO_2, O_3, CO and VOC, were monitored by multiple identical low-cost sensors. To enable cross-comparison with <38>, the median value across identical sensors was used to represent the concentration of the parameters in the evaluation. This dataset and the reference has the temporal resolution of one minute. More detail on this dataset can be found in <39>.

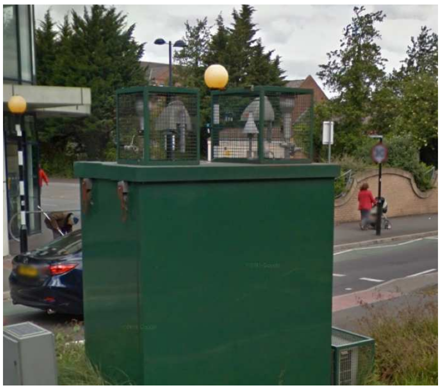

Figure 1: Sensors at Fishergate, York

4.2. Data pre-processing

Regression-based method requires the dependent and independent variables to have the same number of measurements. Since the data from ELM unit has a different temporal resolution to the reference, a pre-processing of the data is needed. Extrapolating the reference data from an hour to 20 seconds may introduce a large uncertainty. Therefore, the data pre-processing aggregates the ELM data (20 seconds) into the same temporal resolution as the reference (hourly). In this work, the hourly aggregation is based on a window from the current whole hour to the next whole hour. For example the value for 12:00:00 is obtained from the samples between (≥) 12:00:00 and (<) 13:00:00. We tested a wide range of windows, the aggregation result did not have a noticeable difference in terms of correlation.

There are many techniques available for data aggregation. Taking the arithmetic mean and median over samples in a window are most commonly used ones. The arithmetic mean is the sum of the received values divided by the number of counts. Thus, the confidence level of the aggregation result is by definition sensitive to the number of samples received in each window. Moreover, the arithmetic mean is sensitive to extreme values. For example, the mean value could be biased if extreme but erroneous value existed in the samples. Even with anomaly detection, these could still exist. However, it does not imply that using the median value is always a better option. As the median value is a single value, it will not be representative of other samples. If spikes are caused by real events, taking the median value would ignore that information. In addition, the median value would still be biased if the percentage of anomalies is more than 50% in the samples. Considering the number of samples received in an hour can be significantly inconsistent <29> and the anomalies are unlikely to be more than 50% of the hourly samples, the median is selected to aggregate the data. The process of the data pre-processing is illustrated in Algorithm 3

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

ALGORITHM 3: Pseudo code for data pre-processing

```
Data:
Dataset from low-cost sensors, $D_{m\times n}$
/* 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.      */
Reference, $R_{r\times 2}$
/* 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 Not
   a Number (NaN).                       */
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.
/* Hourly averaged data for low-cost
   sensors (contains NaN)                */
for $i = 0$ to $m-1$ do
    for $j = 2$ to $n$ do
        /* Determine all values that measured
           within that hour              */
        $D^* = D(\text{find}(t_i \leqslant D(:,1) < t_{i+1}),j)$;
        if $D^*.size < 5$ then
            T(i,j) = NaN;
        else
            /* The nan-median takes median
               without considerin the NaN   */
            T(i,j) = nan-median($D^*$);
        end
    end
end
Join the $R$ with T according to the time-stamp and
  remove all NaN instances in the dataset.
```

data gaps occurred in the reference, the relevant data from the corresponding sensors are removed for consistency.

4.3. Evaluation of the anomaly detection

Obtaining the ground truth for anomalies in the real dataset is practically difficult. Hence, a synthetic dataset was used for the evaluation of anomaly detection. The evaluation was carried out in a similar way to other research determining accuracy, precision and completeness. We firstly introduce how the synthetic dataset was generated in Section 4.3.1. Then, we illustrate how the performance of the method is related to the different threshold values in Section 4.3.2, and cross-compare the proposed method to the one that uses temporal dependence in Section 4.3.3. Finally, in section 4.3.4 we discuss our findings as well as the limitations of the method.

4.3.1. Synthetic data

The synthetic data was constructed by injecting anomalies into a clean dataset. The use of a clean dataset is to avoid unwanted false positives as any anomalies exhibit in the base signal would not be correctly labelled as the anomaly. The base signal of the clean dataset was taken from a reference instrument 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 the gap using linear interpolation. This process maximises the consistency of temporal information which enables a fair comparison to the method that utilises temporal dependency. The clean dataset after this process is free from anomalies and temporally consistent.

Since 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 magnitude of anomalies that will be injected back to the dataset was randomly chosen between 10% to 60% of the maximum values of the clean signal. Furthermore, we decided to inject 8% of the anomalies to the clean dataset as there was 8% of outliers in the low-cost sensor dataset <29>. It is noted that the realism of the injected anomalies is not significant to the final intended outcome, i.e. the dataset allows the evaluation of whether the processing techniques advocated in this paper improve the signal over the state-of-the-art methods. Later in the paper, the overall processing techniques is then evaluated in real datasets with potentially real anomalies to show whether an improvement is also achieved.

The constructed synthetic data for NO_2 is illustrated in Figure 2. In the figure, the clean base signal is in red and the injected anomalies are in blue.

4.3.2. Threshold Value

As discussed in Section 2, the determination of a threshold can be difficult as it often requires expert knowledge and can be data dependent. Therefore in this section we demonstrate how the results of anomaly detection are related to threshold values. The results should assist experts to make more appropriate selection.

The results of the anomaly detection are evaluated in terms of accuracy, precision and completeness, which are normalised in the range from 0 to 1. Those metrics are defined as in Equation 6 to 8.

$$Accuracy = \frac{(\text{\#True Positives} + \text{\#True Negatives})}{\text{\#Samples}} \qquad (6)$$

$$Precision = \frac{\text{\#True Positives}}{\text{\#True Positives} + \text{\#False Positives}} \qquad (7)$$

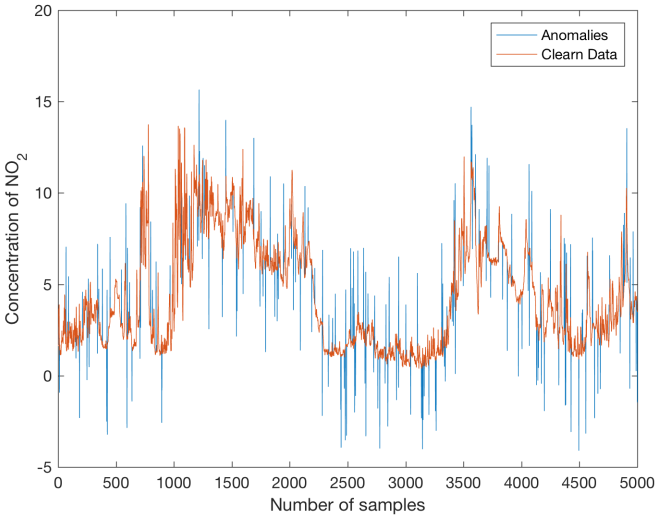

Figure 2: The synthetic dataset

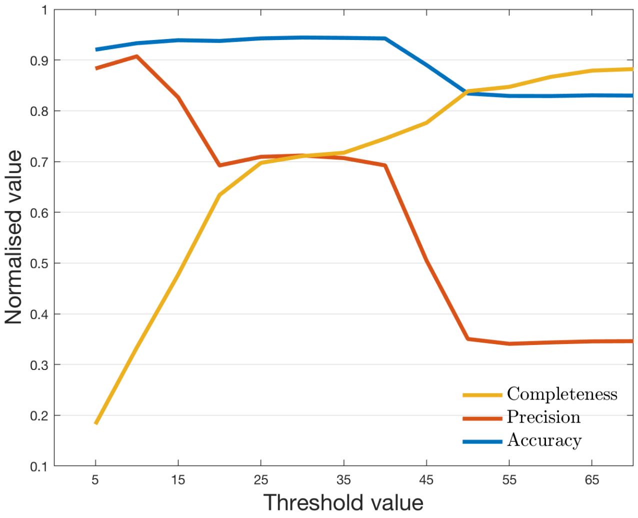

Figure 3: The detection results when using different threshold values

$$Completeness = \frac{\#\text{True Positives}}{\#\text{True Positives} + \#\text{True Negatives}} \quad (8)$$

Figure 3 presents the detection result in terms of accuracy, precision and completeness when the value of the threshold was gradually increase. The x-axis indicates the threshold value and the y-axis presents the normalised results. The figure shows that the accuracy is much less affected by the increase of the threshold comparing to the precision and the completeness. A clear trade-off can be observed between the precision and the completeness as they expose an opposite tend. The completeness starts around 0.2 and gradually increases up to 0.87, whereas the precision starts at a high value and drops to around 0.35. This finding suggests that the selection of a threshold needs to balance the trade-off between precision and completeness. A smaller threshold value tends to obtain a result with better precision but less completeness. Since the purpose of anomaly detection in this work is to improve calibration by removing anomalies, the precision can be more important than the completeness. Therefore, the threshold value of 15 was chosen. In practice, it would be expected users would try different threshold values and determine which ones best meet their needs.

4.3.3. *Evaluation anomaly detection in synthetic data*

For this evaluation, we compare the results from using cross-sensitive parameter against the one using temporal dependency. The same learning process was applied to both methods. Figure 4 shows the results in term of detection accuracy, precision and completeness. One hundred test runs were performed which generate using the approach defined in Section 4.3.1. Each boxplot indicates the variation of the result from the 100 tests. The results in Figure 4 suggest that using cross-sensitive parameters is able to produce a better detection result as the accuracy, the precision and the completeness are on average 2%, 17% and 325% better than the one using the temporal dependency.

In the first experiment, anomalies were only injected into NO_2. As anomalies can affect all parameters, in the following experiment, anomalies were injected in both NO_2 and O_3 data. The percentage and magnitude of the O_3 anomalies were determined in the same way as described in Section 4.3.1. For the second experiment, the injection of NO_2 remained the same as in the previous experiment. 10% of the samples in the clean O_3 data were randomly added by values that had a magnitude in the range of 10% to 60% of maximal O_3.

Figure 5 shows that the accuracy, the precision and the completeness of using the cross-sensitive parameter are still on average significantly better (i.e. 2%, 16% and 283%) than using the temporal information after 10% of anomalies were added into the O_3 data.

4.3.4. *Findings and limitations*

The results for both experiments are summarised in Table 1. These experiments demonstrate that utilising the dependency between cross-sensitive parameters not only can sufficiently detect anomalies in the data, but the detection results are also better than the one that utilises temporal dependency. The results suggest the dependency between cross-sensitive parameters can be used for anomaly detection. This finding can be particularly valuable as it can be used in conditions when spatial and temporal dependencies are weak. In addition, cross-sensitivity of sensors are widely reported, which is not unique for NO_2 and O_3 sensors. Therefore, this method could also apply to other sensors for identifying their data anomalies. Combining cross-sensitive dependency with spatial and temporal dependencies could be investigated as part of future work.

According to lessons that we learnt from our experiment, the limitations of the method for anomaly detection are discussed below:

- The method is built upon the assumption that measurements from a cross-sensitive parameter is able to indirectly

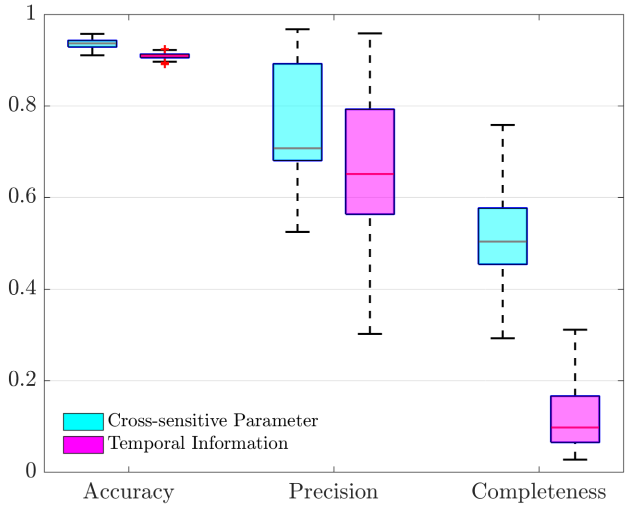

Figure 4: Detection results when anomalies were injected in NO_2

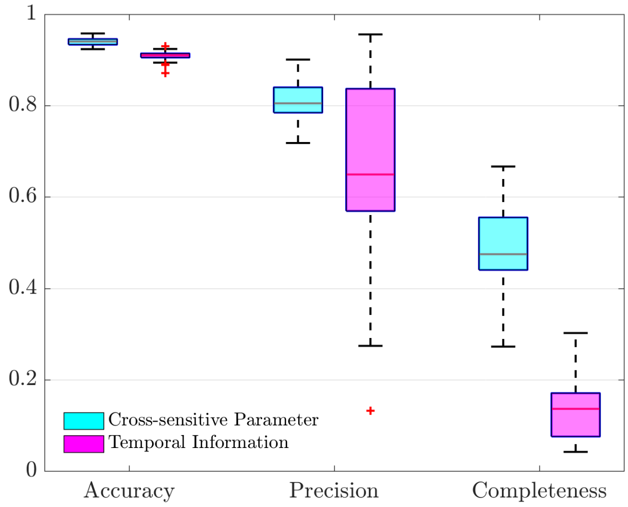

Figure 5: Detection results when anomalies were injected in NO_2 and O_3

indicate the '*normal*' of the target parameter. It would require a strong dependency between target parameter and its cross-sensitive parameter. If the dependency was weak or not strong enough, the anomalies may not be sufficiently identified. Understanding the minimum requirement of such dependency that can be used for sufficiently detecting anomalies is one of important question to answer in the future research.

- Applying an appropriate threshold value is another key step for obtaining a good detection result. This issue is not unique for this method as it would be a general problem for anomaly detection. We demonstrated the relationship between an increased threshold value and the detection result in our experiment, which can be used for selecting *optimal* threshold value. However, such demonstration would only be possible given the fact that the ground truth of anomalies were available. Since it is still an open challenge to obtain ground truth of anomalies in real dataset, it would not be possible to obtain such relationship in real dataset. Even though the smaller threshold value would often indicate a better precision and less completeness, the exact threshold value would require expert judgement to obtain. As a result, the appropriateness of such threshold often be difficult to justify or to evaluate.

- Joint probability table was used to represent the dependency between target parameter and its cross-sensitive parameter. The joint probability is calculated according to the number of samples fall in each bin. The higher/lower the number of sample in each bin indicates the higher/lower probability. The method assumes that the anomalies are the samples in bins that their probability is below a threshold value. A problem we encountered is that some high concentration values have very small number of samples, which are often mis-classified as anomalies. To avoid such situation, we trained our anomalies model for classifying anomalies only for data that span 0 to 99 percentiles. The difference in term of result is shown in Figure 6. Figure 6 again shows the trade-off between completeness and precision as our alternative approach fails to identify higher concentrated anomalies while preserves high concentrated normal measurements.

Even though our method has many practical limitations, our experiment demonstrated an improved result for anomaly detection in comparison to the state-of-the-art method. It suggests that the proposed concept and method is able to sufficiently detect anomalies. In the following sections, we will illustrate how the calibration of low-cost sensors can benefit from applying the proposed anomaly detection.

4.4. *Evaluation of the overall approach on real data*

The overall evaluation of the approach were performed on two real datasets independently. The datasets were introduced in Section 4.1.

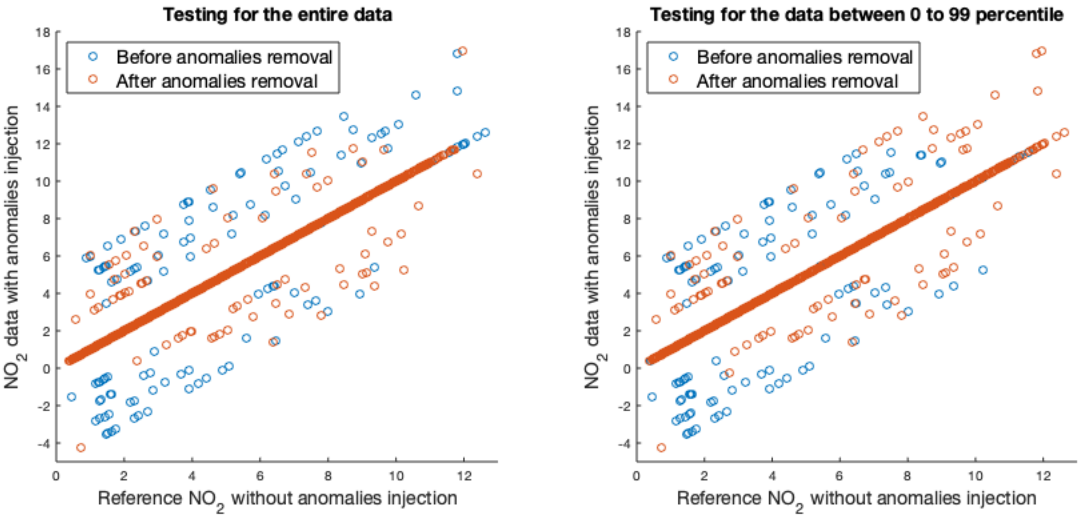

Figure 6: The scatter plot before and after anomalies removals against the ground truth of anomalies.

Table 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

4.4.1. Evaluation on ELM dataset

The ELM dataset was initially pre-processed according to the method in Algorithm 3 in Section 4.2. After pre-processing, the dataset contains around 4,000 samples with the temporal resolution of an hour, and the available parameters are NO_2, O_3, NO, T and H. According to <29>, 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 determining the calibration function, and the calibration function evaluated using the last partition. The results for calibrating NO_2 are represented in Figure 7 and Table 2.

In Table 2, we list the parameters that were used to constructing the calibration function in the first row. We use a wide range of metrics to evaluate the calibration to avoid biased results <29>. The standard deviation and the mean error were calculated from the point-wise, the difference between the reference and the predicted value. A mean error close to zero represents the smaller difference, which implies a better calibration result. The positive and negative sign of the mean error indicate the under and over estimation of the result. The standard deviation is related to the mean value. In general, a smaller standard deviation indicates a better calibration. The RMSE and R are commonly used metrics for the evaluation of sensor calibration, the smaller RMSE and higher R value are often associated with a better calibration. We also calculated a linear function between the reference and the predicted value, which is referred to as *linearity* in the table. The predicted value is expected to be as close to the reference as possible, which implies that the slope and offset of the linearity function need to be close to 1 and 0 respectively.

Figure 7 shows a series of scatter plots between the reference and the data trace from the ELM. Figure 7-(a) shows the correlation between the reference and the raw data from the low-cost sensor. The raw data is ELM data after the pre-processing. The figure shows that the raw data varies from 0 to 200. We emphasis this variation by using the red colour as it is much greater than the reference. A significant number of zero readings can also be observed in the raw data. These are often considered as anomalies <32>. The result confirms that the data from low-cost sensors may not be useful without a proper process.

Figure 7-(b) shows the result of the calibration using the linear uni-variate calibration where only the parameter of NO_2 was used in the calibration. The figure shows that the variation of the ELM data sufficiently reduced as the predicted value varies in the same range as the reference. Table 2 also shows that the RMSE and the standard deviation are all reduced by about 50%. However, the predicted value still contains a large number of constant values (the zero readings in the raw data have been transferred by the calibration), resulting in an irregular data pattern and low correlation (0.77) between the predicted value and the reference. This shows and confirms that the uni-variate calibration is insufficient for the calibration of low-cost sensors, which is in line with <16; 40; 19>.

Figure 7-(c) shows a calibration by the linear multivariate calibration. It is noted that the calibration function in Figure 7-(c) was determined in a different environmental condition than the environment of the operation. As a result, a negative correlation between the predicted value and the reference is observed in the figure. In Table 2, it has the highest errors (i.e. RMSE

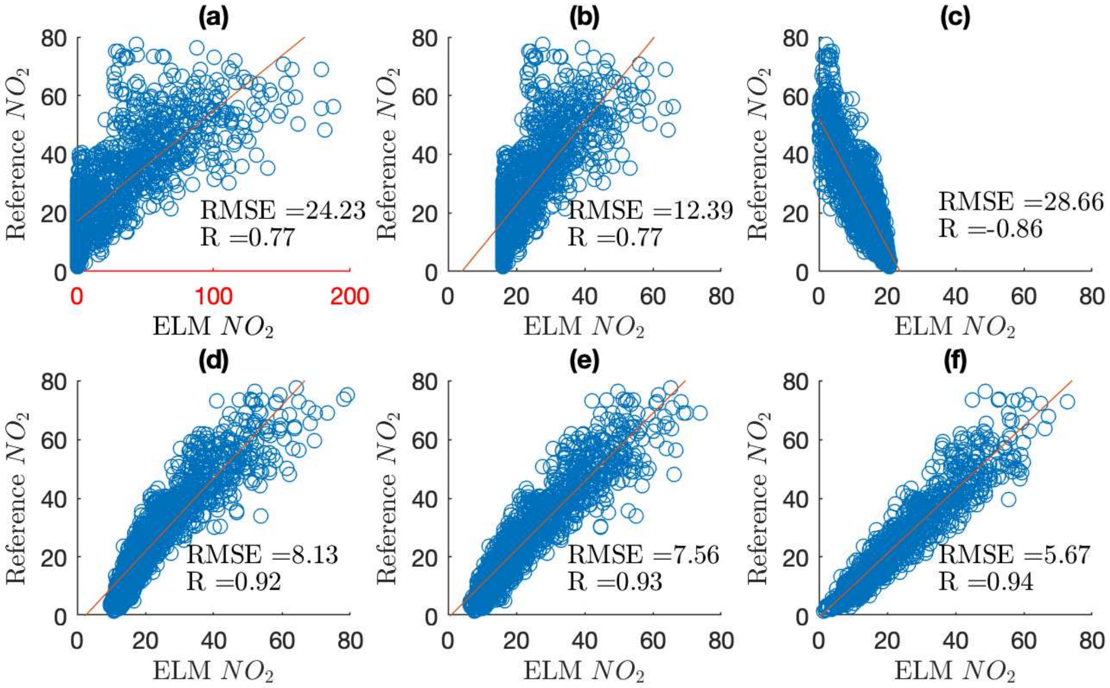

Figure 7: The NO_2 data trace of the low-cost sensor in comparison to the reference data; Figure-(a) shows the correlation between the reference and the raw data from the low-cost sensor; Figure-(b) shows the correlation between the reference and the data of low-cost sensor being calibrated using linear uni-variate calibration; Figure-(c) shows the correlation between the reference and the data of low-cost sensor being calibrated using linear multivariate calibration, where the calibration function was determined in a different environmental condition; Figure-(d) shows the correlation between the reference and the data of low-cost sensor being calibrated using linear multivariate calibration, where the calibration function was determined in-situ; Figure-(e) shows the correlation between the reference and the data of low-cost sensor being calibrated using the proposed calibration; Figure-(f) shows the correlation between the reference and the data of low-cost sensor being treated by the proposed two-phased approach

28.66 and Mean error 17.02) and the worst linearity. The result confirms that the calibration function determined in one place is not necessarily applicable to another place due to the different environmental conditions.

Figure 7-(d) shows the calibration used the same method as Figure 7-(c) with the difference that the calibration function was determined on-site in the environment of operation. In comparison to Figure 7-(c), Figure 7-(d) shows an improved correlation (0.92). Comparing the evaluation metrics in Table 2, the errors are reduced significantly with the respect to the (a), (b) and (c) by at least 35% for mean error and 52% for RMSE. The result shows the importance of using multiple parameters in the calibration of low-cost sensors and the necessity of determining calibration functions in the environment of operation.

Figure 7-(e) shows the result of calibration that used the proposed calibration method without removing anomalies. Comparing the result with the one in Figure 7-(d), the calibration errors and the linearity are further improved. The RMSE improves by 7.5%, and the slope and offset are closer to 1 and 0 respectively. The result indicates that by introducing interaction terms and selecting the use of parameters the calibration result can be further improved.

Figure 7-(f) shows the calibration result that anomalies in the NO_2 data were removed before using the proposed calibration method. In comparison to Figure 7-(e) where the anomalies were not removed, the results of calibration in Figure 7-(f) are further improved. The standard deviation, mean error and RMSE are further reduced by 1.38, 1.73 and 1.89 respectively, which equates to an improvement of 20%, 53% and 25% respectively. The result in Figure 7-(f) has the best calibration among all the methods used, which implies that the calibration result can be further enhanced when the anomalies were removed in advance. The result shows the proposed approach is able to further improve the quality of data than current practices, and suggests the importance of removing random errors before the calibration process.

4.4.2. *Evaluation on Beijing dataset*

The evaluation was also carried out on a different dataset which was used in <38>. To enable a fair cross comparison, we took the best effort to follow their process and the evaluation is presented in the same way as they did. We compare the Gaussian Process (GP) that produced the best calibration result in their paper to our approaches. Our approaches are simplified as *Algorithm 1* and *Algorithm 2* in the following context. *Algorithm 1* calibrates NO_2 directly using the method discussed in Section 3.3 without removing anomalies; whereas *Algorithm 2* applies the process for detection and removal of anomalies before the sensor calibration.

For the first experiment, the first 8490 data instances were separated from the dataset for training and the rest of the data instances (i.e. 24557 samples) were used for testing. After dis-

	(a)	**(b)**	**(c)**	**(d)**	**(e)**	**(f)**
Parameters used	None	NO_2	NO_2,O_3,NO,T,H	NO_2,O_3,NO,T,H	$NO_2,O_3,NO,H,$ $NO_2*O_3,NO_2*H,NO_2*NO,$ O_3*H	$NO_2,O_3,NO,T,H,$ $NO_2*O_3,NO_2*H,NO_2*NO,$ $O_3*T,O_3*H,O_3*NO,$ $T*H$
Standard deviation	24.22	11.64	23.06	7.51	6.84	5.46
Mean error	-1.09	4.23	17.02	3.13	3.24	1.51
RMSE	24.23	12.39	28.66	8.13	7.56	5.67
R	0.77	0.77	-0.86	0.92	0.93	0.94
Linearity	Y = 0.38x + 17.04	Y = 1.42X-5.73	Y = -0.22X+52.24	Y = 1.24X - 2.92	Y= 1.16X - 0.76	Y = 1.09X-0.76

Table 2: This table presents the calibration results in terms of parameters used in the calibration, standard deviation, mean error, RMSE, R and linearity among the different calibrations that demonstrated in Figure 7. The **(a)** to **(f)** are indices that differentiate the calibrations as in Figure 7.

carding the samples with empty values, there were 8437 samples left for training and 23249 samples for testing. We followed their process to implement the GP algorithm that uses the same python library and kernels as they specified in the paper. However, we failed to reproduce their results due to the inconsistent data size between their experiment and the data that we were downloaded. The result is presented in Table 3.

Bin Range	**NRMSE(RMSE/ppb)**			**Number of**	
	GP	***Algorithm 1***	***Algorithm 2***	**Samples**	**Anomalies**
0% - 25%	0.39(7.5)	***0.29(5.4)***	***0.28(5.3)***	22738	197
25% - 50%	0.14(10.3)	***0.13(9.8)***	***0.12(8.9)***	473	105
50% - 75%	0.3(42.4)	***0.28(37.5)***	***0.37(49.5)***	32	24
75% - 100%	0.25(45.97)	***0.13(23.6)***	***N/A***	6	6
Overall RMSE (ppb)	(7.78)	***(5.74)***	***(5.54)***		

Table 3: The bin range indicates the what samples were used to calculate the NRMSE and RMSE. The NRMSEs were shown in the table without brackets and the RMSEs were shown in the brackets as (RMSE/ppb). The Gaussian Process (GP) were implemented according to the description in <38>. The *Algorithm 1* shows the results by using the proposed calibration method only, whereas the anomaly detection and removal was applied before the calibration in *Algorithm 2*. The number of samples in each bin and the number of anomalies identified in each are presented in the last two columns of the table. The overall RMSEs for different methods were presented in the last row of the table and shaded in gray.

The evaluation were taken in the same way as the one in <38> which determines how the data quality is related to different ranges of concentration. The samples in the testing dataset were placed in 4 bins. The 4 bins were determined according the percentage of the maximum concentration of reference NO_2 in the testing dataset, which are bin_1 (0%-25%), bin_2 (25%-50%), bin_3 (50%-75%) and bin_4 (75%-100%). The number of samples in each bin from bin 1 to 4 were 22738, 473, 32 and 6 respectively. The RMSEs between reference NO_2 and calibrated NO_2 in each bin were firstly calculated. The NRMSEs were then normalised accordingly by dividing the RMSEs by the mean concentration for the respective bin. Our results (in italics) are cross-compared to GP in Table 3.

In Table 3, the overall RMSE for *Algorithm 2* is placed the highest. In comparison to the GP, our approaches are more than 30% better for the overall RMSE value. Comparing results between *Algorithm 1* and *Algorithm 2*, the overall RMSE further increased by more than 5% when the anomalies removed before the calibration. Considering the number of sample in bin 3 and bin 4 are extremely small (i.e. 32 and 6 samples respectively), the confidence level of the results in those bins can be significantly lower than results in bin 1 and bin 2. For this reason, the comparison is focused on bin 1 and bin 2 in which the *Algorithm 2* outperforms the *Algorithm 1* by average 4.5% and GP by average 20% respectively.

The samples in bin 4 and majority of samples in bin 3 were identified as anomalies. However, the largest number of anomalies were identified in bin 1 and bin 2. This suggests that the method for anomaly detection is not simply removing extreme value.

This evaluation shows that the proposed method calibration is able to obtain a good calibration result, and these results can be further enhanced when anomalies were identified and removed in advance. Since we could not reproduce their result due to the inconsistent dataset, we extend the comparison further to determine how the calibration result would be different when the training and testing dataset were different.

For this experiment, the Beijing dataset were sequentially divided into 4 partitions with each partition has a similar number of samples (Partition 1 to 4 have samples of 7921 7921 7921 7923 respectively). In the following context, we use partition number to refer what datasets were used for training and testing. Partition 1 + 2 indicates that the dataset is combined from partition 1 and 2. The results of calibration that use different datasets for training and testing are presented in Figure 8 and 9:

Figure 8 shows the calibration result in terms of RMSE values. The result suggests that the RMSE is sensitive the training and testing datasets. This indirectly explains why our reproduced result was not the same as showed in <38> as different training and testing datasets were used. From the figure, the calibration results from our approach consistently better than the GP. The results determined after the removal of anomalies are no worse than the result without anomalies removal, and produce better calibration results in 5 out of 6 experiments.

Figure 9 shows the calibration result in terms of errors. The first plot is the calibration errors. We can observe a large number of outliers in the first plot. As those outliers hinder the variation of errors, we made the second plot that exclude the outliers. The second plot shows a better view on the variation and median value of the errors. From the figure, we can see the error profiles are consistent with the result of RMSE in Figure 8. The number of extreme errors (i.e. outliers point in the fist plot) from the method with anomalies removal (i.e. Algorithm 2) is less than other methods without anomalies removal (i.e. Algorithm 1 and GP). This suggests that the removal of anomalies is able to improve sensor calibration. The variation of errors from the proposed calibrations (i.e. Algorithm 1 and Algorithm

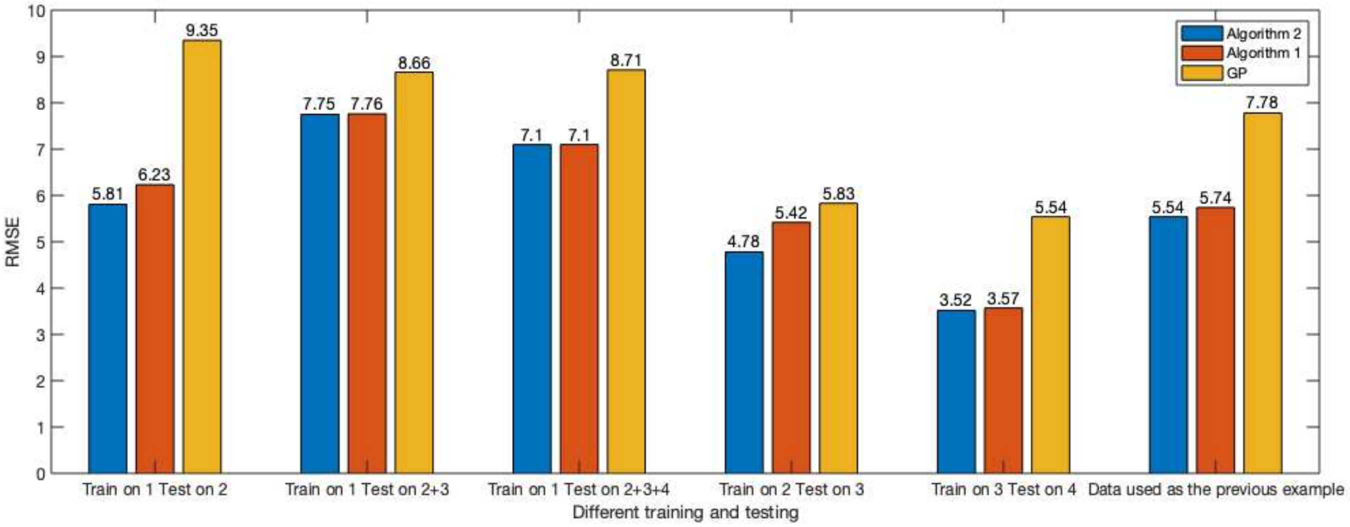

Figure 8: The figure shows the RMSE value for calibrations done by different methods and using different training and testing datasets

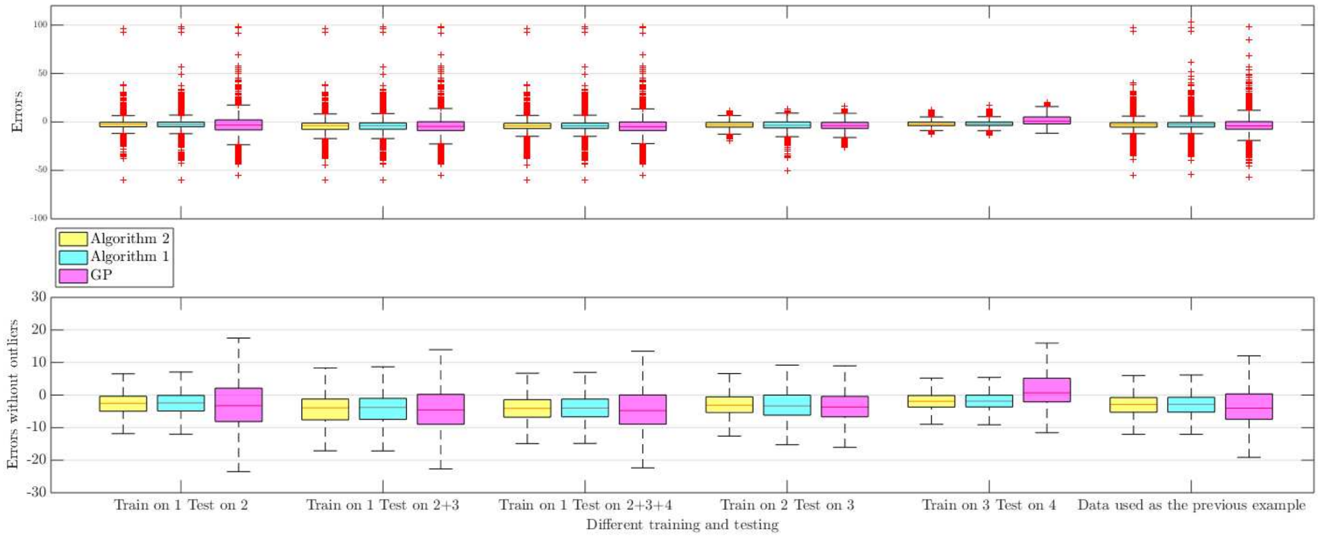

Figure 9: The figure shows errors and errors without outliers for calibrations done by different methods and using different training and testing datasets

2) is generally better than the GP, and the Algorithm 1 and Algorithm 2 has similar variation of errors. This is a reasonable result as applying anomaly detection before the calibration is not expected to change the error profile dramatically.

It is noted that the median errors of GP in some cases is better (i.e. closer to zero) than the ones from the Algorithm 1 and 2 even though its variation is consistently larger than others. It suggests that GP is more unlikely to under or over predict value but it would be sensitive to extreme values. This implies that the GP could produce better results if the environment has a consistent level and the sensors do not have anomalies. However this situation would be extremely unusual. Otherwise the proposed method provides improved results.

These evaluations indicate that all calibrations are sensitive to the use of training and testing datasets. The results show that the proposed calibrations (i.e. Algorithm 1 and Algorithm 2) are able to consistently obtain a better calibration result, and the results can be further enhanced when anomalies were identified and removed before the calibration. This is inline with the result that obtained using the ELM dataset, which suggests that data quality of low-cost sensor can benefit from the proposed calibration method with anomaly detection and removal.

5. Conclusion and future work

This paper presents an improved sensor calibration, which applies a process for detection and removal of anomalies before the sensor calibration. The detection of anomalies utilises the dependency between cross-sensitive parameters, for which the results in terms of detection accuracy, precision and completeness are all better than the state-of-the-art approaches. The method of calibration takes the advantage of stepwise regression and the interaction terms, which provides a good calibration result comparing to the state-of-the-art methods, including machine learning algorithms. The evaluations carried out on two real datasets show a consistent better calibration result when anomalies were removed from the datasets. It suggests that the accurate detection and removal of anomalies will further improve the calibration of low-cost sensors.

For the future work, it is worth to investigate whether combining multiple contextual information would further improve the performance of anomaly detection. In addition, considering many anomalies would be associated with a systematic cause, e.g. communication errors, understanding the root causes of anomalies can be important in the future work.

Acknowledgement

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

References

[1] 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).

[2] Department for Environment Food & Rural Affairs, Monitoring networks (Mar. 2017).
URL `https://uk-air.defra.gov.uk/networks/`

[3] Department for Environment Food & Rural Affairs, Local air quality management (laqm) (2015).
URL `https://laqm.defra.gov.uk/diffusion-tubes/diffusion-tubes.html`

[4] G. Makrai, I. Bate, Analysis of a statistical regression approach for no_2 pollution modelling, in: Proceedings of the International Conference on Distributed Computing in Sensor System, 2017 (2017).

[5] I. Heimann, V. Bright, M. McLeod, M. Mead, O. Popoola, G. Stewart, 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 (2015) 10–19 (2015).

[6] H. Messer, A. Zinevich, P. Alpert, Environmental monitoring by wireless communication networks, Science 312 (5774) (2006) 713–713 (2006).

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

[8] R. Pope, J. Wu, Characterizing air pollution patterns on multiple time scales in urban areas: A landscape ecological approach, Urban Ecosystems 17 (3) (2014) 855–874 (2014).

[9] R. Hijmans, S. Cameron, J. Parra, P. Jones, A. Jarvis, Very high resolution interpolated climate surfaces for global land areas, International Journal of Climatology 25 (15) (2005) 1965–1978 (2005).

[10] I. Elmi, S. Zampolli, E. Cozzani, F. Mancarella, G. Cardinali, Development of ultra-low-power consumption mox sensors with ppb-level voc detection capabilities for emerging applications, Sensors and Actuators B: Chemical 135 (1) (2008) 342–351 (2008).

[11] J. Burgués, S. Marco, Low power operation of temperature-modulated metal oxide semiconductor gas sensors, Sensors 18 (2) (2018) 339 (2018).

[12] X. Fang, 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 (2017).

[13] Y. Cheng, X. Li, Z. Li, S. Jiang, Y. Li, J. Jia, X. Jiang, Aircloud: A cloud-based air-quality monitoring system for everyone, in: Proceedings of the 12th ACM Conference on Embedded Network Sensor Systems, ACM, 2014, pp. 251–265 (2014).

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

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

[16] E. Esposito, S. Devito, M. Salvato, V. Bright, R. Jones, 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 (2016) 701–713 (2016).

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

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

[19] L. Spinelle, M. Gerboles, M. Villani, M. Aleixandre, 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 (2015) 249–257 (2015).

[20] L. Spinelle, M. Gerboles, M. Villani, M. Aleixandre, F. Bonavitacola, Calibration of a cluster of low-cost sensors for the measurement of air pollution in ambient air, in: SENSORS, IEEE, 2014, pp. 21–24 (2014).

[21] S. Devito, M. Piga, L. Martinotto, 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) (2009) 182–191 (2009).

[22] E. Esposito, S. Devito, M. Salvato, G. Fattoruso, G. Difrancia, Computational intelligence for smart air quality monitors calibration, in: International Conference on Computational Science and Its Applications, Springer, 2017, pp. 443–454 (2017).

[23] S. Devito, E. Esposito, M. Salvato, O. Popoola, F. Formisano, R. Jones, G. Difrancia, Calibrating chemical multisensory devices for real world applications: An in-depth comparison of quantitative machine learning approaches, Sensors and Actuators B: Chemical 255 (2018) 1191–1210 (2018).

[24] L. Weissert, K. Alberti, G. Miskell, W. Pattinson, J. Salmond, G. Henshaw, D. Williams, Low-cost sensors and microscale land use regression: Data fusion to resolve air quality variations with high spatial and temporal resolution, Atmospheric Environment (2019).

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

[26] B. Maag, Z. Zhou, L. Thiele, A survey on sensor calibration in air pollution monitoring deployments, IEEE Internet of Things Journal (2018).

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

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

[29] X. Fang, Improving data quality for low-cost environmental sensors, Ph.D. thesis, University of York (2018).

[30] X. Fang, 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 (2017).

[31] P. Clifford, D. Tuma, Characteristics of semiconductor gas sensors i. steady state gas response, Sensors and Actuators 3 (1982) 233–254 (1982).

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

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

[34] O. Saukh, D. Hasenfratz, 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, ACM, 2015, pp. 274–285 (2015).

[35] K. Danzer, L. Currie, Guidelines for calibration in analytical chemistry. part i. fundamentals and single component calibration (iupac recommendations 1998), Pure and Applied Chemistry 70 (4) (1998) 993–1014 (1998).

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

[37] Perkin Elmer, Elm sensor (2015).
URL `https://elm.perkinelmer.com/map/`

[38] K. Smith, P. Edwards, P. Ivatt, J. Lee, F. Squires, C. Dai, R. Peltier, M. Evans, Y. Sun, A. Lewis, An improved low-power measurement of ambient no_2 and o_3 combining electrochemical sensor clusters and machine learning, Atmospheric Measurement Techniques (2019) 1325–1336 (2019).

[39] K. Smith, P. Edwards, Low cost sensor in field calibrations (training and test data) - beijing 2017 (2017).
URL `https://pure.york.ac.uk/portal/en/datasets/low-cost-sensor-in-field-calibrations-training-and-test-data--beijing-2017(1a0c64b0-433b-4eec-b5c7-64d3de0a0351).html`

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