Statistical methods for performance characterisation of lower-cost sensors

J D Hayward, A Forbes, N A Martin, S Bell, G Coppin, G Garcia and D Fryer

1.1 Introduction

This appendix summarises the progress of employing the Breathe London monitoring data together with regulated reference measurements to investigate the application of machine learning tools to extract information with a view to quantifying measurement uncertainty. The Breathe London project generates substantial data from the instrumentation deployed, and the analysis requires a comparison against the established reference methods. The first task was to develop programs to extract the various datasets being employed from different providers and to store them in a standardised format for further processing. This data curation allowed us to audit the data, establishing errors in the AQMesh data and the AirView cars as well as monitor any variances in the data sets caused by conditions such as fog/extreme temperatures.

This includes writing a program to parse the raw AirView data files from the high quality reference grade instruments installed in the mobile platforms (Google AirView cars), programs to communicate with and download data from the Applications Program Interfaces (APIs) provided by both Air Monitors (AQMesh sensor systems) and Imperial College London (formerly King's College London) (London Air Quality Network), and a program that automatically downloads CSV files from the DEFRA website, which contains the air quality data from the Automatic Urban Rural Network (AURN). This part of the development has had to overcome significant challenges including the accommodation of different pollutant concentration units from the different database sources, different file formats, and the incorporation of data flags to describe the data ratification status.

The successful completion of this task now means that data processing and downloading of pollutant measurements from various sources can now be performed with improvements in speed that are orders of magnitude faster than would have been possible using the raw data files. It also allows anyone with an NPL network connection to access them using a standard networked laptop computer, allowing several people to easily access a standardised dataset.

The data is stored in an INFLUX database, which allows for a range of simple analytical functions to be applied that, when queried, can return averages, standard deviations, percentiles across time ranges (seconds to days) and across different measurement sites in the Breathe London Network (fixed and mobile) and the reference networks. This database can also be set to only return categories of measurement data associated with a certain type of flag (for example, the Valid flag that the algorithm from Air Monitors has determined to be reliable data). Utilising this tool means that the data used across different analyses is in a standard format, ensuring reliability across the suite of analysis software and other programs developed in house. This work was necessary not only to be able to quickly access the Breathe London data in a standardised way, but also the LAQN data it is being compared to.

In order to have any confidence in the measurements made by a sensor, we need to know the uncertainty on what they report. Traditional reference instruments require certification, stating their uncertainties. However, low-cost sensor systems are currently not beholden to any kind of regulation, meaning it can be difficult to ascertain the data quality of these instruments in a real

world scenario. Any uncertainty values are also subject to change due to the baseline drifting over time, as well as the degradation of components. To have any confidence in the measurements you not only have to regularly recalibrate these instruments, you also need to regularly assess their associated uncertainties. This is traditionally carried out by moving the sensor systems to either a laboratory or reference monitoring site and comparing the sensor systems to the high quality instruments in a collocation study and/or using reference gases in a laboratory to determine both the accuracy and precision of the measurements. For a high capacity network like Breathe London with over 100 sensor systems, this is very impractical and would incur huge labour costs and large delays as a result of moving the equipment back and forth. To increase the feasibility of low-cost sensor networks, an alternative is needed.

Pollution levels at monitoring sites across an area like London can be reasonably approximated to a sum of two components, local emissions and a regional background. By assessing periods where the pollution levels being recorded are only a product of the regional background, we can ascertain times where sensor systems are seeing the same pollution levels and treat these periods as a pseudo-collocation study, where it is as if the sensor systems are next to each other even if they are several kilometres apart.

However, directly comparing measurements between sensor systems is difficult as they may have drifted apart. Without regular re-calibration it can be difficult to ascertain when sensor systems are seeing the same pollution levels. Unless they have all been calibrated at the same time, low-cost sensor systems tend to have variations in their scaling caused by drifts in their baseline as well as unique defects that cause variation between the same types of components. To be able to determine when two measurements by two different sensor systems are measuring the same thing their scaling has to be adjusted so they report the same measurements. This is done by standard scaling sections of the data (subtracting the mean and dividing by the standard deviation). Two standard scaled measurements from different sensor systems are roughly comparable as long as most of the measurements aren't significantly influenced by local emissions.

1.2 Technical Details

1.2.1 Download and Parse Data

The measurements from individual sensor systems needed to be standardised in order to ensure that they aligned by time. That data was stored in an InfluxDB database, to access it it had to be downloaded from an internal server at the National Physical Laboratory. The data was queried with aid from the InfluxDB library for Python which simplified downloading and parsing the data substantially. The following query was sent to the database:

SELECT {Operator}({Pollutant}) AS "measurement" FROM "{"NPL Server"}"."autogen"."{Breathe London Database}" WHERE time >= {time_start} and time < {time_end} AND {Flag} = {"Valid"} AND {Sensor}=' {Sensor Name} GROUP BY time({Measurement Period}),' {Sensor}'

This query then returned valid measurements made by a sensor for a specific timeframe, the

1.2.2 Standard Scale Data

As 'raw' low-cost sensor measurements provided by the manufacturer often come with an unknown scaling factor, it's very difficult to know exactly how the sensor systems are scaled relative to each other. To be able to determine periods of agreement between a majority of sensor systems, it's important to ensure the measurements of individual sensor systems are scaled to each other, allowing us to isolate trends. This will allow us to separate background measurements from local variances. To standardise the data, it is standard scaled (Also known as z-score normalisation or standardisation). The standard scaled data obtained from **Algorithm 2** will then be fed in to **Algorithm 3** to classify the measurements.

Algorithm 1: Download and Parse Data

Input: Sensor System Type (e.g AQMesh) to Download Data For (x), Measurement Start Date (t_s), Measurement End Date (t_e), Maximum Allowed Value (max), Minimum Allowed Value (min), Slope (m), Offset (c)

s = Sensor systems in x measuring between t_s and t_e
f = Flags assigned to measurements by Air Monitors
for *s sensor systems* **do**
    a = Measurements made by sensor s in x
    for *t measurements in a* **do**
        if *$f_{s,t}$ == 'Valid' AND max > d_t > min* **then**
            $b_{s,t}$ = m * a_t + c
        else
            $b_{s,t}$ = None
dt_r = Measurement timestamps from Influx Database
for *t measurement periods in [15 min (q), 1 hour (h), 1 day (d), 7 days (w)]* **do**
    I_t = Empty list
    Append Index 0 to I_t
    Previous Time = dt_r[0]
    for *Current Time in dt_r* **do**
        if *Current Time - Previous Time >= t Time Range* **then**
            Append Current Time Index to I_t
            Previous Time = Current Time

Output: JSON containing data outputs of algorithm

Code	Key	Description
b	Measurements	Measurements received from InfluxDB database, split by key representing sensor ID (s)
dt_r	Dates	Timestamps of measurements in %Y-%m-%d %H:%M:%S format
I_q	Quarter Hourly Indices	Indices that separate 15 minute measurement periods
I_h	Hourly Indices	Indices that separate 1 hour measurement periods
I_d	Daily Indices	Indices that separate 1 day measurement periods
I_w	Weekly Indices	Indices that separate 1 week measurement periods

Table 1.1: Keys present in JSON Output of **Algorithm 1**

Algorithm 2: Standard Scale Data

Input: Sensor System Measurements (b) from s equivalent spatially disperse sensor systems measuring x pollutants

for *x pollutants* **do**
    for *s sensor systems measuring x* **do**
        Split $b_{s,x}$ in to y sets of measurement period t using indices in I_t
        for *y sets in $a_{s,x}$* **do**
            $c_{s,x,y}$ = ($b_{s,x,y}$ - $\bar{b}_{s,x,y}$) / $\sigma(b_{s,x,y})$ `/* Standard scale data */`

Output: JSON containing data outputs of algorithm

1.2.3 Measurement Classification

The first stage of the measurement classification section takes the standard scaled split measurements and assigns categories based on their values relative to the network. The first check utilises

Code	Key	Description
c	Split Measurements (Standard Scaled)	a_s with standard scaling applied, split in to y measurement periods

Table 1.2: Keys present in JSON Output of **Algorithm 2**

Algorithm 3: Measurement Classification (First Round)

Input: Split Measurements c

for t *sets in* c_y **do**
  for m *minute measurements in* $c_{y,t}$ **do**
    CV = $\bar{c}_{y,t,m}$ / $\sigma(c_{y,t,m})$ `/* Coefficient of Variance */`
    if *CV < 0.5* **then**
      upper_limit = $\bar{c}_{y,t,m}$ + $(*\sigma(c_{y,t,m}))$
      lower_limit = $\bar{c}_{y,t,m}$ - $(*\sigma(c_{y,t,m}))$
      for x *sensor systems measuring* y **do**
        if *upper_limit* $> c_{x,y,t,m} >$ *lower_limit* **then**
          measurement_type_a$_{x,y,t,m}$ classified as "Background"
        else
          measurement_type_a$_{x,y,t,m}$ classified as "Local variance"
    else
      measurement_type_a$_{y,t,m}$ classified as "Poor network agreement"

Output: Classification of measurements at m minutes for x sensor systems measuring y pollutants as background, local variance or poor network agreement

Code	Key	Description
measurement_type_a	Classification of measurements	Measurement labels for measurement in a_s as "Background", "Local variance" or "Poor network agreement"

Table 1.3: Keys present in JSON Output of **Algorithm 3**

the coefficient of variance (CV).If the coefficient of variance is above the user-set limit, it's assumed there are too many variances in the sensor readings to be able to confidently determine the regional background. The measurement is then classified as being in a period of **"Poor Network Agreement"**

If the coefficient of variance (CV) is lower than the user-set limit, the algorithm then decides whether the measurement is of the **"Background"** or is affected by **"Local Variance"**. Local variances can be caused by local emissions of pollution, noise, environmental factors or malfunctions. They should be unique to that sensor and the effects should not be seen across the network. **"Background"** measurements are determined if the standard scaled value exists within a small window around the mean, otherwise the value is assumed to be a **"Local Variance"**.

The second stage of the measurement classification section linearly extrapolates between two background measurements in order to increase the amount of information available. It does this by selecting gaps between two background measurements that don't exceed 60 minutes and fill them with values moving linearly from one measurement to the other. This makes two assumptions, that the regional background doesn't change significantly within an hour and that any change would be largely linear with no sudden step changes. These extrapolated background measurements are then

Algorithm 4: Measurement Classification (Background Extrapolation)

d_y = List of None values of equivalent length to b_y
for t *sets in* b_y **do**
  for m *minute measurements in* $b_{y,t}$ **do**
    if *measurement_type_a*$_{y,t,m}$ == *Background* **then**
      if *Background in measurement_type_a*[m+1:m+60] **then**
        n = Index of next background measurement after m
        d_y[m:n+1] = Linearly extrapolation between background measurements

Output: Extrapolated Background Measurements

Code	Key	Description
d	Extrapolated backgrounds	Linearly extrapolated measurements between two background measurements within an hour of each other

Table 1.4: Keys present in JSON Output of **Algorithm 4**

used in **Algorithm 5** to determine the local variances caused by noise.

Algorithm 5: Measurement Classification (Background Extrapolation)

e_y = List of differences between extrapolated background and measurements
for t *sets in* b_y **do**
  for m *minute measurements in* $b_{y,t}$ **do**
    if $d_{y,t,m}$ *!= None* **then**
      $e_{y,t,m}$ = $b_{y,t,m}$ - $d_{y,t,m}$
if *abs(min*(e_y)*) < abs(max*(e_y)*)* **then**
  max_noise = abs(min(e_y))
  measurement_type_b = measurement_type_a
  for t *sets in* b_y **do**
    for m *minute measurements in* $b_{y,t}$ **do**
      if *measurement_type_b*$_{y,t,m}$ *= "Local Variance"* **then**
        if *abs*($d_{y,t,m}$) *<= max_noise* **then**
          measurement_type_b$_{y,t,m}$ = "Noise"
else
  `Sensor is assumed to be malfunctioning`

Output: Secondary Classifications for Sensor Measurements

Code	Key	Description
measurement_type_b	Classification of measurements	Measurement labels for measurement in b_s as "Background", "Local variance", "Noise" or "Poor network agreement"

Table 1.5: Keys present in JSON Output of **Algorithm 5**

The third and final stage of the measurement classification algorithm separates **"Noise"** measurements from other **"Local Variance"** measurements. The differences between measurements made by the sensor systems and extrapolated background measurements are calculated. If the minimum (most negative difference) is greater in magnitude than the maximum (most positive) difference, the sensor is assumed to be malfunctioning. This is most prominently seen in the $PM_{2.5}$ data

where the laser regularly misfires.

If the minimum difference is lower, it is assumed to be the highest variance caused by noise. Any **"Local Variance"** measurements that lie between the extrapolated background and the calculated noise variance is reclassified as "Noise", increasing clarity between two variances that were intially classified under the same banner.

1.3 Uncertainty Calculation

The same background classification technique cannot be performed on available reference data as there is no currently available reference data recorded at minute frequency. Therefore, we need a well-calibrated set of measurements to compare the other measurements to. Ideally, several sensor systems would be used, however the only long-term collocated ones with usable data were 1505150 and 1506150 which were collocated at the Teddington Bushy Park AURN/LAQN site, which could only be calbrated for $PM_{2.5}$ and PM_{10}. To determine the uncertainty, the following actions are performed:

Calibrate Collocated System Measurements

The collocated sensor systems are calibrated against the reference standard. This is most commonly done with least squares regression, though Bayesian Linear Regression is implemented in this algorithm as it gives more information about the uncertainty between the reference measurements and the low-cost sensor.

$$u_{\text{Collocated}}$$

Uncertainty between reference measurements and low-cost sensor measurements

The uncertainty of the reference instrument is also needed.

$$u_{\text{Reference}}$$

Uncertainty between reference measurements and low-cost sensor measurements

Compare Other Systems to Calibrated Collocated System

The collocated system with the lowest uncertainty against the reference equipment (l) is then chosen to be the "well-calibrated" system, with all other systems then calibrated against it in the established pseudo-collocation study. The errors between the measurements are then determined, first being averaged to 15 minute intervals then averaged again for the month.

$$u_{\text{Pseudo-collocation}}$$

Uncertainty between calibrated sensor measurements and other low-cost sensor measurements via psuedo-collocation study

Calculate Uncertainty on Individual Systems

By summating these uncertainties, as well as the uncertainty of the reference instrument (l), the uncertainty of the individual sensor measurements (s) can be determined.

$$U_s = u_{\text{Reference},l} + u_{\text{Collocated},l} + u_{\text{Pseudo-collocation},s}$$

Uncertainty budget

1.4 Results and Discussion

1.4.1 Background Classification

The background classification algorithm performed well visually. This can be seen in an area of high $PM_{2.5}$ pollution such as Greenwich Church Street (*Figure 1.1*). The algorithm has captured the distinction between local emissions and background levels well, despite the extremes between the two. The same can also be said for environments with less $PM_{2.5}$ pollution such as Triangle Adventure Playground (*Figure 1.2*). The few local emissions present are captured well, showing either extreme can be classified well for $PM_{2.5}$.

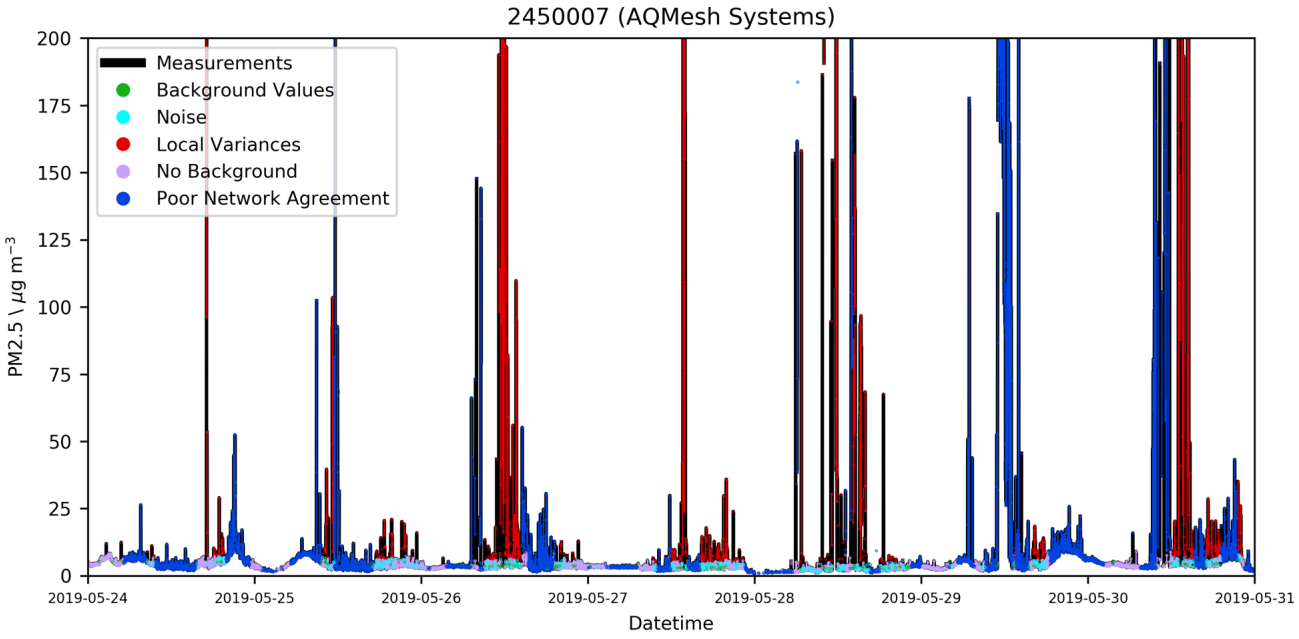

Figure 1.1: $PM_{2.5}$ Background Classifications for Greenwich Church Street

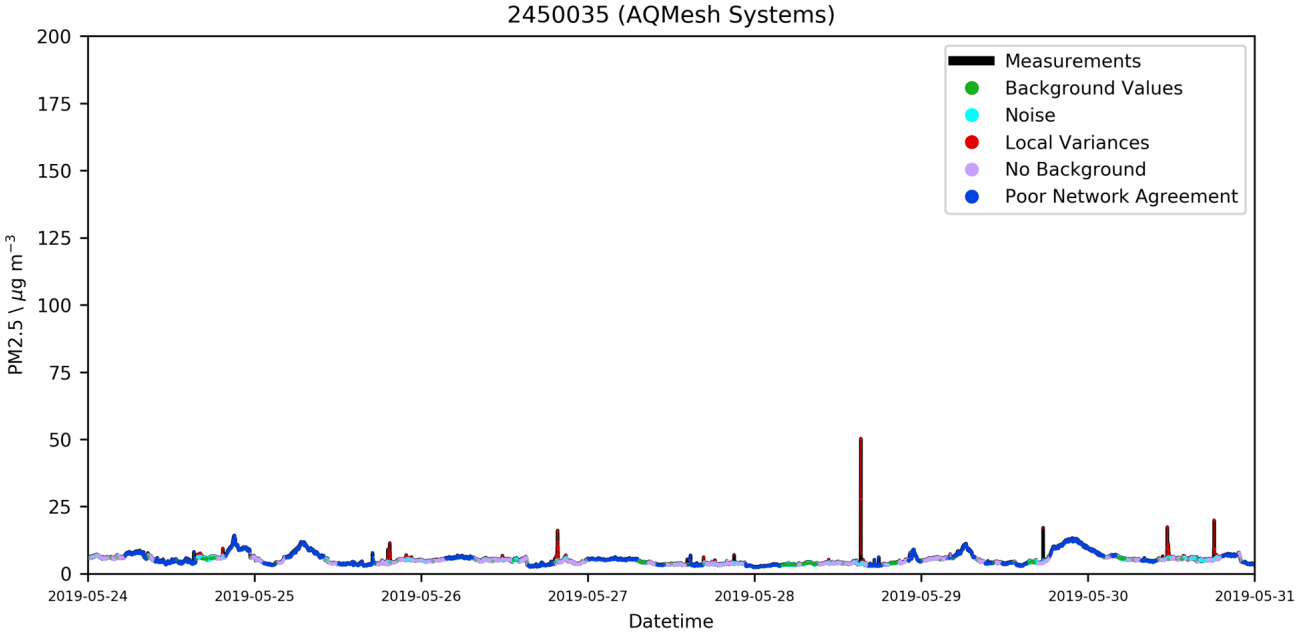

Figure 1.2: $PM_{2.5}$ Background Classifications for Triangle Adventure Playground

The performance of the background classification is more difficult to quantify for NO_2 as the

electrochemical sensors are far more susceptible to noise. This can be seen in the graph for Trianle Adventure Playground (*Figure 1.3*).

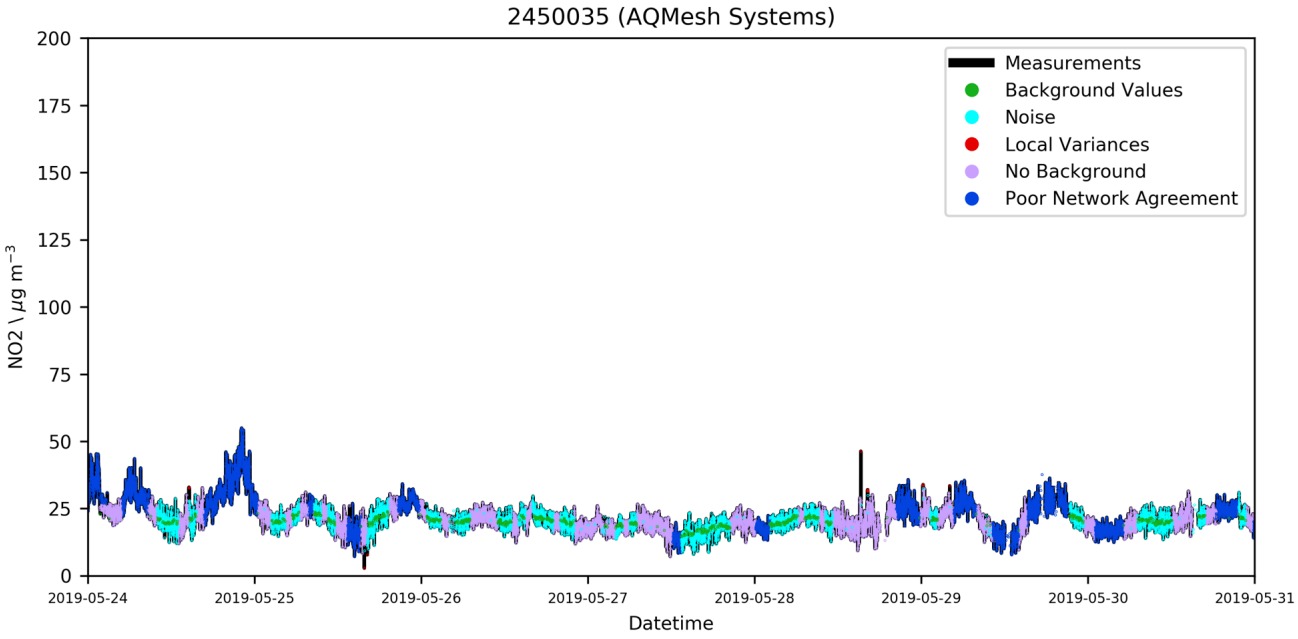

Figure 1.3: NO_2 Background Classifications for Triangle Adventure Playground

1.4.2 Malfunction Flagging

The standard scaled data used to classify backgrounds can also be used to detect poor measurement quality in comparison to the network, caused by a wide range of potential issues. This can be used to detect several different events, as well as probe the long term drift of sensor systems in relation to the network.

The adjusted background graphs are plotted using the first weeks standard scaling parameters (μ and σ) to scale the remainder of the data. All sensor measurements are then centred around zero each minute, with the mean of the network background measurements at that minute subtracted from each individual measurement.

Sensor Change

Changes in low-cost sensing elements cause a slight change in the scaling of the sensor as they each have unique responses to their target gases. This can be difficult to see in real data but is much more easy to see in the standard scaled data (*Figure 1.4*). There is a clear step change in the data as well as an increase in magnitude of the baseline drift, an event corresponding to a change in sensor element.

Sensor Change

Short term malfunctions occasionally occurred in sensor measurements, causing large overreads in the data. This is easy to spot in the measurements but even easier in the standard scaled data (*Figure 1.5*).

Baseline Drift

Long term baseline drifts can also easily be identified with the standard scaled data, including an oddity seen in the following $PM_{2.5}$ data (*Figure 1.6*). The baseline appears to shift in a non linear fashion upwards before decreasing, a trend not seen in other $PM_{2.5}$ sensors which tend to slowly

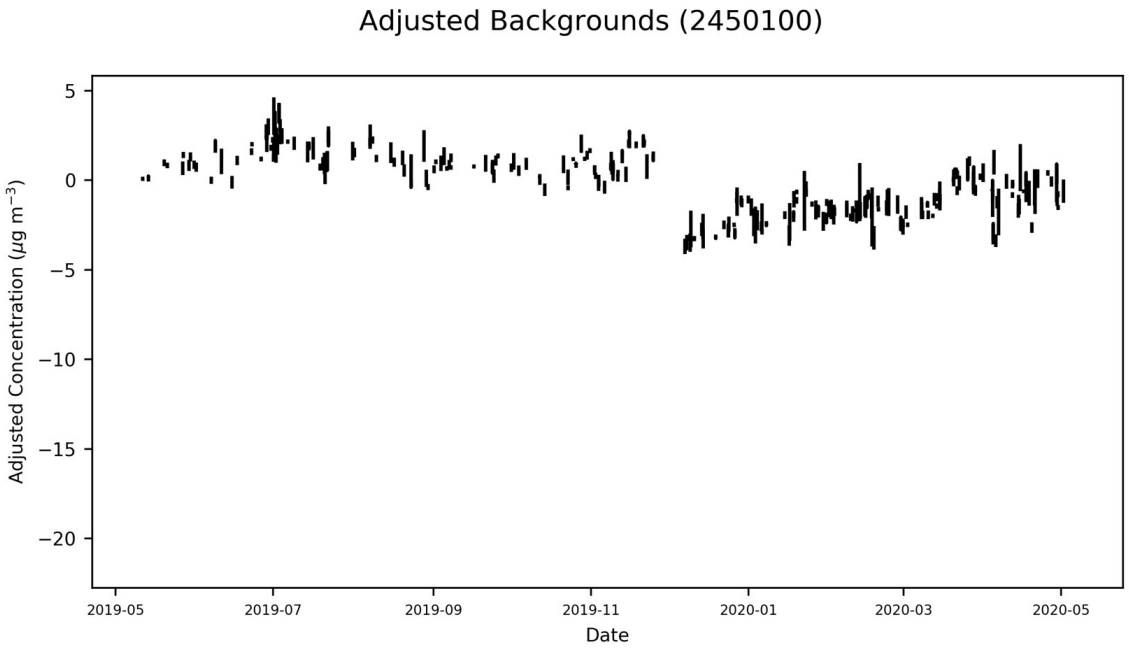

Figure 1.4: NO_2 Sensor Change Event

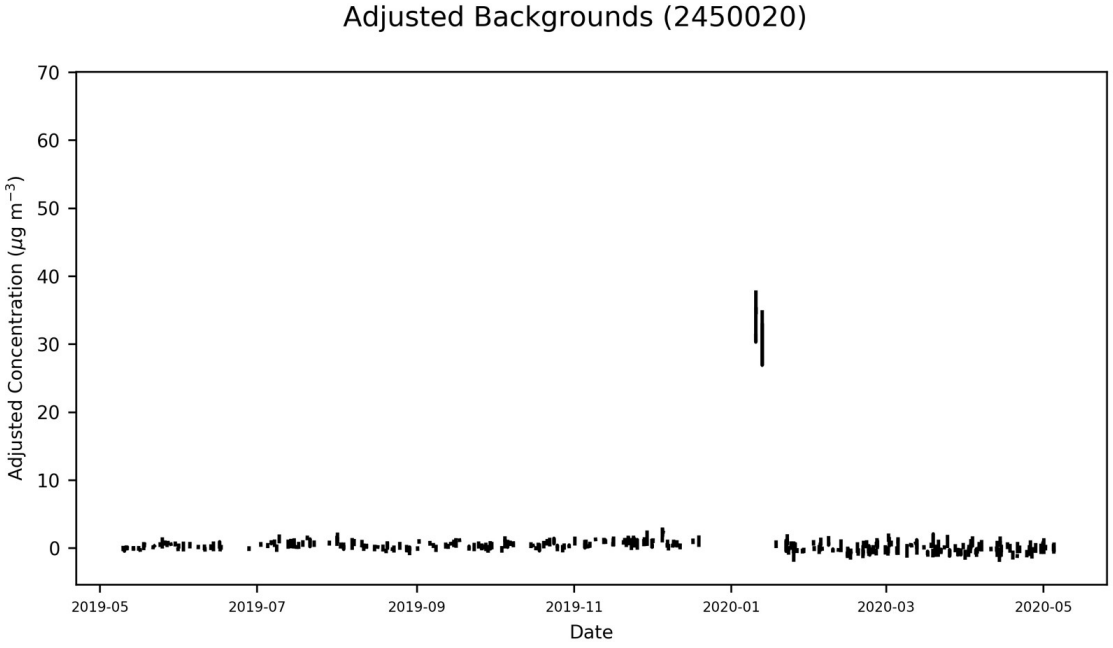

Figure 1.5: NO_2 Malfunction Event

drift in one direction. The cause of this is unknown, but it is much easier to identify when using the standard scaled data.

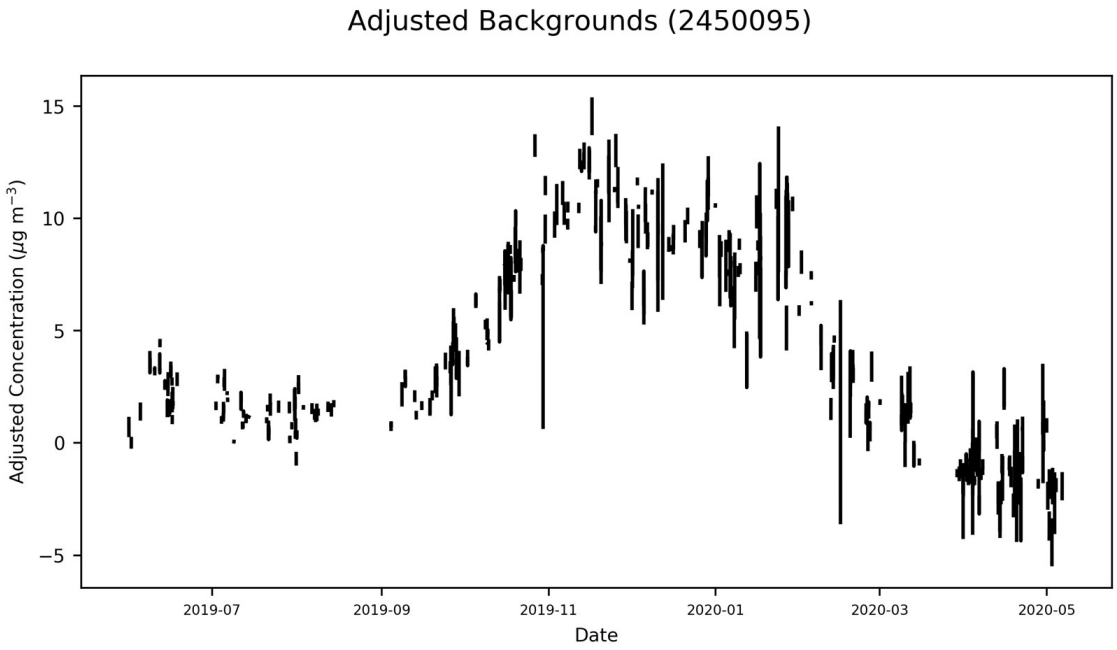

Figure 1.6: $PM_{2.5}$ Baseline Shift

1.5 Conclusion

The background classification algorithm performs well on data from the Breathe London Network as there are many periods of agreement between sensor systems, despite the fact they're far apart. These periods of agreement can then be used in the development of a pseudo-collocation study to determine both the uncertainties of individual systems as well as easily identify events that cause biases and drifts in the measurements. Any uncertainties determined through this method will be higher than if they were calculated via traditional methods as you not only have the contribution from the pseudo-collocation study, but also from calibrating the "well-calibrated" system. The addition of an extra uncertainty component ($u_{\text{Collocated},l}$) increases the overall uncertainty. However, this technique exponentially reduces the labour costs involved with network size as well as network downtime and increases the frequency of calculations that can be performed, trading higher calculated uncertainty for increased flexibility.

The highlighting of malfunctions and other events via standard scaling works especially well with sensor changes, short term malfunctions and long term drifts all highlighted far more clearly than in the measurement data. This makes the automation of detecting malfunctions and calibration changes much simpler as these events stand out far more. This benefits of this increase drastically with network size, particularly if error detection was automated via the adjusted backgrounds.

By classifying individual measurements, the possible pseudo-collocation studies open up a wide range of possibilities, from calculating the uncertainties of individual sensor measurements to determining biases and long term drifts in sensor data. Though these techniques do not have the advantage of accuracy that traditional methods do they significantly reduce the labour involved, allowing for huge increases in network size with minimal overhead on quality control. This is becoming ever more important as awareness around the problems of poor air quality increases.