Breathe London Mobile Monitoring Project: Quality Control/Quality Assurance Protocol

Overview

This document outlines the routine Quality Control/Quality Assurance (QA/QC) protocol for the Breathe London mobile monitoring data collection. The data were processed in stages, with a given stage adding one or more processing steps to the output of the previous stage. All stages are preserved in the Google Cloud BigQuery database tables and source data files are stored in Google Cloud Storage buckets. Data were updated periodically, with each stage checked by semi-automatic inspection of time series. The processing methods in each stage are described below. Additional documents presented in the Appendix of the Breathe London Technical Report present a comprehensive overview of caveats to the full data set.

Summary of Stages	
0	Raw data entry into BigQuery and duplicate removal
1	Time alignment
2	Flag assignments (instrument status, exceptional events, operating limits)
3	Data removal or correction based on flags (replacement with null, bias correction); measurement mode codes added
4	Minimum detection limit and uncertainty parameter calculations

Stage 0

Stage 0 begins with the data in their original format on the Google cloud storage bucket. For wireless-transmitted data, the original format consists of 5-minute periods per data file for each car separately; for manually downloaded data (FIDAS size-resolved PN distributions, AE33 biomass burning % and coefficients) the original data files consist of approximately once weekly downloads.

- The data file format for wireless-transmitted data is described by Air Monitors documentation. The CSV files generated by Google for every 5 min of data from each car (file name contains the vehicle ID) have the following typical entry:

  2018-08-17_11-09-23.000000,Devices:6:o3,127,00000000,24.000000,29317424
  The corresponding header (not included in the files) of column names is:

Timestamp, Node ID, Parameter ID, Status code, Value, ID

Timestamp: is in UTC with format YYYY-mm-dd_HH-MM-SS.000000. Timestamp assigned to wireless-transmitted data originates from the clock on the laptop with data logger software installed. While the car is driving and Google system (known as the JiBox) is on, the laptop is not connected directly to the internet. Prior to 12 April 2019, to avoid time drift on the laptop clock, it was synchronized every hour with the JiBox time which does have internet connectivity. As of 1 February 2019, when the cars are docked at NPL with the JiBox off, drivers are connecting the laptop to WiFi and disconnecting from WiFi when the Jibox is turned on again. When the JiBox was identified as having clock drift greater than one second, a more accurate time synchronization system was implemented beginning 12 April 2019. This system used the timestamp from an additional GPS unit to reset the laptop time ensuring that the laptop and GPS clocks were synchronized.[1]

Node ID: The node id has the following format: Devices:<instrument id>:<parameter name>. Consult the configuration section for an explanation of the instrument and parameter IDs.

Parameter ID: The ID of the parameter according to the configuration settings. Consult the configuration section.

Status Code: This is the status code of the device.

Value: This is the value recorded by the device. The units of the value depend on the parameter. The units are given in the parameter configuration settings. See the configuration section.

ID: Row number incrementing as records are inserted into the laptop database. Not unique as the database may be reset periodically.

- Data recorded by the data logger are at 1 second intervals (or 2 seconds in the case of ozone) as returned from the instruments. Air Monitors has provided separate configuration documentation detailing 158 parameters, 53 of which are recorded automatically by the logger, in addition to a status code for each of the 10 instruments.
    - A subset of these parameters (AE33 intensity values, biomass burning) and some additional parameters (FIDAS size-resolved PN distributions, backup $PM_{2.5}$ and MA350 black carbon) are stored locally on the in-car laptop and manually uploaded weekly to Buckets/street-view-air-quality-london/instrument-data/car_data. The manually logged files are in two formats CSV and a proprietary Palas format filename.promo only readable by their free issue software. The timestamps of manually logged data originating from instrument clocks need to be synchronized to the timestamp of the streaming data. Before 12 April 2019, the timestamps of manually logged data were not synchronized to

[1] The time synchronization issue became apparent when records sorted by timestamp were not in the same order as records sorted by database id (consecutive id incrementing with each row in the order entered until database is reset). To the extent possible, clock uncertainty is flagged beginning in Stage1 and expressed as an uncertainty in space given the speed of the vehicle at the time.

the computer time. There will also be an additional log of the time in UTC when the data is manually downloaded.

- Logs of daily checks and weekly calibrations are stored in spreadsheets that are uploaded to a Google Cloud Storage Bucket (Buckets/street-view-air-quality-london/instrument_and_driving_reports/instrument_reports). The documents include daily tabs with the status of instrument checks and notes on instrument performance. In these tabs daily checks employ a traffic light system. Green means the criterion is working as intended (e.g. there is power to an instrument). Yellow means that there was a problem which was fixed that day (e.g. reboot of the equipment or restart logging software). Red means that the problem could not be fixed that day. Grey means the criterion is not valid for that instrument. Additional tabs notate the weekly vehicle calibration, zero, span, and leak checks performed. After 1 February 2019 this document also included start and end times of zero filter checks.

- Logs of the time shift between the instrument and Jibox are stored in the Google Cloud Storage Bucket (Buckets/street-view-air-quality-london/27533/config/logger_time_shift and Buckets/street-view-air-quality-london/27522/config/logger_time_shift). The files are formatted as <time of sync>,<seconds offset>.

- Driver reports with approximate start and end times of each polygon are included as a spreadsheet which is also uploaded to the Google Cloud Storage on a periodic basis.

- Additional documentation includes GUTS tickets that notate all issues with the vehicle and instrument issues reported to Google. These will be manually reviewed monthly and any relevant issues will be noted for inclusion.

From the files in Cloud Storage, data is extracted and loaded into BigQuery tables. Streaming instrument data are extracted and loaded into an intermediate table using private code created by Google. EDF public code copies from the intermediate table into the Stage 0 table after removal of any duplicate records. Manually logged measurement data from the Palas FIDAS and Magee AE33 and from the daily/weekly instrument checks are extracted and loaded into respective tables using a data pipeline designed by Geocene, code for which is available on request.

Stage 1

The goal for Stage 1 is to align pollutant plumes from the same sample given their different instrument response times, to match measurements with their GPS locations, and to provide efficient access to data in a form that allows creation of basic time series, map visualizations, etc.

- The Google BigQuery output table organizes concatenated data into columns for each parameter based on parameter ID along with their status codes (one per device) and

vehicle ID. Status codes will remain part of the data set throughout each subsequent stage in case there is a need to report them or use for analysis purposes.

- The data are time aligned using adjustments estimated by University of Cambridge based on strike test response times and cross-correlation of pollutant measurements with CO_2. We implement the time alignment in the table below rounded to the nearest second. The Aerodyne CAPS NO_2 instrument particle filter/dead volume was changed on 5 Jan 2019 and new time adjustment values were applied to data after the change. In earliest documentation, car 1 = vehicle ID 27533 and car 2 = vehicle ID 27522. Time adjustment values are updated again after the Palas firmware update on 13 Apr 2019. Another Palas firmware update occurred on 1 Aug 2019.
  Table of Absolute measured fixed time offsets (centroids)

Version: abs_to_aug2020_v1						
Fixed time offsets - absolute						
Channel Name	**27533**			**27522**		
	Post Palas Update (After 13 Apr 2019)	**Post-Filter Change (27533) (5 Jan 2019 to 13 Apr 2019)**	**Pre-Filter Change (27533) (Before 5 Jan 2019)**	**Post Palas Update (After 13 Apr 2019)**	**Post-Filter Change (27522) (5 Jan 2019 to 13 Apr 2019)**	**Pre-Filter Change (27522) (Before 5 Jan 2019)**
Devices:1:pm1	3.9	8.1	8.1	4.2	8.3	8.3
Devices:1:pm2.5	4.0	8.1	8.1	4.2	8.2	8.2
Devices:1:pm10	4.6	8.5	8.5	4.5	7.9	7.9
Devices:1:pmt	5.1	8.6	8.6	5.1	8.2	8.2
Devices:2:LDSA	10.8	10.8	10.8	11.8	11.8	11.8
Devices:3:co2	1.6	1.6	1.6	1.6	1.6	1.6

Devices:3:co2 dry	1.6	1.6	1.6	1.6	1.6	1.6
Devices:4:no	7.5	7.5	9.7	8.6	8.6	11.0
Devices:5:BC all channels	7.0	7.0	7.0	7.1	7.1	7.1
Devices:6:o3	4.7	4.7	4.7	4.7	4.7	4.7
Devices:7:no2	3.8	3.8	4.5	3.9 (except 5.8 between 1 May to 3 Jun 2019)	3.9	4.6
Devices:8:pm	9.4	9.4	9.4	10.2	10.2	10.2

- The data are qualified with a spatial uncertainty in units of meters based on the product of temporal uncertainty (s) and vehicle speed (m/s). There are two sources of temporal uncertainty: 1. Drift between GPS and laptop clocks (prior to implementation of the improved sync system) 2. Time rolling averages of measurement data (if applicable). The larger of these two values is used to estimate spatial uncertainty. For the Palas Fidas at the start of the project it was measuring 10 second rolling averages polled at 2-seconds. Starting 12 April 2019, it was averaging and polling at 1-second. The Magee AE33 uses unknown averaging periods. Tests will be conducted to confirm the averaging time of both devices.

Stage 2

A QAQC flag is assigned to every measurement based on parameter-specific criteria. The QAQC flag value is 0 (valid) unless another flag value is applied. Flagging criteria are applied in the following order:

1. Data are flagged as invalid (QAQC flag 9) or suspect (QAQC flag 2) based on instrument operation issues determined by the status codes and relevant parameters. Three categories of status codes are flagged as shown below. The document with all the status codes and status flags can be found **here**. The status code interpretation logic is documented **here**.

Status code flag 2 - Data is invalid
Status code flag 1 - Non-critical instrument error *e.g flow may be outside range but not affecting data quality*. Data may still be valid with an uncertainty < +/- 10% of measured value. User should review data before using in analysis.
Status code flag 0 – Data is valid (status for information only)

If any status code flag = 2 (invalid), all parameters from this device are flagged as invalid (QAQC flag 9). If any status code flag = 1 (may be invalid), all parameters from this device are flagged as suspect (QAQC flag 2).

2. GPS data (primary and backup inside the AE33) are flagged as invalid (QAQC flag 91) if either latitude or longitude are outside a bounding box for the London sampling domain and were not at Air Monitors for maintenance (i.e. a bounding box that includes greater London and Tewkesbury where AM is located).

3. All data are flagged as invalid (QAQC flag 92) during **MOT** tests.

4. Data are flagged as exceptional event (QAQC flag 10) based on periods denoted as exceptional events/logistical car issues that are logged by project partners. These may apply to an entire car's instruments or one or more specific instruments.

If a data point was not already flagged as invalid, the following flags are applied:

5. Data recorded prior to a check value beyond *action limits* recommended by NPL and after the last check value within limits are flagged as suspect (QAQC flag 21 to 24), depending on the check. Check values are recorded in Excel spreadsheet files in the Daily and Weekly Cal Checks folder on the Google Bucket (Buckets/street-view-air-quality-london/ instrument_and_driving_reports/instrument_reports) and imported into BigQuery tables (UK.instrument_*) using the edf_data_pipeline coded by Geocene (available upon request). Checks to consider:
    a. Zero checks: If the baseline has drifted by more than ± 5.5 ppb for NO_2 and Ozone; ± 5.5 ppm for CO_2 and NO; ± 5.5 µg/m³ for BC and PM; and ± 5.5 µg²/cm³ for LDSA. If the drift was less than this, no adjustments were made, and drifts were simply recorded. Apply QAQC flag 21 to the mobile data collected between the time the zero was measured out of bounds and the last in-bound zero check time.
    b. Flow checks – In all of the gas instruments and some of the PM instruments flow variation of up to 10% is not critical as we are not using size selective inlets (except on the PDR) so there is little or no effect on the data quality unless the flow change is due to a blockage or leak which is picked up in the leak checks. Flow checks are reviewed manually. October flow check data stored at Buckets/street-view-air-quality-london/instrument_and_driving_reports/instrument_reports/Oct 18.
    c. Span checks – If the span concentration is not within +/-10% of the certification value, apply QAQC flag 22 to the mobile data collected between the time the span was measured out of bounds and the last in-bound span check time.

d. Apply QAQC flag 23 to the mobile data at times when both the zero and span checks were measured out of bounds.
e. Data are flagged as suspect (QAQC flag 24) if leak checks failed – These are critical in almost all cases and the severity depends on the position of the leak. E.g. a leak ahead of the detection device would result in the dilution of the sample, a leak after the detection device may only affect the flow readings. Apply QAQC flag 24 to the mobile data collected between the time the leak was detected and the last time prior that no leakage was confirmed. Leaks will almost certainly mean zero and span checks failed, so 24 supersedes flags of 21-23. A known major leak occurred between 2019-04-25 and 2019-06-06 (inclusive) on vehicle 27522 and is flagged as invalid (QAQC flag 9).

6. Data are flagged (QAQC flag 25) if the spatial resolution/uncertainty is > 90 m.

7. Data are flagged (QAQC flag 26) if there was low GPS accuracy due to less than 4 tracking satellites. Data already flagged with another issue in addition to a spatial uncertainty issue are flagged with QAQC flag 27.

8. Unless already flagged with another issue, all remaining unflagged PM data are flagged (QAQC flag 28) to indicate that they have a bias because of mobile optical PM sensing. The bias is determined by analysis documented **here** and corrected in Stage 3. PM data already flagged with another issue are flagged with QAQC flag 27.

9. Data are flagged as warm-up (QAQC flag 4) if taken during a warm-up period. The two instruments that require warm-up flags are 2BTechnologies ozone and Serinus 40 NO_x. The 2B ozone monitor warm-up is ~20 min after start-up (low scrubber temperature is also flagged which would indicate warm up period). The Serinus 40 NO_x Monitor warm-up period (up to 60 min) is determined based on the status code value of bit 16 (warm-up process). There was a period prior to 1 Oct 2019 during which the scrubber temperature flag was erroneously activated on both cars. The data corresponding to the low scrubber temps in this period, where this was the only status flagged and the cars were on-road (i.e. not warming up at NPL), are reverted to valid (QAQC flag 0).

10. Data are flagged as above maximum measured value or below a minimum measured value (a conservative low limit beyond which values are clearly invalid, not the same as the minimum detection limit applied in stage 4) for each instrument (QAQC flag 5) based on the values below. Low limits are based on roughly 2x the 0.1th percentile of the on-road measurements, if the value may have noise less than 0, or 0.

Instrument	Species	Maximum Measured Value	Minimum Measured Value
Thermo PDR $PM_{2.5}$ Nephelometer	$PM_{2.5}$	400 mg m^{-3} (400000 ug/m3)	-1 ug/m^3
Magee AE33 Black Carbon Monitor	Black Carbon (7 wavelength)	100 µg m^{-3} (100000 ng/m3)	-2000 ng/m^3
Serinus 40 NO Monitor	NO	20 ppmv	-0.05 ppmv
Aerodyne CAPS Direct NO2 Monitor	NO_2	3ppm (3000 ppbv)	-60 ppbv
Naneos Partector - nano PM Monitor	Lung Deposited Surface Area	20000 μm^2 cm^{-3}	0 μm^2 cm^{-3}
FIDAS 100 PM Monitor	PM_x, PN	PM_x: 1500 $\mu g/m^3$; PN: 20000 particles cm^{-3}	PMx: 0 ug/m^3 PN: 0 n/cm^3
LiCor Model 7200RS Monitor	CO_2 / H_2O	CO_2: 3000 ppmv (3000 umol/mol) H2O: 60,000 ppmv (60 mmol/mol)	0 umol/mol
2B Tech 211G Ozone Monitor	O_3	500 ppbv	-10 ppbv
GPS	Position	NA	
Accelerometer	Shock	NA	

8. Data are flagged based on relative humidity (RH) above each instrument's operating specification (QAQC flag 7). The FIDAS relative humidity will be used to implement this flag as it is measured in the ambient air under a ventilated Stephenson screen. Maximum RH for each instrument is below:

Instrument / Species	Maximum Relative Humidity
(8) Thermo PDR / PM2.5 [140]	95 %
(5) AE33 / Black Carbon [57,58,59,60,61,62,63]	100% Non-condensing
(4) Serinus 40 / NOx [45, 47]	NA (heated sample prevents condensing)
(7) Aerodyne CAPS / NO2 [133]	100% (non condensing)
(2) Naneos Partector / LDSA [21]	Already indicated by status bit 2 (> 80 %).
(1) FIDAS 100 / PMx, PN	NA (FIDAS already accounts for RH)

[1,2,3,4,5,6]	
(3) LiCor 7200RS / CO_2/H_2O [32,33,34]	95 %
(6) 2B Tech 211G / O_3 [127]	Not specified

9. Data are flagged based on temperatures outside of each instrument's operating specification (QAQC flag 6). Operating temperature ranges are summarized here:

Instrument	**Low**	**High**
Thermo PDR 1500 $PM_{2.5}$ Nephelometer	-10°C	50°C
Magee AE33 Black Carbon Monitor	10°C	40°C
Serinus 40 NO_x Monitor	0°C	40°C
Aerodyne CAPS Direct NO_2 Monitor	0°C	45°C
Naneos Partector - nano PM Monitor	-10°C	50°C
FIDAS 100 PM Monitor	0°C	50°C
LiCor Model 7200RS CO_2/H_2O Monitor	-25°C	50°C
2B Tech 211G Ozone Monitor	10°C	50°C

QAQC flag columns for each instrument are composed of integer values for each timestamp:

QAQC Flag Value Definitions	
0	Valid
2	Suspect (generally based on status)
21	Suspect - Zero check out of bounds
22	Suspect - Span check out of bounds
23	Suspect - Zero and span check out of bounds
24	Suspect - Leak issue
25	Suspect - spatial resolution > 90 m
26	Suspect - low GPS accuracy due to < 4 tracking satellites
27	Suspect - compound issues (more than one spatial issue or at least one spatial issue and one or more general/zero/span/leak/$PM_{2.5}$ issues)

28	Suspect - bias on $PM_{2.5}$ to be corrected at stage 3 with the following factors: PM2.5_corrected = PM2.5 _recorded*f where f = : 27522 27533 Palas 1.10 1.12 PDR 0.83 0.71
4	Warm-up
5	Above maximum detection limit
6	Outside temperature bounds
7	Above relative humidity operating specification
9	Invalid (generally based on status)
91	Invalid - outside London bounding box and not at Air Monitors or null latitude or longitude
92	Invalid - MOT testing
93	Invalid - NO2 baseline includes substantial ambient air
94	Invalid - confirmed leak or pump malfunction
10	Exceptional event

Stage 2.1: Stage 2 data are checked through semi-automated or manual procedures to determine whether valid and suspect flagged data are valid. At this point, suspect flags are changed to either valid (flag 0) or invalid (flag 9). The original status codes are retained (see above) such that the suspect flag information is preserved.

Stage 3

Concentration and particle number data are replaced with null (removed) if flag value = 4, 5, 6, 7, 9. 91, 92 or 10 for each instrument/parameter separately. $PM_{2.5}$ measurements that are not invalid for any other reasons are replaced with a bias corrected value.

- For instruments measuring more than one species (LiCOR CO_2 / H_2O and FIDAS PM_x and PN), only data above maximum value for that species (flag 5) would be replaced with null. Additional (non-concentration) parameters for each instrument will also be treated differently. For example, FIDAS meteorological parameters would not have above detection limit, above RH flags applied.

- Bias in otherwise valid $PM_{2.5}$ data is corrected with the following factors:

  PM2.5_corrected = PM2.5_recorded*f
  where f = :
  27522 27533
  Palas 1.10 1.12
  PDR 0.83 0.71

  Analysis used to determine this correction can be found in the Appendix of the Breathe London Technical Report.

- A measurement mode column is added to denote the type of measurement manually or incorporated from supplementary files that will lag the instrument data by days or weeks. Calibration periods (mode=3) are identified by times when the cars are at NPL and there is a spike in the CO_2 concentration over 800 ppm (assumed to be a span check). The period extends 1-hour prior to the span until the cars leave NPL based on GPS outside a buffer around the facility. Calibration periods (mode=3) also include maintenance periods during the days indicated **here**.

  The zero periods (mode=4) include both the weekly zeros over a 5-minute period and longer gas zero calibration periods with analytical air. The dates and times of the longer gas zero calibration periods are provided in the table below:

Gas zero calibration periods			
Vehicle	Devices	Start time	Stop time
27533	All but LiCor (1,2,4,5,6,7,8)	2019-05-23 11:04 UTC	2019-05-23 12:47 UTC
27533	LiCor (3)	2019-05-23 11:04 UTC	2019-05-23 11:49 UTC
27522	Particulate analyzers (1,5,8)	2019-06-05 10:02 UTC	2019-06-05 11:41 UTC
27522	Gas analyzers (2,3,4,6,7)	2019-06-06 09:18 UTC	2019-06-06 10:57 UTC

Measurement mode codes are summarized in the table below:

Measurement Mode Codes	
1	Mobile monitoring (based on time leaving/returning to NPL or lat/lon coordinates) representative of on-road conditions in London
2	Stationary at NPL (e.g. overnight, off-day) based on GPS-coordinate bounding box around NPL and/or driver logs
3	Calibration (excluding zero checks), span, flow, leak check period at NPL (for one or more instruments)
4	Zero periods (scrubber/filter time series) (includes both zero checks and longer gas zero calibrations). This mode is instrument-specific.
5	Other uncategorized. May be close to NPL but unclear whether stationary in the parking lot. May be useful for stationary analyses though GPS coordinates uncertain.

Stage 4

Minimum detection limits, instrument accuracy, and instrument precision are quantified at this stage and provided in the Mobile Uncertainty Documentation (Appendix 3), including:

- **Instrument Uncertainty Documentation Section 1 (NO_2, NO, and CO_2)**
- **Instrument Uncertainty Documentation Section 2 (O_3, LDSA, and Black Carbon)**
- **$PM_{2.5}$ sampling losses aboard the Google Street View cars in London**
- **$PM_{2.5}$ instrumental uncertainties aboard the Google Street View cars in London**

The documents provide details on the methods for calculating uncertainty parameters as well as recommended values. Data users should apply an appropriate minimum detection limit and other uncertainty bounds as they choose based on analysis needs.