Data Dictionaries

# About this Document

This document contains data dictionaries for the data in this repository, as described in the README document.

# Stop-Level Data

Data file:

* stop\_data.csv

Table 1. Stop-Level Data Dictionary

|  |  |
| --- | --- |
| **Variable** | **Definition** |
| runname | A string composed of the date and route on which a sample was collected. The ‘runname’ and ‘time’ variable may be different if a drive day was ran past midnight. |
| date | The sample collection date |
| time | The sample collection date and location arrival time (sampling stops were approximately 2 minutes in duration). |
| location | The stop name (N=309). ‘MS’ prefixes indicate sampling locations representative of the ACT cohort (n=304). ‘MC’ prefixes indicate regulatory agency collocation sites (n=5). |
| stop\_id | A unique identifier for every stop in the campaign |
| instrument\_id | A unique identifier for every instrument used in the campaign consisting of the pollutant measured and a number. Instruments sampled black carbon (‘BC\_’), carbon dioxide (‘CO2\_’), PM2.5 (PM25\_; indirectly with a nephelometer), nitrogen dioxide (‘NO2\_’) and ultrafine particulates (UFP). UFPs were sampled with multiple instruments: NanoScan (‘PMSCAN\_’), DiSCmini (‘pmdisc\_’), screened (‘PMPT\_’) and unscreened (‘PMPTSCREEN\_’) P-TRAK (see Appendix Table A1 for details). Multiple instruments were used to sample the same pollutant, or ‘variable’. |
| variable | The pollutant measured and its concentration units: (information following the colon is the variable name in the dataset)   * PNC (pt/cm3): ns\_total\_conc, pmdisc\_number, pnc\_noscreen, pnc\_screen * BC (ng/m3): ma200\_ir\_bc1 * PM2.5 (µg/m3): neph\_bscat * NO2 (ppb): no2 * CO2 (ppm): co2\_umol\_mol   See Appendix Table A1 for details. PNC was measured with multiple instruments. PM2.5 was estimated from nephelometer readings (see Blanco et al. 2022 reference in the README for details). |
| primary\_instrument | Whether the measurement was collected with the campaign’s primary or backup instrument. Most measurements were collected with a primary instrument. |
| mean\_value | The mean 2-minute pollutant concentration |
| median\_value | The median 2-minute concentration. These values have been quality controlled (QC) more extensively than the 'mean\_value' data (e.g., calibrated against other instrument median readings, compared against agency readings, etc.) since median 2-min concentrations were our summary measures of interest. QC details are described in ACT-TRAP Campaign paper (preprint: https://doi.org/10.1101/2021.09.18.21263522). |

# Annual Average Site Estimates and Predictions for Mobile Monitoring Sites

Data file:

* annual\_data\_and\_predictions.csv

Table 2. Data dictionary for annual average site estimates and predictions for mobile monitoring sites.

|  |  |
| --- | --- |
| **Variable** | **Definition** |
| variable | The pollutant measured and its concentration units:   * PNC (pt/cm3): ns\_total\_conc, pmdisc\_number, pnc\_noscreen, pnc\_screen * BC (ng/m3): ma200\_ir\_bc1: * PM2.5 (µg/m3): pm2.5\_ug\_m3 * NO2 (ppb): no2 * CO2 (ppm): co2\_umol\_mol   See Appendix Table A1 for details. PNC was measured with multiple instruments. PM2.5 was estimated from nephelometer readings (see Blanco et al. 2022 reference in the README for details). |
| location | The stop name (N=309). ‘MS’ prefixes indicate sampling locations representative of the ACT cohort (n=304). ‘MC’ prefixes indicate regulatory agency collocation sites (n=5). |
| longitude | Location longitude |
| latitude | Location latitude |
| annual | Annual summary estimation method. The first word indicates how stop-level concentrations were summarized. The last word indicates how 2-min stops were summarized.  These dataset may be restricted to the primary pitcp,es, which use the mean of winsorized medians (mean\_of\_win\_medians) approach. In this approach, two-minute medians were calculated for each stop. The bottom and top 5% concentrations for each site were set to those respective thresholds (winsorized) to reduce the influence of outlier concentrations. The mean of these winsorized medians was used to summarize site annual averages. |
| value | Pollutant concentration estimate using a given “annual” approach (above). |
| prediction | Pollutant concentration UK-PLS model prediction using a given “annual” approach (above). |

# Predictions at New Locations

Data Files:

* grid\_predictions.csv
* census\_block\_predictions.csv

Table 3. Data dictionary for predictions at new locations

|  |  |
| --- | --- |
| **Variable** | **Definition** |
| Location\_id, native\_id | Unique location identifiers for each prediction point |
| block\_key, blockgroup\_key, tract\_key, msa | Spatial identifiers for the location\_id/native\_id according to the 2010 Census geographical divisions. See census reference for more information: <https://www.census.gov/programs-surveys/geography/guidance/geo-identifiers.html>  These variables are useful when merging to Census files. |
| longitude | Location longitude |
| latitude | Location latitude |
| in\_monitoring\_area | TRUE if the location\_id/native\_id was within the monitoring area, FALSE otherwise. Locations not in the monitoring area may have spatial extrapolation concerns, depending on how far away they are. |
| variable | The pollutant measured and its concentration units:   * PNC (pt/cm3): ns\_total\_conc, pmdisc\_number, pnc\_noscreen, pnc\_screen * BC (ng/m3): ma200\_ir\_bc1: * PM2.5 (µg/m3): pm2.5\_ug\_m3 * NO2 (ppb): no2 * CO2 (ppm): co2\_umol\_mol   See Appendix Table A1 for details. PNC was measured with multiple instruments. PM2.5 was estimated from nephelometer readings (see Blanco et al. 2022 reference in the README for details). |
| annual | Annual summary estimation method. The first word is whether means or medians were used to summarize stop-level concentrations. The last word is whether means or medians were used to summarize 2-min stops.  The primary analysis uses the mean of winsorized medians (mean\_of\_win\_medians) approach. In this approach, two-minute medians were calculated for each stop. The bottom and top 5% concentrations for each site were set to those respective thresholds (winsorized) to reduce the influence of outlier concentrations. The mean of these winsorized medians was used to summarize site annual averages.  You may alternatively be interested in simple site averages (mean\_of\_means) or medians (median\_of\_medians). |
| prediction | Pollutant concentration UK-PLS model prediction using a given “annual” approach (above). |

# Geographic Covariates

Data files:

* dr0311\_mobile\_covariates.csv
  + covariates where mobile monitoring measurements were collected
* dr0311\_grid\_covariates.csv
  + covariates for grid locations where predictions may be of interest
* wa\_block10\_covariates.csv
  + covariates for Census blocks where predictions may be of interest

These files contain geographic covariates, some of which are summarized below. Details can be found in the MESAAirDOOP\_20190501.pdf file.

Table 4. geographic covariate labels and descriptions

|  |  |  |  |
| --- | --- | --- | --- |
| Kind | Covariate | Buffers | Description |
| airports | log\_m\_to\_airp | 0 | log meters to closest airport |
| airports | log\_m\_to\_l\_airp | 0 | log meters to closest large airport |
| bus | bus\_s | 100, 150, 300, 400, 500, 750, 1000, 1500, 3000, 5000 | sum of bus routes |
| bus | log\_m\_to\_bus | 0 | log meters to bus route |
| coast | log\_m\_to\_coast | 0 | log meters to closest coastline |
| columnar NO2 | no2\_behr | 0 | columnar NO2, mean from 2005-2007 |
| commercial and services | log\_m\_to\_comm | 0 | log meters to closest commercial and services area |
| elevation | elev\_above | 1000, 5000 | number of points (out of 24) more than 20 m and 50 m uphill of a location for a 1000 m and 5000 m buffer, respectively |
| elevation | elev\_at\_elev | 1000, 5000 | number of points (out of 24) within 20 m and 50 m of the location' elevation for a 1000 m and 5000 m buffer, respectively |
| elevation | elev\_below | 1000, 5000 | number of points (out of 24) more than 20 m and 50 m downhill of a location for a 1000 m and 5000 m buffer, respectively |
| elevation | elev\_elevation | 0 | elevation above sea level in meters |
| elevation | elev\_stdev | 1000, 5000 | standard deviation of elevation of 20 points surrounding the location |
| imperviousness | imp\_a | 50, 100, 150, 300, 400, 500, 750, 1000, 3000, 5000 | average imperviousness |
| land use | rlu\_decid\_forest\_p | 500, 750, 1000 | proportion of deciduous forest |
| land use | rlu\_dev\_hi\_p | 300, 400, 500, 750, 1000, 3000, 5000 | proportion of highly developed land (e.g., commercial and services; industrial; transportation, communication and utilities) |
| land use | rlu\_dev\_lo\_p | 50, 100, 150, 300, 400, 500, 750, 1000, 3000, 5000 | proportion of low developed land (e.g., residential) |
| land use | rlu\_dev\_med\_p | 50, 100, 150, 300, 400, 500, 750, 1000, 3000, 5000 | proportion of medium developed land (e.g., residential) |
| land use | rlu\_dev\_open\_p | 150, 300, 400, 500, 750, 1000, 3000, 5000 | proportion of developed open land |
| land use | rlu\_evergreen\_p | 400, 500, 750, 1000 | proportion of evergreen forest |
| land use | rlu\_mix\_forest\_p | 500, 750, 1000, 5000 | proportion of mixed forest |
| NDVI | ndvi\_q25\_a | 250, 500, 1000, 2500, 5000, 7500, 10000 | NDVI (25th quantile) |
| NDVI | ndvi\_q50\_a | 250, 500, 1000, 2500, 5000, 7500, 10000 | NDVI (50th quantile) |
| NDVI | ndvi\_q75\_a | 250, 500, 1000, 2500, 5000, 7500, 10000 | NDVI (75th quantile) |
| NDVI | ndvi\_summer\_a | 250, 500, 1000, 2500, 5000, 7500, 10000 | average summer time NDVI |
| NDVI | ndvi\_winter\_a | 250, 500, 1000, 2500, 5000, 7500, 10000 | average winter time NDVI |
| population | pop10\_s | 500, 1000, 1500, 2000, 2500, 3000, 5000, 10000, 15000 | 2010 population density |
| port | log\_m\_to\_l\_port | 0 | log meters to closest large port |
| railroads, rail yards | log\_m\_to\_rr | 0 | log meters to closest railroad |
| railroads, rail yards | log\_m\_to\_ry | 0 | log meters to closest rail yard |
| roads | intersect\_a1\_a3\_s | 3000 | number of A1-A3 road intersections |
| roads | intersect\_a3\_a3\_s | 500, 1000, 3000 | number of A3-A3 road intersections |
| roads | ll\_a1\_s | 1500, 3000, 5000 | length of A1 roads |
| roads | ll\_a2\_s | 5000 | length of A2 roads |
| roads | ll\_a23\_s | 100, 150, 300, 400, 500, 750, 1000, 1500, 3000, 5000 | length of A2 and A3 roads |
| roads | ll\_a3\_s | 100, 150, 300, 400, 500, 750, 1000, 1500, 3000, 5000 | length of A3 roads |
| roads | log\_m\_to\_a1 | 0 | log meters to closest A1 road |
| roads | log\_m\_to\_a1\_a1\_intersect | 0 | log meters to closest A1-A1 road intersection |
| roads | log\_m\_to\_a1\_a2\_intersect | 0 | log meters to closest A1-A2 road intersection |
| roads | log\_m\_to\_a1\_a3\_intersect | 0 | log meters to closest A1-A3 road intersection |
| roads | log\_m\_to\_a123 | 0 | log meters to closest A1, A2 or A3 road |
| roads | log\_m\_to\_a2 | 0 | log meters to closest A2 road |
| roads | log\_m\_to\_a2\_a2\_intersect | 0 | log meters to closest A2-A2 road intersection |
| roads | log\_m\_to\_a2\_a3\_intersect | 0 | log meters to closest A2-A3 road intersection |
| roads | log\_m\_to\_a23 | 0 | log meters to closest A2 or A3 road |
| roads | log\_m\_to\_a3 | 0 | log meters to closest A3 road |
| roads | log\_m\_to\_a3\_a3\_intersect | 0 | log meters to closest A3-A3 road intersection |
| stack emissions | em\_CO\_s | 3000, 15000, 30000 | sum of CO stack emissions |
| stack emissions | em\_NOx\_s | 15000, 30000 | sum of NOx stack emissions |
| stack emissions | em\_PM10\_s | 15000, 30000 | sum of PM10 stack emissions |
| stack emissions | em\_PM25\_s | 15000, 30000 | sum of PM2.5 stack emissions |
| stack emissions | em\_SO2\_s | 15000 | sum of SO2 stack emissions |
| truck routes | log\_m\_to\_truck | 0 | log meters to closest truck route |
| truck routes | tl\_s | 750, 1000, 1500, 3000, 5000, 10000, 15000 | length of truck routes |
| water | log\_m\_to\_waterway | 0 | log meters to closest waterway |
| water | rlu\_water\_p | 1000, 3000, 5000 | proportion of water |

# UK-PLS Model Performances for Annual Average TRAP from Mobile Monitoring

Data Files:

* model\_performance.csv

Table 5. Data dictionary for UK-PLS model performances for annual average TRAP from mobile monitoring.

|  |  |
| --- | --- |
| **Variable** | **Definition** |
| variable | The pollutant measured and its concentration units:   * PNC (pt/cm3): ns\_total\_conc, pmdisc\_number, pnc\_noscreen, pnc\_screen * BC (ng/m3): ma200\_ir\_bc1 * PM2.5 (µg/m3): pm2.5\_ug\_m3 * NO2 (ppb): no2 * CO2 (ppm): co2\_umol\_mol   See Appendix Table A1 for details. PNC was measured with multiple instruments. |
| annual | Annual summary estimation method. The first word is whether means or medians were used to summarize stop-level concentrations. The last word is whether means or medians were used to summarize 2-min stops.  The primary analysis uses the mean of winsorized medians (mean\_of\_win\_medians) approach. In this approach, two-minute medians were calculated for each stop. The bottom and top 5% concentrations for each site were set to those respective thresholds (winsorized) to reduce the influence of outlier concentrations. The mean of these winsorized medians was used to summarize site annual averages.  You may alternatively be interested in simple site averages (mean\_of\_means) or medians (median\_of\_medians). |
| out\_of\_sample | Sites used to estimate performance statistics |
| no\_sites | Number of sites used to calculate performance statistics |
| RMSE | Root mean square error |
| MSE\_based\_R2 | MSE-based R2 |
| Reg\_based\_R2 | Regression-based R2 (typical approach) |

# Appendix

Table A1. Instrumentation used to measure air pollutants

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Parameter** | **Instrument** | **Manufacturer** | **Measurement Range** | **Limit of Quantification** | **Time Resolution** |
| **Particles (pt)** |  |  |  |  |  |
| UFP |  |  |  |  |  |
| 10-420 nm  (13-bin PSDa) | NanoScan 3910 | TSI | 102 - 106 pt/cm3 | 10 pt/ cm3 | 60 sec |
| 10-700 nm | DiSCmini | Testo | 103 - 106 pt/cm3 | 500-2,000 pt/cm3 b | 1 sec |
| 20-1,000 nm | PTRAK 8525 | TSI | 0 - 5x105 pt/cm3 | 1 pt/cm3 | 1 sec |
| 36-1,000 nm | PTRAK 8525, with diffusion screen | TSI | 0 - 5x105 pt/cm3 | 1 pt/cm3 | 1 sec |
| BCb | microAeth MA200 | AethLabs | 0 - 106 ng/m3 | 30 ng BC/m3 c | 10 sec |
| Light scattering nephelometer (PM2.5) | M903 | Radiance Research | 0 - >1 km-1 | 10-6 m-1 | 10 sec |
| **Gases** |  |  |  |  |  |
| NO2 | CAPS NO2 | Aerodyne Research, Inc. | 0 – 2x103 ppb | 2 ppbv | 1 sec |
| CO2 | LI-850 | Li-Cor | 0-5x103 ppm (vol) | 100 ppmv | 1 sec |

a Only total counts are provided in these data

b Black carbon was assessed by Micro-aethalometer (MA200), which measures the concentration of light absorbing particles on a paper tape which advances automatically, allowing it to run without human intervention. It has 5 different analytical channels measuring wavelengths 880 nm (infrared/IR – the measure presented in these data), 625 nm (red), 528 nm (green), 470 nm (blue), and 375 nm (ultraviolet/UV), with the 800 nm interpreted as black carbon, and the 375 nm interpreted as ultraviolet particulate matter. At each time point, the instrument reports the absolute reading of the reference (REF) area of the tape (where no particles are deposited), the attenuation (ATN) of the collection area, and the concentration value (BC) for the relevant timescale, which is derived by the instrument's internal software.