Steps for Creating Annual Averages

Calculating Derived Local Conditions

Air quality samples in the Air Toxics Archive (ATA) are identified as being collected using either local conditions or standard conditions.

Calculating local conditions/standard conditions ratios

1. Subset archive data to those samples that have both standard and local conditions available.
2. Within a given year, calculate the average standard conditions and the average local conditions by pollutant/POC/site/day/sampling duration.
3. Within a given year, average the ratios calculated above by site/day. Note: this is averaged across pollutants and sampling durations.
4. Within a given year, average the ratios calculated above by site/quarter.
5. Within a given year, average the ratios calculated above by site/year.
6. From the ratio by site/year calculated above, average the ratios for the site across all available years.

Pre-processing the data

Subsetting data
1. Data are removed that are not relevant or appropriate for the calculation of annual averages (e.g., a sampling duration of 1 month and integrated 2-weeks samples, pollutants that are not HAPs, etc.).

Applying the local conditions/standard conditions ratios to derive local conditions
1. Identify all the samples collected using standard conditions but not local conditions.
2. If the ratio exists for that pollutant/POC/site/day/sampling duration, derive local conditions by multiplying the sample collected using standard conditions and the ratio.
3. If the above ratio does not exist but the ratio exists for that site/day, derive local conditions by multiplying the sample collected using standard conditions and the ratio.
4. If the above ratio does not exist but the ratio exists for that site/quarter, derive local conditions by multiplying the sample collected using standard conditions and the ratio.
5. If the above ratio does not exist but the ratio exists for that site/year, derive local conditions by multiplying the sample collected using standard conditions and the ratio.
6. If the above ratio does not exist but the ratio exists for that site, derive local conditions by multiplying the sample collected using standard conditions and the ratio.
7. If the ratio by site is missing for the sample collected using standard conditions, assume that the ratio=1 (i.e., local conditions equals standard conditions).

Samples in either local conditions or derived local conditions are carried forth to calculate daily averages unless otherwise noted.

Calculating daily averages of the non-NOAA data
The non-NOAA data are collected from multiple sampling durations.
1. Separate the data into those samples collected using minute sampling durations and those collected using hourly sampling durations.

Calculating daily averages from the minute sampling durations
1. Determine the number of samples per pollutant/site/POC/day/hour/sampling duration. If this number meets or exceeds the minimum number of samples needed, average up to the hour. This constitutes a valid hour. If the minimum number of samples is not met, these samples are removed.
2. From all the valid hours, determine the number of samples per pollutant/site/POC/day/sampling duration. If there are at least 18 VALID HOURS in a day, averaged up to the day. This constitutes a valid day from the minute data.

Calculating daily averages from the hourly sampling durations
1. Determine the number of samples per pollutant/site/POC/day/sampling duration. If this number meets or exceeds the minimum number of samples needed, average up to the day. If the minimum number of samples is not met, these samples are removed. Note: 90 MINUTES and 150 MINUTES are considered to have an hourly sampling duration.

Averaging across POCs
1. If there are multiple POCs per pollutant/site/day/sampling duration, remove all samples equal to zero if at least 50% of samples are NOT zero.
2. If there are multiple POCs per pollutant/site/day/sampling duration, remove all samples NOT equal to zero if more than 50% of the samples ARE zero.
3. After the appropriate valid collocated daily averages are removed (if necessary), average the remaining valid collocated daily averages by pollutant/site/day/sampling duration across POCs.
Note: Removing some collocated POCs ensures that values that are mostly zero will result in a zero daily average and values that are mostly NOT zero will result in a daily average that does not contain zeros.

Calculating daily averages of the NOAA data
1. Average the NOAA data by pollutant/site/day/sampling duration. Note: The NOAA data do not require a minimum number of minute/hourly samples. The NOAA data may have multiple collocated POCs by pollutant/site/day/sampling duration.

Calculating annual averages from the daily averages
An annual average is calculated by averaging together valid daily averages by pollutant/site/year/sampling duration meeting the criteria discussed below.

Valid quarters
1. Annual averages are calculated for a given pollutant/site/year/sampling duration only if there are at least 3 valid quarters for a given pollutant/site/year/sampling duration. A valid quarter is defined as a quarter having at least 7 daily averages per pollutant/site/quarter/sampling duration. If there are less than 3 valid quarters, all daily averages for a given pollutant/site/year/sampling duration are removed and an annual average is not calculated. Note: It is possible for a pollutant/site/year/sampling duration to contain the minimum number of valid quarters and still contain one “invalid” quarter that contains data. The data in this “invalid” quarter are included to calculate an annual average by pollutant/site/year/sampling duration. Note: a daily non-detect average assigned as zero can be included in the as part of the calculation of a valid quarter.

Calculating annual averages
Annual averages and corresponding statistics are calculated by pollutant/site/year/sampling duration.

Annual averages from all data
Averages and statistics are calculated from all valid daily averages using local conditions or derived local conditions from an LC/STD ratio.

Annual statistics defining non-detects as zero
The following annual statistics are calculated from valid daily averages using non-detects as zeros assuming the minimum threshold of valid quarters is met: arithmetic mean, variance, maximum, 10th percentile, 25th percentile, median, 75th percentile, and 90th percentile.

Annual statistics using censored daily averages
Annual statistics are also calculated from valid daily averages using regression on order statistics (ROS) that allows for censored values (i.e., daily averages that are non-detects). Due to the stability of the ROS, these statistics are only calculated if less than 80% of the valid daily averages are below MDL. The censored value assigned to each daily value is $$\min\{MDL, \min \ (day \ in \ year)\}$$, where $\min \ (day \ in \ year)$ is the minimum daily non-zero value for a given pollutant/site/year/sampling duration.

The following annual statistics are calculated from valid daily averages using ROS assuming the minimum threshold of valid quarters is met: arithmetic mean, 10th percentile, 25th percentile, median, 75th percentile, and 90th percentile.

Note: The MDLs are in standard conditions while the original samples may be collected using local conditions. If there were multiple MDLs used in the construction of the daily average (e.g., through multiple POCs or through a sub-daily sampling duration), the average MDL is taken.

Annual averages only from data collected using standard conditions
Annual averages and corresponding statistics are calculated by pollutant/site/year/sampling duration for data only originally collected using standard conditions.

Annual statistics defining non-detects as zero
The following annual statistics are calculated from valid daily averages using non-detects as zeros on samples originally collected using standard conditions assuming the minimum threshold of valid quarters is met: arithmetic mean, variance, maximum, 10th percentile, 25th percentile, median, 75th percentile, and 90th percentile.

**Annual statistics using censured daily averages**
The following annual statistics are calculated from valid daily averages using ROS on samples originally collected using standard conditions assuming the minimum threshold of valid quarters is met: arithmetic mean, 10th percentile, 25th percentile, median, 75th percentile, and 90th percentile.

**Q&As**

**Why are local conditions derived?**
Local conditions are derived from standard conditions to utilize data for toxics applications where local conditions are required (e.g., model evaluation and risk applications).

**Why are daily averages calculated differently for NOAA and non-NOAA data?**
Daily averages are calculated differently for NOAA and non-NOAA data. For non-NOAA data, sub-daily data depending on the sampling duration require a minimum number of samples to calculate a daily average. This ensures that daily averages are representative of the day. However, NOAA data do not have that requirement. For example, for NOAA data it is possible for a daily average to be represented by one 5-minute sample. This criterion is loosened for the NOAA data because the pollutants and monitored locations typically represent background concentrations that do not substantially fluctuate at sub-daily time scales.

**Why are annual averages separated by sampling duration?**
Annual averages are separated by pollutant/site/year/sampling duration to account for broad method differences among different sampling durations.

**How is the minimum number of samples of the different sampling durations determined when calculating the daily averages?**
A minimum number of samples are needed to calculate a daily average when the sampling duration is sub-daily. The minimum number of samples vary determining on the scale of the sampling duration (i.e., minute or hourly). These are found under the “Durations” tab for ATA2018_lookups.xlsx. If the duration description is under 60 minutes, the minimum count is defined as \[0.75 \times \frac{60}{\text{DUR\_DESC}}\]. If the duration description is 60 minutes or above, the minimum count is defined as \[0.75 \times \frac{24}{\text{DUR\_DESC}}\], where the duration description is defined in hours. In other words, the minimum count is rounding up the minimum number of samples needed to have 75% completeness for an hour (or day).

**Table 1. Minimum sampling count needed to construct an hourly or daily average by sampling duration.**

<table>
<thead>
<tr>
<th>Duration Description</th>
<th>Average To</th>
<th>Minimum Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>5 MINUTES</td>
<td>HOURLY</td>
<td>9</td>
</tr>
<tr>
<td>10 MINUTES</td>
<td>HOURLY</td>
<td>5</td>
</tr>
<tr>
<td>15 MINUTES</td>
<td>HOURLY</td>
<td>3</td>
</tr>
<tr>
<td>30 MINUTES</td>
<td>HOURLY</td>
<td>2</td>
</tr>
<tr>
<td>150 MINUTES</td>
<td>DAILY</td>
<td>8</td>
</tr>
<tr>
<td>90 MINUTES</td>
<td>DAILY</td>
<td>12</td>
</tr>
<tr>
<td>1 HOUR</td>
<td>DAILY</td>
<td>18</td>
</tr>
<tr>
<td>2 HOUR</td>
<td>DAILY</td>
<td>9</td>
</tr>
<tr>
<td>3 HOURS</td>
<td>DAILY</td>
<td>6</td>
</tr>
<tr>
<td>4 HOUR</td>
<td>DAILY</td>
<td>5</td>
</tr>
<tr>
<td>5 HOUR</td>
<td>DAILY</td>
<td>4</td>
</tr>
</tbody>
</table>
Where was the information for the tabs in the look up tables obtained?

- AQS_CAS: [https://www3.epa.gov/ttnairs1/airsaqsorig/manuals/Parms%20by%20CAS.xls](https://www3.epa.gov/ttnairs1/airsaqsorig/manuals/Parms%20by%20CAS.xls)

What is the minimum number of daily averages required to create a valid quarter?

A minimum of 7 daily averages is required per pollutant/site/sampling duration to create a valid quarter.

What is the number of valid quarters needed to calculate an annual average?

A minimum of 3 valid quarters in a year are needed per pollutant/site/sampling duration to create an annual average. This is to ensure that the annual average is representative of the year. All daily averages are used in the calculation of an annual average. Therefore, it is possible for a daily average to be contained within an “invalid quarter” but still be used in the calculation of an annual average if the 3 valid quarter criterion is met.

Code Descriptions

- ATA2018_analysis.R: Sources and calls all the functions used to create annual statistics.
- ATA2018_LC2STDRatio.R: Creates an average LC to STD ratio based on an average across days, quarters, years, or all years for that site.
- ATA2018_preprocessing.R: Pre-processes the data including removing some select sites and parameters and applies the LC/STD ratio to data. These ratios are used to estimate LC where the final standardized value is STD due to a lack of temperature/pressure data for that site/day which would be used to convert STD to LC.
- ATA2018_duration2daily.R: Calculates daily averages by pollutant/site/day/sampling duration of all the non-NOAA data from all the possible sub-daily and sub-hourly sampling durations based on a set of criteria.
- ATA2018_duration2daily_NOAA.R: Calculates daily averages by pollutant/site/day/sampling duration of all the NOAA data from all the possible sub-daily and sub-hourly sampling durations based on a set of criteria.
- ATA2018_daily2annual.R: Calculates annual averages by pollutant/site/day/sampling duration of the non-NOAA and NOAA data based on a set of criteria.
- ATA2018_file.R: Creates the final annual excel file.

Quality Assuring the Methodology

Method Differences from the Previous Archive

The methods are slightly difference between the ATA version 2017 and the ATA version 2018.

1. Samples with the a sampling collection frequency code identified with a letter (e.g., “J” = “EVERY 3RD DAY: 1-24 HR – PAM,” “A” = “DAILY: 24 – 1 HR SAMPLES -PAMS”) are now included in the ATA.
2. Collocation by pollutant/site/day/sampling duration of the NOAA data are now differentiated by POC.
3. The minimum count needed to construct a valid hourly/daily average changed for some sampling durations. Namely, 10 MINUTES (4 to 5), 30 MINUTES (1 to 2), 150 MINUTES (6 to 8), and 90 MINUTES (9 to 12) changed. 5 HOUR criterion was added (minimum count of 4). These changes had no measurable impact on the ATA version 2018 annual averages due to how infrequently samples are collected for these sampling durations.
Quality Assuring the Air Toxics Archive

QA of the ATA version 2018 was performed by comparing the annual averages by pollutant/site/year/sampling duration from the 2018 version to the 2017 version using the updated method outlined above. The QA criteria and results are presented in the bulleted list below. In short, the QA procedures of the ATA were split into two time periods: (1) years 1990-2017 and (2) year 2018. The years 1990-2017 encompass the time period for the ATA version 2017 and all annual averages from 2018 are new. Because the ATA version 2018 still pulls data from the previous time period (i.e. 1990-2017), it is possible for the annual averages from this previous time period to be altered from the ATA version 2017. From 1990-2017, annual averages can be (1) identical, (2) different, (3) added, or (4) removed. The criteria for identifying potentially suspect annual averages are outlined in the bulleted list below.

QA Criteria

Comparing annual averages by pollutant/site/year/sampling duration from ATA version 2017 to ATA version 2018

- Annual average from the ATA version 2018 from 1990-2018 (267,936 annual averages; 100%*)
- Time period 1990 – 2017 (256,550 annual averages; 95.8%)
  - Annual averages are the **SAME** (222,524 annual averages; 83.1%)
    - Suspect criteria: NA
    - **Suspect number:** NA
  - Annual averages are **DIFFERENT** (30,616 annual averages; 11.4%)
    - Annual average exists within a valid TS from ATA version 2017 (29,809; 11.1%)
      - Suspect criteria: (1) annual average > 10 x max{TS version 2017} AND (2) percent difference from version 2018 to version 2017 > 10
      - **Suspect number:** 10; **0.004%**
    - Annual average does NOT exist within a valid TS from ATA version 2017 (807; 0.3%)
      - Suspect criteria: (1) annual average > 95th percentile of pollutant/year/sampling duration across sites from version 2017 AND (2) percent difference from version 2018 to version 2017 > 10
      - **Suspect number:** 1; **0.0004%**
  - Annual averages are **REMOVED** (780 annual averages)
    - Suspect criteria: NA
    - **Suspect number:** NA
    - Action: characterize removed annual averages by pollutant, site, year, sampling duration, supporting agency, number of “NA” samples, and number of invalided samples. Note: 780 is excluded from 256,550.
  - Annual averages are **ADDED** (3,410 annual averages; 1.3%)
    - Annual average exists within a valid TS with a sufficient number of years (1,337; 0.5%)
      - Suspect criteria: (1) max{TS version 2018} > 10 x max{TS version 2017} AND (2) number of years in TS from version 2018 < 2 x number of years in TS from version 2017
      - **Suspect number:** 13; **0.005%**
    - Note: additional criteria to be added in subsequent ATA versions.
  - **Total suspect number from 1990-2017:** **24; 0.01%**
- Time period 2018 (11,386 annual averages; 4.2%)
  - Annual average for 2017 exists (9,349; 3.5%)
    - Suspect criteria: (1) annual average > 10 x annual average for 2017 from ATA version 2018 AND (2) annual average > max{TS up to year 2017 from version 2018}
    - **Suspect number:** 58; **0.02%**
  - Annual average for 2017 does NOT exist AND valid TS exists (1,256; 0.5%)
    - Suspect criteria: annual average > 10 x max{TS up to year 2017 from version 2018}
    - **Suspect number:** 4; **0.001%**
Annual average for 2017 does NOT exist AND valid TS does NOT exist (781; 0.3%)

- Suspect criteria: annual average > 95th percentile of pollutant/year/sampling duration across sites from version 2018
- **Suspect number: 45; 0.02%**

**Total suspect number from 2018: 107; 0.04%**

^All percentage are calculated from the total number of annual averages from ATA version 2018
&A time series is considered a “valid” time series if there exist annual averages across more than one year by pollutant/site/sampling duration where the range of the annual averages is greater than zero.
*TS = time series

**Characterizing Removed Annual Averages**

Removed annual averages have no comparison set, so they are evaluated through a characterization given their known data and data sources. Given that most of the QA is done are the level of annual averages and not the individual datum, it is possible that an annual average exists for the previous version of the ATA and a few samples were removed causing it to fall just below the threshold for creating an annual average. Therefore, it is necessary to evaluate individual samples of the removed annual averages. For each annual average the following information is determined: (1) the number of days of data in the previous version of the ATA, (2) the number of valid days of data, (3) the number of days of data (including NA or invalidated data), (4) the percentage of data invalidated, (5) the percentage of NA data. Note: it is possible that the annual averages removed simply do not exist in the newest archive.

### Table 2. Characterizing the 780 annual averages removed from ATA version 2017 to ATA version 2018.

<table>
<thead>
<tr>
<th>Number of annual averages removed</th>
<th>Site/pollutant/year/duration description</th>
<th>Data file missingness description</th>
</tr>
</thead>
<tbody>
<tr>
<td>22</td>
<td>Naphthalene (AQS code 45850)</td>
<td>100% data now NA</td>
</tr>
<tr>
<td>35</td>
<td>ATA site code 42-049-004 in 2016</td>
<td>&gt;50% data now NA</td>
</tr>
<tr>
<td>14</td>
<td>ATA site code 06-019-008 in 2012 and 2013</td>
<td>100% invalidated</td>
</tr>
<tr>
<td>705</td>
<td>5 MINUTE NOAA (all years removed/invalidated)</td>
<td>38 were 100% invalidated (all of these were Carbon Tet) and 667 were removed from data file</td>
</tr>
<tr>
<td>2</td>
<td>ATA site code 48-257-0020 in 2015 and 2016</td>
<td>large % NA</td>
</tr>
<tr>
<td>2</td>
<td>5 MINUTE NOAA (15001NMLO for 1,1,1-Trichloroethane in 1996 and 55099NLEF for Tetrachloroethylene in 2007)</td>
<td>---</td>
</tr>
</tbody>
</table>

### Quality Assurance Assumptions

- It is assumed that the ATA version 2017 is a valid comparison set to the ATA version 2018.
- The suspect annual averages are those that are abnormally high, but not those that are abnormally low.
- The suspect annual averages are those that are abnormally high, but the QA procedures do not necessarily identify the reasons why the annual averages are abnormally high.