# U.S. EPA 2018 Air Trends Report Methodologies

## Overview

EPA maintains an annual air trends report in the form of an interactive web application (ex. <https://gispub.epa.gov/air/trendsreport/2018/>). The online report features a suite of visualization tools that allow the user to:

* Learn about air pollution and how it can affect our health and environment.
* Compare key air emissions to gross domestic product, vehicle miles traveled, population, and energy consumption back to 1970.
* Take a closer look at how the number of days with unhealthy air has dropped since 2000 in 35 major U.S. cities.
* Explore how air quality and emissions have changed over time for each of the common air pollutants.
* Check out air trends where you live.

Users are also able to share this content across social media, with one-click access to Facebook, Twitter, Pinterest, and other major social media sites. Data, source code and documentation, including this document, are available for download at the air trends report GitHub repository <https://github.com/USEPA/Air-Trends-Report>.

This document details the methodologies used in compiling the National Ambient Air Quality Standards (NAAQS), PM2.5 speciation, visibility and toxics trends data found in the online annual air trends report.

### National Ambient Air Quality Standards (NAAQS) Data

#### Methodology for assessing trends in criteria pollutant concentrations

1. Query the EPA's Air Quality System (AQS) for the relevant summary statistics for each criteria pollutant, along with any information necessary for assessing annual completeness. Include data flagged as exceptional events. PM2.5 is stored under multiple parameter codes in AQS - for trends purposes, query only the Federal Reference and Equivalent Method data (parameter code 88101 in AQS). Annual summary lead concentrations for lead are rounded to two decimals places in accordance with the national ambient air quality standard. Lead summary statistics below 0.005 μg/m3 are reported as zero.
2. Assess completeness of the annual summary statistic. Valid years must meet the following criteria:

|  |  |  |
| --- | --- | --- |
| **Pollutant** | **Annual Summary Statistic** | **Completeness Criteria** |
| CO | Second max non-overlapping 8-hour concentration, ppm | >= 3285 hourly obs |
| Pb | Max rolling 3-month average concentration, μg/m3 | >= 9 valid rolling 3-month averages in the year |
| NO2 | Arithmetic mean concentration, ppb | >= 4380 hourly obs |
| 98th percentile of daily max 1-hour concentrations, ppb | >= 75% hours in a day>= 75% days in a qtr for all 4 qtrs |
| Ozone | Fourth highest of daily max 8-hour concentrations, ppm | >= 50% of the hourly obs(annual\_obs\_pct >= 50%) |
| PM10 | Second max 24-hour concentration, μg/m3 | >= 30 daily obs |
| PM2.5 | Weighted annual mean concentration, μg/m3 | >=11 daily obs for all 4 calendar quarters |
| 98th percentile 24-hour concentration, μg/m3 | >=11 daily obs for all 4 calendar quarters |
| SO2 | 99th percentile of daily max 1-hour concentrations, ppb | >= 75% hours in a day>= 75% days in a qtr for all 4 qtrs |

1. Require at least 75 percent of valid years during the trend period. For example, 12 valid years would be required for the 15-year period 1990-2004. In addition, sites must not be missing more than two consecutive years of data to be considered a “trend site” or “national stats site” factored into the summary statistics (national average etc.).
2. If there are multiple monitors at the same site, then use only the most complete (by number of valid years), lowest Parameter Occurrence Code (POC; 1,2,3, etc.) monitor, in that order, to represent the site.
3. Use linearly interpolation to fill in for any missing years. Missing end years are replaced with the value of the nearest year.
4. Use the interpolated data set for assessing national and regional trends. Use the uninterpolated data set for assessing trends at individual sites.
5. For the interpolated data sets, determine whether the data indicates a statistically significant trend up or down, or no trend, using a nonparametric method commonly referred to as the Theil test (M. Hollander and D.A. Wolfe, Nonparametric Statistical Methods, John Wiley and Sons, Inc., New York, NY, 1973.) The test provides a slope estimate, indicating the direction of the trend. For the purposes of this trend assessment, use a p-value of 0.05 to determine significance.

### PM2.5 Speciation Data

#### Methodology for generating figures for the PM2.5 chemical composition

1. Query AQS quarterly summary tables for all PM2.5 chemical speciation sites that are part of the Interagency Monitoring of Protected Visual Environments (IMPROVE) network, Chemical Speciation Network (CSN), and the NCore Multipollutant Monitoring Network.
2. Download the relevant summary statistics for chemical species, along with the number of observations per calendar quarter for assessing completeness.
3. Require at least 11 valid samples for each quarter for each chemical species
4. Compute the PM2.5 chemical components (all units in μg/m3) from the species information following the table below:

|  |  |  |
| --- | --- | --- |
| **PM2.5 chemical component** | **Chemical species included** | **Parameter Code** |
| Sulfate | sulfate | 88403 |
| Nitrate | nitrate | 88306 |
| Organic Carbon | 1) TOR OC,2) (TOT OC-1.57432)/1.15101 | 88320 (TOR OC), 88370 (TOR OC), 88305 (TOT OC) |
| Elemental Carbon | 1) TOR EC,2) (TOT EC-0.104071)/0.92462 | 88321(TOR OC), 88380(TOR OC), 88307 (TOT EC) |
| Crustal | 2.2×Al+2.49×Si+1.63×Ca+2.42×Fe+1.94×Ti | 88104 (Al), 88165 (Si), 88111 (Ca), 88126 (Fe), 88161 (Ti) |
| Sea-salt | 1) Chloride×1.8,2) Chlorine×1.8 | 88115 (Chlorine), 88203 (Chloride) |

1. If one component is missing for the quarter, the stacked bar chart for that quarter is not shown for any component.
2. If one quarterly stacked bar chart is missing, the stacked bar chart of the annual averages for that year is not shown for any component.

### Visibility Data

#### Methodology for generating data for the visibility graphics

1. Download the latest Regional Haze Rule daily values summary file at http://vista.cira.colostate.edu/Improve/rhr-summary-data/
2. Remove years from the analysis that do not meet the Regional Haze Rule requirements for a complete year (All four quarters of that year should be at least 50% complete, and overall the year should be 75% complete. Additionally, there cannot be more than 10 missing sampling days in a row at any time during the calendar year.)
3. Identify the 20% clearest and 20% most impaired days for each complete year. The 20% clearest days are those with the lowest deciview values while the identification of the 20% most impaired days is described by the draft guidance document for the 2016 update to the Regional Haze Rule (https://www.epa.gov/sites/production/files/2016-07/documents/draft\_regional\_haze\_guidance\_july\_2016.pdf).
4. Identifying the 20% most impaired days requires some information about the estimated natural conditions of the site. For sites without this estimate, EPA used the natural conditions estimates for the nearest IMPROVE site with an estimate of natural conditions.
5. For national trends, selection of sites was limited to those with at least 75% valid years over the trend period (2000-2016). In addition, sites must not be missing more than 2 consecutive years of data to be considered a “trend site” or “national stats site” factored into the summary statistics (national average).

### Air Toxics Data

#### Overview and Ambient Data Used

Air toxics are measured across the country at the 27 National Air Toxics Trends Sites (NATTS) and hundreds of other sites operated by state, local, tribal agencies or other organizations that are not part of the NATTS network. The NATTS sites were created to generate long-term, quality assured, standardized ambient air toxics data to identify trends in air toxic concentrations, support model evaluation and other air toxics program needs. EPA used both NATTS and non-NATTS data to develop site specific and national aggregate trends to maximize the monitors used. Each point on the map provides a trend of the annual mean concentration between the years 2003 and 2016 for a site (NATTS or other) and selected air toxics pollutant. Many air toxics were monitored earlier than 2003, but 2003 was chosen since that was when most of the NATTS began operating. For some pollutants the trend years are different due to the lack of data. For example, the polycyclic aromatic hydrocarbon (PAH) trends start in 2008 and hexavalent chromium trends end in 2012 (though a few sites continued monitoring after 2012). Annual means were computed, as described below, from the daily or sub-daily concentrations provided in the Air Toxics Archive, Phase XII. The database is available at <https://www3.epa.gov/ttn/amtic/toxdat.html#data>; in addition to the data, it includes a data dictionary and technical memorandum documenting the compilation.

Most of the daily or sub-daily data in the Air Toxics Archive used to develop these trends were obtained from the Air Quality System; in some cases, where data were not reported to AQS, they were obtained directly from the organization collecting the data. A new source of air toxics data added to the Phase XII are the [National Oceanic and Atmospheric Administration (NOAA)’s Global Monitoring Division](https://www.esrl.noaa.gov/gmd/) measurements of trace gases at remote and regional background concentrations at monitoring stations across the country.

#### Computing Site Annual Means

The first step to compute the annual mean was to compute daily values. Most of the data used for trends utilize 24-hour averaging times, but sub-daily data are also used. Sub-daily or daily data from multiple monitors at the site were averaged if they used the same monitoring averaging time (e.g., 1 hour, 3 hour or 24-hour). If, however, multiple values for the same day included at least one value that was non-detect (ND), this averaging was not done. Instead, if 50% or less of the values were ND, the ND’s were removed such that only the non-ND values were averaged. If more than 50% of the values were ND, then the value for that day was considered ND. For the NOAA monitors, the 5-minute averages were treated as 24-hour averages because these monitors are located in remote areas or represent regional background that tends to represent the lowest level concentrations. Data not already in local conditions were converted to micrograms per cubic meter (ug/m3) in local conditions if temperature and pressure data were available. Otherwise they were converted to ug/m3 at standard conditions. The NOAA monitors were converted ug/m3 at standard temperature and pressure. One NATTS site, La Grande Oregon (410610119), moved in the middle of the year 2016 to a different location in the same county (site 410610123). Data from the two sites were combined and assigned to "410610119".

To compute an annual mean at a specific site, the daily data needed to meet 70% completeness.  Site-pollutant-year combinations that did not meet the completeness criteria were excluded from the trend. For 1-in-12-day data, at least 6 daily averages are needed in a quarter, for 1-in-6 day data, at least 11 were needed, and for 1-in-3 day data, at least 22 measurements were needed.  For the NOAA data, which was approximately 1-in-7 days, 9 measurements were needed. For all monitors, at least 3 complete quarters of data per year were required. For monitors that met the 70% completeness criteria, all days were used to compute the annual mean.

Data below MDL were used as-is. Data that were flagged as non detect (“ND”) that were reported as 0 were not used as-is, but rather were imputed via a Regression-on-Order Statistics (ROS) approach. The ROS Routine was used in the NADA R-package (<https://cran.r-project.org/web/packages/NADA/NADA.pdf> ) to estimate annual means. This is a semi-parametric procedure which imputes data that are below detection, or “censored” data, combines with the uncensored data, and computes the annual statistics. In this approach, a linear regression is formed using the plotting positions of the uncensored observations and their normal quantiles. This model is then used to estimate the concentration of the censored observations as a function of their normal quantiles. Because we did not have the detection limit below which data were flagged as “ND”, the minimum value reported for the year as the “censor” value was used.  This value is input into the NADA “cenros” function. Although method detection limits (MDLs) were available, they were only used as the “censor” value when below all other data reported for that site-pollutant in that year. An annual mean was not computed if more than 80% of the data were “ND”.

For some sites and years, there were sufficient data to compute two separate annual means reflecting sufficient data completeness for two different averaging times; for example, 1 hour and 24-hour data at the same site. If this occurred at NATTS sites, the annual mean from the 24-hour data was used since only the 24-hour data are associated with the NATTS program. At other (non-NATTS) sites, the annual mean for the site was computed as the average of the two separate annual means.

The below table shows the air toxics covered and the years included. EPA selected the air toxics required at the NATTS, other than acrolein. Acrolein does not have sufficient data for trends due to sampling/analysis issues. Additional air toxics that come from the same types of sources as the NATTS pollutants were included such as toluene and ethyl benzene. Except at NATTS, there is not a consistent set of air toxics pollutants measured at all sites that measure air toxics.

The table also shows the number of sites used for computing the national statistics (10th 50th and 90th percentiles) shown on the charts. The eligibility for a site to be used for the national statistics is described below.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **AQS\_****PARAMETER\_****CODE** | **AQS\_PARAMETER\_NAME** | **start year** | **end year** | **NATTS Required?a** | **Number of Sites Used for the national trend** |
| 14115 | Chromium VI (LC) | 2005 | 2012 |  | 20 |
| 17141 | Naphthalene (total tsp & vapor) | 2008 | 2016 | X | 27 |
| 17242 | Benzo(a)pyrene (total tsp & vapor) | 2008 | 2016 | X | 20 |
| 43218 | 1,3-Butadiene | 2003 | 2016 | X | 98 |
| 43502 | Formaldehyde | 2003 | 2016 | X | 79 |
| 43503 | Acetaldehyde | 2003 | 2016 | X | 80 |
| 43802 | Methylene chloride | 2003 | 2016 | X | 102 |
| 43803 | Chloroform | 2003 | 2016 | X | 107 |
| 43804 | Carbon tetrachloride | 2003 | 2016 | X | 112 |
| 43815 | Ethylene dichloride | 2004 | 2016 |  | 75 |
| 43817 | Tetrachloroethylene | 2003 | 2016 | X | 113 |
| 43824 | Trichloroethylene | 2003 | 2016 | X | 41 |
| 43860 | Vinyl chloride | 2003 | 2016 | X | 14 |
| 45201 | Benzene | 2003 | 2016 | X | 150 |
| 45202 | Toluene | 2003 | 2016 |  | 139 |
| 45203 | Ethylbenzene | 2003 | 2016 |  | 119 |
| 85103 | Arsenic Pm10 Lc | 2003 | 2016 | X | 20 |
| 85105 | Beryllium Pm10 Lc | 2004 | 2016 | X | 17 |
| 85110 | Cadmium Pm10 Lc | 2003 | 2016 | X | 16 |
| 85128 | Lead Pm10 Lc | 2003 | 2016 | X | 21 |
| 85132 | Manganese Pm10 Lc | 2003 | 2016 | X | 23 |
| 85136 | Nickel Pm10 Lc | 2003 | 2016 | X | 19 |
| a The requirement forChromium VI was discontinued after 2012 |

#### Determining Site Trends

Once the annual means were computed, the Spearman rank correlation and significance test was run by site and pollutant to determine the relationship between the mean annual concentration and the year. For this determination all available data between 2003 and 2016 were used. No missing years were gap filled. A p-value of less than or equal to 0.05 and a negative Spearman coefficient was considered a decreasing trend, a p-value of less than or equal to 0.05 and a positive Spearman coefficient was considered a decreasing trend, and a p-value greater than 0.05 was considered to have no trend (“none”).  A trend could not be determined for site-pollutant pairs with less than 4 years of data, and are therefore classified as “undetermined.” These sites may have had 80% or more ND in some years, and insufficient data to meet the completeness criteria in other years such that there were less than 4 years for which an annual mean could be computed. Three of the NATTS locations moved during the time period: Bronx, NY; a rural site in Wisconsin and a rural site in Kentucky. For these sites, the trend was determined based on combining annual means across the 2 sites as follows, as it was assumed the locations were capturing similar air sheds:

|  |  |  |
| --- | --- | --- |
| NATTS Site | Years | Trend based on Combination of |
| Grayson Lake, KY (210430500) | 2009-2015 | 210430500 & 211930003 |
| Hazard, KY (211930003) \* | 2003-2007 | 210430500 & 211930003 |
|  |
| Bronx1, NY (360050110) \* | 2003-2009, 2013-2015 | 360050080 & 360050110 |
| Bronx2, NY (360050080) | 2011 | 360050080 & 360050110 |
|  |
| Mayville, Wisconsin (550270007) \* | 2003-2009 | 550270007 & 550270001 |
| Horicon Wisconsin (550270001)  | 2010-2015 | 550270007 & 550270001 |
| \*This is the site id displayed on the map for the combined tend |

#### Aggregating Site Trends to Produce National Pollutant Trends

Individual site data were aggregated to produce a national aggregate trend by pollutant. For a site to be included in the aggregate trend (i.e., “national stats site”), it needed to have annual averages for at least 75% of the years. Most pollutants’ trends were 2003 to 2016, therefore, at least 11 years were required. The annual means for the NATTS sites that moved (as described above- Bronx, NY, WI and KY) were combined across years prior to applying the 75% criterion as it was assumed they were capturing a similar air shed. The combined sites allowed their use in the national aggregate trends; individually they would not have met the completeness criterion. In addition to the 75% completeness criterion, a site was required to have no more than 2 missing consecutive years. Missing years were filled in by interpolation. A missing end year used the year prior with no adjustment. If the begin-year was missing, an earlier year was interpolated, if available. If not, the begin year equal to the next year lowest year (e.g., the 2004 value would be used for 2003) was set. Once all missing data were filled in, the 10th, 90th and 50th percentiles by year and pollutant were computed.