APEX FormattR: User Guide

APEX Task 15

# Purpose

This program was created to bring greater accessibility to the APEX model for those that may have been previously unable to format the EPA Air Quality System (AQS) data due to lack of time, resources, or familiarity with a language able to efficiently format the data to be accepted by the APEX model.

Based on user inputs this tool fetches, formats, and fixes the AQS data to be readily accepted by APEX. The following sections provide greater detail on both the requirements and the processing incorporated into this tool.

# Using APEX FormattR

## Requirements

This tool requires the user to have R installed and accessible but does not require any other sort of interpreter to interact with R. Users can download R from [this website](http://archive.linux.duke.edu/cran/), or any of the other links on [this page](https://cran.r-project.org/mirrors.html).

This tool will require the user to input the location of the R executable, which can be easily found by searching “R.exe” in the “Start” menu, right clicking the program, and selecting “Open file location”. This file path can then be copied and pasted directly when prompted.

**Note: File paths should not contain spaces. Please replace any spaces in file names with hyphens or underscores before running the program.**

After installing R the program is ready to run!

1. Double click the “APEX\_FormattR\_runnR” batch file.
2. Input all file paths when prompted.
   1. Note: To paste into the command line, right click and select “Paste”. Keyboard commands will fail.
3. Define and save all user inputs required in the text file opened by the program.
4. Press any key to complete the program.

**Note: This program will install all dependent packages. Depending on the permissions of the user, this process may be limited by a third party. If you are experiencing permissions errors, and have administrator privileges, try running the file as an administrator by right clicking the file, and selecting “Run as administrator”.**

## User Defined Inputs

The user will be asked to provide the following:

1. **Data-year**: This field directs the program to the year(s) of interest when downloading the required data from the AQS website. Multiple years can be included by adding a new line in the same format (including variable name).
2. **Start-date**: This field should be filled in the “YYYY-MM-DD” format. This field must be enclosed by quotation marks.
3. **End-date**: This field should also be filled in the “YYYY-MM-DD” format and must be after the start date field. This field must be enclosed by quotation marks.
4. **FIPS-code**: Users should provide county FIP codes in this field. Multiple codes can be included by adding a new line in the same format (including variable name). This field must be enclosed by quotation marks.
5. **AQS-chem**: This field will direct the program to the correct AQS file. The current options are:
   1. "Ozone"
   2. "SO2"
   3. "CO"
   4. "NO2"
   5. "PM2.5 FRM/FEM"
   6. "PM2.5 non FRM/FEM"
   7. "PM10 Mass"
   8. "PM2.5 Speciation"
   9. "PM10 Speciation"
   10. "HAPs"
   11. "VOCs"
   12. "NONOxNOy"
   13. "Lead"

Availabiliity of data varies by chemical and year. Users should verify that the data exists on AQS prior to entering the data for download. This field must be enclosed by quotation marks.

1. **Chemical**: The chemical of interest within the AQS file must be included. For file types j through l above, there may be several chemical options. Users should specify which chemical of interest will be assessed. For all other chemical options the “Chemical” field will be the same text as is entered under “AQS-chem”.

Currently the options for “HAPs”, “VOCs”, and “NONOxNOy” are:

1. HAPs
   1. “1,3-Butadiene”
   2. “Benzene
2. VOCs
   1. “Total NMOC (non-methane organic compound)”
3. NONOx
   1. “Nitric oxide (NO)”
   2. “Oxides of nitrogen (NOx)”
   3. “Reactive oxides of nitrogen (NOy)”

**Note: As of August, 2019 the AQS data files for “Lead” are empty. Users should visit the AQS website to verify there is data for the chosen chemical before proceeding with the use of this tool**.

1. **MaxFrac**: Users should define the maximum allowable fraction of missing data (in decimal form). Data will be evaluated prior to processing to determine if it meets this criteria. For example, if MaxFrac = 0, data with any amount of missing data will be processed. If MaxFrac = 1.0, only complete sets of data will be processed.
2. **MaxD**: This field allows the user to define how far away from their FIPS, data should be allowed to be collected from. As described in the “Process” section of this document, missing data may be filled with data from the next nearest station with data available for a given hour. If there is a situation in which there is no data for a given hour within the bounds of the study area, sites beyond that threshold will be searched, prioritizing the nearest station. This search is limited to 3 times the user defined distance, or 300 kilometers, whichever is less.
3. **MaxCumMiss**: The maximum cumulative missing data variable defines the threshold between what the user would like to consider “short” gaps of missing data, and “large. The processing for each is described in the “Process” section of this document. Based on the diurnal cycle of many air pollutants it is recommended that this threshold be no greater than 12 hours.

**Note: Spelling and capitalization are important in this instance. Additionally, all non-numeric user defined inputs (including dates) should be enclosed in quotation marks (ex. “Ozone”), as seen above.**

# Process

This tool opens a session of R in the command line to run the R scripts without the user having to interact with the script directly. The user is prompted to fill in a text file with the desired inputs, which the R script will then pull and utilize.

## Overview

The user should start by opening the APEX\_FormattR\_runnR.bat file. This will function as the user interface. Users will be prompted to provide file paths

## Guide to the Scripts

The body of this tool is broken into five scripts, which function as described below. The scripts are listed in order of execution with each script being chained to the others as they are needed.

**Apex\_Util\_ControlStart**

This script functions as the link between the batch file and the main body of the processing. It holds the functions which, install any necessary packages, download the data, and import the user inputs. This script sets up the data inputs for the “Main” script, which is called at the end of the script.

**Apex\_Util\_Main**

This script compiles all of the functions and data entered into the tool to be called by the rest of the scripts. Each of the remaining scripts act as a function, which are called in this script. First, the data is processed, to prepare it for the rest of the processing. This happens in the “PreprocessData” script.

After the data is processed into a format ready for blank values to be replaced, the data-fixing function within the “FixData” script is called.

**Apex\_Util\_PreprocessData**

R reads in and filters the site description information for that information relevant to the user provided FIPS. The distances between the sites in the provided FIPS code(s) and the neighboring sites (up to three times the user set maximum distance or 300km, whichever is less) are calculated for future use in the data fixes. Monitors in Canada and Mexico are excluded regardless of distance from the sites of interest.

The chemical data is read in and filtered for the chemical of interest after the site distances are calculated. Here the data is also converted into the appropriate formats for use in the rest of the scripts. At this point in the script all negative pollutant measurement values are replaced with 0. For any missing hours or days of pollutant concentrations are created and filled with NAs, to be filled in later when the data is “fixed”. Each hour of each day at each station must have a pollutant concentration to be accepted by APEX.

**Apex\_Util\_FixData**

The “fixdata” script defines several functions, which work in conjunction to replace any blank (“NA”) values in the data with an estimate of the pollutant concentration at that day and time. The user defines the number of consecutive missing hours of data that should considered a “small” data gap, to be fixed with linear interpolation. The process of linear interpolation involves estimating the value of missing data by solving for the required value using the equation of the line connecting the two points on either side of the missing value.

The gap threshold can be set to 0 for all of the data to be fixed with data from nearby stations. The maximum recommended threshold is 12 hours, to account for any diurnal patterning of pollutant concentrations.

Gaps in the data greater than the user defined threshold (“large”) are filled in with the corresponding data from the closest station with available data to provide. Site data is filled in with only the closest available data, so gaps may be filled in with data from multiple stations.

For those cases where the first value of the data is missing and the gap of missing data is considered “small”, the values are interpolated between the first available measurement and the last available measurement of the same day. In essence this assumes that the data from the day prior was the same as the start of the data, and interpolates the values that are missing. For cases where the last hours of the day are missing, the same process is used to interpolate between the last value of the day and the first value of the “next” day (assumed).

The first pass of data fixing is performed according to user specifications. If there are any remaining blank values the data (due to large data gaps that did not have data across any of the sites within the user defined radius of interest), the data fixing process will be repeated with a radius equal to three times the user defined distance up to 300km. This allows for any large gaps that may have become “small” to be passed through linear interpolation and any remaining “large” gaps a chance to be filled with data not previously preferred.

After these fixes are applied, any stations with remaining blank values are excluded from the final output.

**Note: Lower thresholds separating “small” and “large” gaps will increase the run time of the tool.**

**Apex\_Util\_FormatDataForOutput**

Finally, the fixed data is formatted for use in APEX. Among other things, APEX requires the input to read the day’s data horizontally rather than vertically, as it comes out of AQS and is worked with in this tool.

**Apex\_Util\_PrintMetaData**

The last script called by the “Main” script, collects and formats descriptive data for printing in the metadata excel spreadsheet. Summaries of each station are provided separately in individual sheets of the excel spreadsheet.

# Final results

The final resulting product of the tool outlined in this document is are the following files of data:

1. The data specified data formatted for immediate use in APEX.
2. A metadata file, which describes the data to be input into APEX. This file provides the following information for each station:
   1. Data pulled from AQS: The user should verify that the data processed was correct before moving forward with its use in APEX.
   2. Chemical: This will print the chemical of interest as input by the user.
   3. Dates included: This will return the dates requested for processing by the user.
   4. Fraction missing data: This value is the amount of data actually missing from the dataset.
   5. A summary of the pollutant concentrations.
3. A log file of R warnings and errors. Users should review this file prior to use of the data to evaluate any warnings or in the case of a failed run.

# User notes

Because of the flexibility of the tool, runtimes may vary based on several factors including but not limited to: specific year of AQS data requested, MaxCumMiss threshold, site density in a given FIPS code, or length of time requested. External limitations may also apply (such as, user’s RAM availability, computer or program bit availability etc.). If users encounter issues processing the amount of data requested, it may be beneficial to reevaluate the breadth of the parameters defined and/or run the data in batches.