Removal of Smoke Events from Particulate Background Concentration Data

Wildfires and other smoke events have been on the rise in recent years. These smoke events cause an increase in the concentrations that are measured by ambient monitors and thereby affect the background concentrations that are used in dispersion modeling analyses. This document describes the method used by the Iowa DNR to identify these extreme smoke events and remove their influence from the data used to calculate background concentrations.

Identification of Smoke Events

The DNR uses a two-step approach to identifying days that will be excluded from background data used in dispersion modeling.

1. Identify outliers in the ambient monitoring data
2. Correlate the outliers with smoke observations

Identifying Outliers

Traditionally, outliers in a data set are often determined using a certain number of standard deviations from the mean. However, both the mean and the standard deviation are heavily influenced by extreme outliers, like those from smoke events. An alternate approach is to use the Median Absolute Deviation (MAD). This technique is less influenced by outliers, making it a more ideal measure for identifying smoke events in the ambient monitoring data. MAD is calculated by finding the median of the absolute values of the differences between the data set median and each data point. In the case of heavily skewed data, a double-MAD method can be used to identify outliers using separate scales that are specific to each side of the distribution[1]. In the case of ambient monitoring data, where we are interested in finding high concentrations caused by smoke events, we only need to calculate the MAD of those values greater than or equal to the median of the ambient data.

Once the MAD has been determined for a data set it is multiplied by 1.4826, a constant linked to the assumption of normality of the data, disregarding the abnormality induced by outliers[2]. Finally, an outlier threshold is determined by choosing the acceptable number of deviations from the median. This decision is somewhat subjective and commonly ranges between 2-3 times the MAD. The DNR chose to use two deviations, which results in a lower outlier threshold. This method was chosen because the threshold itself is not the only criterion that will be used to determine if each data point will actually be treated as an outlier, allowing for the identification of less extreme smoke events.

Correlation of Data

Smoke data was retrieved from NOAA's Hazard Mapping System (HMS)[3] for every day of the period being evaluated. An initial cursory review of each day was conducted to determine if a smoke plume was present over any part of the state on each day. For days where smoke was present somewhere in the state a more detailed review was conducted for each ambient monitor location. Each day with a smoke plume present over a specific monitor was recorded for use in the next step.

Specific days at each monitor were flagged as a smoke event if both: 1) the observed concentration exceeded the outlier threshold, and 2) a smoke plume was present at that monitor, or if it occurred on the 4th of July (fireworks). Using these criteria in tandem is important because high concentrations can occur in the absence of a smoke event, and because the smoke data only indicates that a smoke plume was observed in the column of air above the monitor, not that it was actually impacting the monitor at the surface. If both criteria are true, it is likely that the observed data are in fact

[1] https://aakinshin.net/posts/harrell-davis-double-mad-outlier-detector/

[2] Leys, C., et al., Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median, Journal of Experimental Social Psychology (2013), http://dx.doi.org/10.1016/j.jesp.2013.03.013

[3] https://www.ospo.noaa.gov/Products/land/hms.html

caused by the smoke event. Each day that was flagged as a smoke event was then filtered out of the data for each specific monitor. This resulted in the removal of approximately 4% of the days for $PM_{2.5}$ and 3% for PM_{10}, on average.

Evaluation of the Results

After filtering out the smoke events we compared the distribution of the raw data with the filtered data by evaluating the skewness and excess kurtosis of each data set (see Table 1). Skewness and kurtosis are descriptive statistics that provide insight about how closely a data set resembles a normal distribution.

Skewness indicates if the data are more heavily weighted to one side of the distribution. A negative skewness indicates that the data are more heavily weighted to the left, and a positive skewness indicates the data are more heavily weighted to the right. A symmetrical distribution, such as a normal distribution, will have a skewness equal to zero. We can expect ambient monitoring data impacted by smoke events to be skewed to the right because the smoke events create outliers on the right.

Kurtosis provides a measure of the shape of the tails of the distribution as compared to a normal distribution. The kurtosis of a normal distribution is 3. Excess kurtosis is kurtosis minus 3, resulting in a normal distribution having an excess kurtosis of zero. A negative excess kurtosis indicates fewer and less extreme outliers in the data than a normal distribution, and a positive excess kurtosis indicates more outliers than a normal distribution. We can expect ambient monitoring data impacted by smoke events to have a positive excess kurtosis because the smoke events create outliers.

Table 1. Summary Statistics for Raw and Filtered Ambient Monitoring Data

		Raw Data		**Filtered Data**	
		Average	Range	Average	Range
PM2.5	Skewness	5.1	2.1-8.1	1.3	1-1.6
	Excess Kurtosis	52.1	8.4-114.9	3.1	2-5.7
PM10	Skewness	1.6	1.1-2.3	1.2	0.9-1.8
	Excess Kurtosis	5.4	1.9-11.4	4.3	2-9.8

The filtered data is still skewed toward higher concentrations, but by a smaller degree. These results seem reasonable because the data are bound on the left by zero (concentrations cannot be negative), and high concentrations are sometimes observed in the absence of smoke events. Filtering the data had a greater impact on $PM_{2.5}$ than it did PM_{10}, which makes sense because smaller particles are more likely to stay suspended long enough to impact the monitors. Outliers remain in both the $PM_{2.5}$ and PM_{10} data sets, but the number and magnitude of them has been dramatically reduced. This is to be expected because we are specifically targeting only the outliers caused by smoke events.