This document is based on data that predates the current background concentrations provided by the DNR. While the data used is historical, the methodologies, analyses, and conclusions presented within this document remain valid and are considered reliable.

Removal of Smoke Events from Particulate Background Concentration Data

Smoke events cause an increase in background concentrations used in dispersion modeling analyses that are not representative of normal conditions. In their April 2019 memo "Additional Methods, Determinations, and Analyses to Modify Air Quality Data Beyond Exceptional Events" EPA suggests that it may be appropriate to modify monitoring data for use in modeling analyses where the data are not representative to characterize background concentrations. 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 the 2021-2023 background concentrations.

Note: This method does not apply to the calculation of AQI or monitor design values

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 a consistency constant. Most commonly, 1.4826 is used, which makes MAD a consistent estimator of the standard deviation in normal distributions[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 it allows for the identification of less extreme smoke events and because the threshold itself is not the only criterion that will be used to determine if each data point will actually be removed.

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. This is not to say that the ambient data should be distributed normally, but rather: a normal distribution provides a consistent metric to compare both the filtered and unfiltered data to.

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

		Average		Range	
		Raw Data	**Filtered Data**	**Raw Data**	**Filtered Data**
Skewness	$PM_{2.5}$	5.1	1.3	2.1-8.1	1-1.6
	PM_{10}	1.6	1.2	1.1-2.3	0.9-1.8
Excess Kurtosis	$PM_{2.5}$	52.1	3.1	8.4-114.9	2-5.7
	PM_{10}	5.4	4.3	1.9-11.4	2-9.8

Conclusion

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 from distant wildfires are more likely to stay suspended long enough to impact the monitors. The number and magnitude of outliers has been dramatically reduced, but some remain in both the $PM_{2.5}$ and PM_{10} data sets. This is to be expected because we are specifically targeting only the outliers caused by smoke events.

Applying this method to the last six years results in average background concentrations that better correlate with industrial emissions trends during those years. At this time the proposed method appears to provide a more representative estimate of what background concentrations should be in the absence of extreme events like wildfires. However, the DNR will continue to monitor the prevalence of smoke events and will revisit this method for adjusting the background as needed.