Notes on Processing Pollutant Data

This note gives a brief description of how the pollutant data series were constructed from the raw monitor data.

1. Each city has associated with it a number of monitors for a given pollutant. The possible pollutants are PM_{10}, $PM_{2.5}$, SO_2, O_3, NO_2, and CO.

2. Let $X^c_{j,t}$ be the raw pollutant value for monitor j in city c at time/day t. The detrended value $\tilde{X}^c_{j,t}$ is defined as

$$\begin{aligned} M^c_{j,t} &= \frac{1}{365} \sum_{\ell=-182}^{182} X^c_{j,t-\ell} \\ \tilde{X}^c_{j,t} &= X^c_{j,t} - M^c_{j,t} \end{aligned}$$

   The values $\tilde{X}^c_{j,t}$ are the detrended "residuals" from the raw pollutant series.

3. If a city only has one monitor, then the series $\tilde{X}^c_{j,t}$ is the final result and is used for analysis.

4. If a city has 2 monitors, then a final $\bar{X}^c_t$ is computed for each time point t as

$$\bar{X}^c_t = \frac{1}{2}\left(\tilde{X}^c_{1,t} + \tilde{X}^c_{2,t}\right)$$

   If a city has more than 2 monitors, then a 10% trimmed mean of the $\tilde{X}^c_{j,t}$'s is taken for each day. That is, if there are J monitors in a city, then for each timepoint t

$$\bar{X}^c_t = \text{TrimmedMean}_{10\%}\left[\tilde{X}^c_{1,t}, \ldots, \tilde{X}^c_{J,t}\right]$$

   If there are fewer than 10 monitors, the lowest and highest values for each day are still always discarded.

   One can see now why the detrending must be done first in Step 2. If a particular monitor has a higher overall level, then it will consistently be discarded when the trimmed mean is taken.

5. The series $\bar{X}^c_t$ is used as the pollutant measurement for city c on day t. In each city dataframe, this series is given the name `*tmean` where "`*`" is either `pm10`, `pm25`, `so2`, `o3`, `no2`, or `co`.

6. The median of the 365-day moving averages are also computed, that is

$$\bar{M}^c_t = \text{Median}\left[M^c_{1,t}, \ldots, M^c_{J,t}\right]$$

   This series is given the name `*mtrend` in each dataframe where "`*`" is the name of a pollutant.

The dataframes do not contain the original monitor data, but if one wishes to examine a series that is reminiscent of a true pollutant series, one can add the `*tmean` series to the `*mtrend` series. For example, to construct a PM_{10} series, one can add the `pm10tmean` and `pm10mtrend` variables.