An Introduction to PM2.5s, their Importance, and a Cluster Methodology to Analyze their Meteorological Dynamics

Rickie Xian, MSc Candidate, r4xian@uwaterloo.ca; Dr. Dylan Jones

Introduction:

The influence of human activity on the earth’s atmospheric composition has never been more pronounced. Anthropogenic pollution is in fact the largest effector of the observed evolving atmospheric composition (Wallace, 2006). PM2.5 is a class of particulate matter pollutants of notable interest due to their significant driving of chemical, atmospheric change, their wide-scale, global circulations, and their malignant effects on human health. PM2.5 are particulate matter pollutants with a diameter of less than 2.5 microns; their source is largely combustion of organic materials, including most notably fossil fuel combustion necessary to power our automobiles, homes, and industrial processes (Wallace, 2006), as well as the burning of plant matter in forest fires (Newman, 2007). The gases released in these combustion reactions then condense in the atmosphere, undergoing gas to particle conversion, and resulting in the atmospheric presence of PM2.5s. Particulate matter (PM) pollutants are harmful to human health in all diameter scales (Araujo, 2011). They’re largely responsible for the overall global recent increase in human morbidity and mortality: virtually all nations have seen an increase in mortality rates, which can be attributed to the increased presence of atmospheric particulate matter pollutants over the past few decades (Araujo, 2011).

The health risks of PM2.5 in particular are troubling due to their small size, which facilitates their permeability in the respiratory system and ready diffusion into the bloodstream. The mechanism by which PM2.5 induces pathologies involves free radical reactions at the cellular level, inducing cellular oxidative stress, and triggering a variety of disease, cascade reactions (Zhang, 2013). The atmospheric presence of PM2.5 has been shown to increase the risk of
developing many serious respiratory and cardiovascular pathologies, including but not limited to ischaemic heart disease, respiratory infections, and lung cancers to name a few (Araujo, 2011). Once PM2.5 manifest in the atmosphere, they circulate on a larger scale due to atmospheric circulation patterns. An easterly travelling, meandering airflow from alternating high and low pressure centres, transports PM2.5s on a global scale, resulting in particulate matter from Asia transported to the Americas, for example (Wallace, 2006). In addition to the easterly traveling airflow, circulations around pressure centres also arise, with clockwise circulations taking place around high pressure centres, and counter clockwise circulations taking place around low pressure centres (Wallace, 2006). Geopotential heights are also closely related to pressure surfaces, wherein the wind fields flow parallel to surfaces of constant Geopotential height surfaces (Wallace, 2006). These circulations all influence the formation and transportation of PM2.5s, due to their slow gas to particle conversion which are on timescales comparable to the aforementioned atmospheric circulations (Wallace, 2006).

Though government-enacted air quality measures have reduced the average PM2.5 levels in North America, pollution episodes still cause localized, acute PM2.5 exposure. The purpose of this project was to analyze PM2.5 mean concentration across America to identify and quantify any pollution episodes, as well as try to explain their dynamics using large scale, meteorological processes. What characterizes the nature of these PM2.5 pollution episodes? Do any pollution episodes occur annually on a consistent basis across regions? Which meteorological events influence the observed national distribution of pollution episodes in America?

**Methods:**

Efforts were taken to analyze Air Quality Systems (AQS) MDA (mean daily average) data for PM2.5 from various stations across USA to identity extreme pollution episodes and investigate the large-scale weather patterns which may be responsible for their dynamics. The AQS PM2.5
data was analyzed using cluster analysis consisting of seven steps: AQS file processing, subseasonal filtering, station-level extreme threshold calculation, cluster analysis, cluster quality metric examination, PM2.5 episode identification, and finally meteorological analysis. In terms of AQS file processing, PM2.5 data from 1999-2019 was downloaded, sequentially read, then appended into a single .csv file. Following file processing, subseasonal filtering was performed to filter out slow processes with time scales longer than the duration of a single season, which is defined as 90 days. Before a subseasonal filter can be applied, the original data must be processed to ensure continuity such that any seasonal gaps in data are filled in with randomly sampled data points from the same season. The discontinuities were resolved with points from the same season to avoid for artifacts of incorrectly interpolating results from inappropriate seasons in a single season, which would ultimately compromise the accuracy of the data. Once the data is made continuous, it can be passed through a low pass filter to subtract the slower processes with timescales longer than 90 days from faster, subseasonal processes with timescales shorter than 90 days. The subseasonal processes include day-to-day weather variations, as well as synoptic weather systems, such as cold fronts and warm fronts, connected to longer time scale atmospheric patterns.

Proceeding subseasonal filtering, extreme PM2.5 emission days were identified using station-level extreme threshold calculation. A rolling 95th percentile was calculated across all stations to sufficiently capture the magnitude of the variable standard deviations as the time series evolved such that anomalies above the 95th percentile were properly identified at each time step. These anomaly calculations were done for each site in parallel to identify each site’s extreme values for each day. The extreme anomalies were then grouped together into a larger data set for the next portion of data processing: cluster analysis and cluster quality metric examination. The main goal of cluster analysis was to combine extreme data from separate sites into larger regions to observe any similarities across larger number of stations. Because the weather systems of interest had length sales of 10^3 km, larger scale clusters of similar
length scales were expected to yield data more telling of the role of large-scale weather systems in these extreme pollution episodes. Jaccard distance calculation was done for every possible pair of sites wherein the higher the amount of overlapping extreme days, the more similarities there are between the two sites. Thus, stations were gradually grouped together by similarity of the specific day extreme PM2.5 pollution episodes occurred.

Following this step, the clustered data was then ready for cluster quality metric analysis. The Calinski-Harabaz Index (C-H score) was calculated for the clustered data, wherein this C-H score is directly proportional to the sum of the squares of the Jaccard distances between groups (BG), and inversely proportional to the same of the squares of the Jaccard distance within members of the same group (WG) (Wen, 2017). Because members within a group are desired to be similar (lower WG) and members between groups are desired to be different (Higher BG), a higher CH score corresponds to better separation between group characteristics. Plotting CH score versus the total number of clusters, any local maxima are candidates for interesting changes to cluster patterns. Data resulting from CH scores of 23, 34, and 45 were found and displayed in the resulting regions calculated from the data. Different dataweightmin parameters were also tested, wherein the parameter corresponds to the minimum proportion of observed, not randomly filled in, data a station must have to be counted in cluster analysis. Once the stations were grouped into larger clusters, PM2.5 episode identification was executed. The 95th percentile across all 21 years for all stations on each day were compared to reveal any seasonal consistencies in episodic weather patterns. Pollution episodes were classified as any three consecutive days with PM2.5 extreme pollution episodes. These pollution episodes were then submitted for meteorological analysis, in which selected regions from total cluster numbers 23 and 45 of these episodes were submitted for geopotential high anomaly calculations, to try to elucidate some of the mechanisms behind the observed PM2.5 pollution episode distribution using the large-scale circulations mentioned in the introduction.
Results and Discussion:
45 Total Clusters:
23 Total Clusters:

The included plots for the California time series of both the 21 year mean, as well as the 21 year summer mean demonstrates the overall decreasing mean PM2.5 concentration as expected, as well as the erratic variability resulting from pollution episodes and forest fires.

Looking at the subseasonal filtering figure, the continuous time series of the data, the black line, is passed through the orange line, the low pass filter. This results in the blue line of station-level extreme threshold calculation containing only subseasonal processes, which is the original data minus the seasonal processes. The red line indicates the rolling 95th percentile, wherein the red dots above the 95th percentile are classified as PM2.5 anomalies. It should be noted that the years of 2001-2003 and 2018 did not have enough data points to satisfy the minimum threshold of the 95th percentile window, thus anomaly calculations couldn’t be done for these four years. Following cluster analysis and cluster quality metric examination, the local maxima in the CH score versus total cluster number can be seen at total cluster numbers 23, 34, and 45.

For a total cluster number of 23 and dataweightmin of 0.8, four distinct regions are visualized. A cluster number of 34 yields six regions with dataweightmin 0.8, and a cluster number of 45 total clusters with dataweightmin 0.5 yields eight regions. What’s important to note is the higher the number of total clusters, the higher the number of regions. As well, regions in the central US seem to be less robustly defined due to scarcity of available observational data.
Looking at PM2.5 episode identification, only years consisting of enough data to again fulfill the 95th percentile window were included. The visualization demonstrates episode identification done on a total cluster number of 45 with dataweightmin 0.5, with episode analysis run for each region in the cluster plots. A distinction of episodic analysis to highlight is the increased data density of the Eastern and Northwestern regions, displaying pollution episodes across the entire season. The Southwestern and Northeastern regions didn’t have as high data density, but do have a consistent pollution episode across almost all years in the beginning of July, which may be attributed to the influx of particulate matter from July 4th fireworks.

Finally, looking at the geopotential height anomaly figures, we see clear correlations between positive geopotential height anomalies (the purple region) and pollution episodes. Geopotential is roughly defined as the work that must be done against Earth’s gravitational field to raise a parcel of air from sea level to some Geopotential height (Wallace, 2006). This height approximates the actual height of a pressure surface above mean sea level (Wallace, 2006). Purple regions indicate positive (higher than usual) Geopotential height anomalies, and green regions indicate negative (lower than usual) Geopotential height anomalies. From the included analysis of selected regions from 45 total clusters, an alternating pattern from west to east of low and high Geopotential anomalies is observed, which corresponds to the alternating high and low pressure centre patterns observed in large-scale airflow. As well, the pollution episodes all fall into positive anomalies, which corresponds to pressure surfaces higher than what is expected.

For the 23 cluster number analysis, there still exists an observed positive correlation between pollution episodes and positive Geopotential height anomalies. As expected, pollution episodes are associated with higher pressure surfaces, and the Northeastern region conserves the high-low anomaly pattern in Geopotential height. It should be noted that for the West Coast
Pacific Northwest region, the observed alternating high-low anomaly pattern is slightly smeared and less defined. This is thought to be due to lower data density in this region, as well as its wider span, which blurs characteristics between clusters and thus the clearly defined alternating pattern. Positive Geopotential height anomalies indicate less dense, warmer pressure surfaces with higher altitudes than expected (Wallace, 2006). The increased temperatures and pressures of these air masses not only make sense in the context of the greenhouse effects of these pollution episodes, but also in the rapid advection and circulation of these pollutants, as hotter air circulates faster and more turbulently than cooler air (Wallace, 2006). As well, descent or subsidence is created directly under high pressure systems, which traps PM2.5 in the boundary layer, increasing the concentration of these particulates and allowing for the observation of a pollution episode (Wallace, 2006).

Conclusion:

Thus in conclusion, the cluster analysis, episode identification, and meteorological analysis efforts have shown to reveal hidden dynamical mechanisms of PM2.5 pollution episodes in America, revealing a correlation between positive Geopotential height anomalies and thus higher pressure surfaces and pollution episodes. This high pressure system and large scale pattern is similar to what was seen in ozone pollution episode analysis. The high-low Geopotential height anomaly and thus pressure surface is consistent with what’s seen on a large scale in terms of the location of alternating high and low pressure centres and the overall meandering wind pattern. These large-scale patterns in geopotential height anomalies might offer a means of predicting PM2.5 pollution episodes. The means by which meteorology can be used to predict PM2.5 dynamics, distribution, and pollution episodes spurs further work on what exactly drives the behaviour of these dangerous particulate matter pollutants.
References:

Araujo, Jesus A. "Particulate air pollution, systemic oxidative stress, inflammation, and atherosclerosis." *NCBI*, vol. 1, 4 Mar. 2011, pp. 79-93, doi:10.1007/s11869-010-0101-8.

Cohen, Arthur, et al. "Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: an analysis of data from the Global Burden of Diseases Study 2015." *Lancet*, vol. 389, no. 1008, 19 Apr. 2018, pp. 1907-18, doi:https://doi.org/10.1016/S0140-6736(17)30505-6.

Hossen, Bipul. *A comparative study of Clustering for Gene expression data in Bioinformatics*, 5 Dec. 2013, www.slideshare.net/mbipu/a-comparative-study-of-clustering-for-gene-expression-data-in-bioinformatics.

Newman, John. "Forest fire." *Britannica*, 2007, www.britannica.com/science/forest-fire.

Staff, RT. "AIR POLLUTION INCREASES COPD RISK, SPEEDS AGING PROCESS." *RT Magazine*, 30 Sept. 2019, www.rtmagazine.com/disorders-diseases/chronic-pulmonary-disorders/copd/air-pollution-increases-copd-risk-speeds-aging-process/.

Wallace, John M., and Peter V. Hobbs. *Atmospheric Science An Introductory Survey*. Vol. 92, Elsevier Inc, 2006, pp. 14-184.

Wen, xi & Yang, Sizhong & Horn, Fabian & Winkel, Matthias & Wagner, Dirk & Liebner, Susanne. (2017). Global Biogeographic Analysis of Methanogenic Archaea Identifies Community-Shaping Environmental Factors of Natural Environments. *Frontiers in Microbiology*. 8. 1339. 10.3389/fmicb.2017.01339.
Zhang, HH, et al. "Physical and chemical characteristics of PM2.5 and its toxicity to human bronchial cells BEAS-2B in the winter and summer." *NCBI*, vol. 4, 5 Dec. 2013, pp. 317-26, doi:10.1631/jzus.B1700123.