Data Analytics
Fall 2019
Oleg Yakovets

K-Means Analysis of Pollutants and Geographic Influence: Southern Albany

Abstract

In order to study the influence of topography coupled with wind speed and direction on pollution rates, a PEJA located in the South End of Albany, NY was chosen for study due to a moderately large hill and measuring station in the neighborhood. The data for the clustering study was collected from a year-long, hourly-averaged data set provided by the New York State Department of Environmental Conservation on a public access website, with the date interval

ranging from 24 Sept. 2018 to 24 Sept 2019. The data comes from the measuring station located in the Ezra Prentice neighborhood as part of a network of statewide measuring stations. The main method of processing the directional correlation was K-Means Clustering. While the Nanoparticle distributions proved there is good mixing in the region, the K-Means clustering algorithm, using 15 clusters per chart, pointed towards a Northeastern correlation for origin of PM 2.5 and Black Carbon particles. However, further study is required in order to test for topography with a hillside with multiple measuring stations.

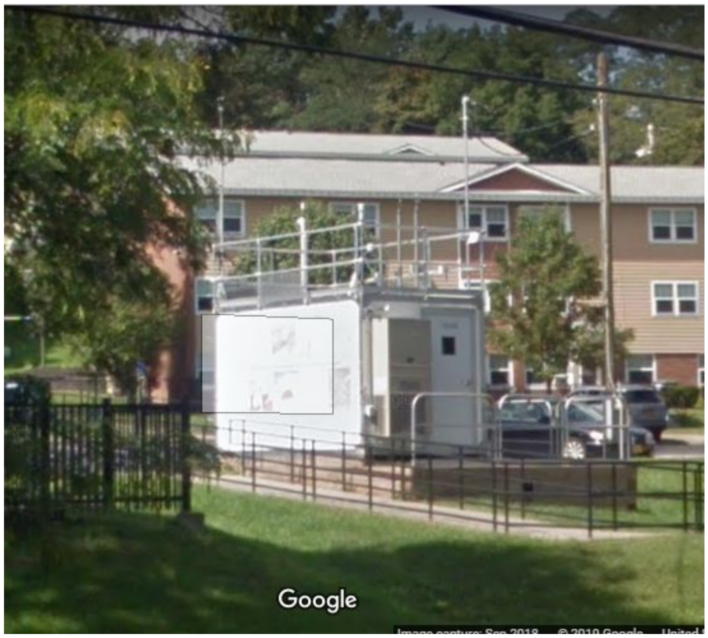
Introduction

Air pollution is a major concern for human health and environmental quality. The burning of fossil fuels releases not only CO_2 and SO_X, but also particles of various sizes. These particles are classified from the perspective of how far they go into the human lung, as well as how easily they are absorbed into the bloodstream. Their formation is often tied to incomplete combustion, especially Black Carbon, or soot[1]. Nanoparticles are linked quite closely to pulmonary diseases[2], while larger particles on the scale of 10μm will only get trapped in the oropharynx. Particles that are smaller than 5μm start to invade alveoli[3].

While the EPA currently regulates particles on the scale of 2.5μm [4], regulations have been increasing over the years and it is important for government agencies to begin collecting data on health effects and sources of these pollutants before legislation is passed in order to prepare for setting regulation standards, as well as what to expect for emissions from different sources.

A particular neighborhood in the South End of Albany was investigated as a Potential Environmental Justice Area(PEJA). A PEJA is an area where lower socioeconomic groups may experience disproportionately more pollution, and are investigated for possible sources of pollutants. The particular community, named Erza Prentice, is located on South Pearl St between Mt. Hope Drive and the Center for Disabilities Services bussing parking lot, and goes up Mount Hope Ave towards the West(Figure 1)(Appendix 1). The investigation would be conducted, at least in part, to see if emissions of the area were up to State and Federal regulations, but also compared to surrounding areas with much more green space and less traffic.

The surrounding area has a lot of industry, including waste management, shipping, oil distribution, and metals recycling. The main roadway through the neighborhood is a designated highway access for trucks onto Interstate 787, which can be classified as High Emitting Vehicles. The combination of industry and major road access makes for high traffic of multi-axle vehicles, as well as high traffic of cars, is cause for concern. There is a fixed monitoring station that is close to S. Pearl St, as well as most of the neighborhood being located on a hillside. There is already work being conducted on urban environments, and looking at how buildings and wind have an impact on pollution dissipation[5]. The work presented in this report will focus on exploring if there is any correlation between wind direction at the fixed monitoring station in the Ezra Prentice neighborhood and the measured pollutants, as well as any possible influence on the measurements from the hill in the neighborhood.



The measuring station.

Figure 1: Image of the fixed monitoring station and outline of the target area. North is denoted by the black arrow, and the majority of the pollutant origin direction is denoted by the pink arrow.

Data Description

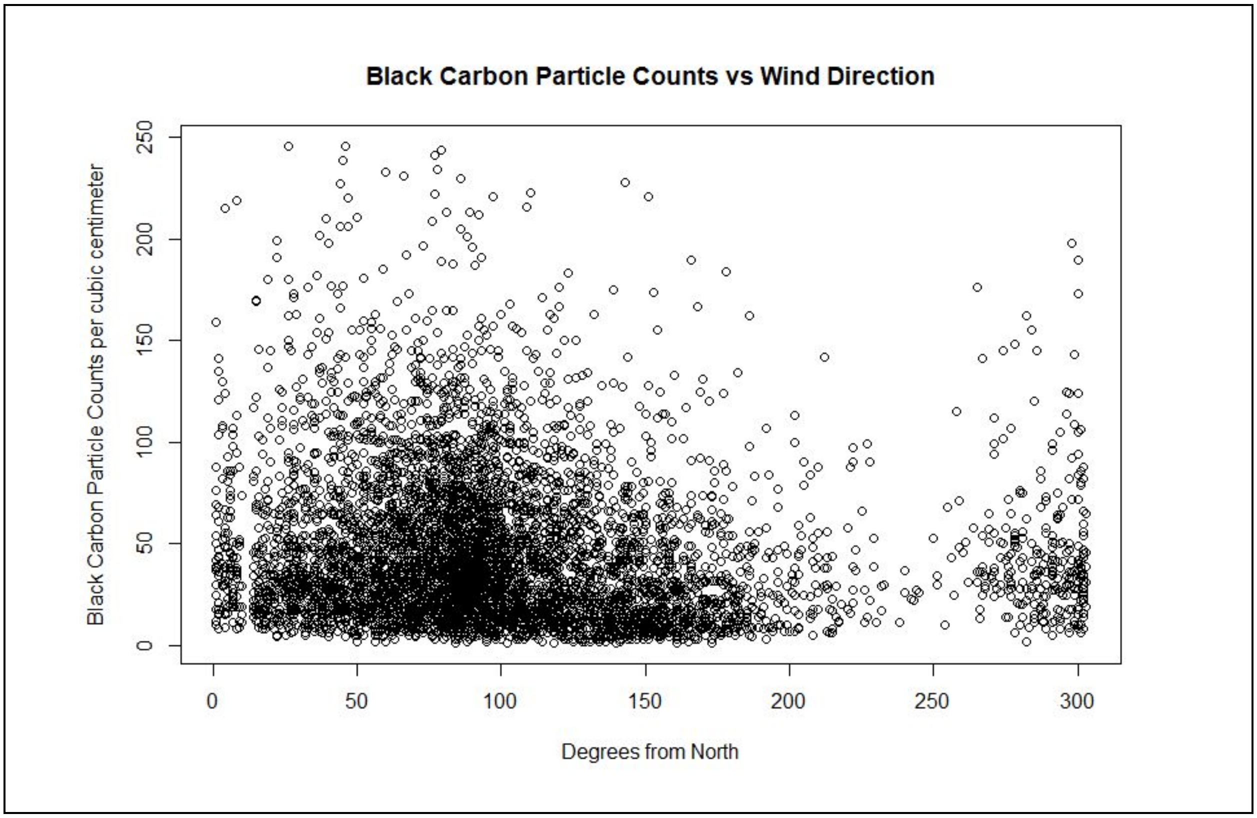
The data set used for the project was eventually decided upon using a public, 1-hr averaged data set the DEC provides on a website associated with fixed monitoring stations throughout the state of New York[7]. Common practices of analyzing include time-based averages to get rid of noise, so the data set was deemed acceptable. The date interval selected was from September 24th, 2018 to September 24th 2019. While data was collected since 2015, most year-long files were sparsely populated until 2018.

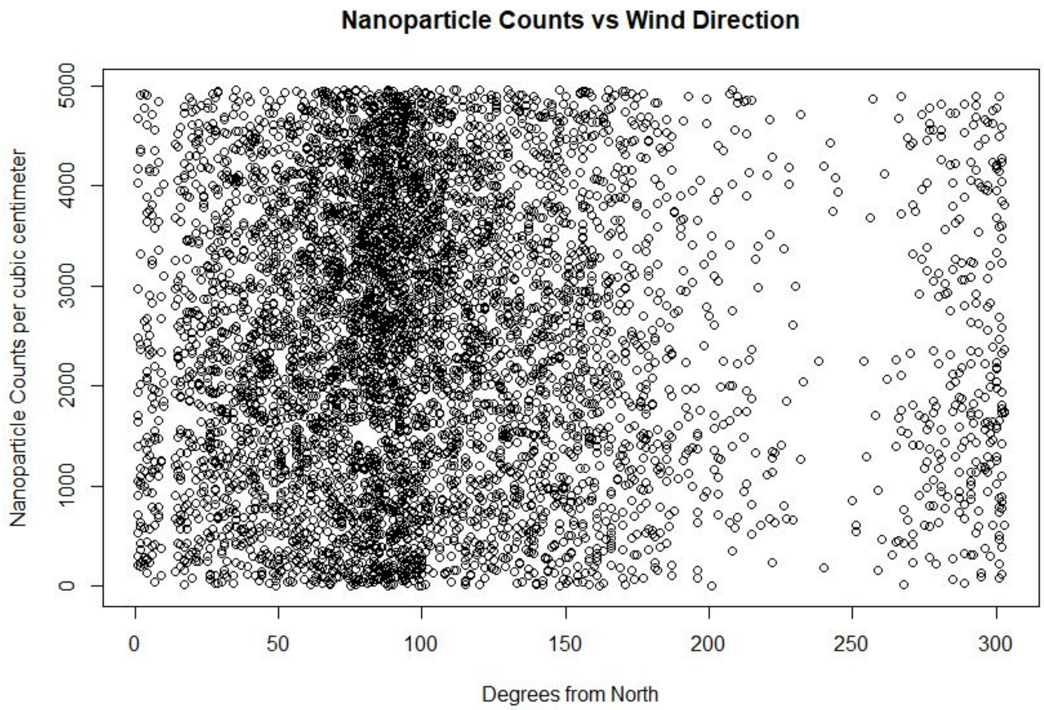
Analysis

The data was cleaned only to keep a few key columns. Wind Speed, Wind Direction, 2.5 micron particle concentration of $\mu g/m^3$(PM 2.5), Nanoparticle $\#/cm^3$(Particle Count, PC), and Black Carbon Particle concentration of $\mu g/m^3$(BC). The munging of the data set revolved around removing any observations that were missing values with na.omit() and any dates of reported instrument failure. The removed dates were from July 20th, 2019 to August 18th, 2019. The NO_2 column was removed entirely as it was never populated. The data set ended up being 5.2k observations.

It should be noted that wind direction references where the wind is coming from. 0 degrees is a wind coming from North, and going South, and the total range of collected wind directions is [0,360). Flavors of linear regression and multiple regression do not do the dataset justice due to the fixed range of the X-axis of the collected data. Larger grouping systems are required and Kmeans clustering was most-fitting as done by other researchers(P. Govender and

S. Sivakumar). Wind Direction data was not deemed suitable for analysis due to data points appearing to be highly quantized, but appear in Appendix 2.





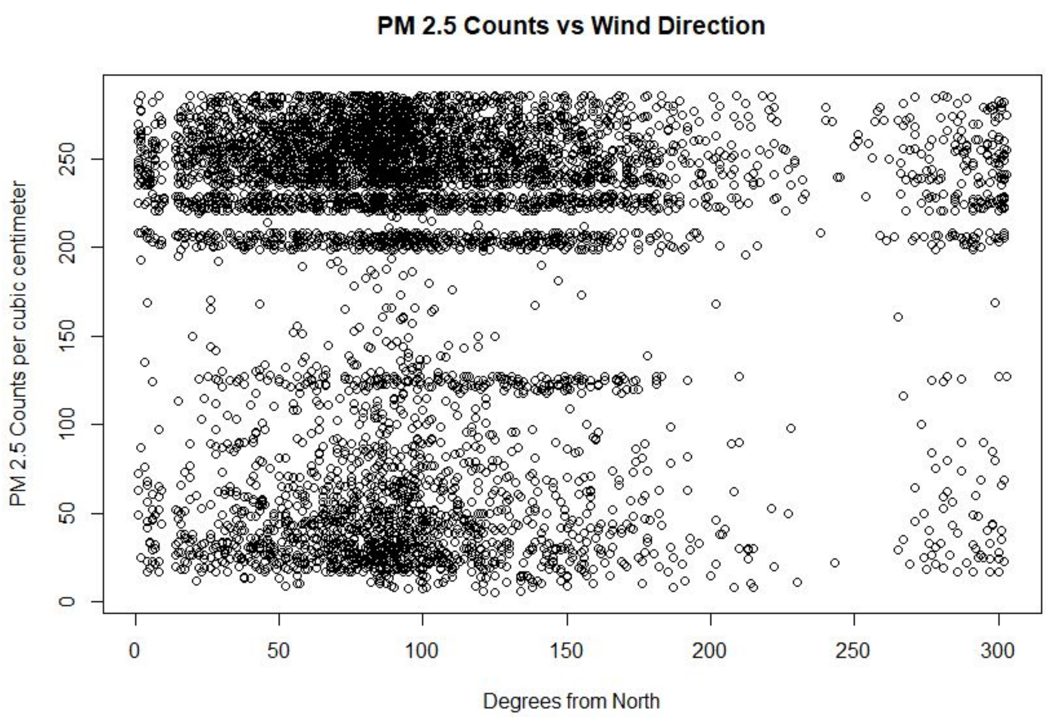


Figure 2: Initial data plots of the 3 pollutants vs. wind direction.

Initial plotting shows a dead-zone of pollution readings around 250 degrees from North, which could have something to do with the previously-mentioned hill, as shown above in Figure 2. Traveling West up Mt. Hope Drive 0.2 miles yields a rise in 105ft in elevation, which relates back to the main question for this study if topography has any influence of where pollutants come from.

The sample function in R was used to pick 5000 random data points in order to run Shapiro Testing. The results of the Shapiro test for normality and lognormality of the 3 pollutants have p-values of less than 2e(-16), so it is assumed that none of the data are normal or lognormal.

The nature of the data set does not lend itself to easy isolation of variables. While pollution dissipation increases with higher wind speeds[1], car traffic is also a hidden variable as no traffic counter data was used. Because of the time-averaged data set, the downside is that there is no way to see any localized peaks that may occur for further exploration in the data set.

Model Development and Application

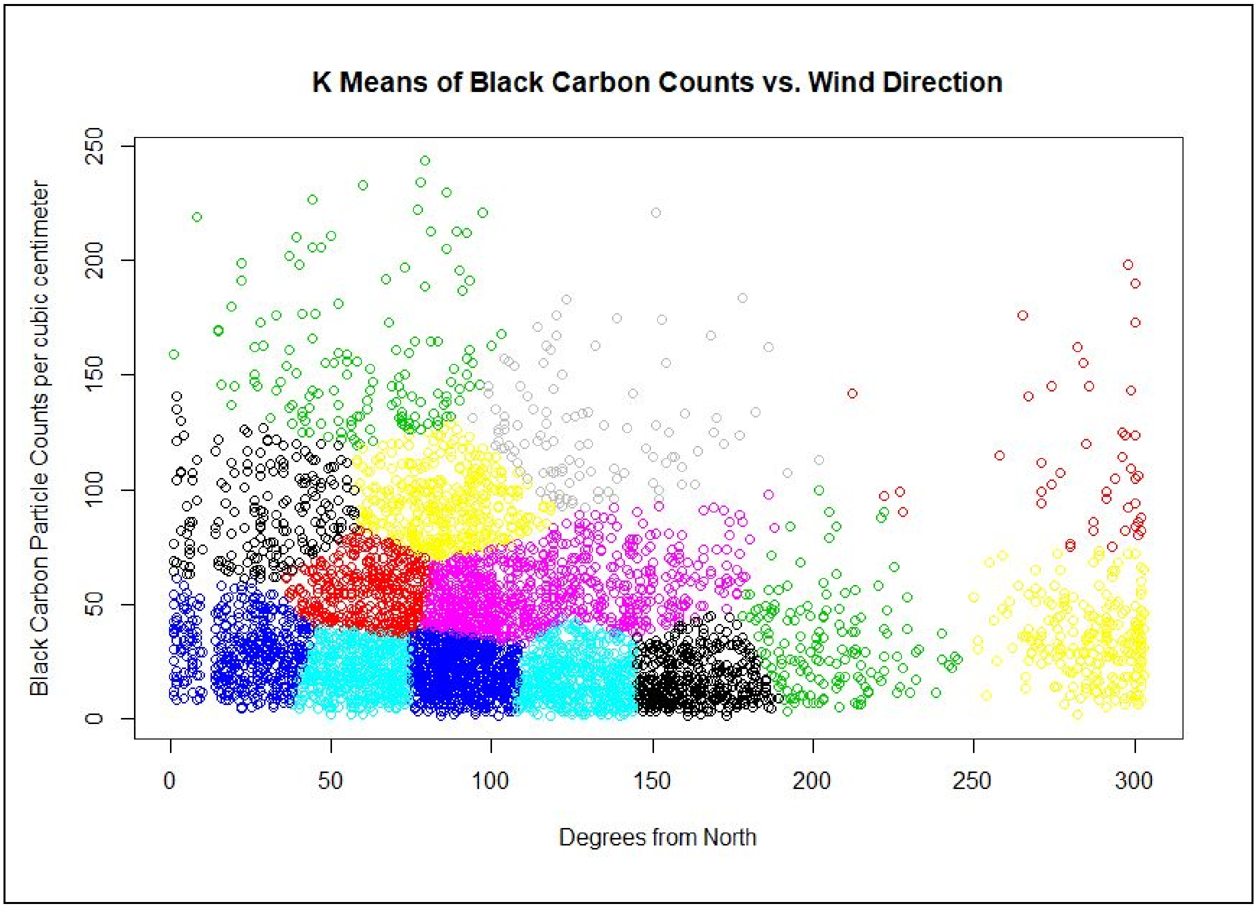
The Kmeans clustering was created using RStudio 3.6. The data was split into dataframes of pairs containing each of the 3 pollutants measured compared to wind speed and the vectorized X and Y values of the wind speed(WS) and wind direction(WD). The vectors were acquired using basic trigonometry by multiplying the wind speed by the sine or cosine of the angle and are displayed below.

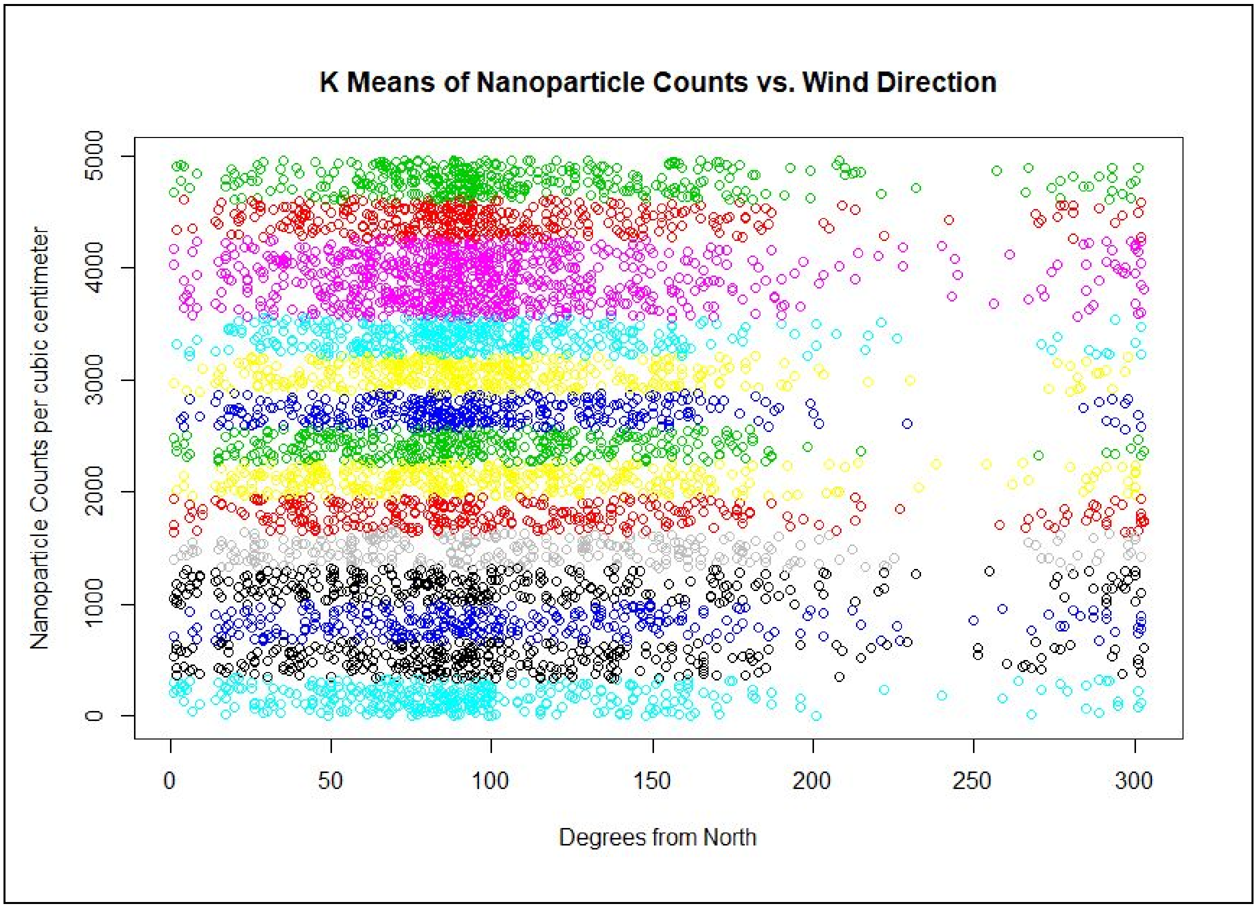
X-Vector Adjustment	WS * cos(WD/180*pi)	EQN 1
Y-Vector Adjustment	WS * sin(WD/180*pi)	EQN 2

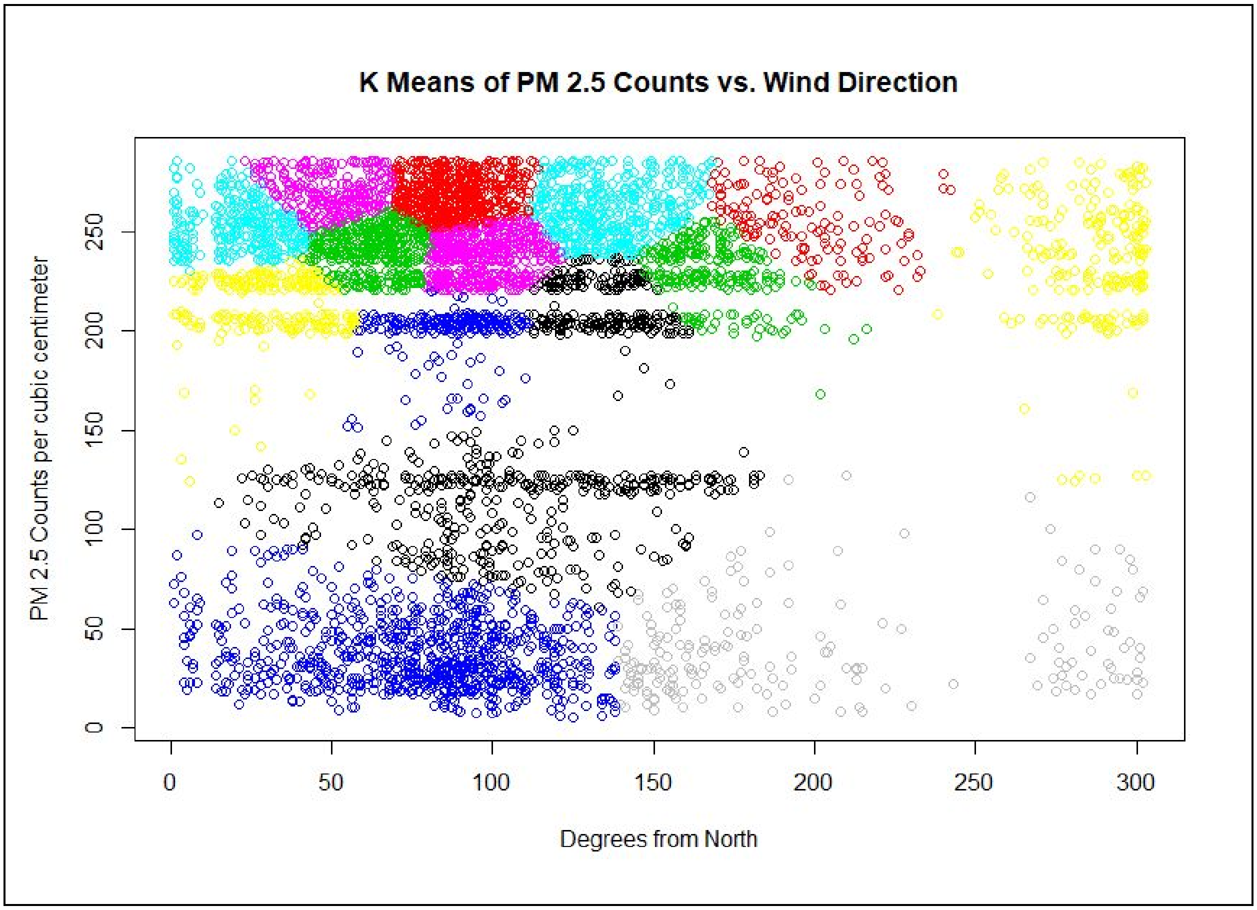
The purpose of getting the X and Y vectors is an attempt to normalize the data and isolate the North/South winds from the East/West winds and compare them to the data that weren't normalized. Comparing centers of these clusters is important for figuring out if there is any target direction that can be drawn to associate higher and lower readings, and if that wind vector has anything to do with the large hill in the neighborhood.

15 clusters were chosen as the result of initial testing based on Black Carbon results, and were capped off when cluster centers started to stop moving, but also keep some reasonable wind angle for a range if there were to be a wind rose for comparison. At 15 clusters, there would theoretically be 24 degrees per slice. The full cluster data is in Appendix 4, and the wind-speed-only clustering data set is located in Appendix 3.

Analysis of PC, BC, and PM 2.5 without vector-based adjustment




K Means of Nanoparticle Counts vs. Wind Direction
Nanoparticle Counts per cubic centimeter
0
1000
2000
3000
4000
5000
0
50
100
150
200
250
300
Degrees from North



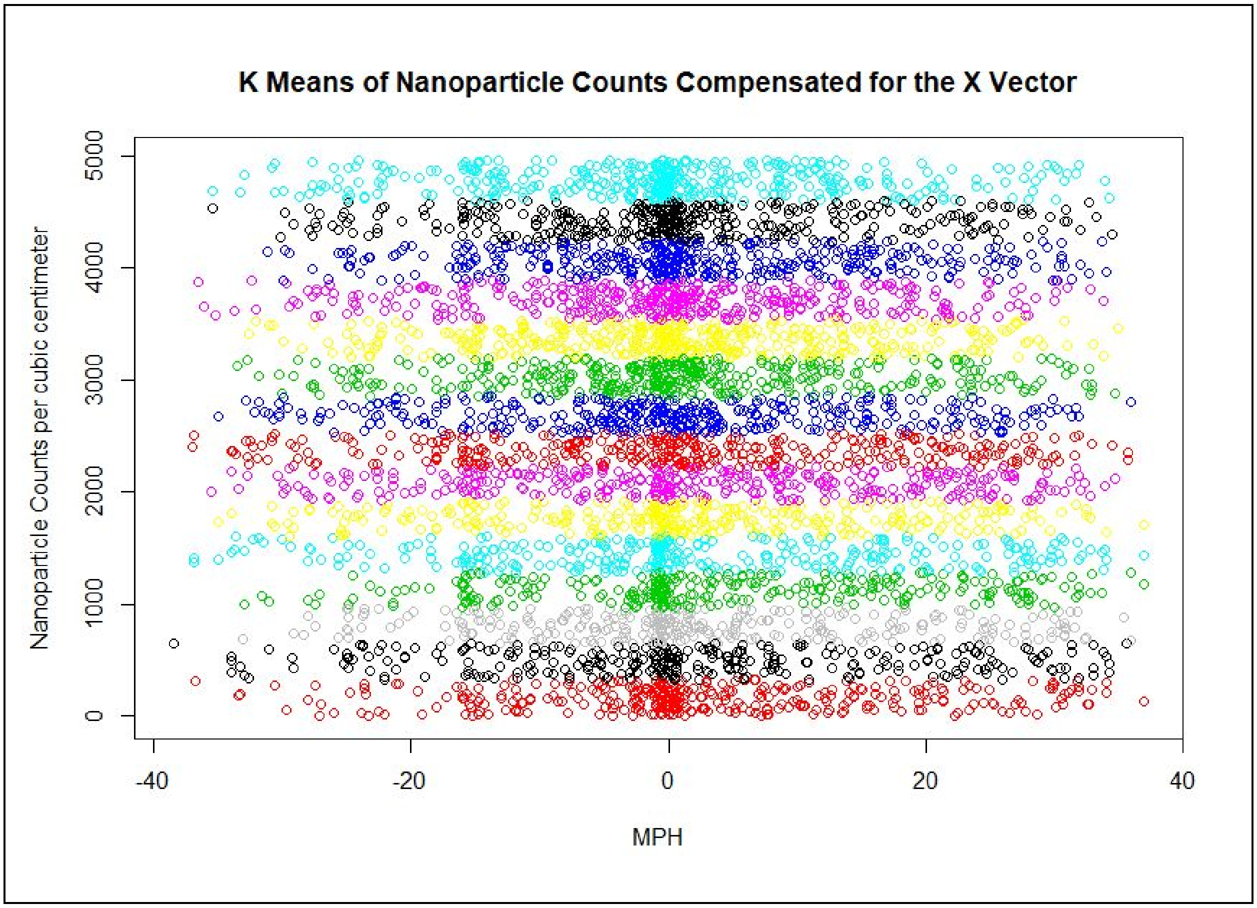
The most noticeable result of the Kmeans clustering was that of the Nanoparticle counts, and various levels of pollutants were present at every coordinate. The cluster centers were also centered around 90 degrees. While this means an east-bound reading was favored, the lack of isolated groupings along the X-axis implies the data cannot be pointed at having any directional correlation. This means there is an even distribution of fine nanoparticles in the air, and therefore will most likely need a different model to predict if wind direction has an impact on readings. It is also possible that nanoparticles are already evenly-mixed in the atmosphere from a larger scope of sources beyond the neighborhood that a more remote location is required to test this theory.

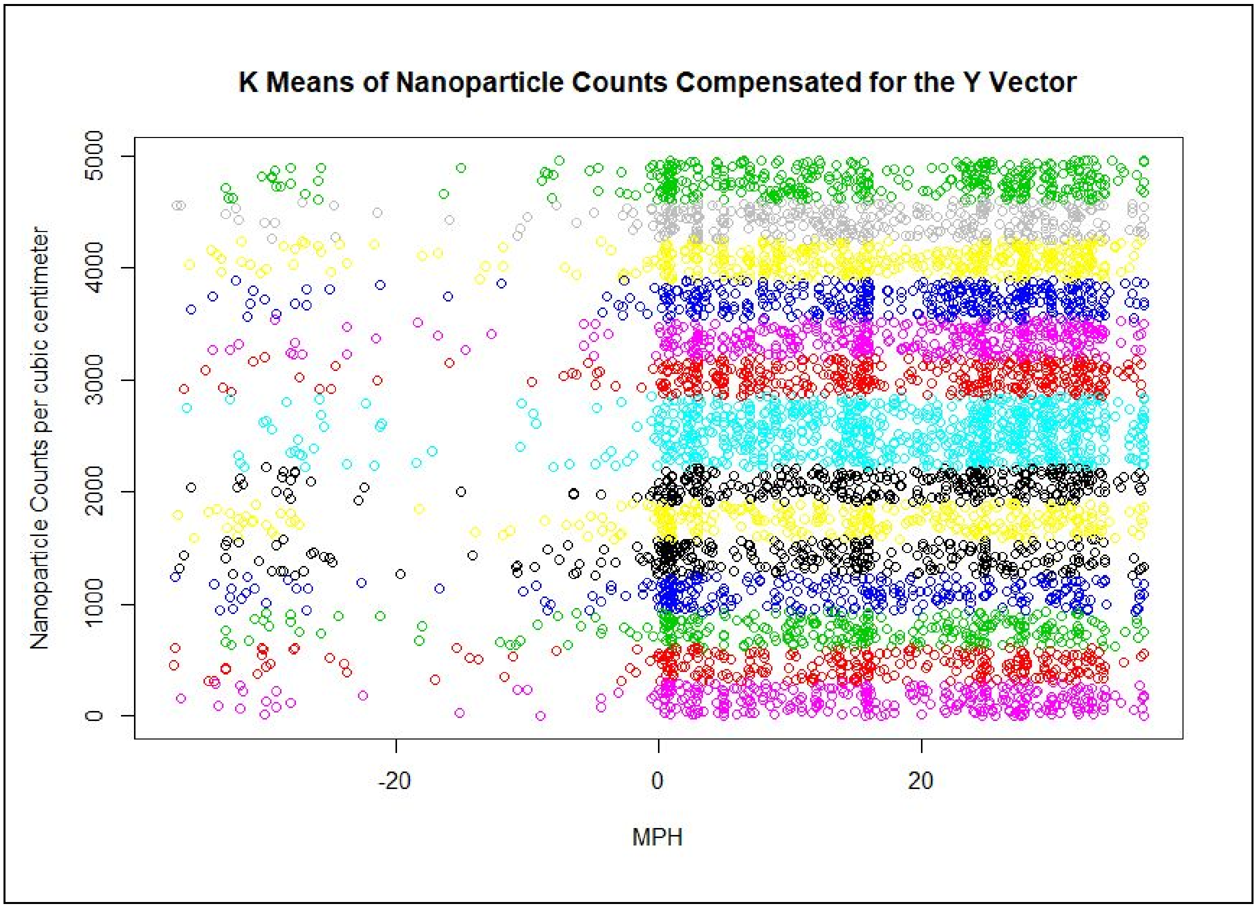
The PM 2.5 and Black Carbon results show much better clustering, with 13 out of 15 clusters being within 10 degrees of each other if sorted in ascending order. 7 out of the 15 clusters for BC are within 0-90 degrees, and no clusters occur after 286 degrees. Black Carbon emissions are based off of unburned fuel[1], so this would make sense as some Northeastern wind was blowing pollutants from the road.

Analysis of PC, BC, and PM 2.5 with vector-based adjustment

The vector based adjustment uses equations one and two in order to normalize the wind speed and direction of wind. As such, positive Y-values are winds coming from the North, while positive X-values are coming from the East. Due to the varying results, they will be analyzed individually.

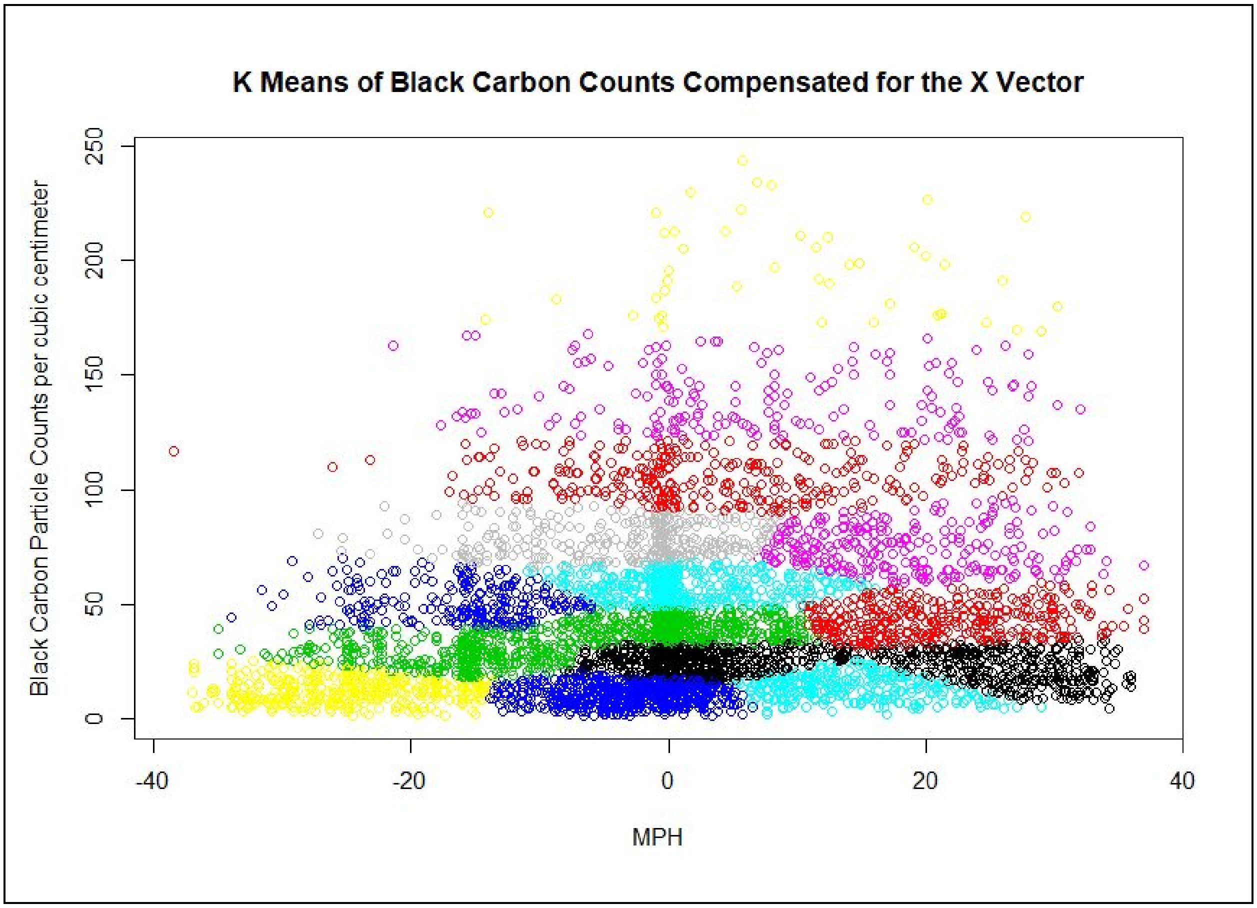
Nanoparticles


K Means of Nanoparticle Counts Compensated for the X Vector
Nanoparticle Counts per cubic centimeter
0
1000
2000
3000
4000
5000
-40
-20
0
20
40
MPH

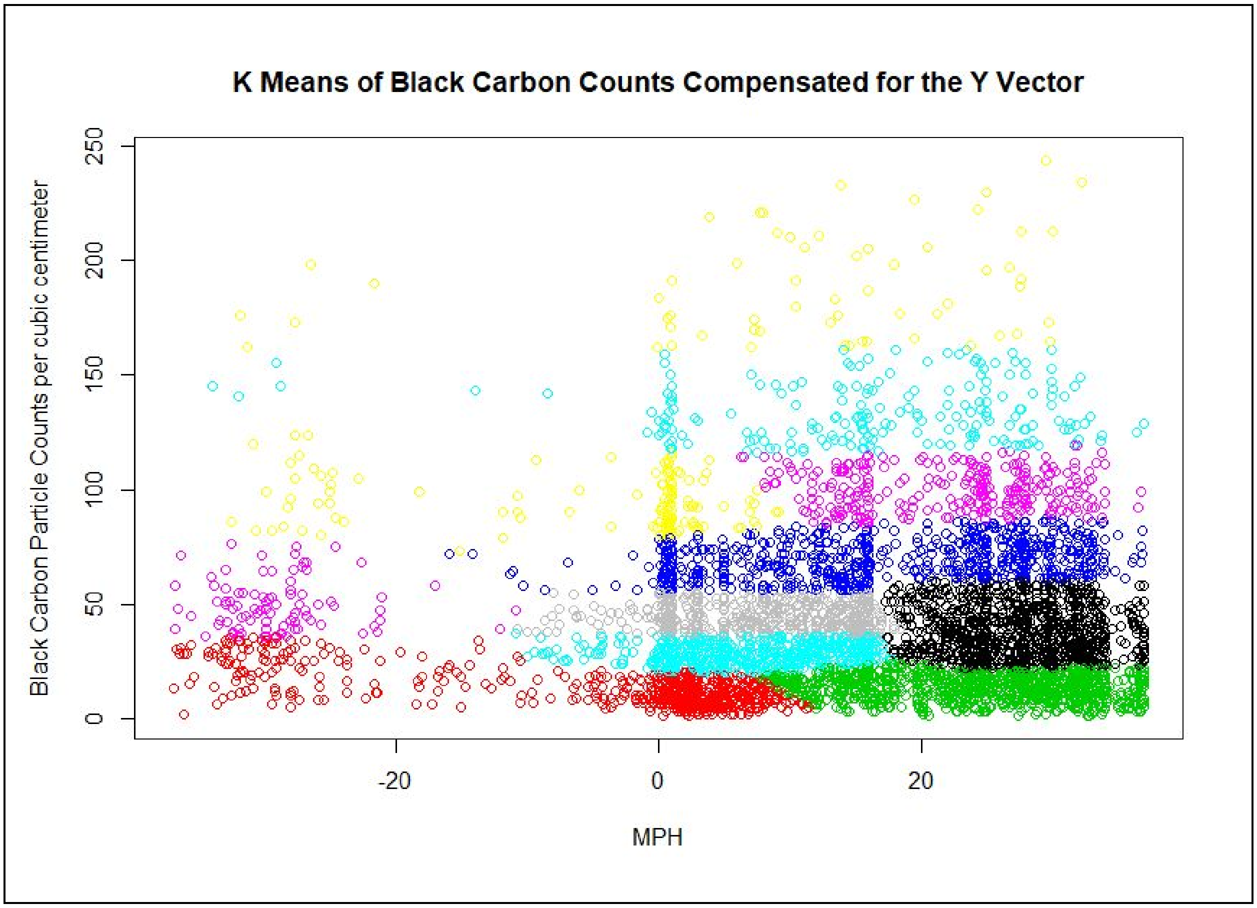


Due to the even spread along the Y- axis for both plots, the only conclusion that can be drawn is that most nanoparticle pollutants come from the North. While technically a reasonable statement coming from a roadway intersection that is North of the measuring station, this information is not enough to conclude anything.

Black Carbon

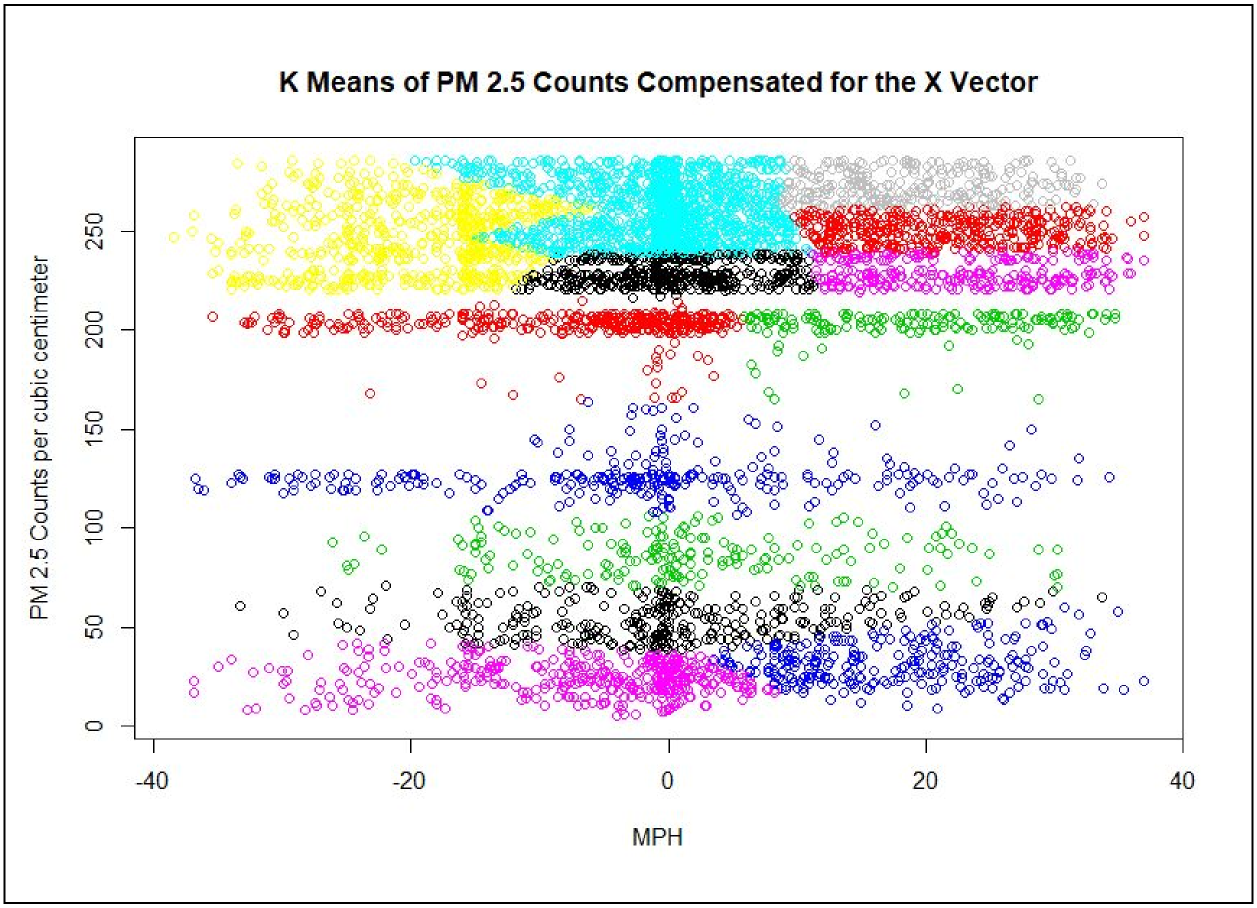


K Means of Black Carbon Counts Compensated for the X Vector
Black Carbon Particle Counts per cubic centimeter
0
50
100
150
200
250
-40
-20
0
20
40
MPH

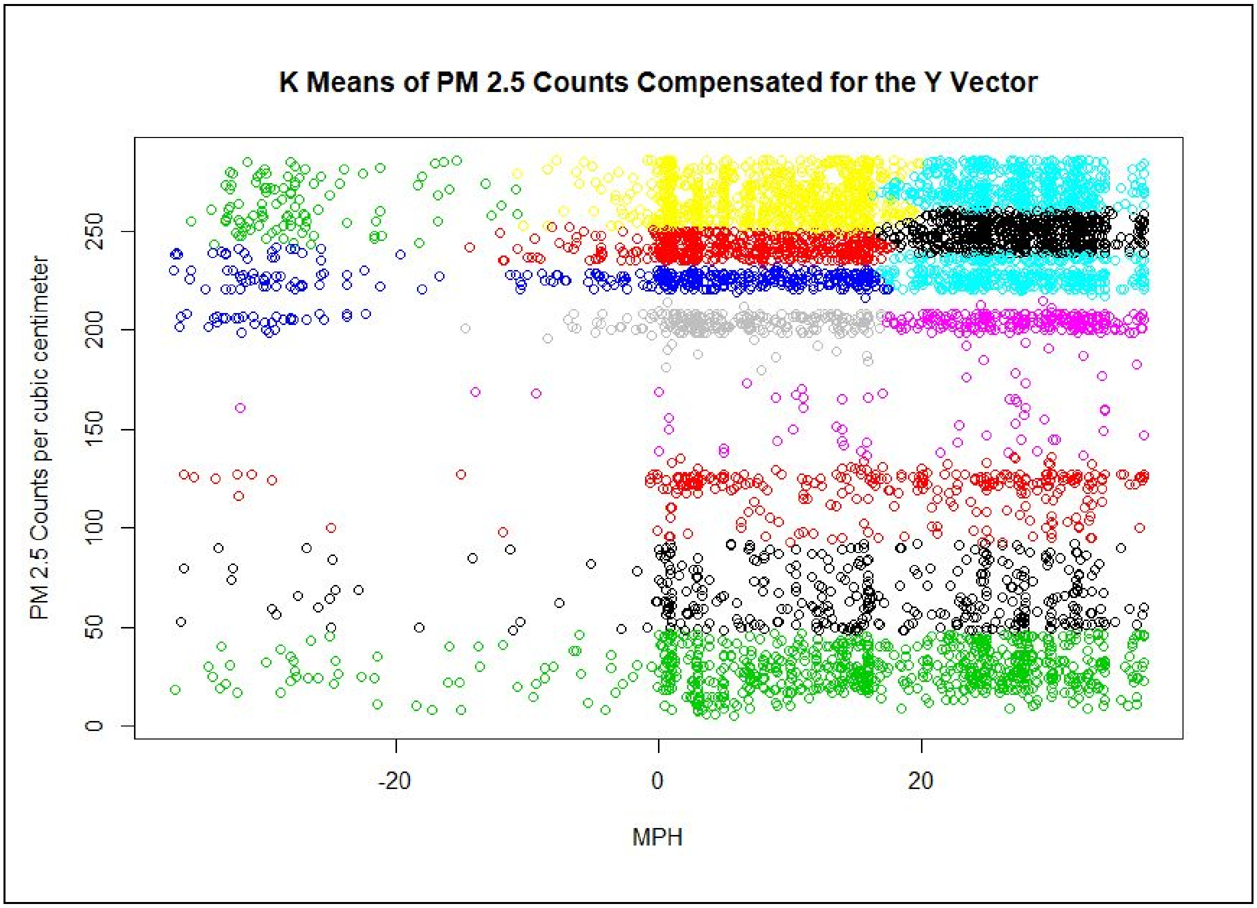


Upon initial inspection, the clustering appears to be similar to the nanoparticle analysis, with Northbound clustering in the Y Vector and a roughly even distribution in the X-Vector. However, most X-Vector clusters are in the positive end of the axis, implying a stronger Northeastern correlation than a purely Northern correlation.

PM 2.5



K Means of PM 2.5 Counts Compensated for the X Vector
PM 2.5 Counts per cubic centimeter
0
50
100
150
200
250
-40
-20
0
20
40
MPH



The PM 2.5 clustering is the least reliable telling of the data set as there is little recorded between 125p/cm^3 and 200p/cm^3. However, the clustering centers of the X-Vector point to having more negative centers, implying a stronger Northwestern correlation, which contradicts the BC readings.

Conclusions and Discussion

Depending on the data organization chosen, as well as visual inspection of the direct plots of data, it is reasonable to conclude there is some form of Northeastern influence in the pollution levels for all types analyzed. This makes sense due to how the roadway runs past the neighborhood and how there are very few measurements in the Southern and Western directions.

Because of the even distribution of nanoparticles from the of the non-vectorized data, there is reason to believe the mixing of nanoparticles is quite uniform. If anything, this analysis proved that they are everywhere, at least when averaged out on an hourly basis.

Comparisons of K-Means for wind speed were not considered at all as the speeds were quantized to whole values and there is missing data between 16 and 25 MPH. Sustained winds over 30MPH for an hour-long average, and an average wind speed of 23.34MPH are also reasons to be skeptical of the data. It is therefore difficult to create some sort of correlation factor in order to compensate for wind direction and create a “feels like” rating similar to wind chill due to a huge doubt in wind speed distribution.

A source of comparison for this work is a report published by the New York State Department of Environmental Conservation: Division of Air in October of 2019 with regards to emissions in the Ezra Prentice neighborhood that started the PEJA investigation[6].

From the work conducted by Frank et. al, the wind roses point to a Southern-favoring wind, which is contrary to the dataset presented. The DEC report also claims a minute-based average wind speed of 8.8 miles per hour overall during the experiment when conducted from June 1st, 2018 to July 31st, 2018. When exporting the hour-based data from the public DEC data set on the same time interval, the average was 4.4MPH. Regardless of how the DEC data set was time averaged, versus how the public data set is recorded and averaged automatically, 20+MPH is, by these comparisons, appears to be quite off.

With all of the sources of error considered, the fact that most clustering occurred with Northeastern tendencies implies that there is a directional correlation with pollution. With the road being East and an intersection Northeast of the measuring station, the overall directionality of the data makes sense. It is recommended, however, that a hill that is surrounded by residential areas with multiple measuring stations and sources of pollutants would probably be best to test the theory of if topography has any impact on pollution distributions.

Acknowledgements

The author of this study would like to thank Dr. Brian Frank, Marilyn Wurth, Gil LaDuke, Tom Giorgio, Temple Bailey, and the rest of the Mobile Sources group for the experience over the Summer of 2019, as well as the ability to further contribute to the public for the Fall 2019 term. The author has learned much on the subject of data analysis of the real world, as well as experimental design.

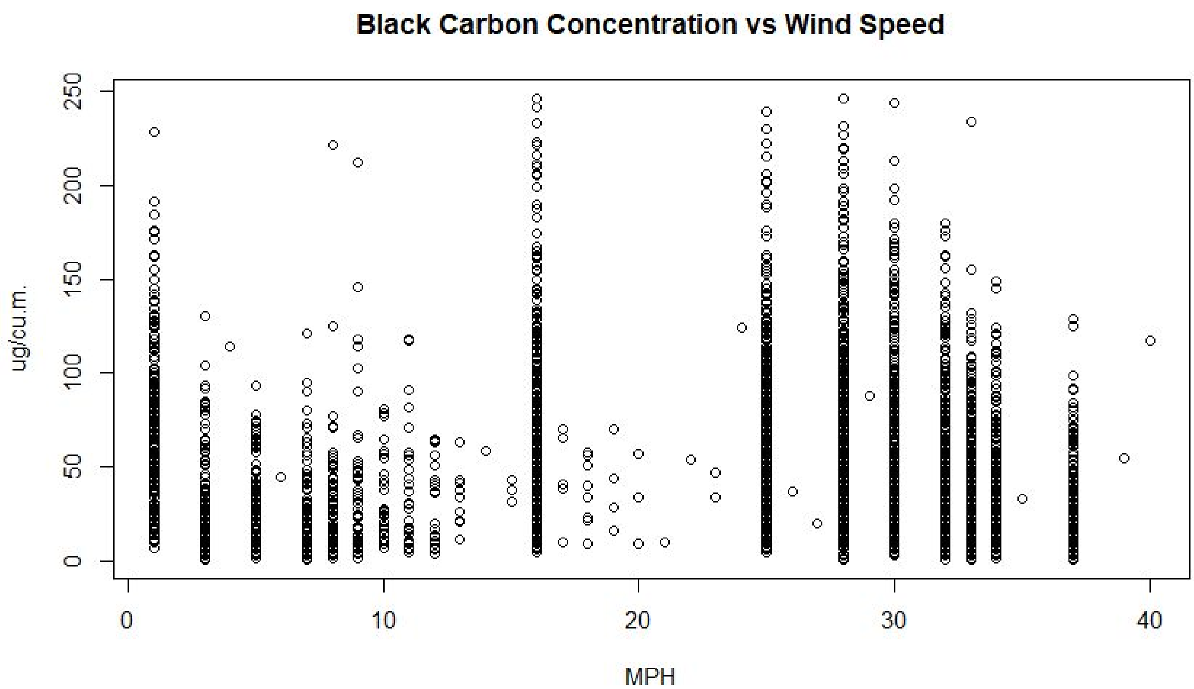
References

1) Neeldip Barman and Sharad Gokhale, “Urban black carbon - source apportionment, emissions and long-range transport over the Brahmaputra River Valley”, *Science of The Total Environment*, Volume 693, 2019
2) Lu, X., Zhu, T., Chen, C., & Liu, Y. (2014). Right or left: the role of nanoparticles in pulmonary diseases. *International journal of molecular sciences*, *15*(10), 17577–17600. doi:10.3390/ijms151017577
3) Ana Fernández Tena, Pere Casan Clarà, “Deposition of Inhaled Particles in the Lungs”, *Archivos de Bronconeumología (English Edition)*, Volume 48, Issue 7, 2012.
4) “What are the Air Quality Standards for PM?” Retrieved November 27, 2019, from https://www3.epa.gov/region1/airquality/pm-aq-standards.html.

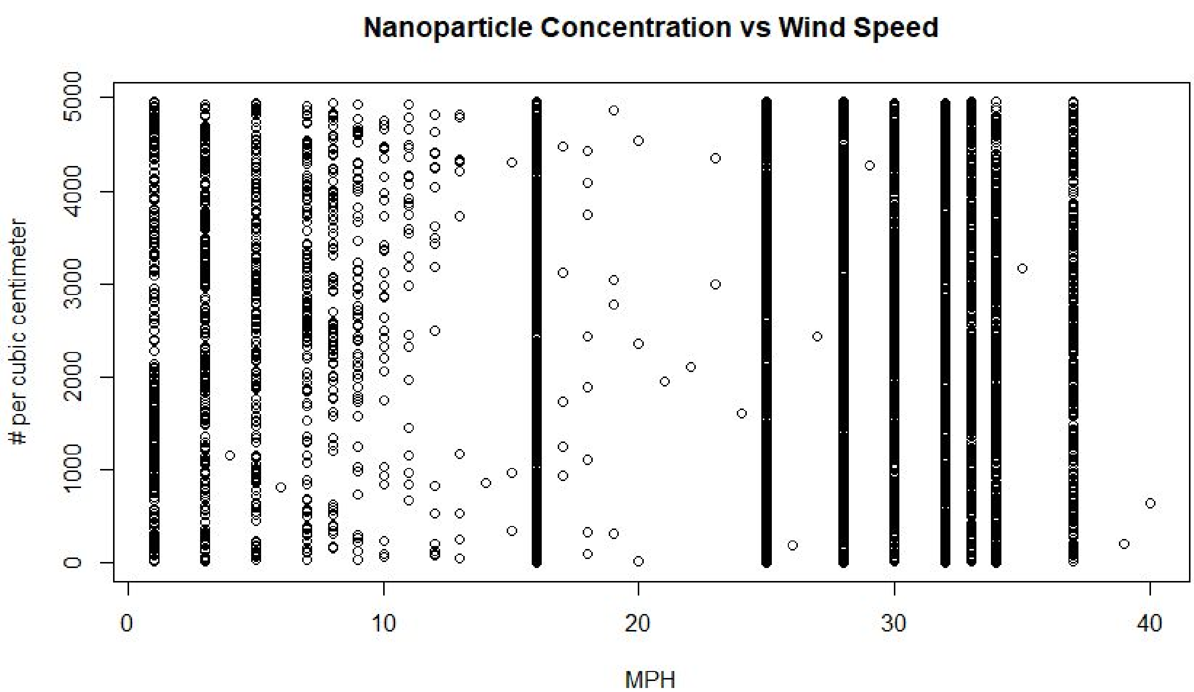
5) Junyan Yang, Beixiang Shi, Yi Shi, Simon Marvin, Yi Zheng, Geyang Xia, "Air pollution dispersal in high density urban areas: Research on the triadic relation of wind, air pollution, and urban form", *Sustainable Cities and Society*, 2019.
6) New York State Department of Environmental Conservation (NYS DEC), Division of Air Resources. *Albany South End Community Air Quality Study*. October 2019.
7) *nyaqinow.net*. [Online]. Available: http://www.nyaqinow.net/. [Accessed: 15-Oct-2019].

Appendix 1: Industrial Overview of the Area

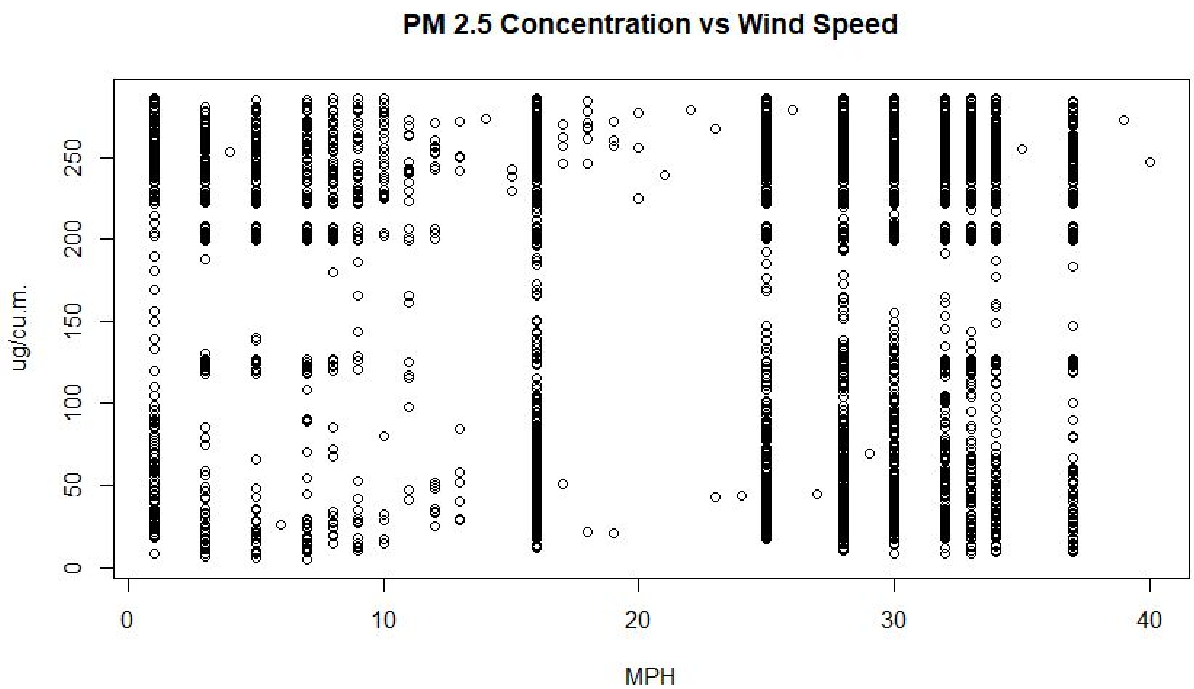
Appendix 2: Wind Speed Plotting versus Pollutants



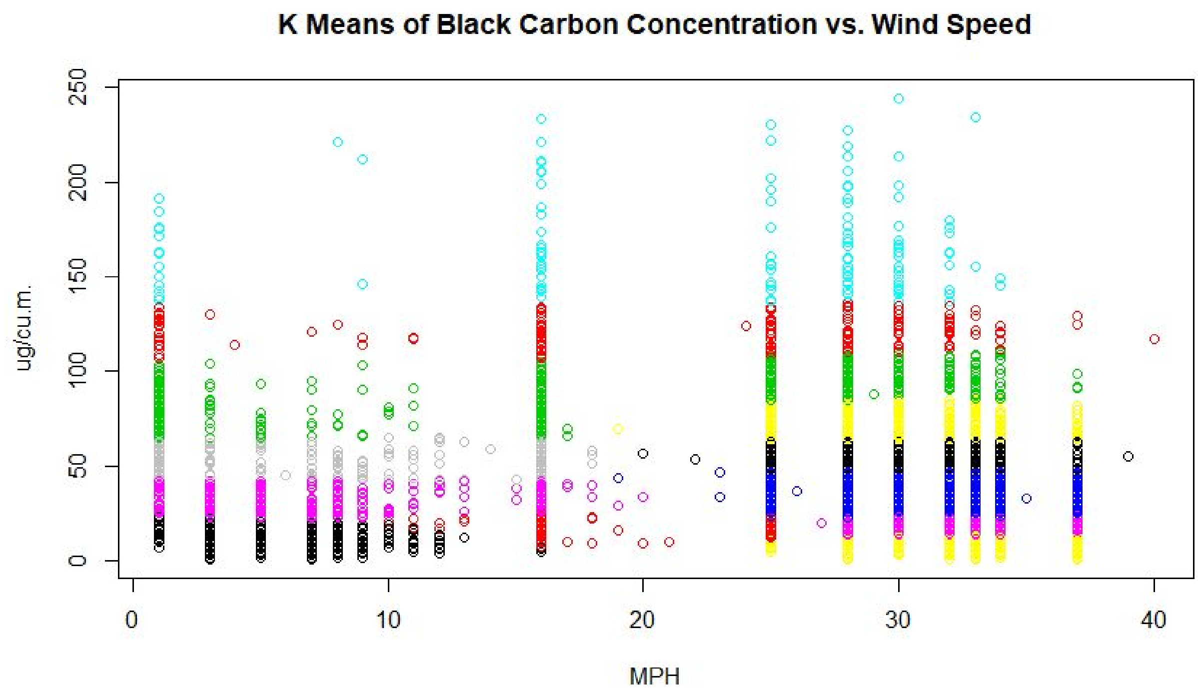
Black Carbon Concentration vs Wind Speed
ug/cu.m.
0
50
100
150
200
250
0
10
20
30
40
MPH

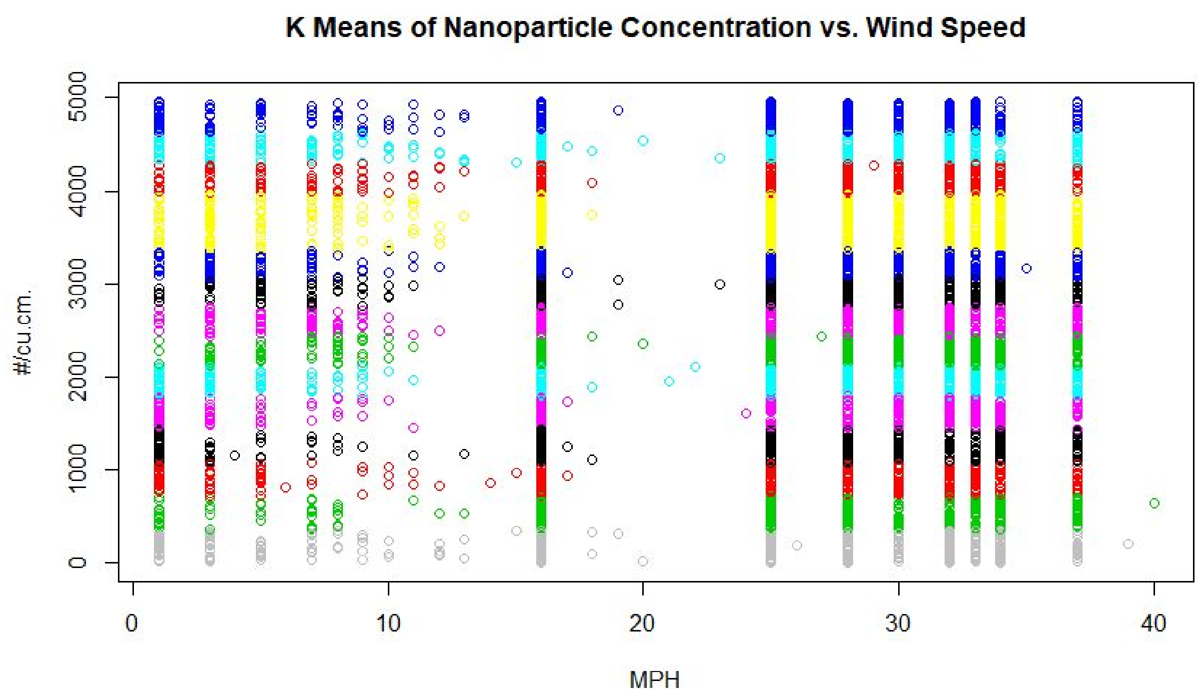


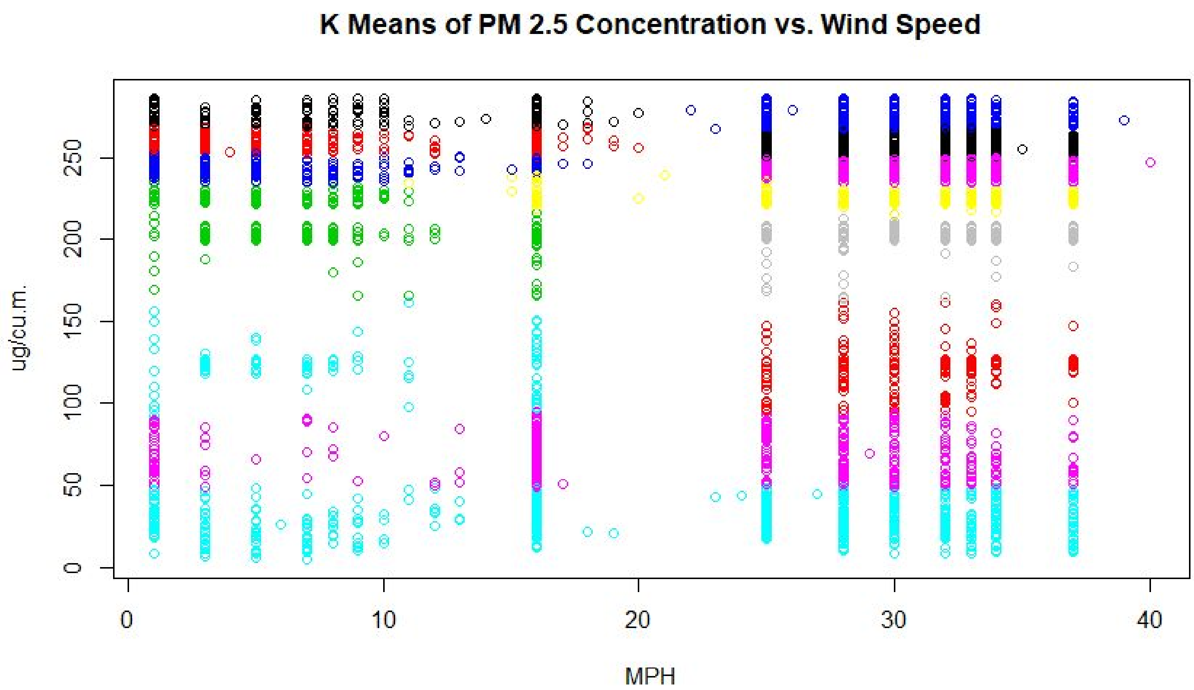
Nanoparticle Concentration vs Wind Speed
per cubic centimeter
0
1000
2000
3000
4000
5000
0
10
20
30
40
MPH


PM 2.5 Concentration vs Wind Speed
ug/cu.m.
0
50
100
150
200
250
0
10
20
30
40
MPH

Appendix 3: Clustering Results Compared to Wind Speed


K Means of Black Carbon Concentration vs. Wind Speed
ug/cu.m.
0
50
100
150
200
250
0
10
20
30
40
MPH


K Means of Nanoparticle Concentration vs. Wind Speed
#/cu.cm.
0
1000
2000
3000
4000
5000
0
10
20
30
40
MPH



Appendix 4: Clustering Data Table Results

Black Carbon

WD(deg)	BC(ug/cu.m.)
113.84007	17.99081
30.18127	22.18725
151.73592	16.35146
27.41991	63.1342
61.4386	186.2807
285.58182	33.41364
193.37288	26.5678
44.97041	119.04142
61.09318	42.125
82.07036	79.80597
278.26	110.02
111.07772	121.42487

79.5928	17.75485
95.98065	45.72258
142.46302	62.08039
WS(MPH)	BC(ug/cu.m.)
30.55298	28.703642
30.613226	40.589178
15.870307	30.515358
26.029197	96.678832
21.415842	120.450495
10.302949	51.927614
8.982833	76.51073
31.030534	18.058015
21.969697	205.606061
22.960784	153.196078
32.475285	8.847909
7.131579	10.892713
30.007634	54.664122
4.514523	30.435685
29.621849	72.885154
Xvec(MPH)	BC(ug/cu.m.)
-12.263542	13.57792
9.6677934	196.63043
18.8880325	73.2619
6.4661164	139.48571
4.3869083	105.51768
22.8289904	17.22654
-2.1343988	79.07365
0.2629168	10.49502
22.4293624	39.73892
2.6455581	25.15556
1.1734674	39.64403
-27.0783315	12.96321
-14.7485727	54.87013
2.6524	56.98186

-16.1127814	32.24784
Yvec(MPH)	BC(ug/cu.m.)
28.83326	12.75516
24.375806	95.8502
15.379802	119.90521
14.413437	206.40625
14.302888	18.63457
6.918774	53.86937
28.021073	29.50181
26.413078	47.04555
6.984848	33.47115
26.200043	69.55645
-24.896906	86.30769
-28.41037	30.98964
12.805632	154.12121
6.314175	78.93333
3.064273	11.02353

PM 2.5

WD(deg)	PC(#/cu.cm.)
98.6034	3648.847
117.85	1673.6765
106.81341	929.6647
98.35	3329.6553
99.11176	3950.4029
98.47895	180.7974
97.89031	2672.6964
107.78655	1304.9766
102.4345	4229.4121
105.59683	4522.9429
102.04023	558.8736
102.2252	3008.9464

100.59838	2347.4528
109.00949	4811.693
99.60054	2026.3164
WS(MPH)	PC(#/cu.cm.)
23.45133	107.5796
23.89815	559.7778
23.39545	323.65
22.60087	3216.4252
22.35443	798.2954
22.44358	1072.3541
23.61384	3996.0848
22.40811	4764.6563
23.82589	3607.1295
22.48355	1376.2533
24.32266	2061.3448
23.43987	1711.0222
24.90646	2430.1269
24.43421	2816.2346
21.7524	4379.1538
Xvec(MPH)	PC(#/cu.cm.)
-0.3256587	4813.0671
0.8869697	2340.4553
2.703353	178.3147
1.1326431	2665.967
1.2718114	1662.5274
2.4501071	3951.2588
0.7160519	3005.3799
2.7435585	2019.9812
0.8577805	4231.2603
2.8539083	549.2059
4.4880096	1288.0754
0.9932328	3329.231
0.9123675	4525.7143
0.6019875	3649.274

2.0105096	913.3127
Yvec(MPH)	PC(#/cu.cm.)
16.44396	2340.0136
16.95842	3329.231
16.15902	3951.2588
16.57693	3005.3799
10.89043	1296.2412
13.9294	4525.7143
15.78277	4231.2603
15.0056	180.7974
11.6399	1666.2471
15.2279	2020.938
13.24334	557.8121
14.8098	4813.0671
17.06209	3649.274
17.08459	2665.5544
12.81057	924.7396

Nanoparticle

WD(deg)	PM25(ug/cu.m.)
153.64527	217.26351
86.7619	241.75198
116.46246	239.48949
58.34921	269.64762
58.9434	239.10782
31.52119	211.77966
93.49378	271.02282
93.34828	203.47586
193.66839	247.74093
101.48214	112.6199
145.48328	262.18237
286.37619	235.96667

199.75385	40.3641
77.62909	37.45818
23.24026	255.09416
WS(MPH)	PM25(ug/cu.m.)
10.760684	277.25214
29.910603	29.97297
7.496933	243.95706
23.503145	122.80503
7.983871	201.30108
5.21374	226.06107
31.616022	202.90331
10.450704	28.39085
31.105096	242.64809
30.534247	258.02226
31.47619	225.04989
17.983146	231.3427
10.146758	259.68601
29.717213	275.82992
22.898551	68.56232
Xvec(MPH)	PM25(ug/cu.m.)
-18.3345156	259.84906
-6.65437404	26.1179
21.9826325	247.15493
-0.54008192	242.41391
-1.56601137	124.34333
-20.0589343	203.50806
0.58298088	225.32764
-2.93545398	277.69596
2.66741386	259.32072
1.64318195	201.31845
-21.03222776	231.93548
19.56378967	272.96679
12.84117208	38.60759

24.29910323	217.95161
-0.02621354	75.16727
Yvec(MPH)	PM25(ug/cu.m.)
17.157298	125.32982
27.090113	228.21077
6.237059	229.60677
12.810003	83.51759
6.787989	249.5129
27.998403	202.94024
27.234635	249.26917
13.904945	24.66608
-24.951654	275.13462
-28.165756	249.13333
18.775821	49.86898
-29.35017	215.16667
8.625958	272.56504
27.123577	272.47107
6.234962	201.61423