Clustering Algorithms for Understanding Air Pollution

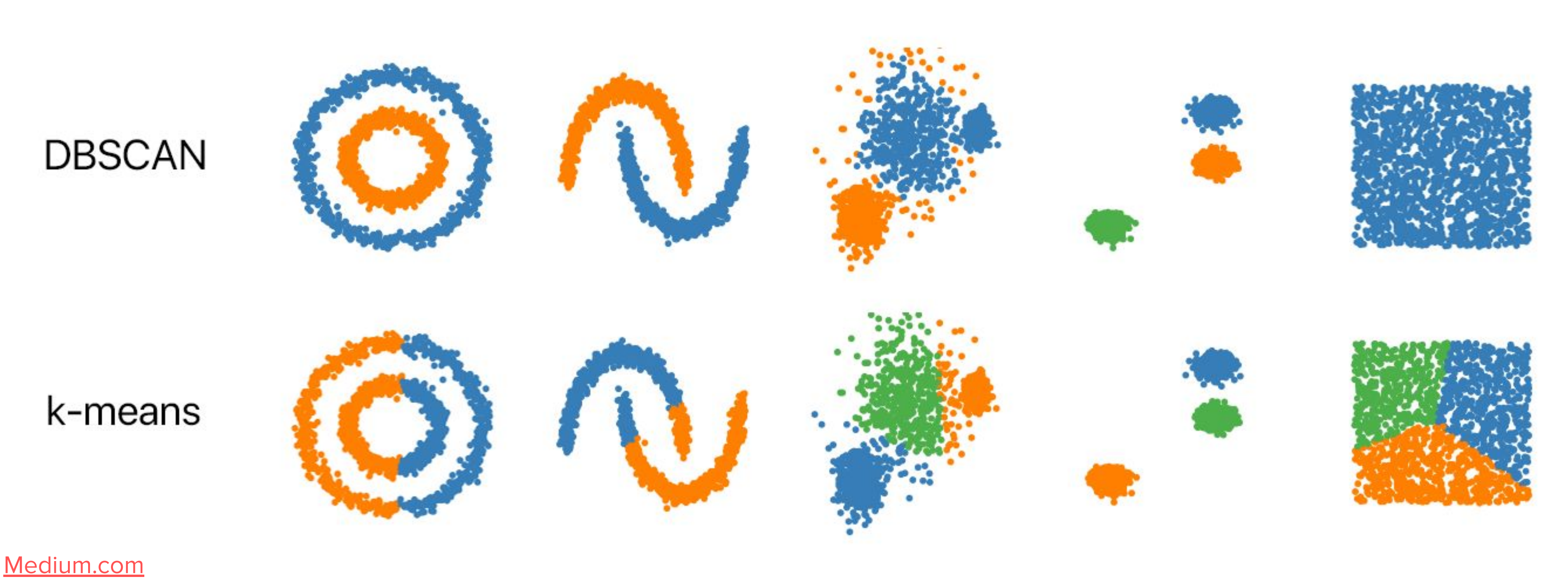
Toby Armstrong, Dr. Heather Holmes, Wilkes Center for Climate Science and Policy

HDBSCAN

- Density based clusters
- Clusters can form arbitrary shapes
- Can label points as outliers/noise
- 2 hyperparameters
  - Minimum cluster size
  - Minimum samples

K-MEANS

- Distance based clusters
- Clusters are spherical or convex
- Sensitive to outliers/noise
- 1 hyperparameter
  - Number of clusters



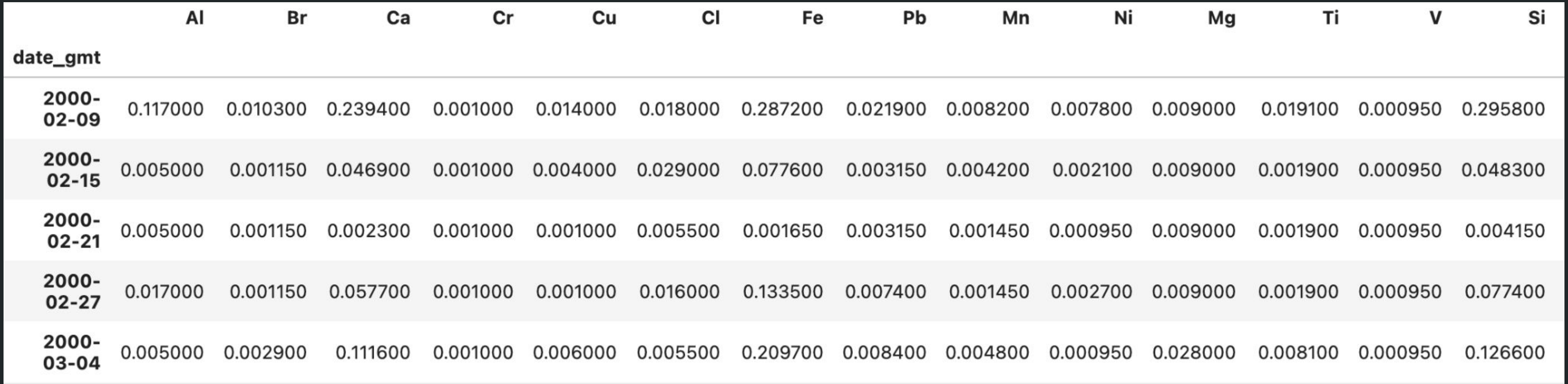
Medium.com

Why Use Clustering for Air Pollution

- The results of various clustering algorithms can promote the identification of pollution patterns and sources, allowing for more effective countermeasures.
- Comparing the results of clustering algorithms to existing source apportionment studies allows us to determine the efficacy of these newer methods.
    - Furthermore, conducting comparisons between different clustering algorithms (K-MEANS, DBSCAN, etc…) highlights which algorithm is best suited for meteorological/pollutant data.
- The goal of this project was to compare the efficacy of 2 algorithms (K-MEANS and HDBSCAN) to previous source apportionment studies within the same region to see if they show promise for future applications within the field.

Clustering for Air Pollution: The Process

- Parameters
  - Pollutant Concentrations
  - Meteorological Measurements
  - Derived Values
- Considerations
  - Correlation Confusion
  - Normalization/Scaling
  - Which Algorithm to Choose
- Outputs
  - Cluster Assignment
  - Tuning Algorithm
  - Re-running

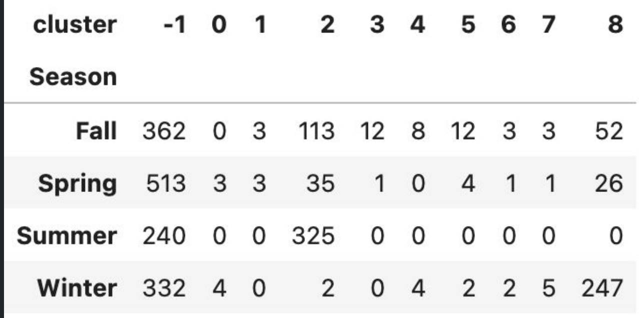


	Al	Br	Ca	Cr	Cu	Cl	Fe	Pb	Mn	Ni	Mg	Ti	V	Si
date_gmt														
2000-02-09	0.117000	0.010300	0.239400	0.001000	0.014000	0.018000	0.287200	0.021900	0.008200	0.007800	0.009000	0.019100	0.000950	0.295800
2000-02-15	0.005000	0.001150	0.046900	0.001000	0.004000	0.029000	0.077600	0.003150	0.004200	0.002100	0.009000	0.001900	0.000950	0.048300
2000-02-21	0.005000	0.001150	0.002300	0.001000	0.001000	0.005500	0.001650	0.003150	0.001450	0.000950	0.009000	0.001900	0.000950	0.004150
2000-02-27	0.017000	0.001150	0.057700	0.001000	0.001000	0.016000	0.133500	0.007400	0.001450	0.002700	0.009000	0.001900	0.000950	0.077400
2000-03-04	0.005000	0.002900	0.111600	0.001000	0.006000	0.005500	0.209700	0.008400	0.004800	0.000950	0.028000	0.008100	0.000950	0.126600

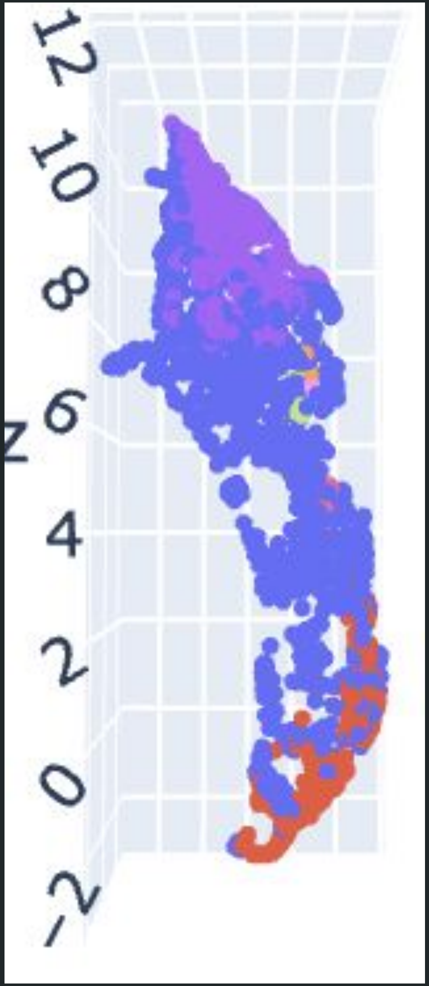
The Data

- All data was collected from the Hawthorne Elementary School monitoring station.
  - 1675 S 600 E, Salt Lake City, Utah, 84105
- Data spans from 2000-2021 and includes the following parameter concentrations (collected every third day):
  - PM2.5 Chemical Speciation Network (CSN)
  - PM10
  - Ozone
  - Carbon Monoxide
  - Sulfur Dioxide
  - Nitrous Oxides
- Hourly wind speed and direction measurements from the same location were used to create windrose plots for clusters.

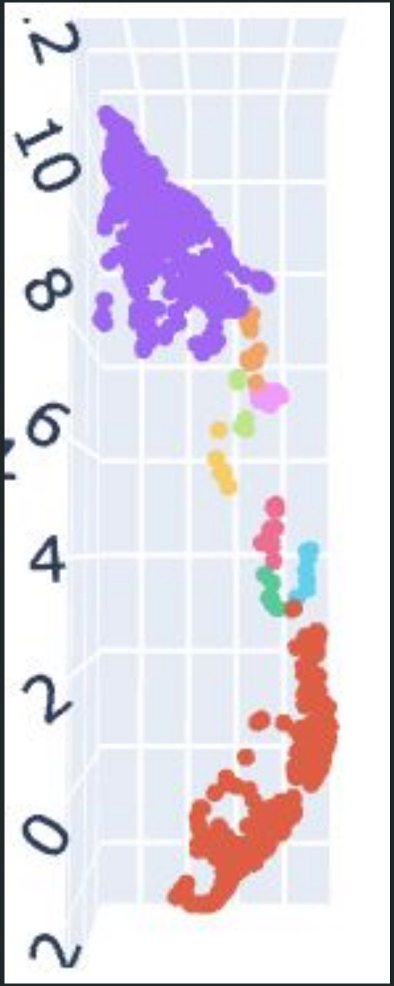
HDBSCAN Results Without PM2.5 Concentrations



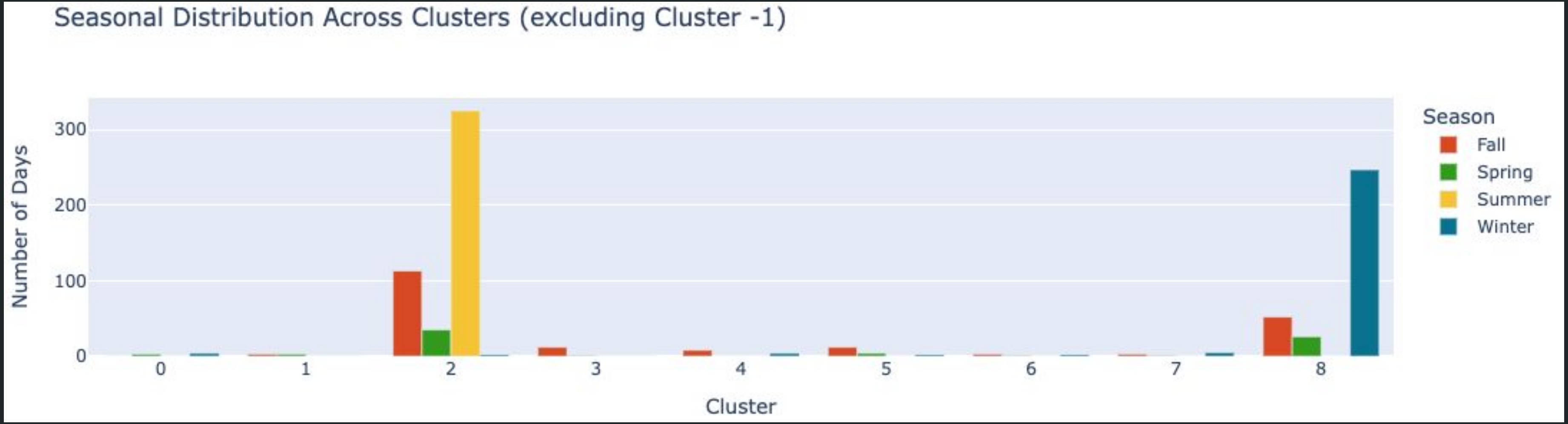
cluster	-1	0	1	2	3	4	5	6	7	8
Season										
Fall	362	0	3	113	12	8	12	3	3	52
Spring	513	3	3	35	1	0	4	1	1	26
Summer	240	0	0	325	0	0	0	0	0	0
Winter	332	4	0	2	0	4	2	2	5	247



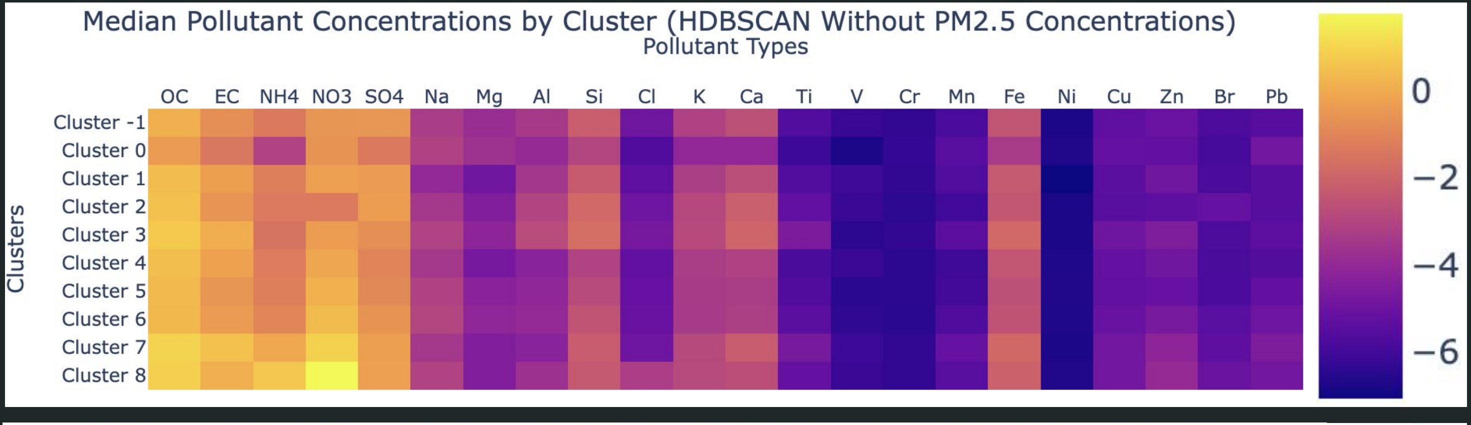
Removing Noise



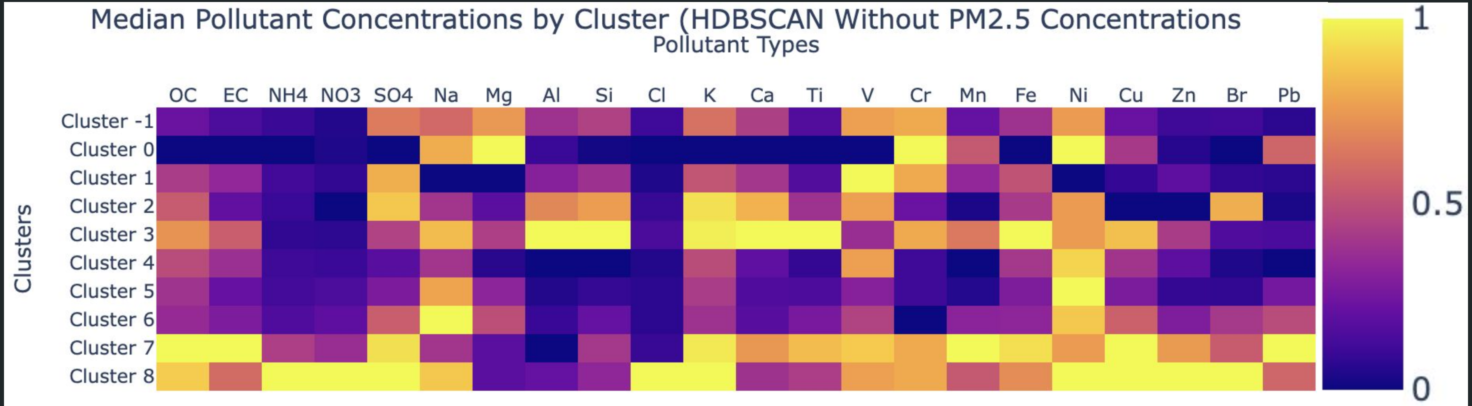
HDBSCAN Results Without PM2.5 Concentrations



HDBSCAN Results Without PM2.5 Concentrations

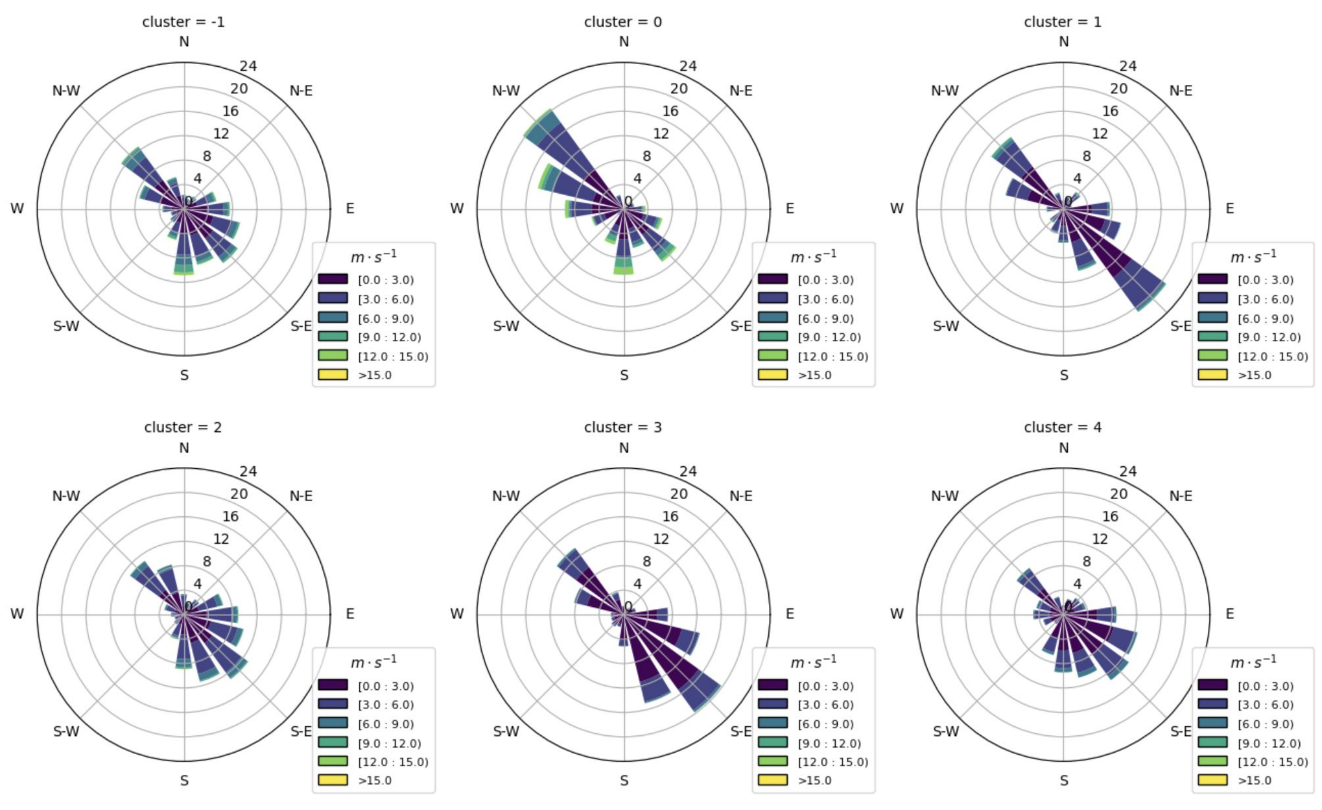


Log-Scaled Median Concentrations

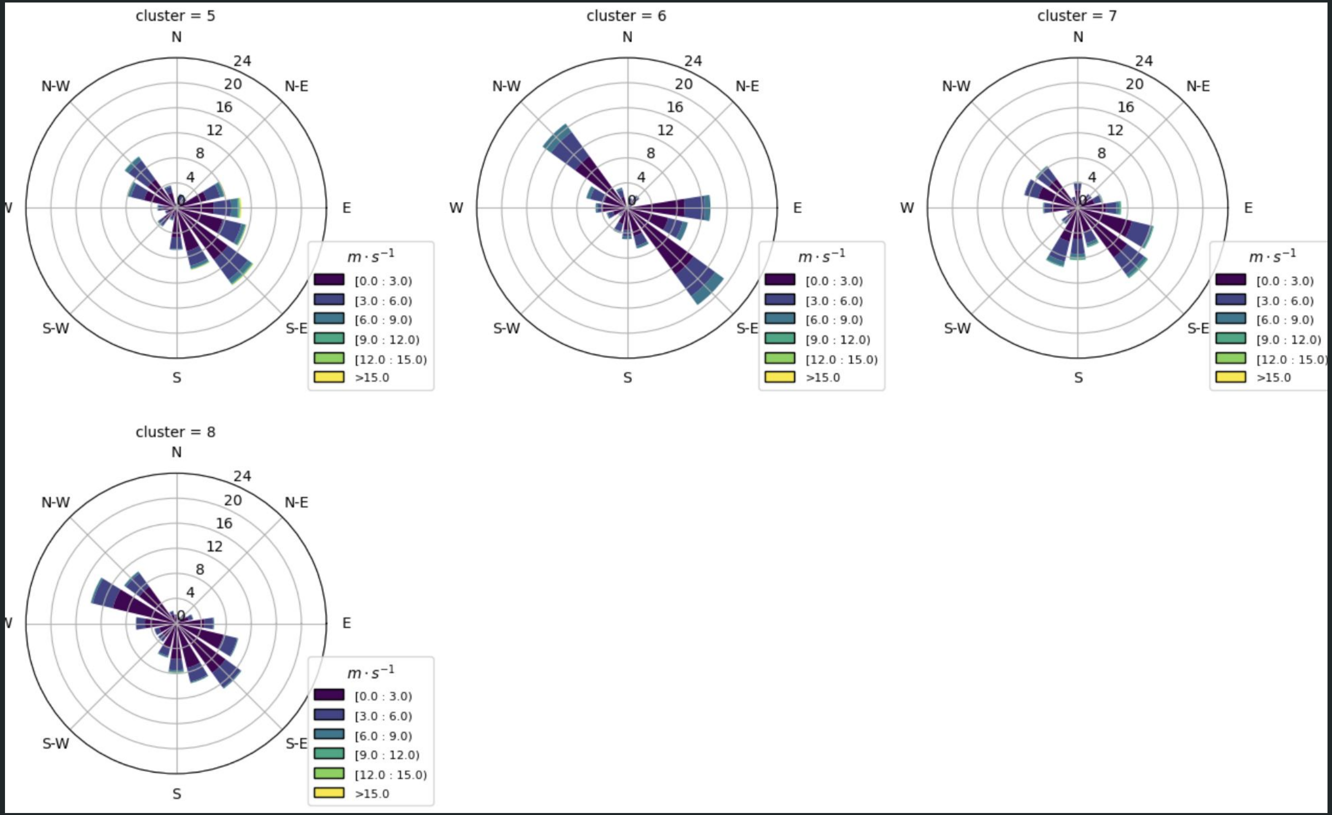


Normalized Median Concentrations

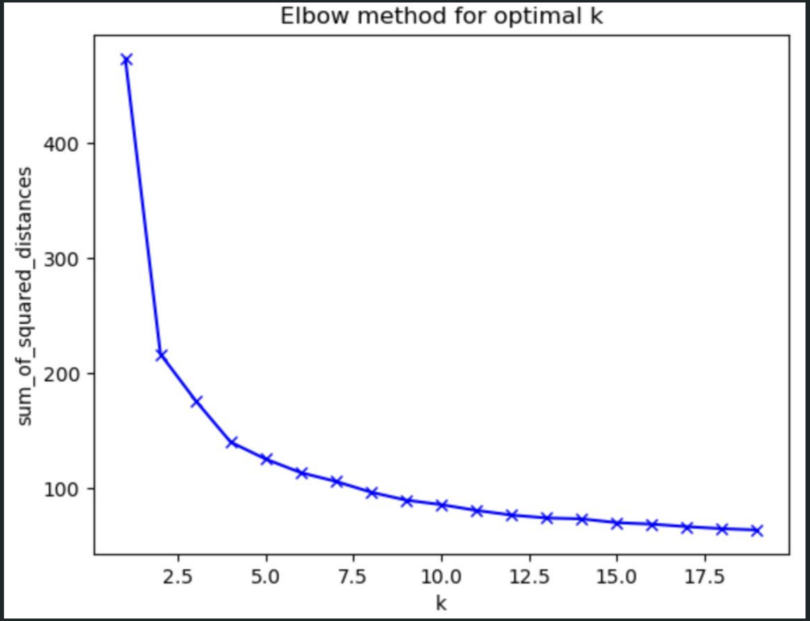
HDBSCAN Results Without PM2.5 Concentrations

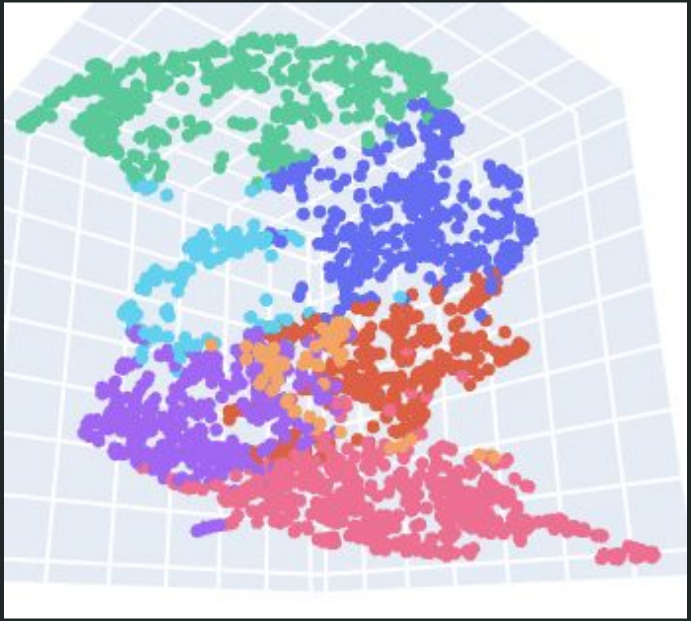


HDBSCAN Results Without PM2.5 Concentrations

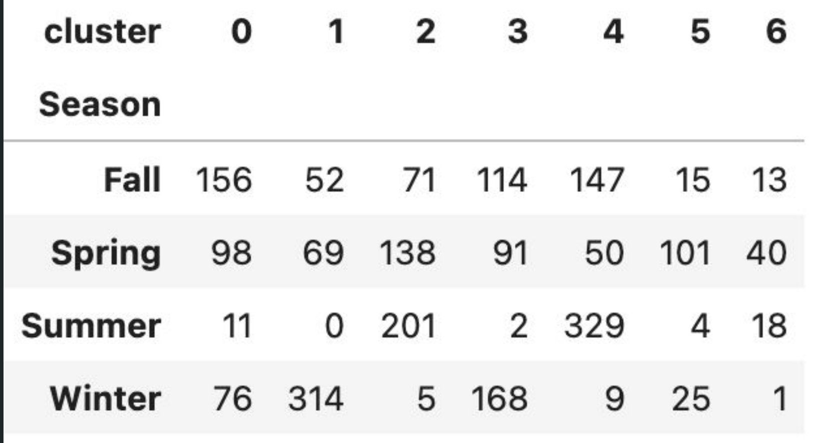


K-MEANS Results Without PM2.5 Concentrations

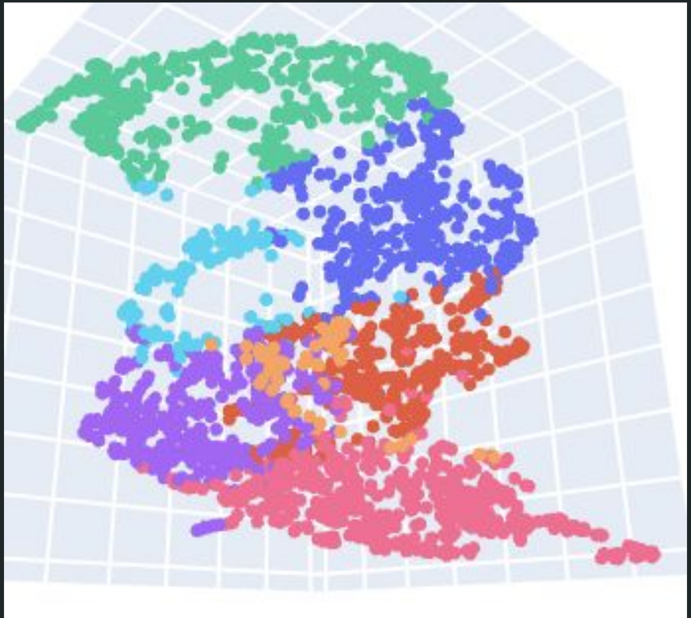




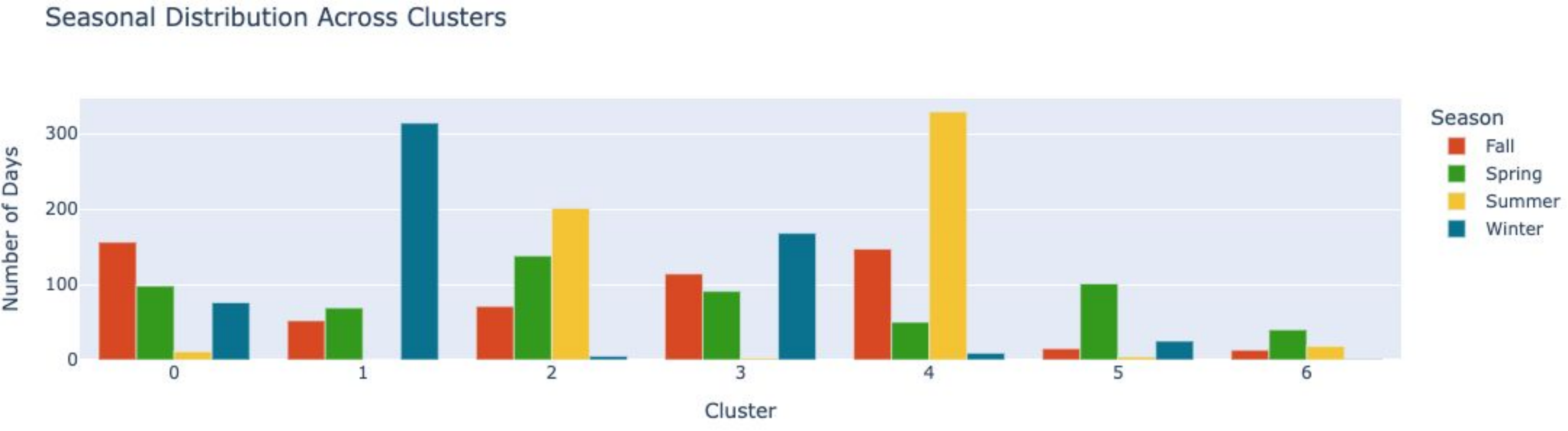
K-MEANS Results Without PM2.5 Concentrations



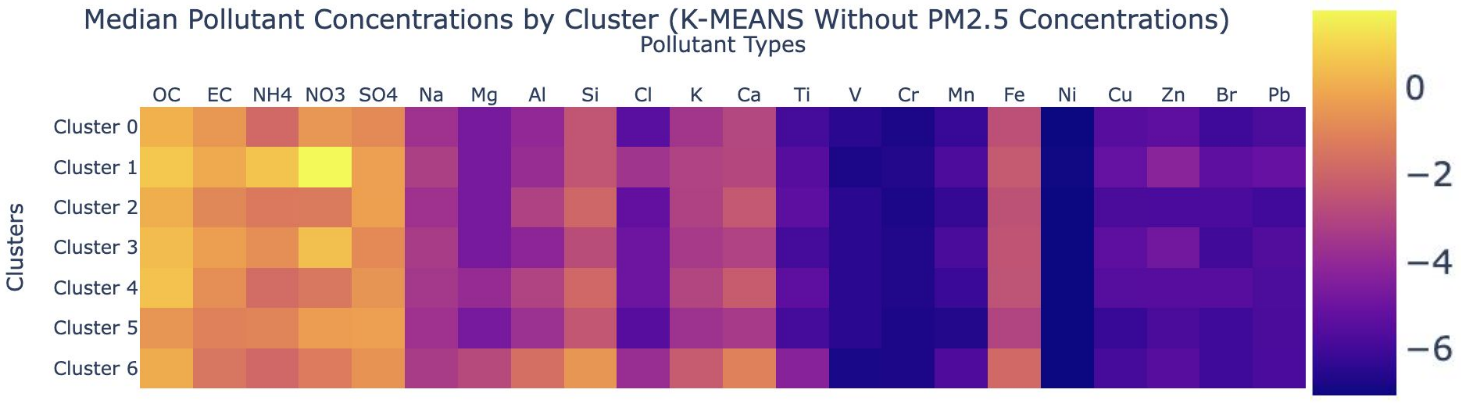
cluster	0	1	2	3	4	5	6
Season							
Fall	156	52	71	114	147	15	13
Spring	98	69	138	91	50	101	40
Summer	11	0	201	2	329	4	18
Winter	76	314	5	168	9	25	1



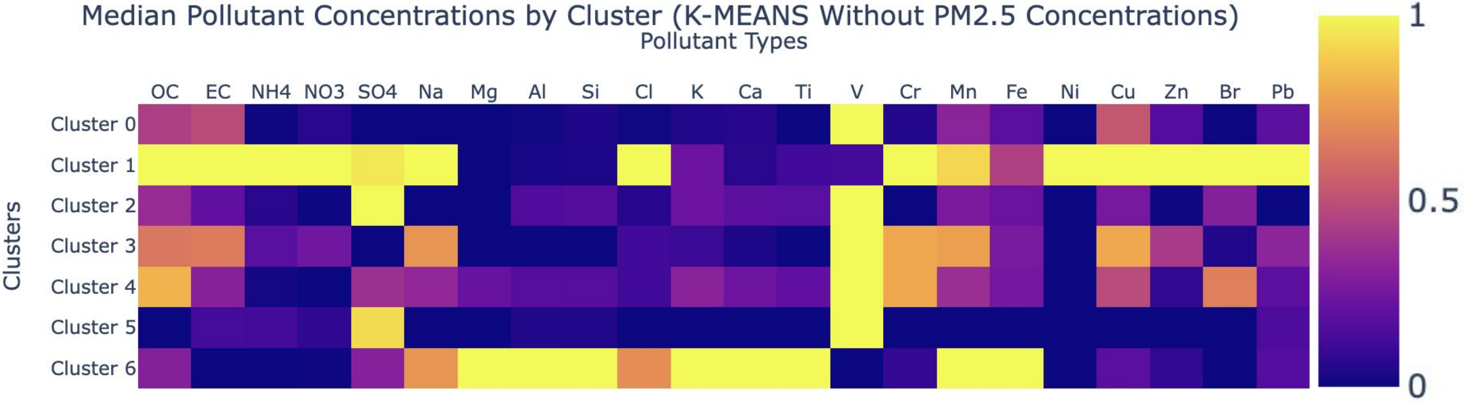
K-MEANS Results Without PM2.5 Concentrations



K-MEANS Results Without PM2.5 Concentrations

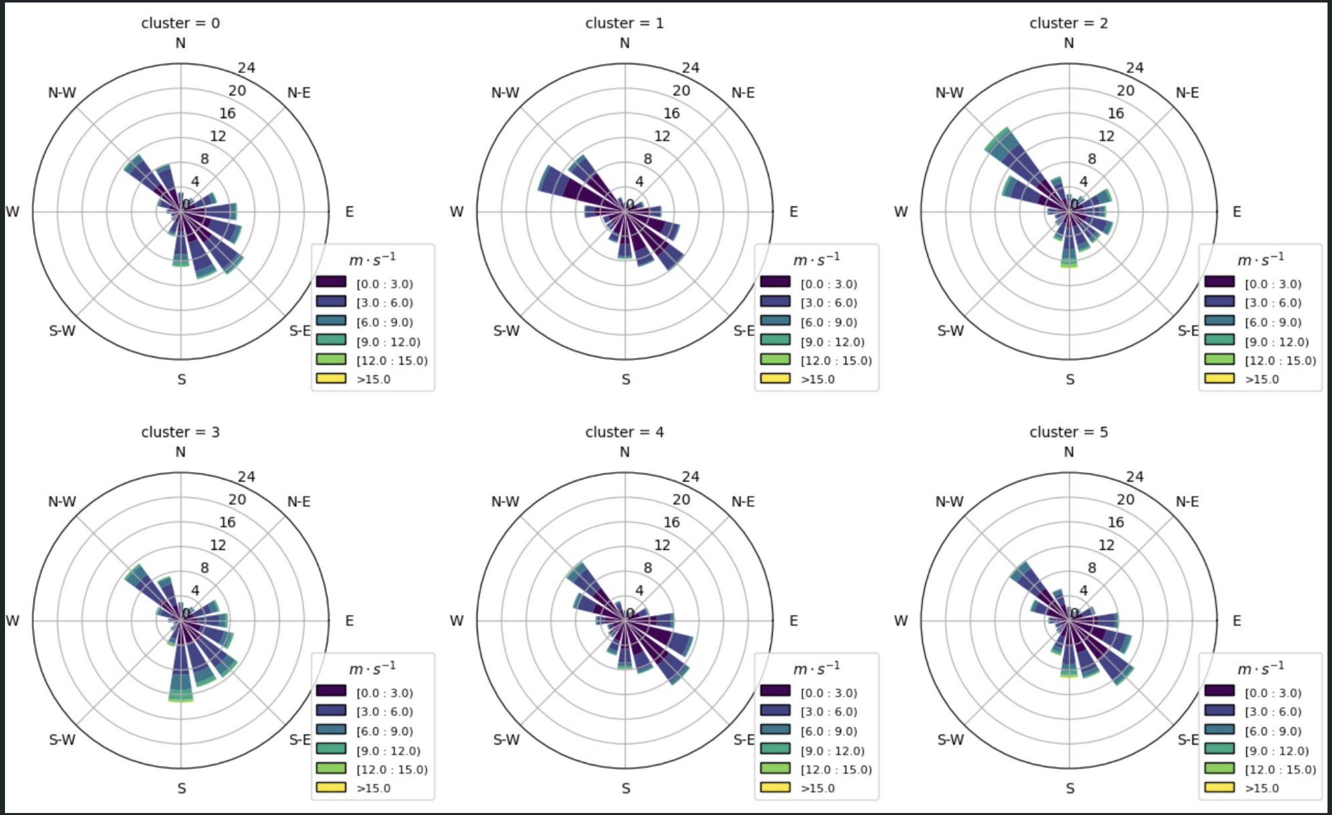


Log-Scaled Median Concentrations

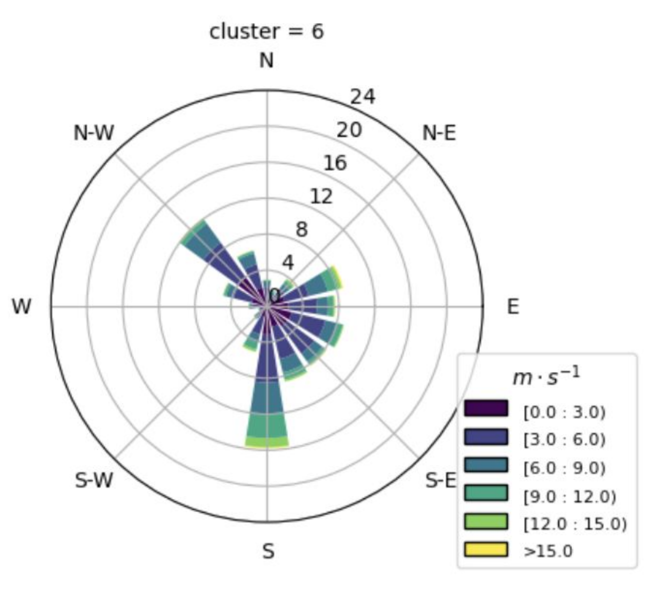


Normalized Median Concentrations

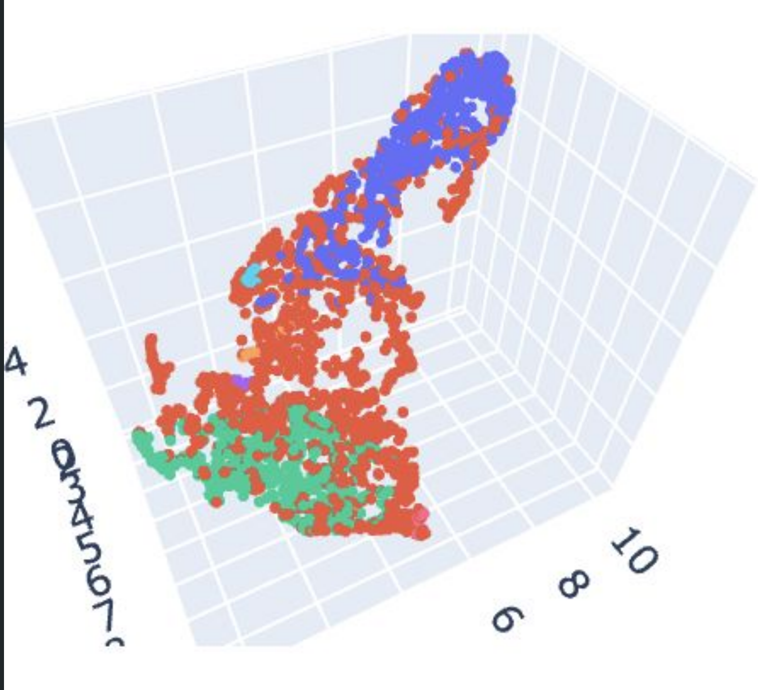
K-MEANS Results Without PM2.5 Concentrations



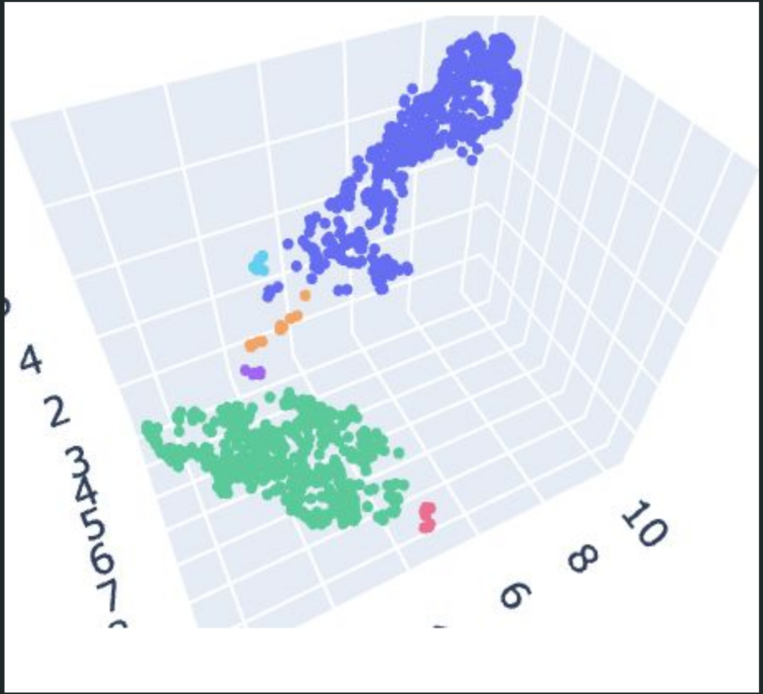
K-MEANS Results Without PM2.5 Concentrations



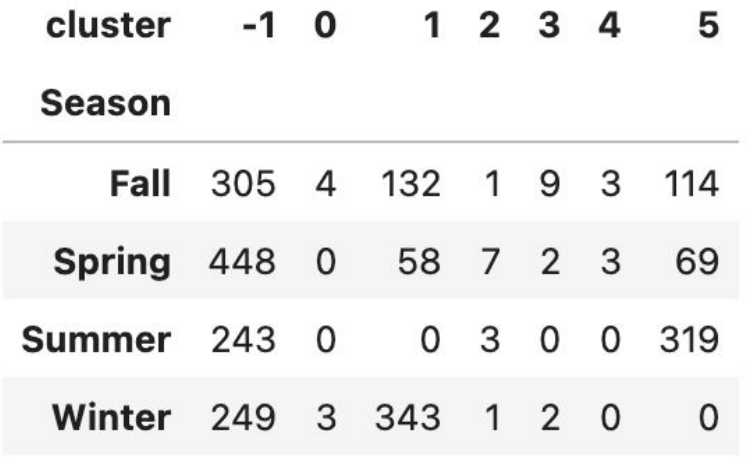
HDBSCAN Results With PM2.5 Concentrations



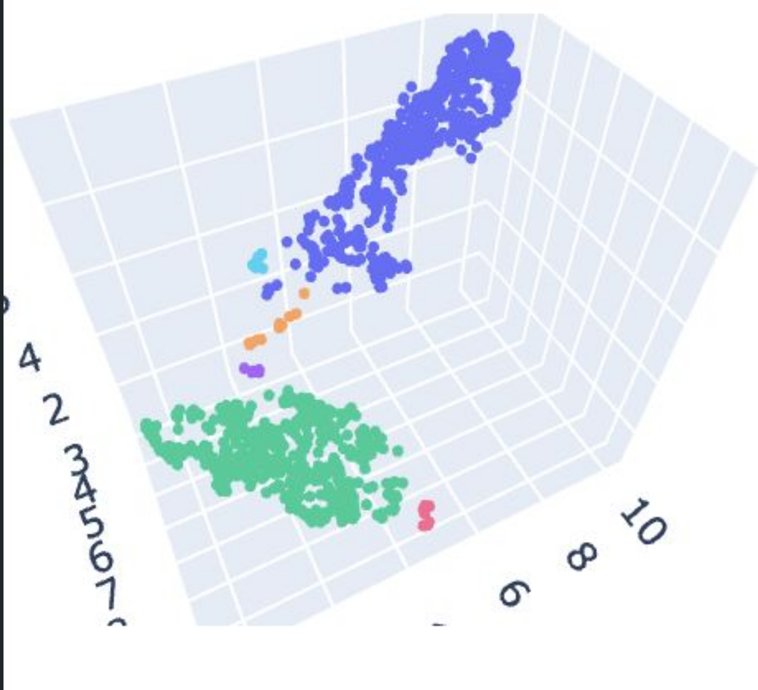
Removing Noise



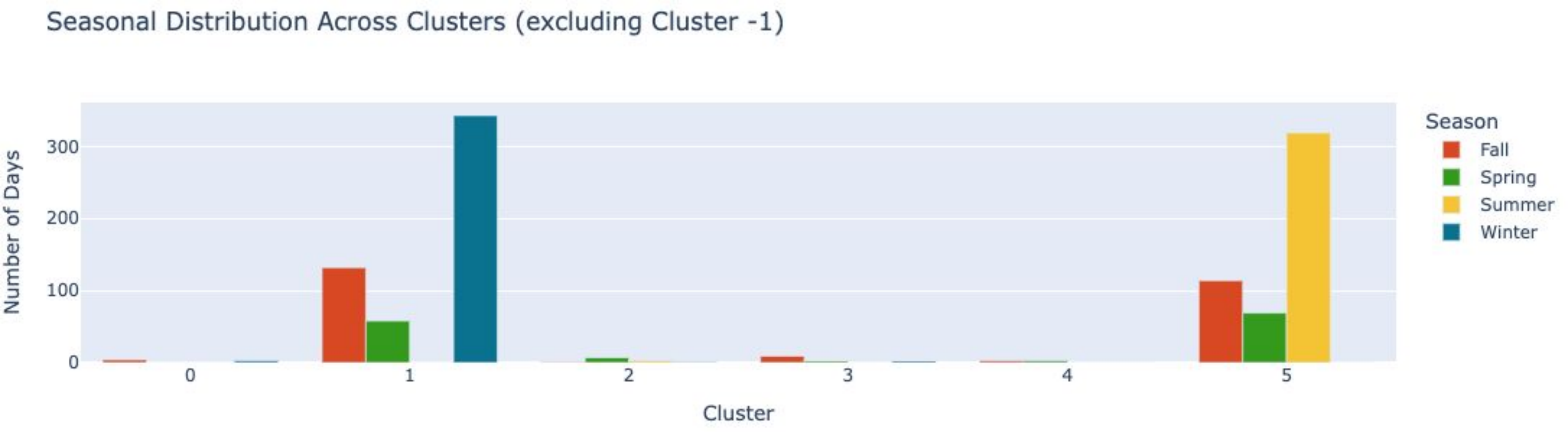
HDBSCAN Results With PM2.5 Concentrations



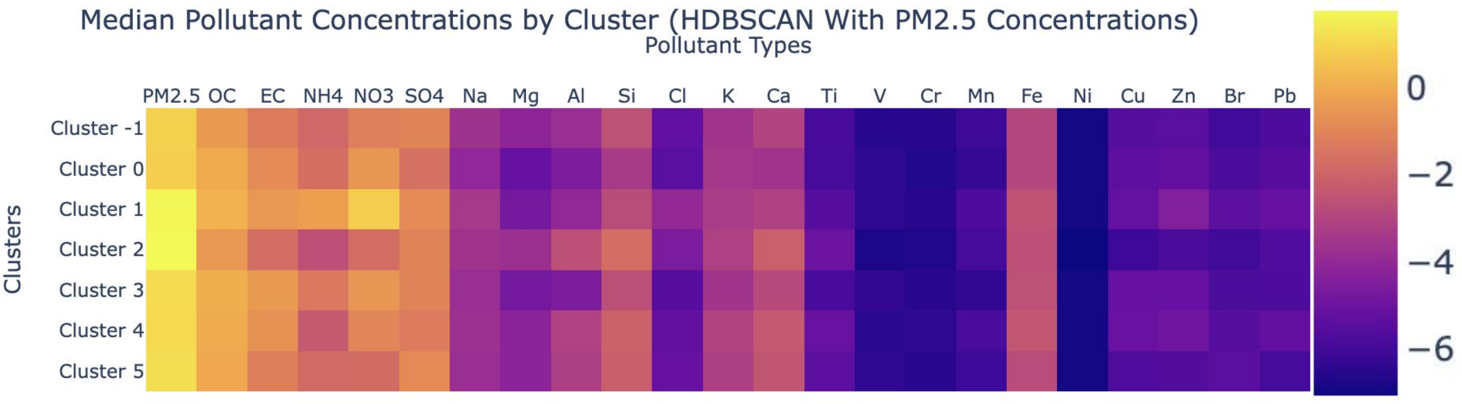
cluster	-1	0	1	2	3	4	5
Season							
Fall	305	4	132	1	9	3	114
Spring	448	0	58	7	2	3	69
Summer	243	0	0	3	0	0	319
Winter	249	3	343	1	2	0	0



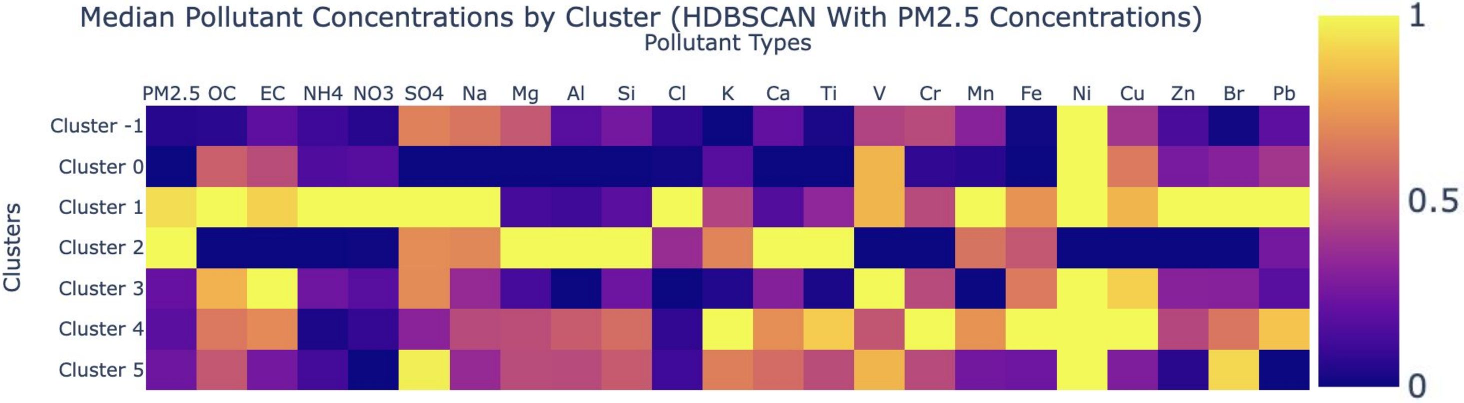
HDBSCAN Results With PM2.5 Concentrations



HDBSCAN Results With PM2.5 Concentrations

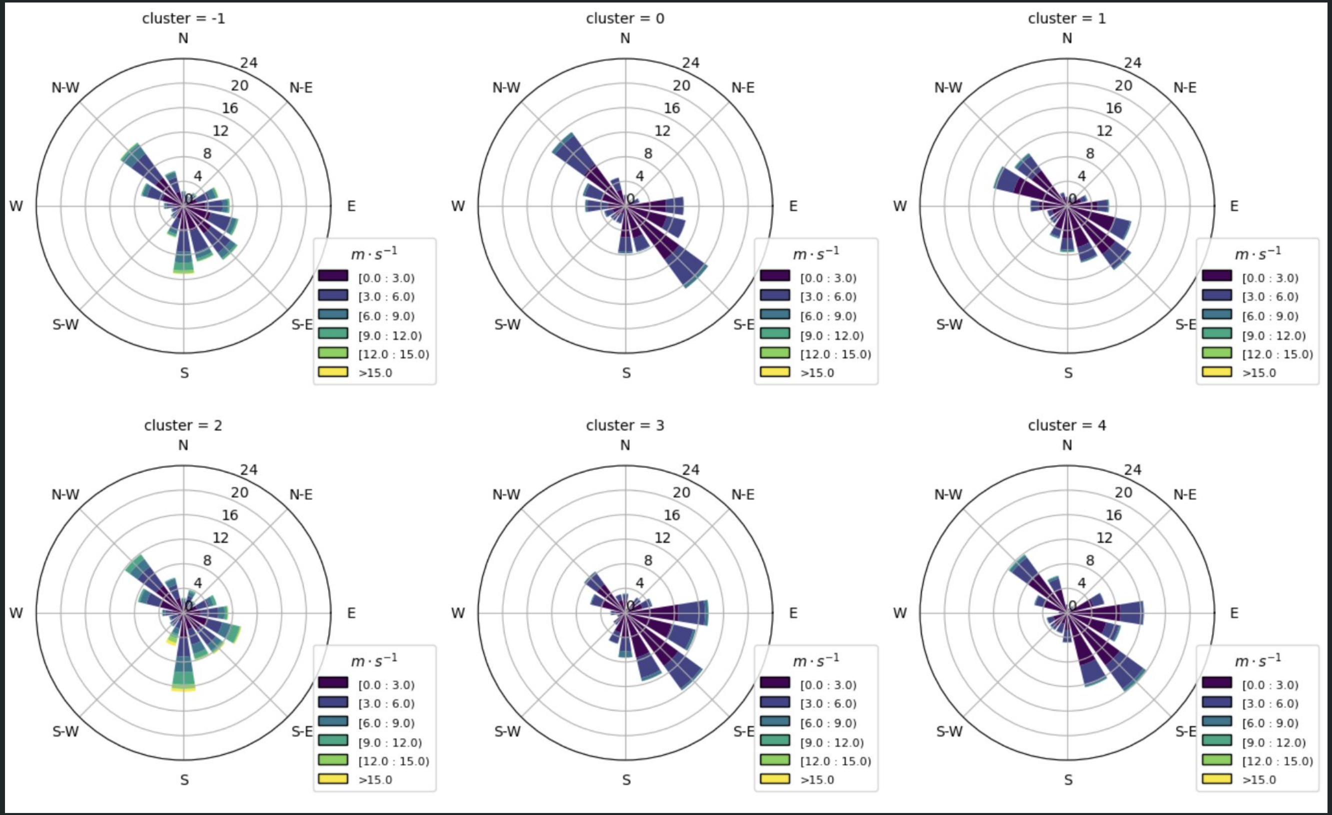


Log-Scaled Median Concentrations

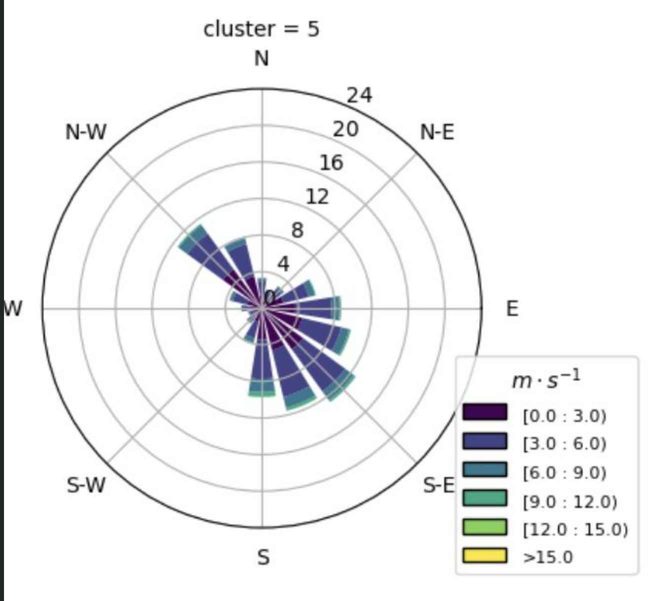


Normalized Median Concentrations

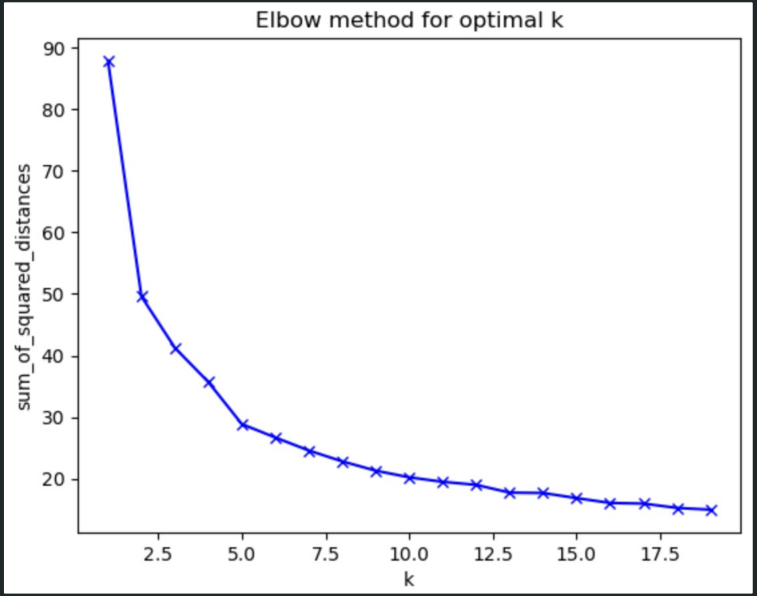
HDBSCAN Results With PM2.5 Concentrations

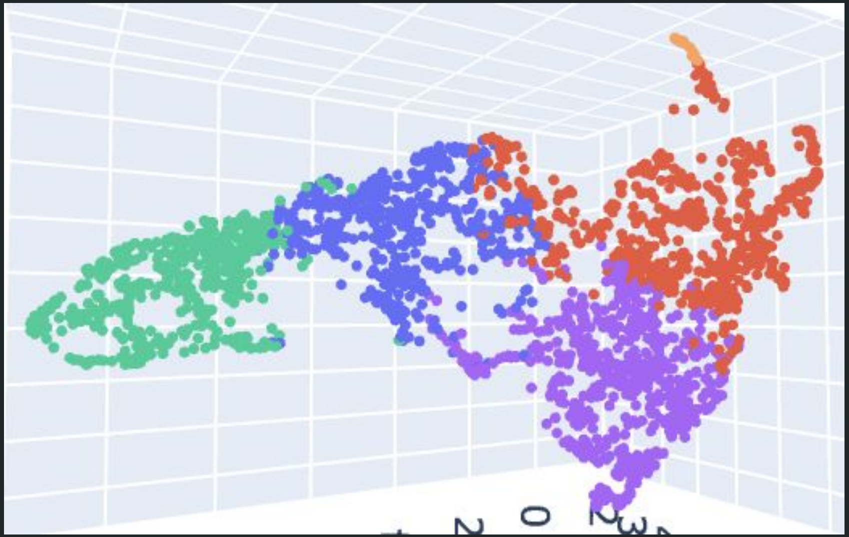


HDBSCAN Results With PM2.5 Concentrations

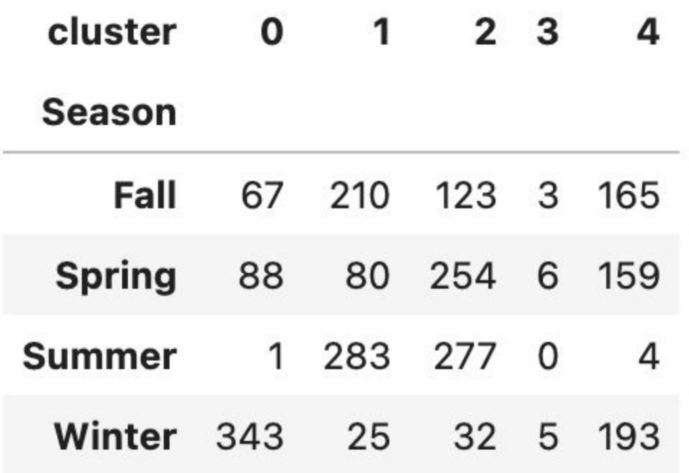


K-MEANS Results With PM2.5 Concentrations

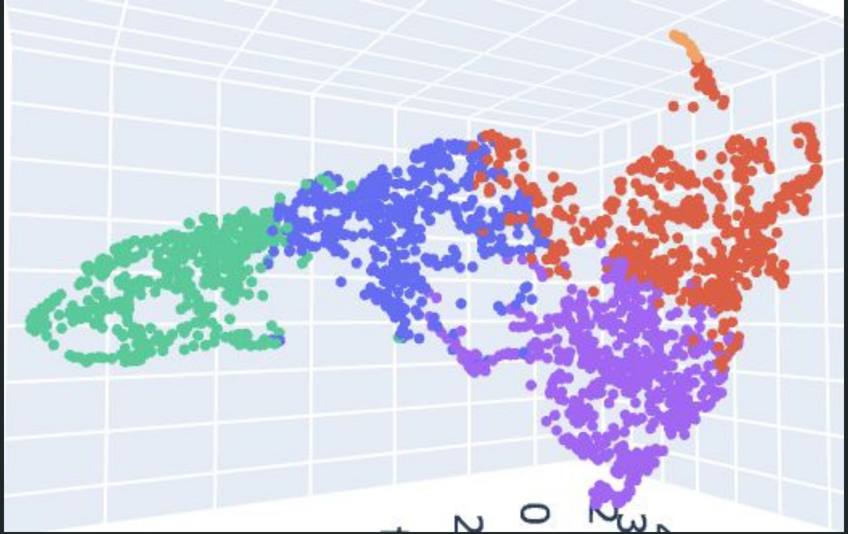




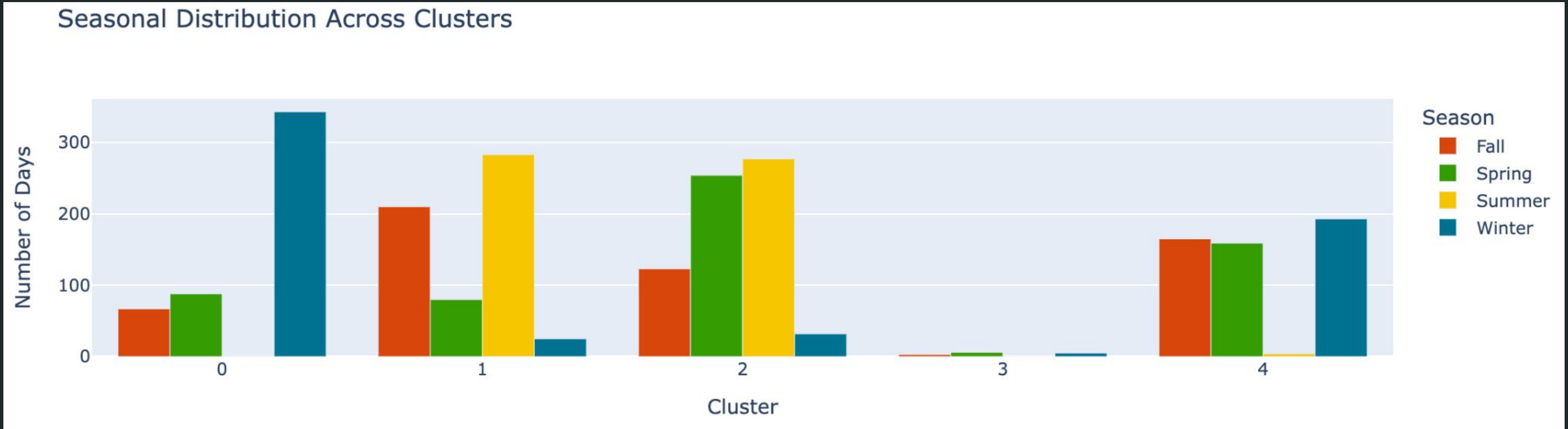
K-MEANS Results With PM2.5 Concentrations



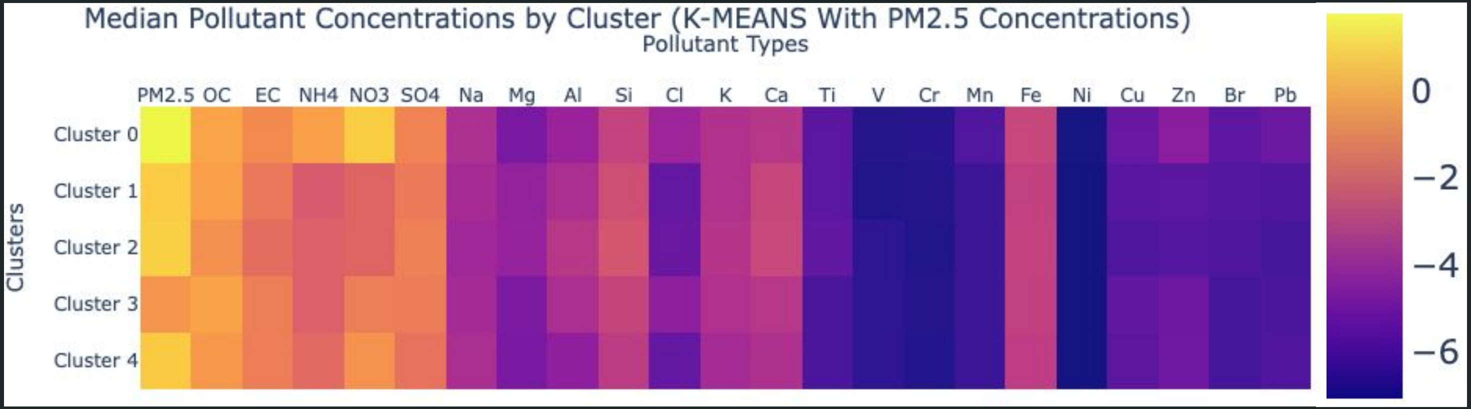
cluster Season	0	1	2	3	4
Fall	67	210	123	3	165
Spring	88	80	254	6	159
Summer	1	283	277	0	4
Winter	343	25	32	5	193



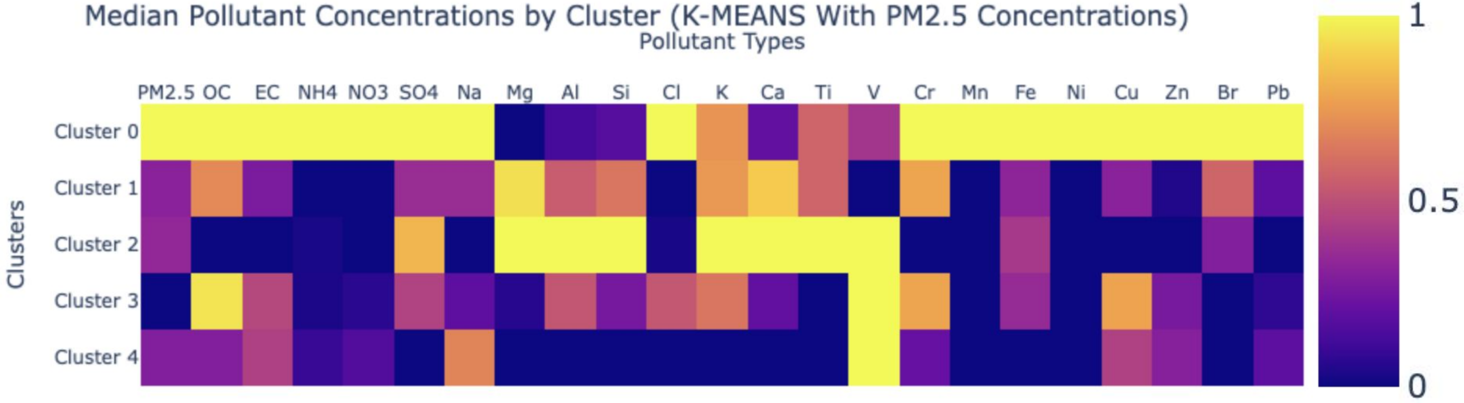
K-MEANS Results With PM2.5 Concentrations



K-MEANS Results With PM2.5 Concentrations

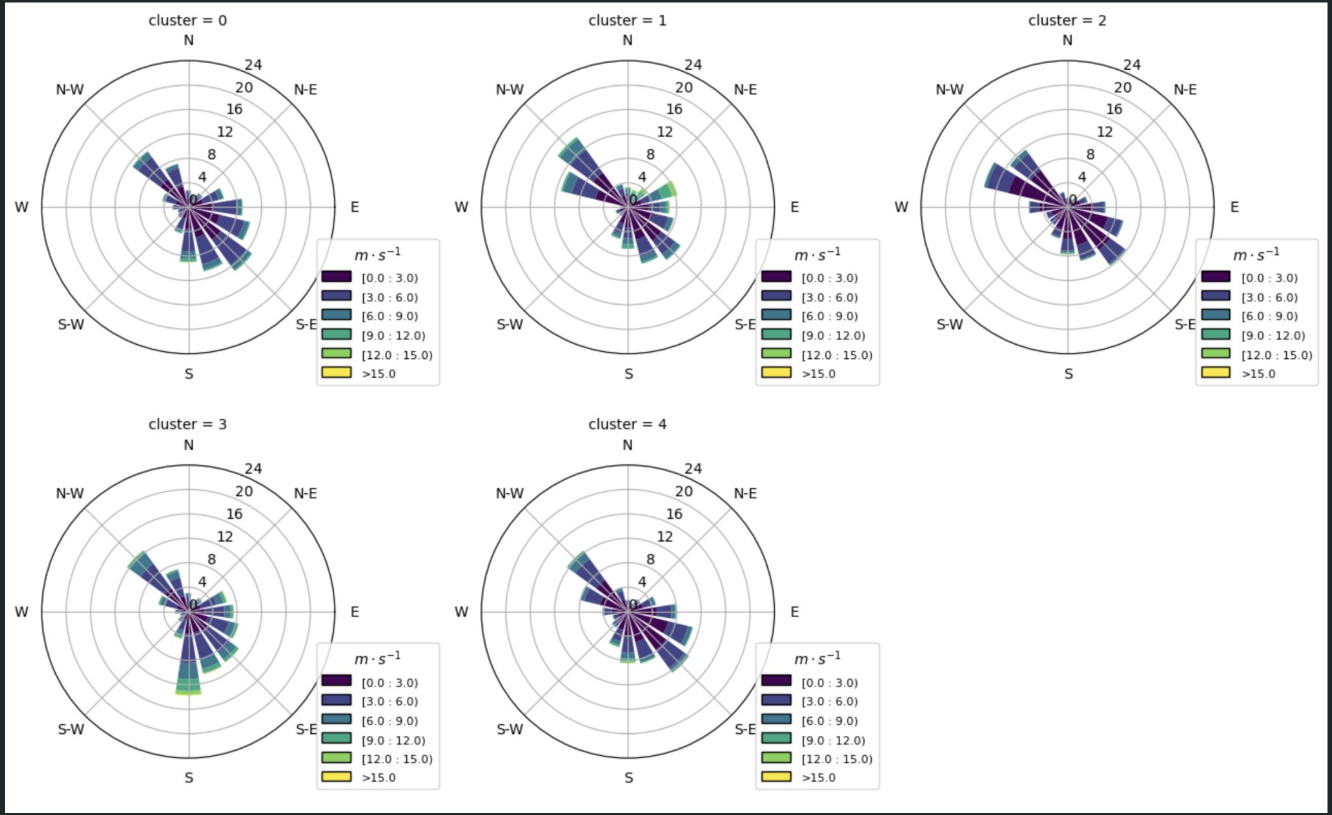


Log-Scaled Median Concentrations



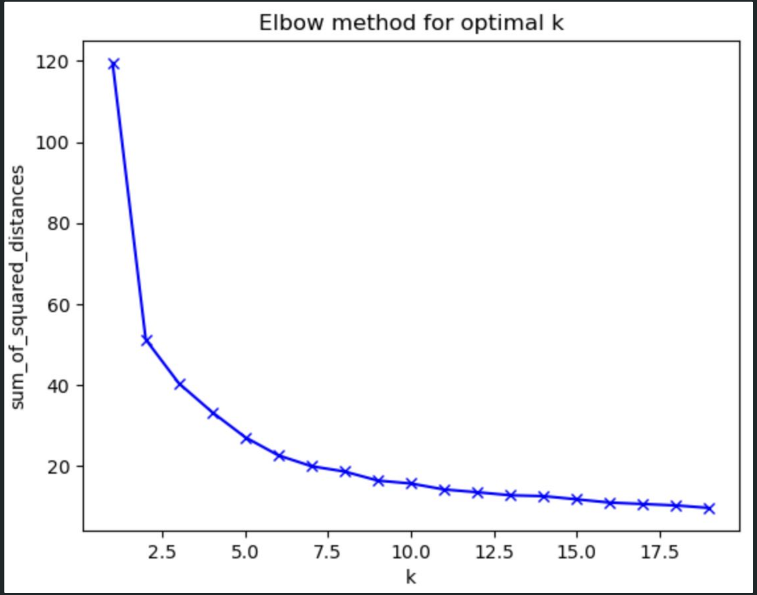
Normalized Median Concentrations

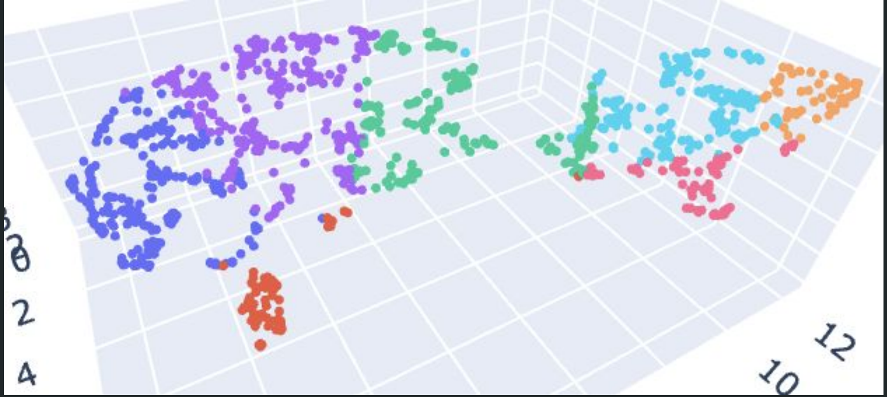
K-MEANS Results With PM2.5 Concentrations



K-MEANS on Filtered PM2.5 Species and Other Pollutants

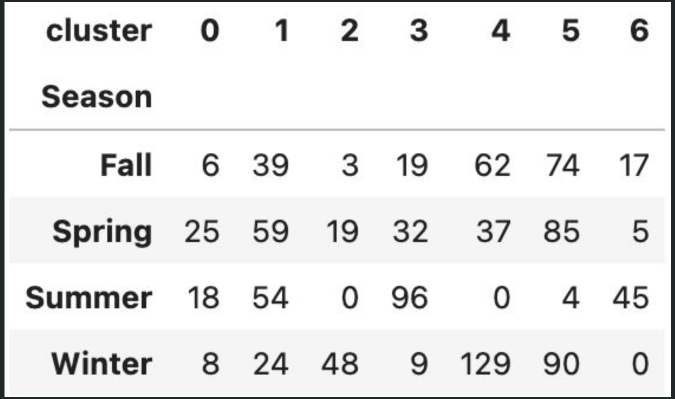
- PM2.5 Concentrations, Species Concentrations, PM10, Carbon Monoxide, Ozone, Nitrous Oxides, and Sulfur Dioxide



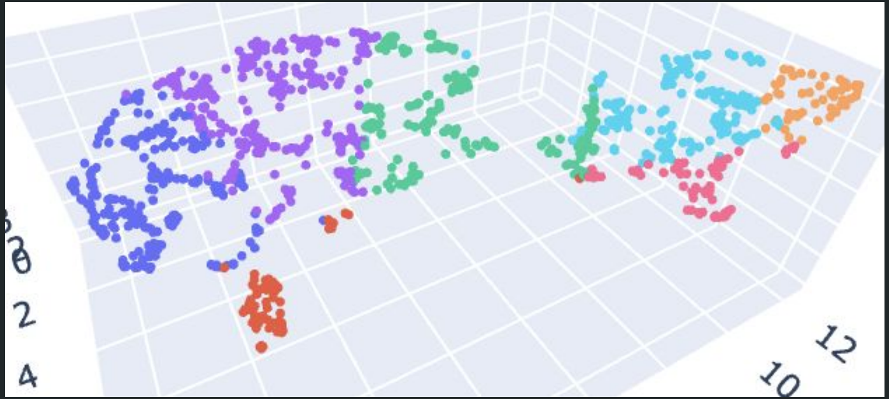


K-MEANS on Filtered PM2.5 Species and Other Pollutants

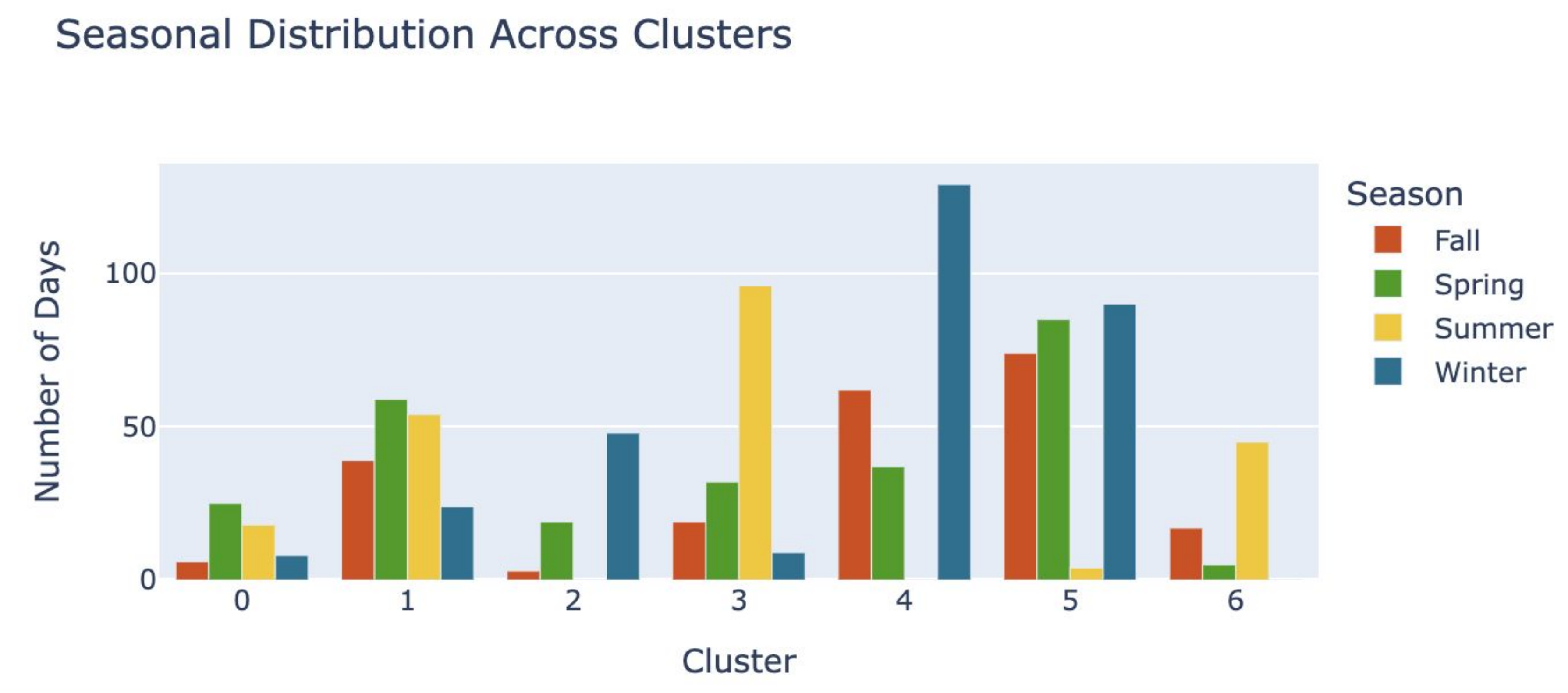
- PM2.5 Concentrations, Species Concentrations, PM10, Carbon Monoxide, Ozone, Nitrous Oxides, and Sulfur Dioxide



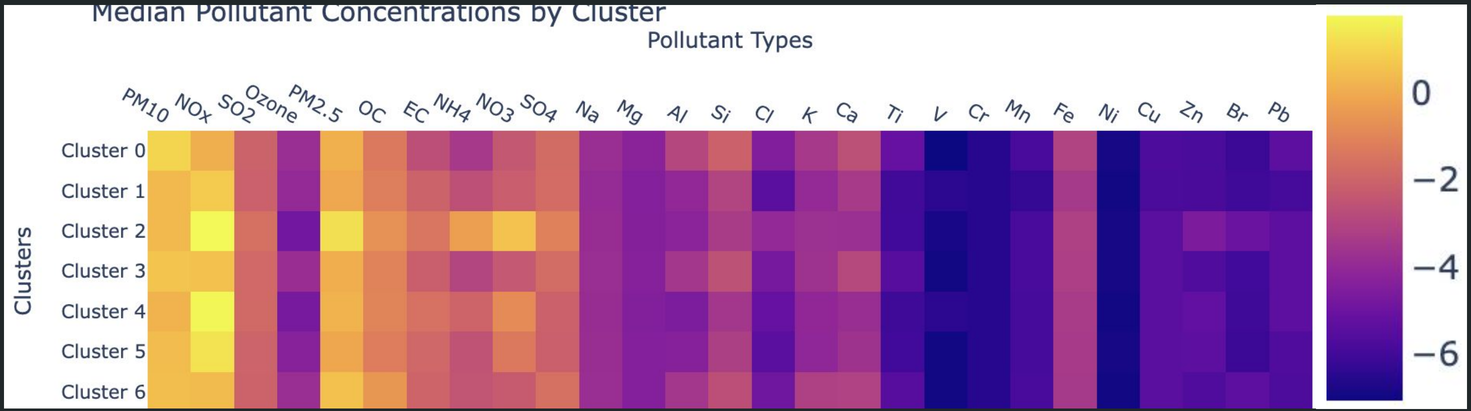
cluster	0	1	2	3	4	5	6
Season							
Fall	6	39	3	19	62	74	17
Spring	25	59	19	32	37	85	5
Summer	18	54	0	96	0	4	45
Winter	8	24	48	9	129	90	0



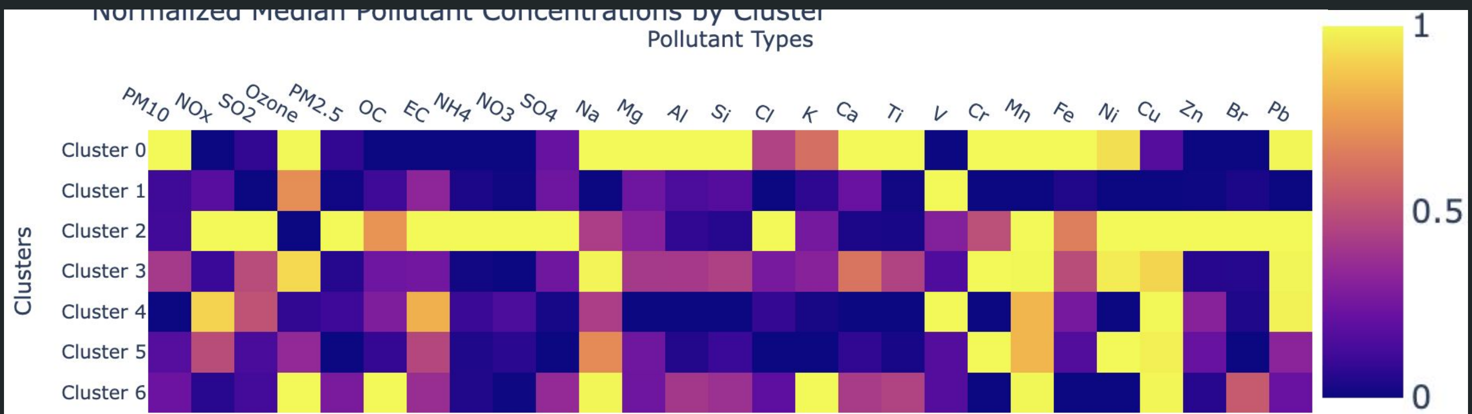
K-MEANS on Filtered PM2.5 Species and Other Pollutants



K-MEANS on Filtered PM2.5 Species and Other Pollutants

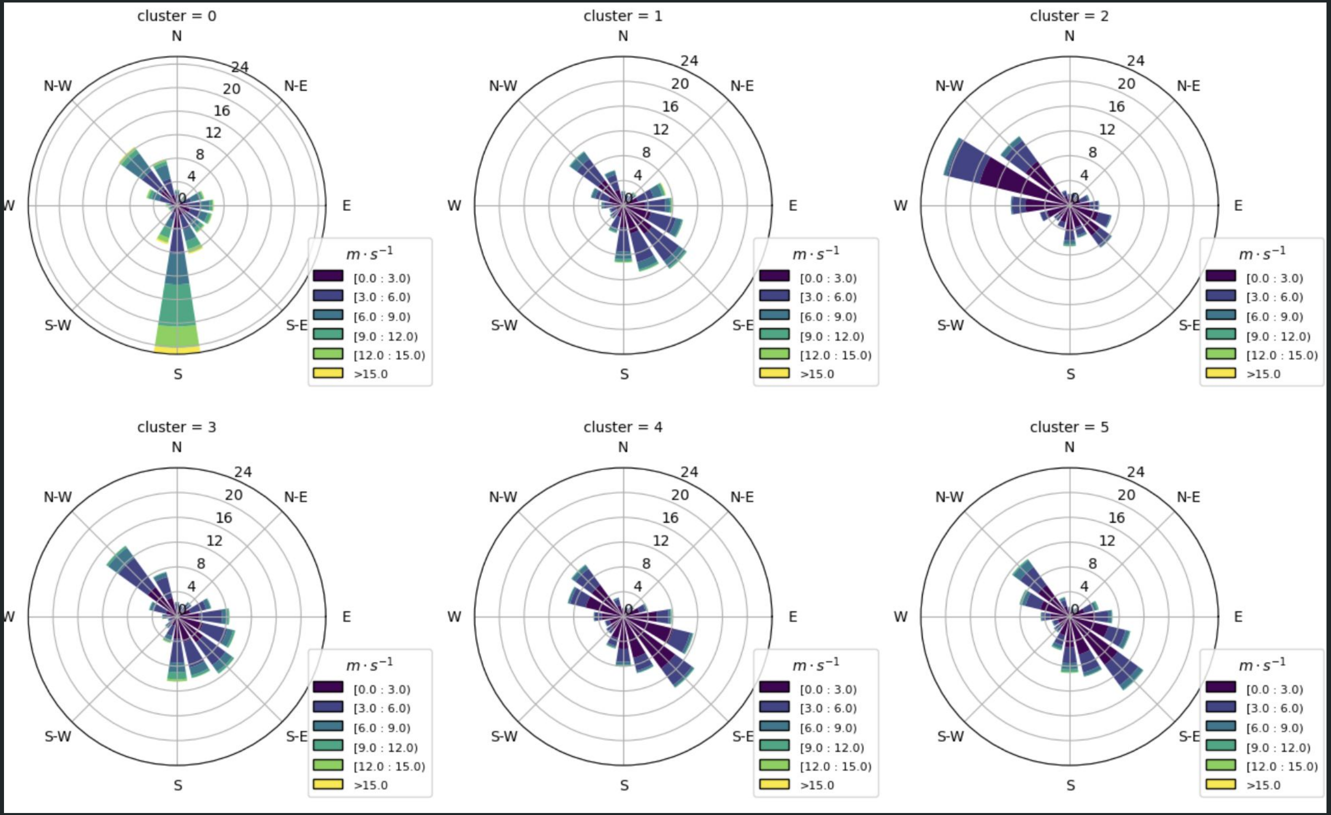


Log-Scaled Median Concentrations

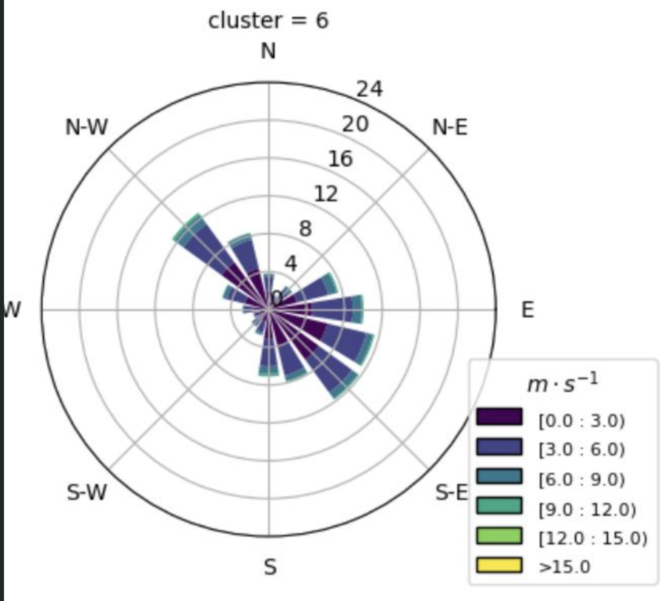


Normalized Median Concentrations

K-MEANS on Filtered PM2.5 Species and Other Pollutants



K-MEANS on Filtered PM2.5 Species and Other Pollutants


cluster = 6
N
N-E
E
S-E
S
S-W
W
N-W
$m \cdot s^{-1}$
[0.0 : 3.0)
[3.0 : 6.0)
[6.0 : 9.0)
[9.0 : 12.0)
[12.0 : 15.0)
>15.0