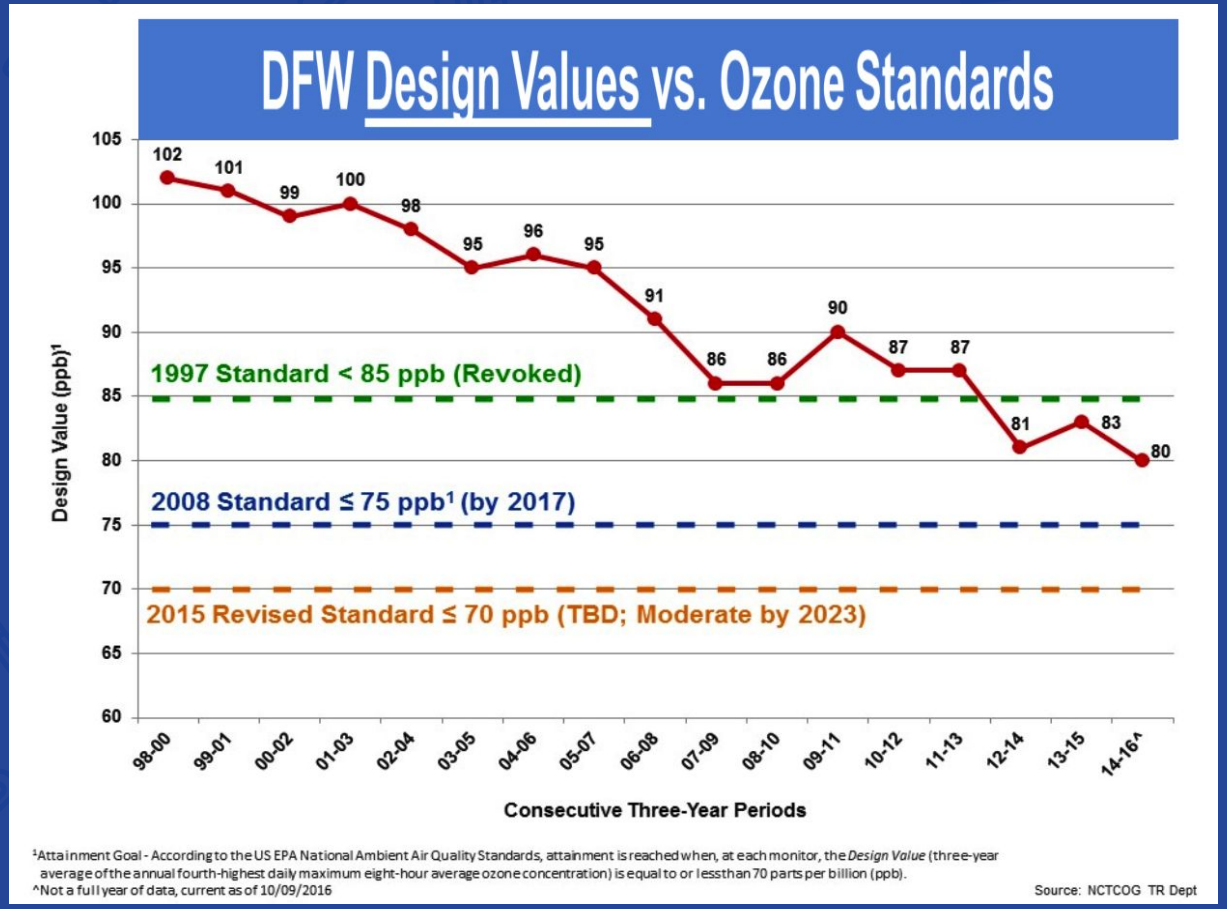
Spatial Analysis of Air Pollutants that Impact the Respiratory Health of North Central Texas by the Application of Principal Component Analysis and Hierarchical Clustering

Nguyen Cao

Andrew Hunt, PhD

RATIONALE AND OBJECTIVES



Rank	City	Avoided Deaths	Avoided Morbidities	Fewer Impacted Days
1	Los Angeles (Long Beach-Glendale), CA	1,341	3,255	2,892,029
2	Riverside (San Bernardino-Ontario), CA	808	1,416	1,321,762
3	New York (Jersey City-White Plains), NY-NJ	282	977	818,666
4	Phoenix (Mesa-Scottsdale), AZ	283	598	636,730
5	Pittsburgh, PA	285	533	281,858
6	Fresno, CA	260	672	390,551
7	Bakersfield, CA	241	333	220,722
8	Houston (The Woodlands-Sugar Land), TX	229	661	636,211
9	Cleveland (Elyria), OH	196	487	231,859
10	Cincinnati, OH-KY-IN	173	298	192,989
11	Dallas (Plano-Irving), TX	142	431	572,502

Non-attainment Counties for Ozone in the North Central Texas in 2016 (10 out of 16 counties)

DATA

Adult Asthma Discharge Dx (NCT hospitals/ Period 2010 – 2014)

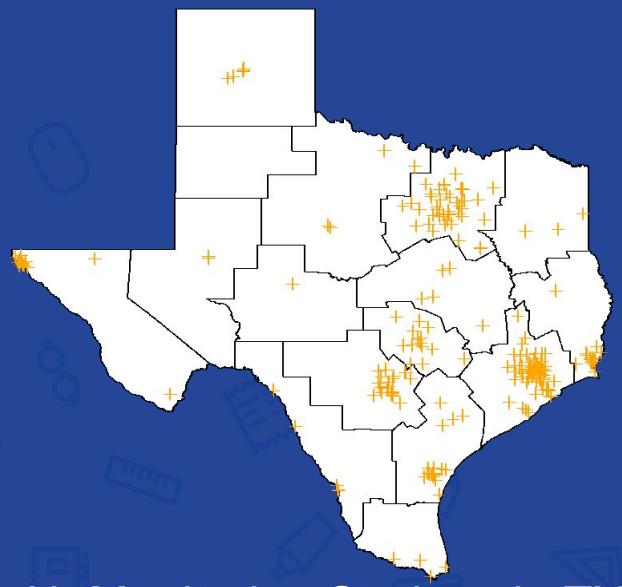
GIS Shapefiles

- Texas/ Region 4/ Active Monitoring Sites
- Obtained from the North Central Texas Council of Governments' Regional Data Site

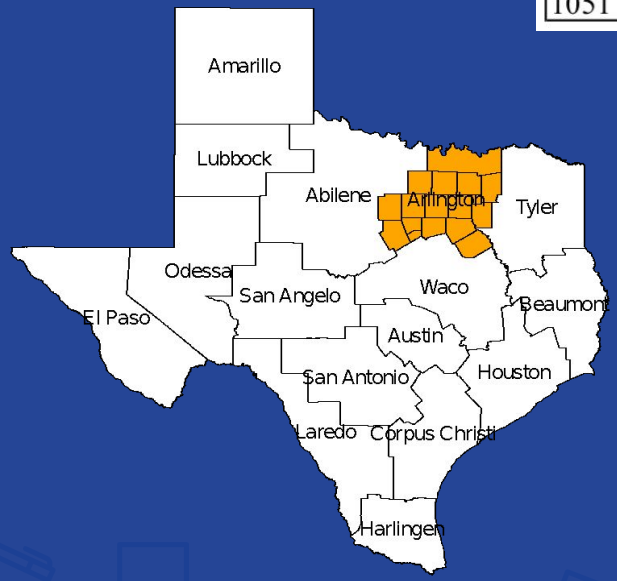
Air Pollution and Meteorological Data

- Obtained from 20 Continuous Ambient Monitoring Sites (CAMS), operated by the Texas Commission on Environmental Quality (TCEQ)
- Hourly recorded values of NO, NO2, O3, PM2.5, Outdoor Temperature, and Windspeed
- Period: 2010 – 2014

Station ID	Station Name	County	Lat	Long
13	Ft. Worth	Tarrant	32.8058183	-97.3565675
17	Keller	Tarrant	32.9224736	-97.282088
31	Firsco	Collin	33.1324003	-96.7864188
52	Midlothian OFW	Ellis	32.4820829	-97.0268987
56	Denton Airport South	Denton	33.219069	-97.1962836
60	Dallas Hinton	Dallas	32.8200608	-96.8601165
61	Arlington Municipal Airport	Tarrant	32.6563574	-97.0885849
63	Dallas North #2	Dallas	32.9192056	-96.8084975
70	Grapevine Fairway	Tarrant	32.9842596	-97.0637211
71	Kaufman	Kaufman	32.5649684	-96.3176873
75	Eagle Mountain Lake	Tarrant	32.9878908	-97.4771754
76	Parker County	Parker	32.8687727	-97.9059308
77	Cleburne Airport	Johnson	32.3535945	-97.4367419
310	Haws Athletic Center	Tarrant	32.7591432	-97.3423337
312	Convention Center	Dallas	32.7742622	-96.7976859
402	Dallas Redbird Airport Executive	Dallas	32.6764506	-96.8720596
1006	Greenville	Hunt	33.1530882	-96.1155717
1032	Pilot Point	Denton	33.4106476	-96.9445903
1044	Italy	Ellis	32.1754166	-96.8701892
1051	Corsicana Airport	Navarro	32.0319335	-96.3991408



Air Monitoring Stations in TX



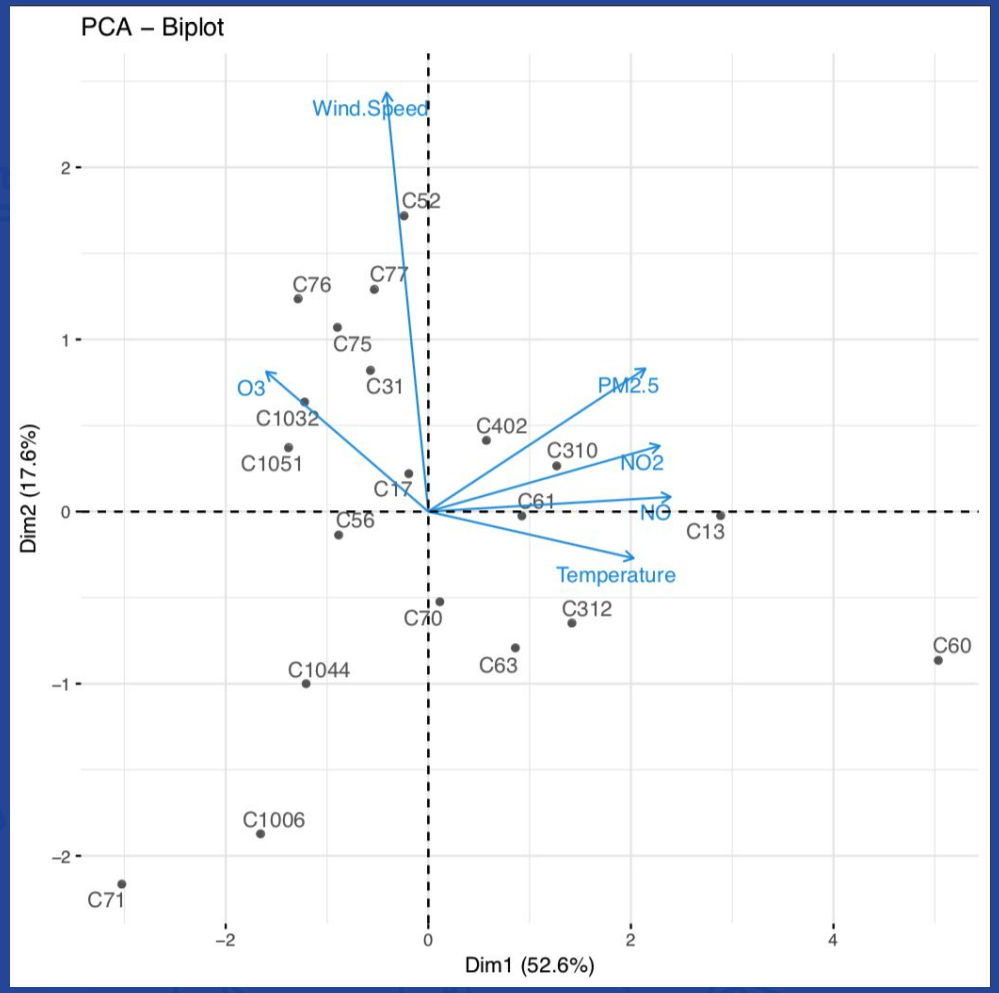
TCEQ Regions

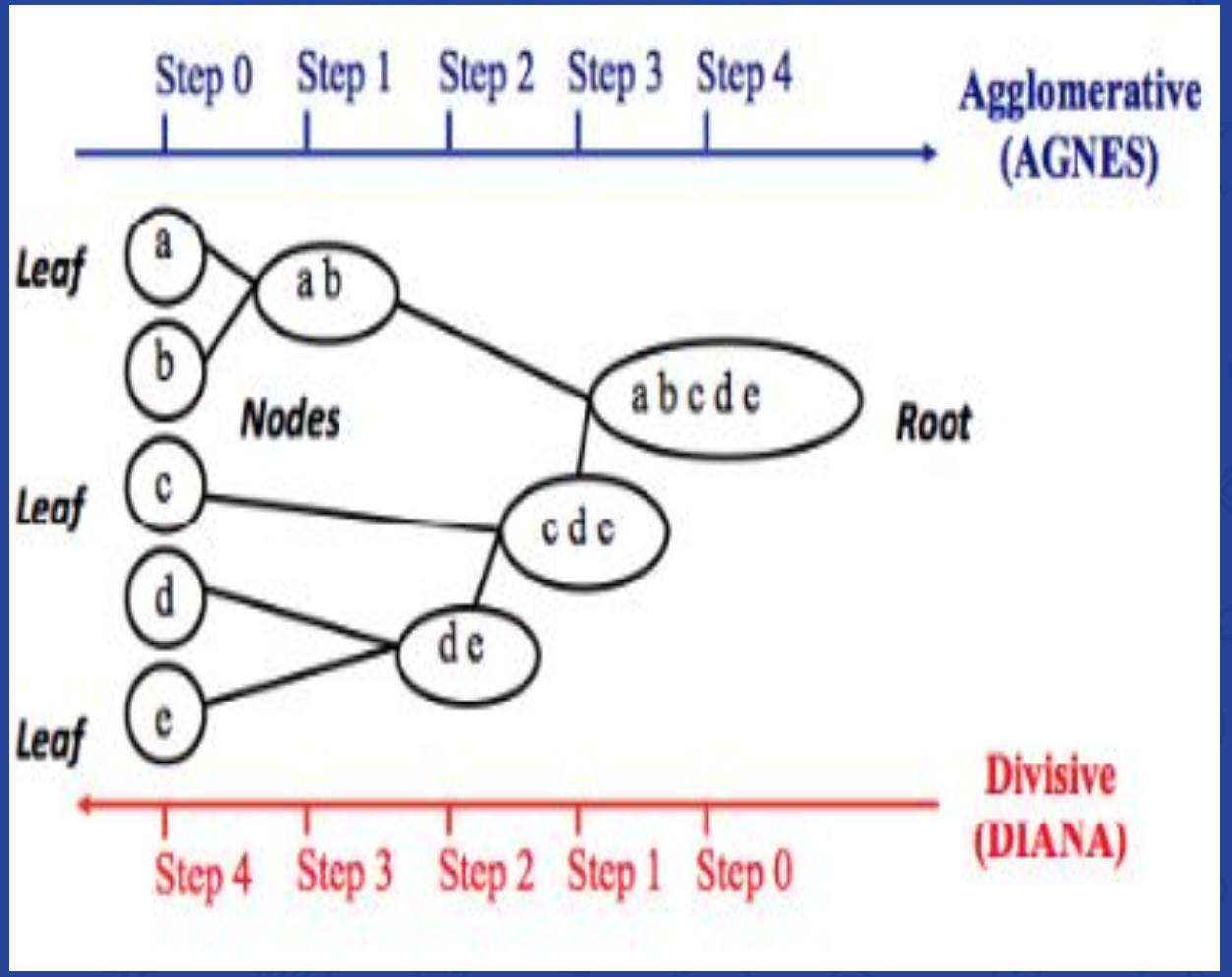
Region 4 and NCT Counties

METHODS

Multivariate Analysis

- Principal Component Analysis (PCA)
- Hierarchical Agglomerative Cluster Analysis (HACA)





Principal Component Analysis (PCA)

- Allows for the extraction of important information from a multivariate data table and the expression of such information in a set of new variables called “principal components.” These new variables are the linear combination of the original variables
- The information in the original data set corresponds to the “total variance” it contains. PCA helps identify the direction along which the variance is maximal
- In other words, PCA reduces the dimensionality of the original data set into two or three principal components (PCs), representing the majority of the total variance, with minimal loss of information
- The first PC represents the maximum of the total variance; the second PC, uncorrelated with the first PC, accounts for the maximum of the remaining variance, and so on

Principal Component Analysis (PCA)

1) Determining the Number of Principal Components to Keep

Step 1: Standardize each variable. This is indispensable because if one variable has a much larger variance compared to the rest, this variable will dominate the first PC, skewing the test

Step 2: Perform PCA. The results consist of eigenvectors and eigenvalues. The eigenvectors (dimensions) with the largest eigenvalues correspond to the dimensions having the strongest correlation in the dataset

```
eig_vals = get_eigenvalue(pca)
eig_vals
```

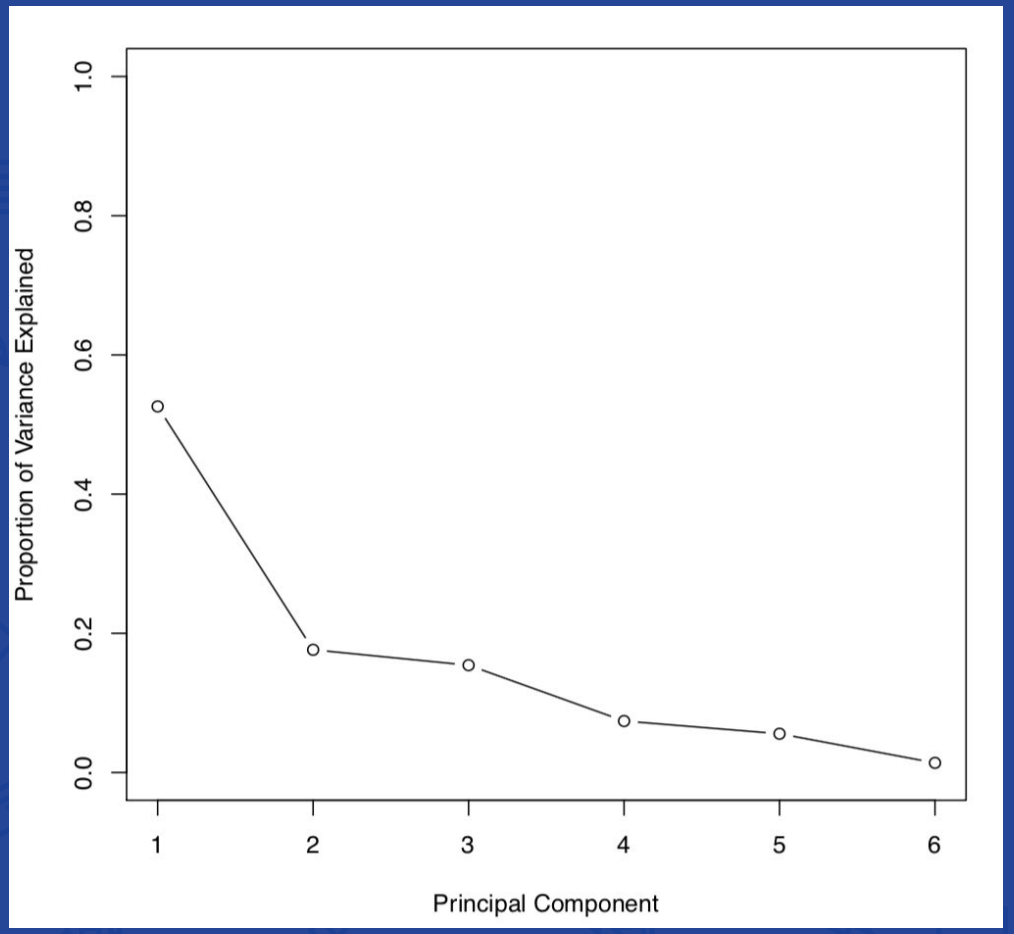
	eigenvalue	variance.percent	cumulative.variance.percent
Dim.1	2.99552695	52.596972	52.59697
Dim.2	1.00410634	17.630605	70.22758
Dim.3	0.87814205	15.418861	85.64644
Dim.4	0.42132599	7.397854	93.04429
Dim.5	0.31723260	5.570130	98.61442
Dim.6	0.07891207	1.385578	100.00000

- 2.9955/6 = 0.525969 or 52.60% => about 52.60% of the total variance is explained by the first eigenvalue
- 1.0041/6 = 0.1763 or 17.63% => about 17.63% of the total variance is explained by the second eigenvalue
- Together, the first two dimensions (eigenvalues) explain 70.23% of the total variance

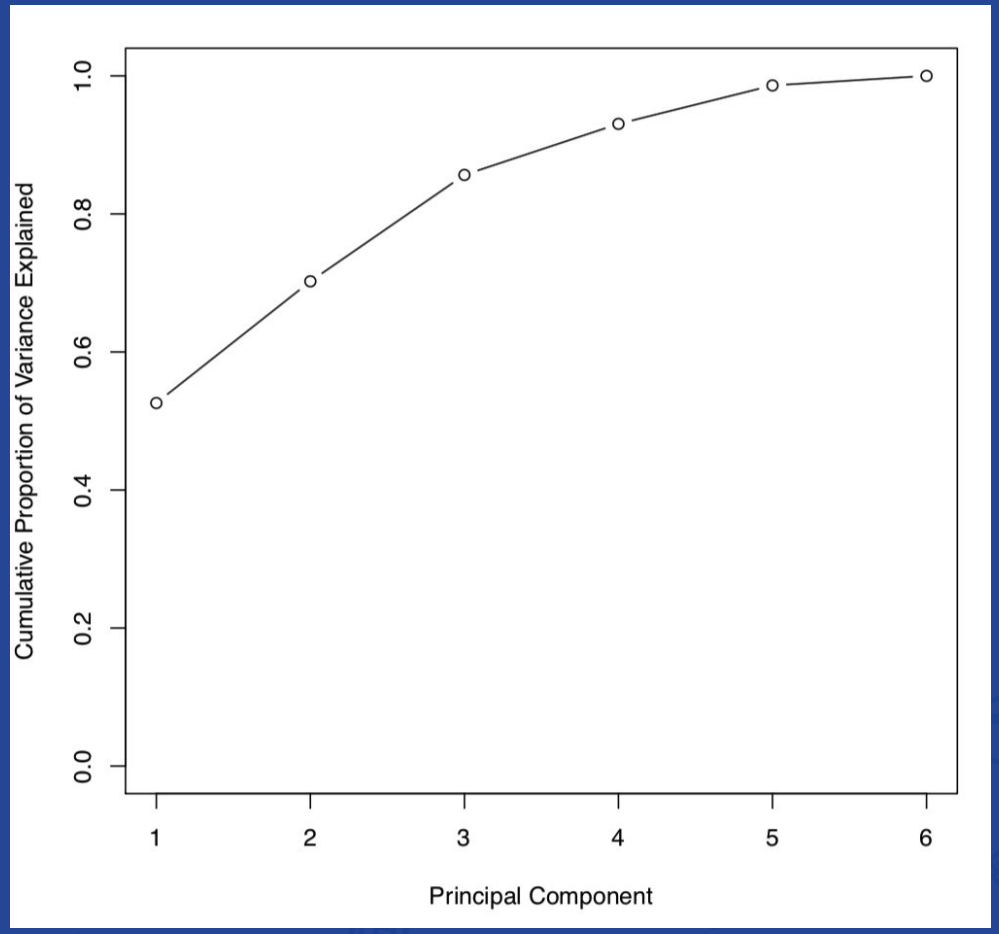
Principal Component Analysis (PCA)

1) Determining the Number of Principal Components to Keep

Step 3: Perform Scree Test. When a proper number of principal components is obtained, the residual variance will level off, indicated by the sudden drop in slope



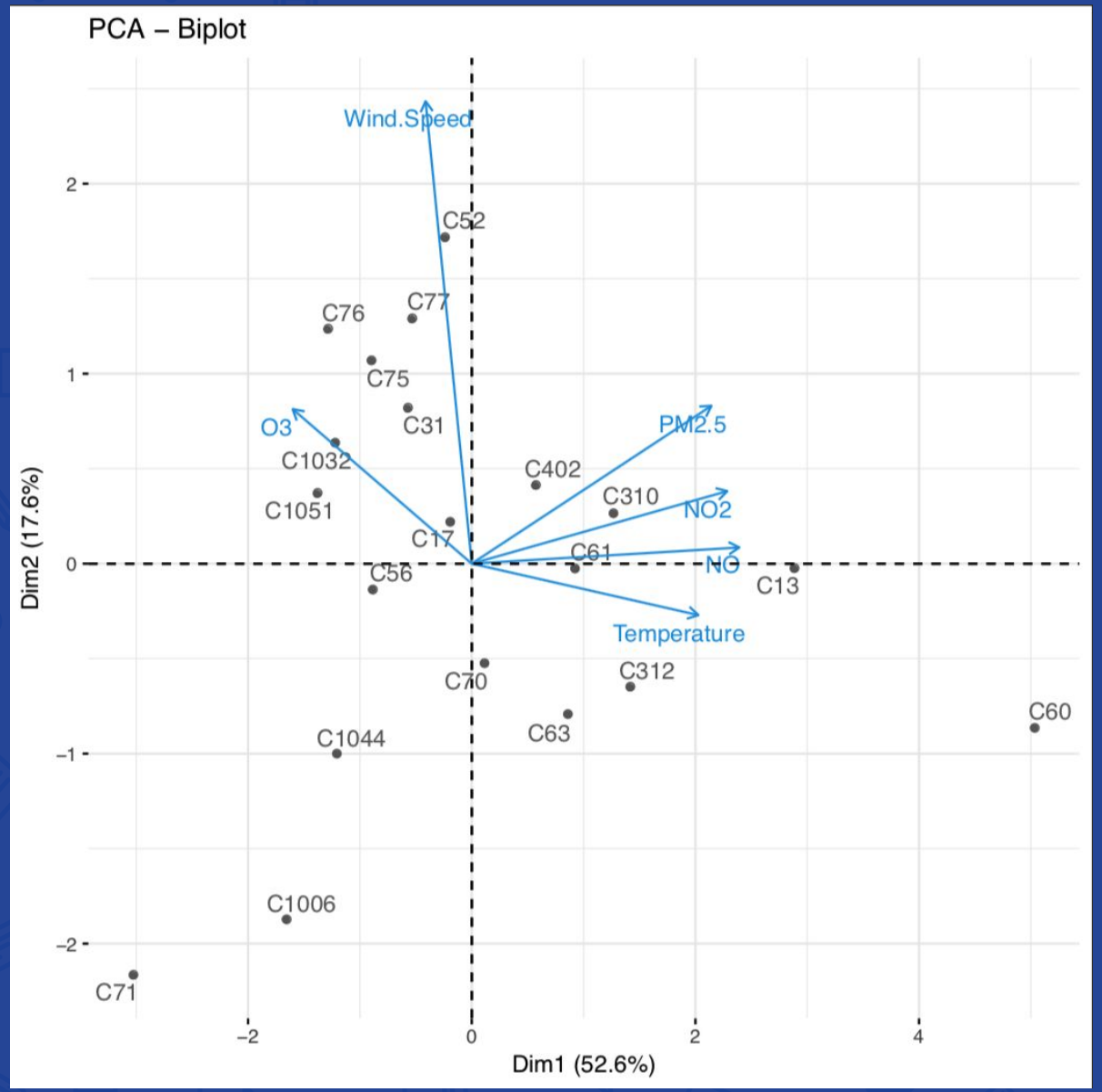
Proportion of Variance Explained



Cumulative Proportion of Variance Explained

Principal Component Analysis (PCA)

2) Interpretation of Principal Components



- NO, NO2, PM2.5, and Temperature are highly correlated
- O3 is more highly correlated with windspeed then it is temperature and the rest of the pollutants
- NOx and O3 are in opposite directions as the appearance of one variable will favor the disappearance of the other variable (NOx plays a key role in the formation of tropospheric O3)
- The first loading vector (1st dimension) places approximately equal weight on NO2, NO, PM2.5, and temperature
- The second loading vector (2nd dimension) corresponds to wind speed

HACA: Groups data into clusters with low intra-group dissimilarity and high inter-group dissimilarity

1) Computing 5-year average of all variables

Station/ Pollutant	NO	NO2	O3	PM2.5	Wind.Speed	Temperature
C13	3.91	9.25	27.2	10.29	7.68	67.13
C17	1.22	7.83	31.08	9.45	7.67	66.37
C31	1.74	7.31	33.36	9.32	8.17	66.32
C52	1.48	5.5	30.16	9.78	10.52	66.88
C56	1.71	6.5	29.85	8.63	7.73	65.51
C60	4.73	11.91	27.78	11.28	5.6	68.34
C61	2.52	7.57	28.19	9.15	8.15	66.77
C63	2.18	8.09	30.17	9.62	6.17	66.75
C70	1.42	7.31	30.45	9.51	6.69	66.63
C71	0.6	3.83	31.49	5.85	5.85	66.29
C75	0.85	6.15	29.25	9.41	9.62	65.83
C76	1.8	7.31	35.18	9.25	8.29	65.58
C77	1.74	6.72	30.66	9.28	9.62	66
C310	2.09	7.61	28.46	10.38	8.03	66.872
C312	1.81	7.08	28.96	10.71	6.44	67.36
C402	1.49	8.14	28.7	9.83	8.35	66.42
C1006	0.75	4.67	29.7	9.198	4.82	64.952
C1032	1.74	7.36	33.04	9.184	7.76	65.04
C1044	0.62	3.96	28.75	8.88	7.01	66.432
C1051	0.51	3.6	30.48	9.1	8.87	66.97

← 5-year Averages

2) Scaling the results

Station/ Pollutant	NO	NO2	O3	PM2.5	Wind.Speed
C13	2.06	1.21	-1.49	0.84	0.02
C17	-0.5	0.48	0.47	0.04	0.01
C31	-0.01	0.22	1.62	-0.08	0.36
C52	-0.25	-0.71	0.01	0.36	1.99
C56	-0.03	-0.2	-0.15	-0.73	0.05
C60	2.84	2.57	-1.19	1.78	-1.42
C61	0.74	0.35	-0.99	-0.24	0.35
C63	0.41	0.62	0.01	0.2	-1.03
C70	-0.31	0.22	0.15	0.1	-0.67
C71	-1.09	-1.56	0.68	-3.37	-1.25
C75	-0.85	-0.38	-0.45	0	1.36
C76	0.05	0.22	2.54	-0.15	0.44
C77	-0.01	-0.08	0.26	-0.12	1.36
C310	0.33	0.37	-0.85	0.92	0.26
C312	0.06	0.1	-0.6	1.24	-0.84
C402	-0.24	0.64	-0.73	0.4	0.48
C1006	-0.95	-1.13	-0.22	-0.2	-1.96
C1032	-0.01	0.24	1.46	-0.21	0.07
C1044	-1.07	-1.49	-0.7	-0.5	-0.45
C1051	-1.18	-1.68	0.17	-0.29	0.84

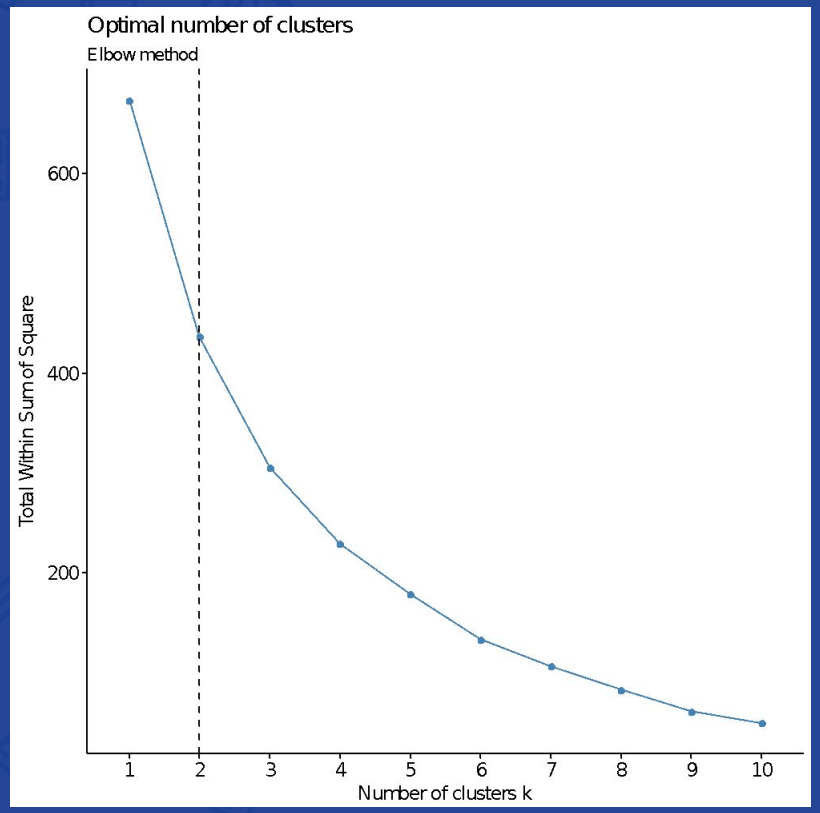
← Scaled 5-year Averages

3) Computing Euclidean distance based on the distance (values of pollutant concentrations/ wind speed/ temperature) between every pair of stations in the dataset

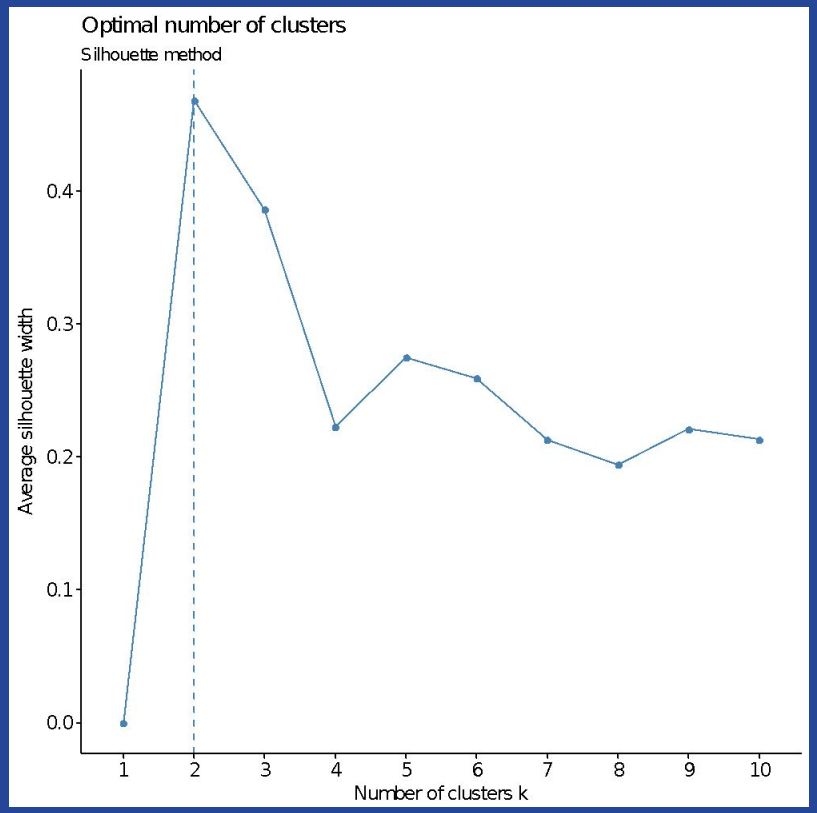
	C13	C17	C31	C52	C56	C60	C61	C63	C70	C71	C75	C76	C77	C310	C312	C402	C1006	C1032	C1044	C1051
C13	0	3.53	4.11	3.93	3.83	2.79	2.05	2.66	3.27	6.54	4.15	5.11	3.7	2.07	2.65	2.72	5.34	4.67	4.53	4.86
C17	3.53	0	1.33	2.47	1.68	5.42	2.03	1.55	0.88	4.22	2	2.42	1.63	1.92	2.3	1.37	3.21	2.02	2.47	2.56
C31	4.11	1.33	0	2.62	2.2	5.88	2.78	2.29	1.87	4.32	2.59	1.31	1.76	2.78	3.13	2.45	3.79	1.63	3.21	2.83
C52	3.93	2.47	2.62	0	2.87	6.22	2.48	3.37	2.85	5.18	1.71	3.56	1.53	2.36	3.2	2.23	4.73	3.51	2.97	1.89
C56	3.83	1.68	2.2	2.87	0	6.16	2.1	2.31	1.87	3.64	1.79	2.81	1.63	2.57	3.21	1.96	2.55	1.84	2.16	2.78
C60	2.79	5.42	5.88	6.22	6.16	0	4.52	4.21	5.02	8.31	6.6	6.8	6.09	4.26	4.04	4.96	7.14	6.61	6.62	6.96
C61	2.05	2.03	2.78	2.48	2.1	4.52	0	1.81	1.9	4.74	2.41	3.89	2.07	1.25	2.2	1.31	4.01	3.36	2.76	3.08
C63	2.66	1.55	2.29	3.37	2.31	4.21	1.81	0	0.93	4.54	3.14	3.33	2.73	1.74	1.57	1.86	3.32	2.89	2.85	3.41
C70	3.27	0.88	1.87	2.85	1.87	5.02	1.9	0.93	0	4.08	2.48	2.98	2.23	1.75	1.69	1.56	2.91	2.52	2.17	2.64
C71	6.54	4.22	4.32	5.18	3.64	8.31	4.74	4.54	4.08	0	4.61	4.68	4.59	5.41	5.38	4.98	3.78	4.38	3.29	3.85
C75	4.15	2	2.59	1.71	1.79	6.6	2.41	3.14	2.48	4.61	0	3.33	1.17	2.42	3.33	1.72	3.59	2.72	2.33	2.13
C76	5.11	2.42	1.31	3.56	2.81	6.8	3.89	3.33	2.98	4.68	3.33	0	2.53	3.92	4.29	3.52	4.1	1.33	4.09	3.73
C77	3.7	1.63	1.76	1.53	1.63	6.09	2.07	2.73	2.23	4.59	1.17	2.53	0	2.24	3.22	1.7	3.86	2.15	2.78	2.39
C310	2.07	1.92	2.78	2.36	2.57	4.26	1.25	1.74	1.75	5.41	2.42	3.92	2.24	0	1.38	1.02	4.02	3.46	2.87	3.06
C312	2.65	2.3	3.13	3.2	3.21	4.04	2.2	1.57	1.69	5.38	3.33	4.29	3.22	1.38	0	2.05	3.87	3.94	2.89	3.27
C402	2.72	1.37	2.45	2.23	1.96	4.96	1.31	1.86	1.56	4.98	1.72	3.52	1.7	1.02	2.05	0	3.68	2.92	2.63	2.85
C1006	5.34	3.21	3.79	4.73	2.55	7.14	4.01	3.32	2.91	3.78	3.59	4.1	3.86	4.02	3.87	3.68	0	3.12	2.48	3.83
C1032	4.67	2.02	1.63	3.51	1.84	6.61	3.36	2.89	2.52	4.38	2.72	1.33	2.15	3.46	3.94	2.92	3.12	0	3.48	3.62
C1044	4.53	2.47	3.21	2.97	2.16	6.62	2.76	2.85	2.17	3.29	2.33	4.09	2.78	2.87	2.89	2.63	2.48	3.48	0	1.72
C1051	4.86	2.56	2.83	1.89	2.78	6.96	3.08	3.41	2.64	3.85	2.13	3.73	2.39	3.06	3.27	2.85	3.83	3.62	1.72	0

Euclidean Distance Matrix

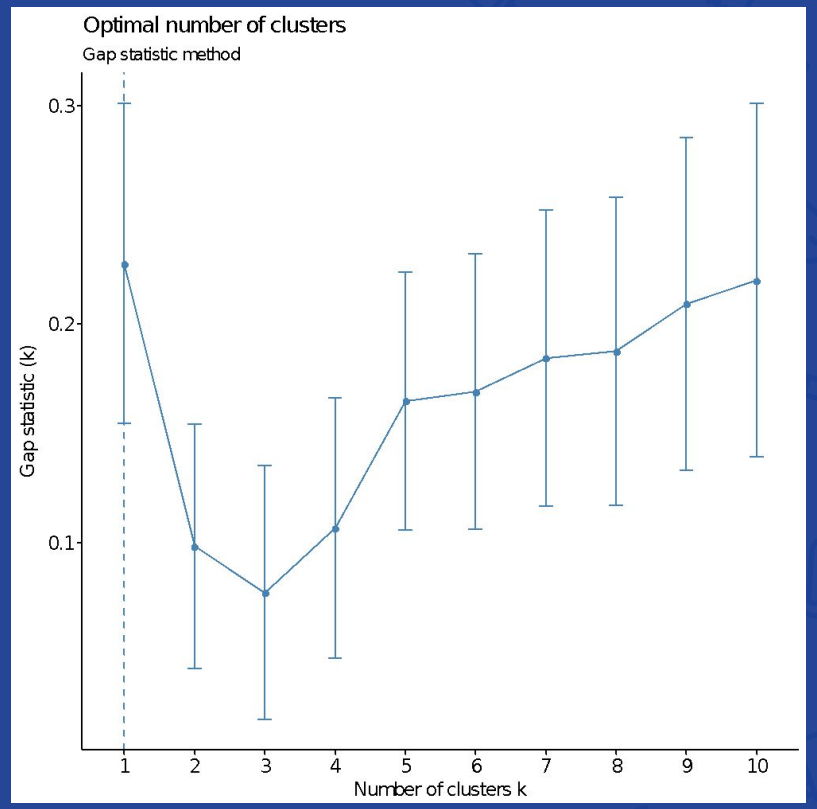
4) Determining the best linkage function to group stations into a dendrogram and the optimal number of clusters based on the majority result of different tests (for example: elbow, silhouette, and gap statistics)



Elbow Method

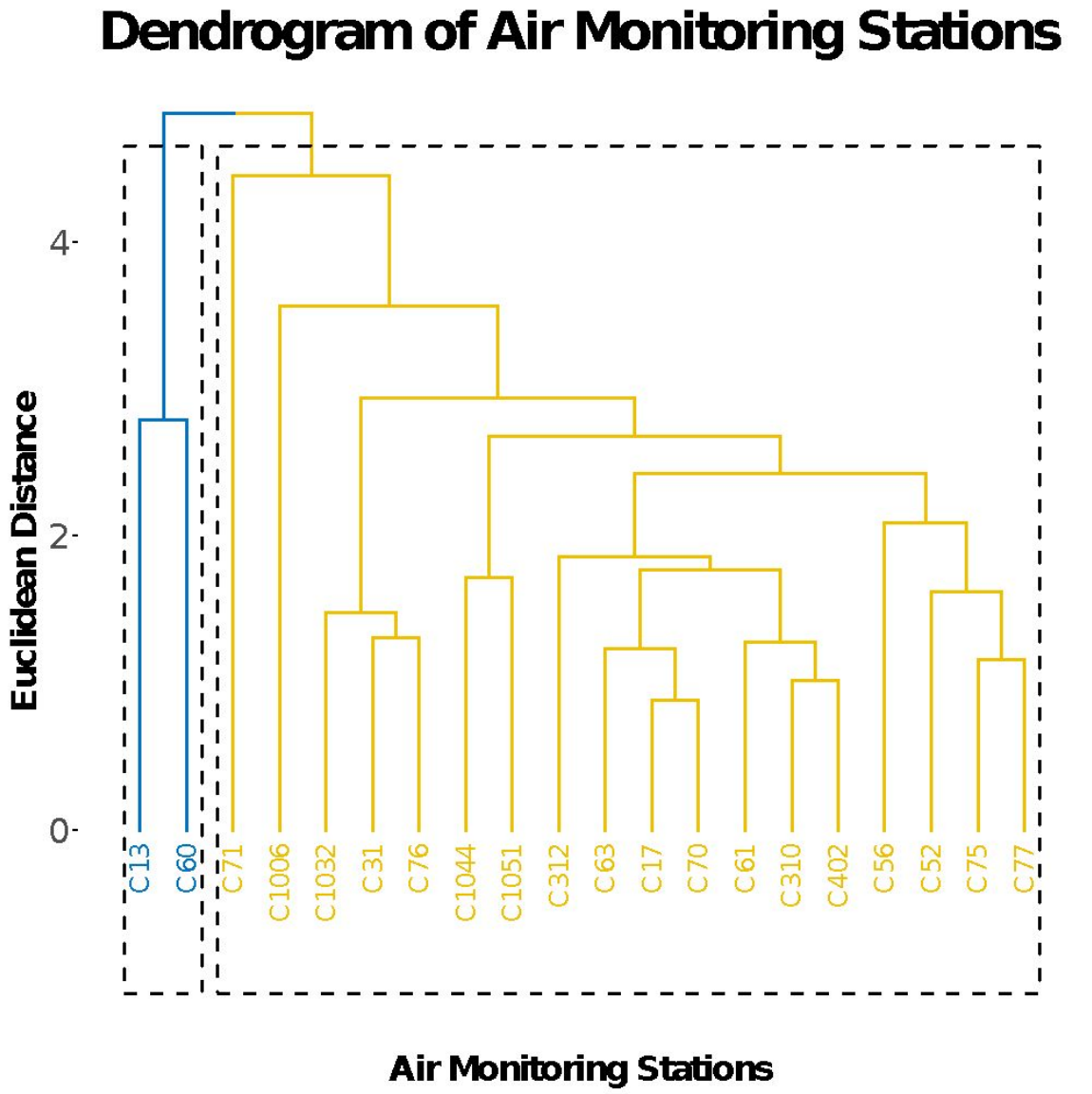


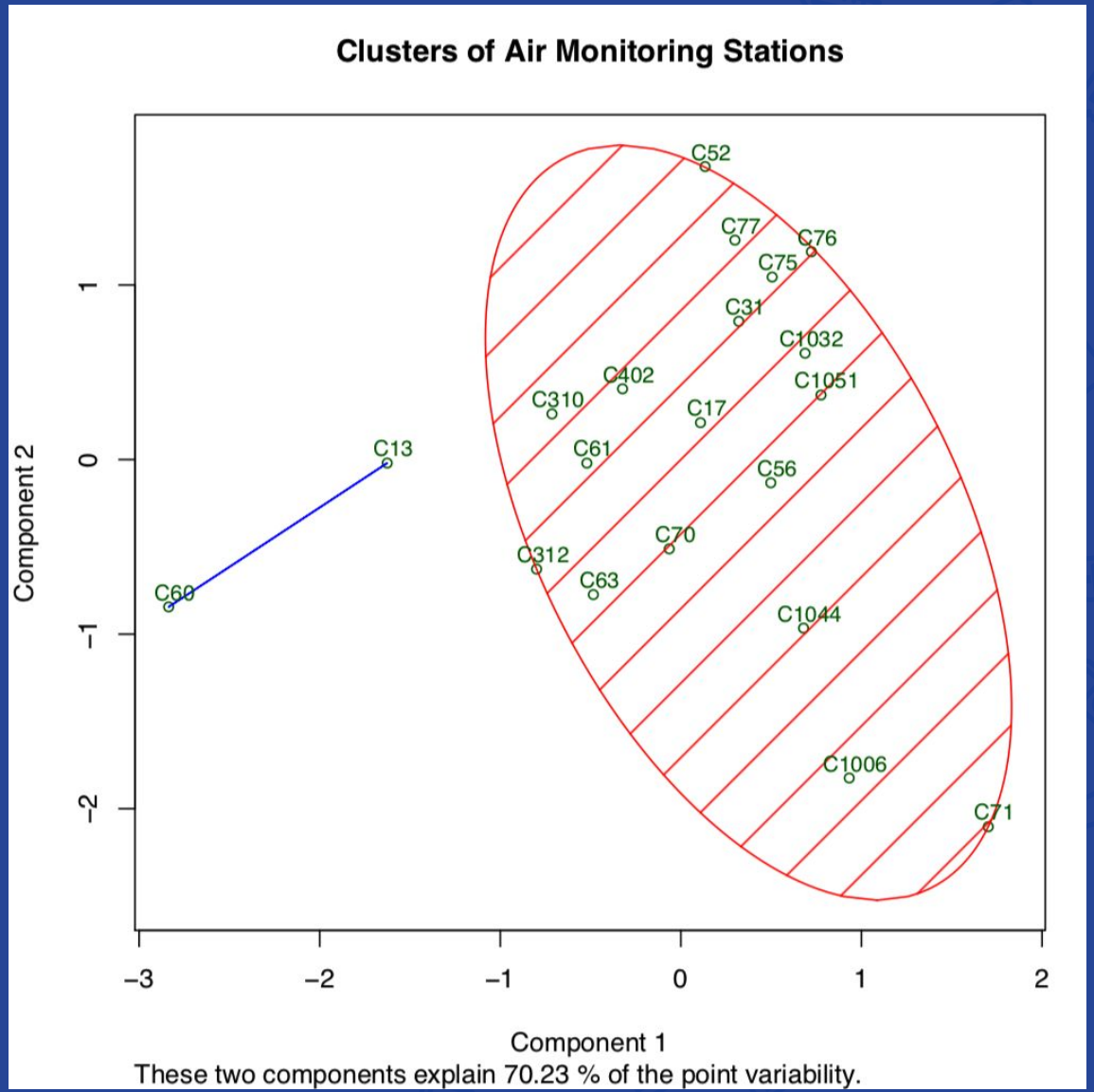
Silhouette Method



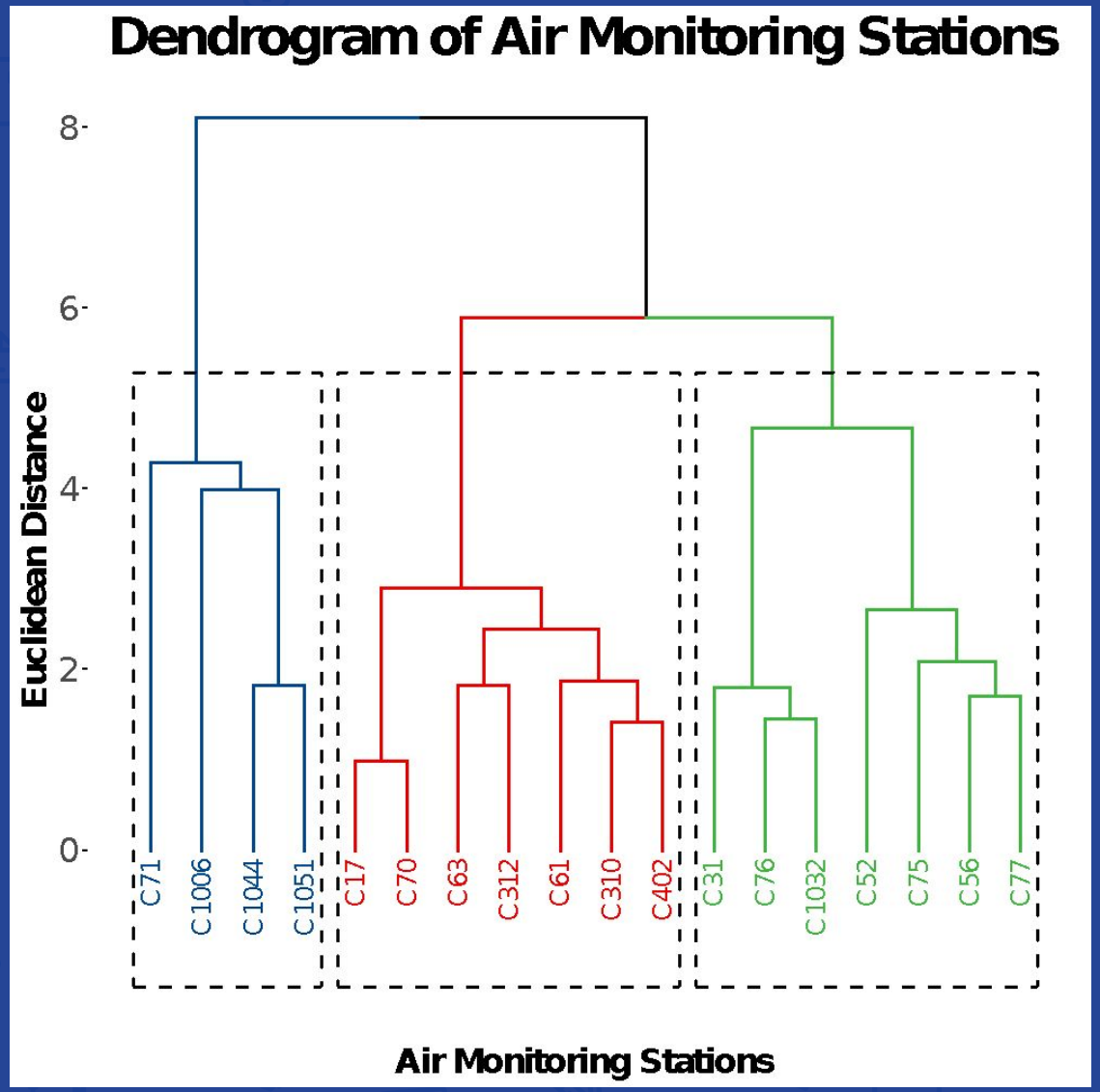
Gap Statistics Method

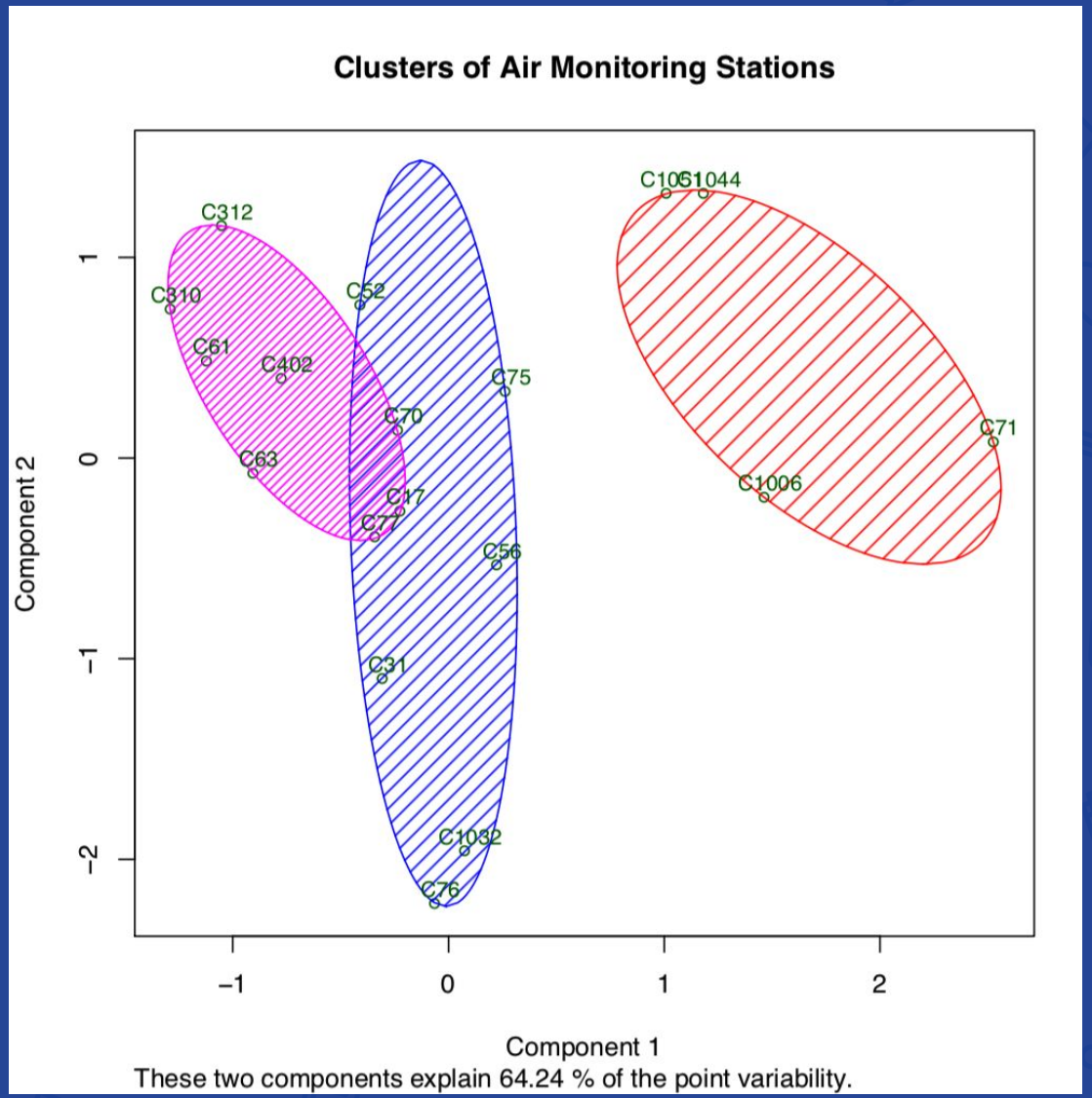
5) Constructing the Dendrogram and Plotting the Corresponding Clusters





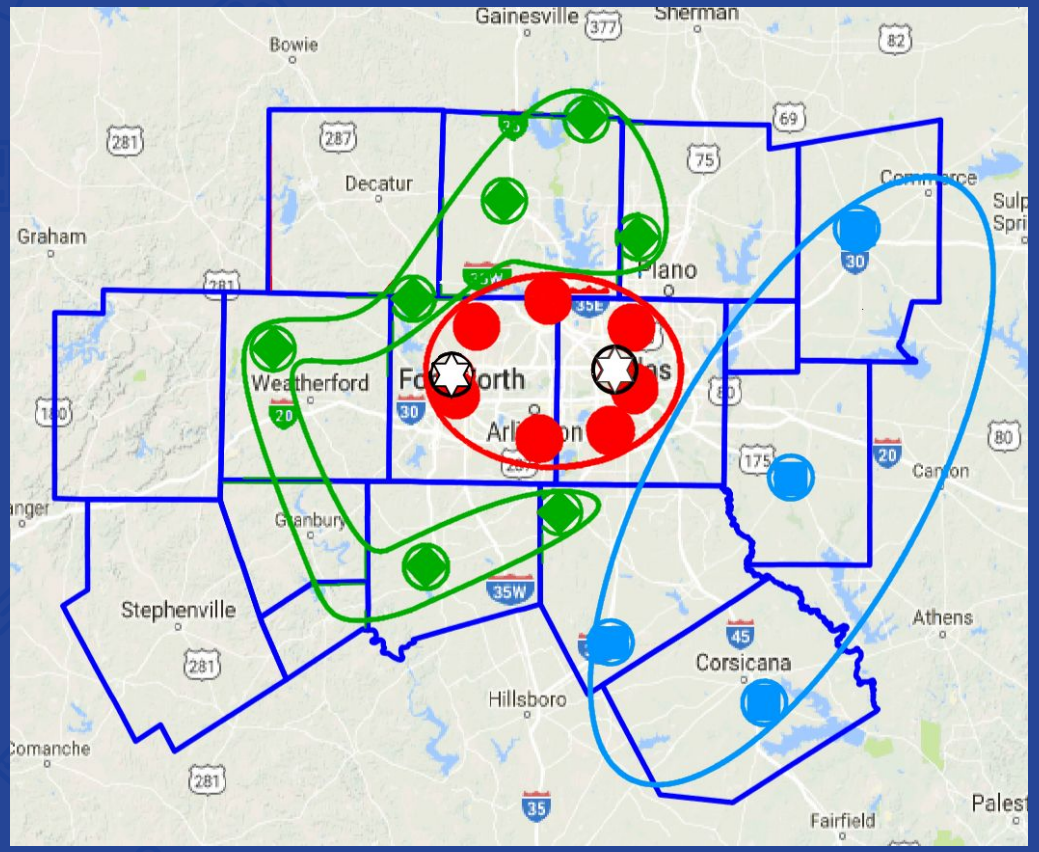
6) For further look into the second cluster which comprised 18 stations and a more precise interpretation of the result, we followed the same approach to construct a new dendrogram and clusters, utilizing data from the 18 stations instead of the 20 initial stations

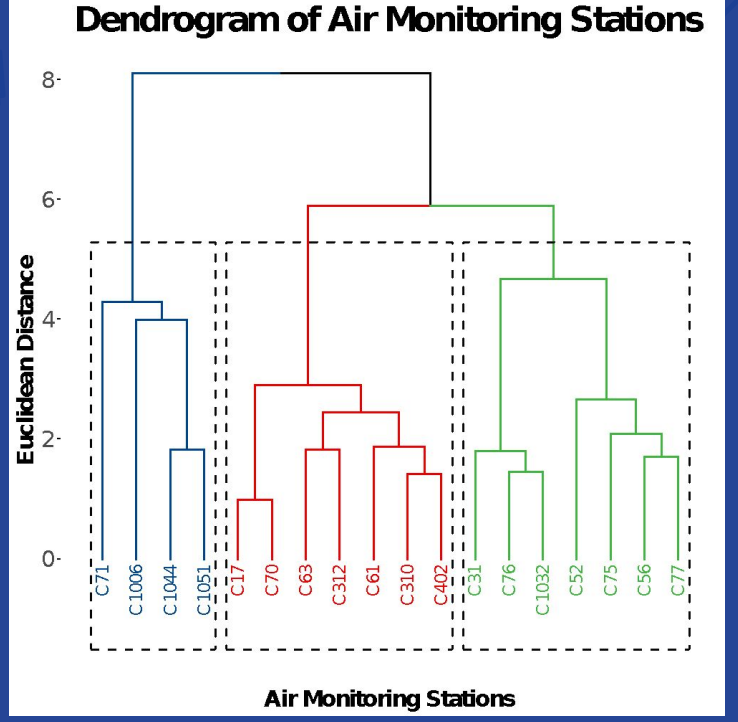


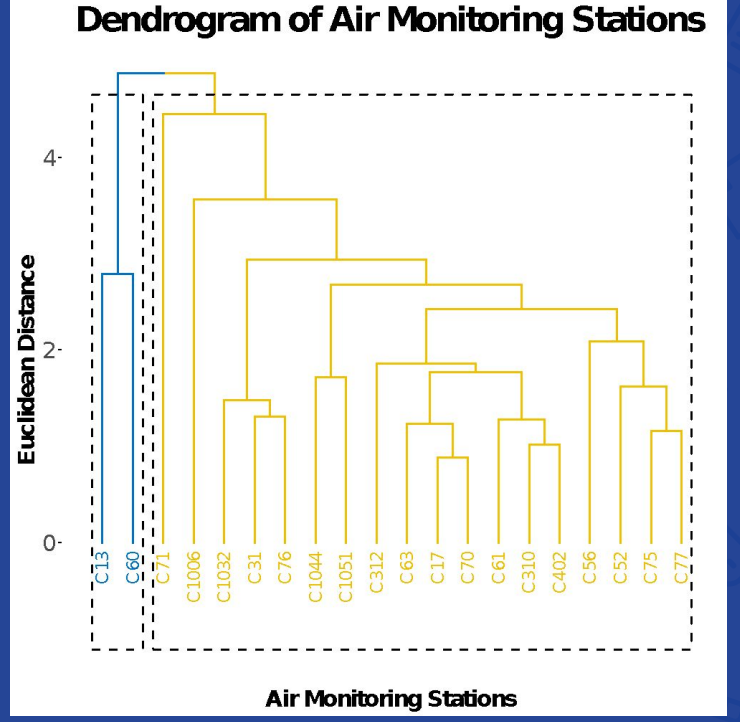


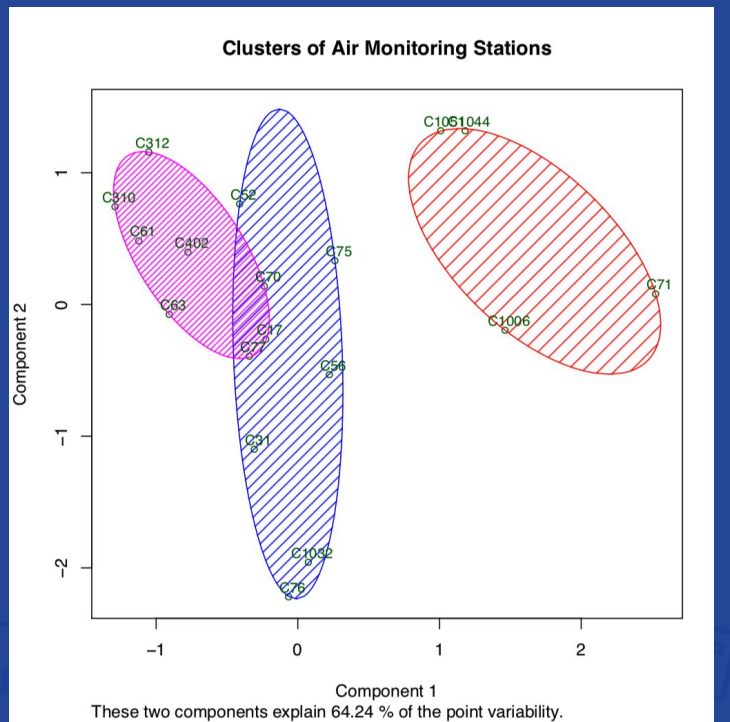
CONCLUSION

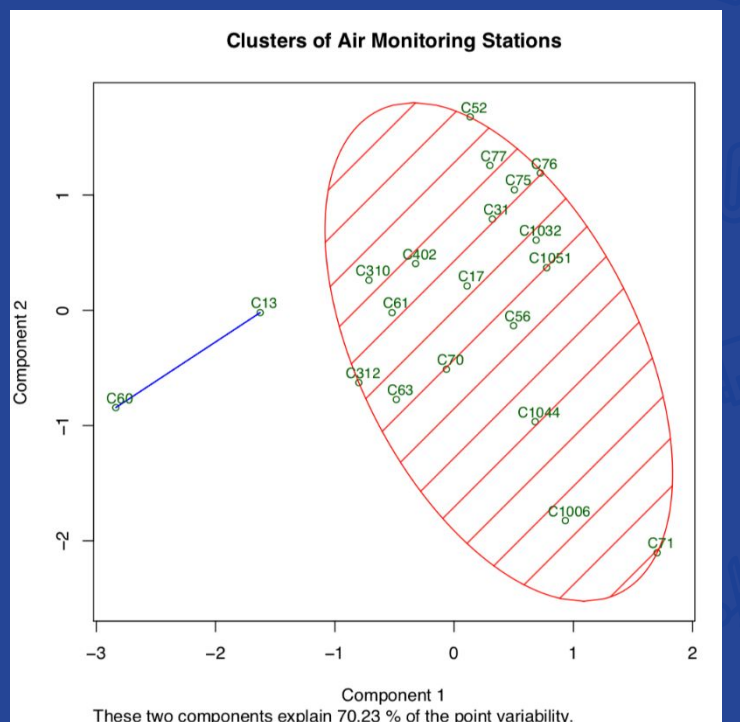
1) Spatial Analysis







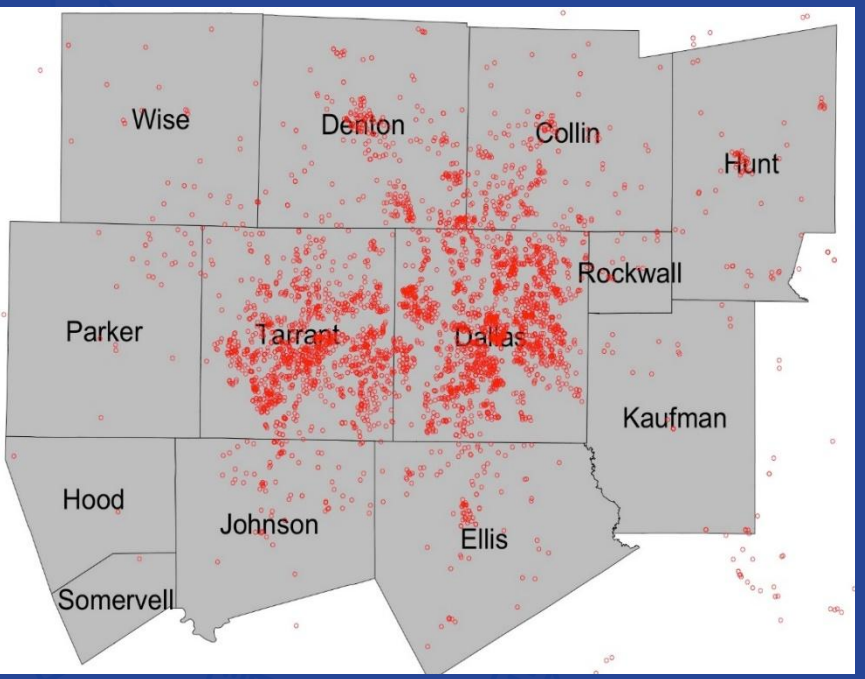


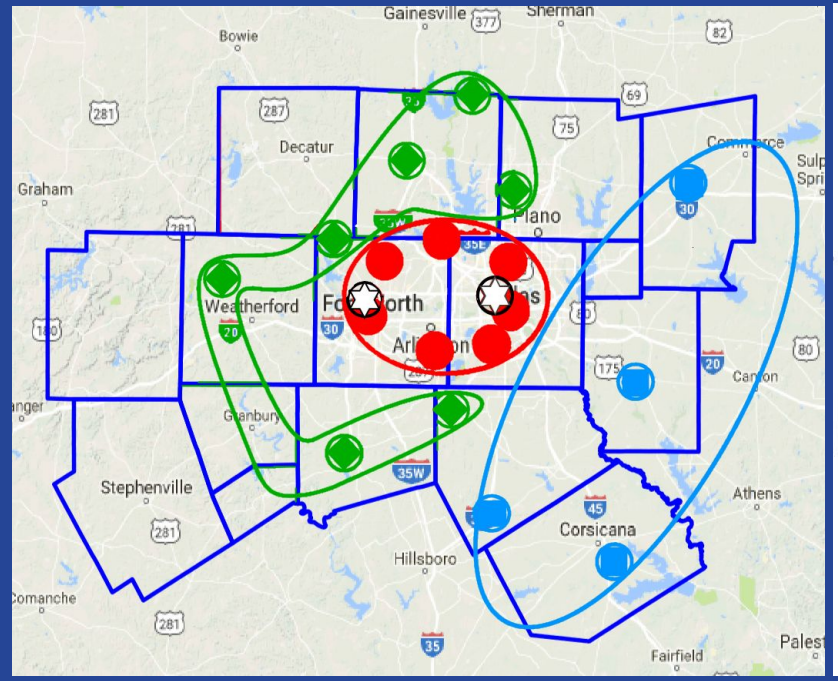


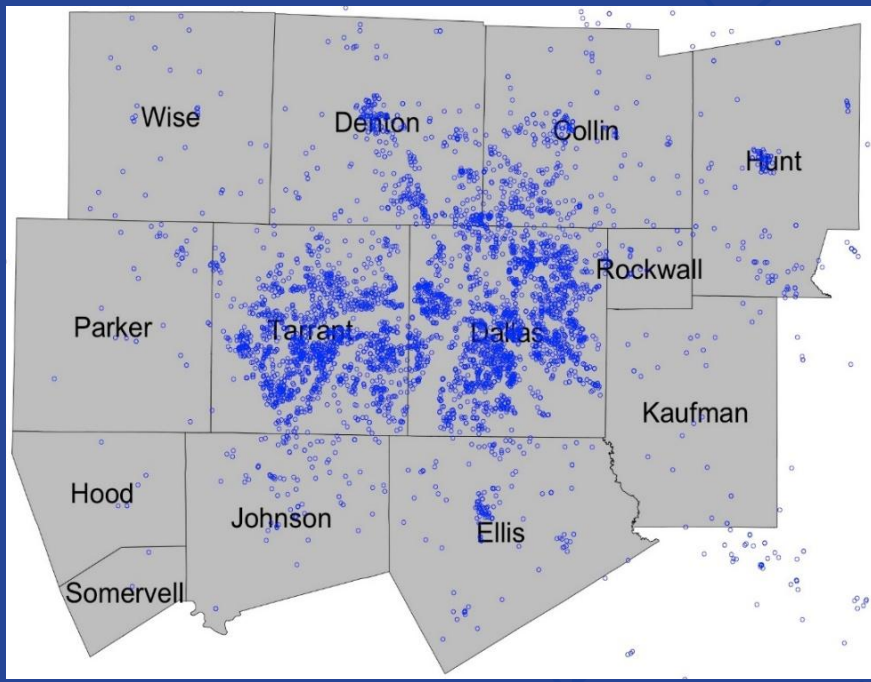
CONCLUSION

2) Implications regarding asthma

- Clusters can generally be used to explain asthma pattern. For example, the cluster of rural stations (blue cluster) corresponds to areas with extremely low asthma visits, regardless of the season. PCA results indicates that all stations included exhibit low concentration of air pollutants and low temperature (due to its being a rural cluster). On the other hand, the red cluster comprises stations with the highest level of NO, NO2, and PM2.5. The green cluster consists of stations with high level of O3. These two clusters correspond to the high asthma cluster spreading across mostly three counties Tarrant, Dallas, and Denton.







Future Research

- Spatial Interpolation of Unrecorded Data (explore different approaches)
- Incorporation of more variables (e.g., CO, SO2, relative humidity)
- Multivariate analysis for each pollutant
- Analysis of wind patterns
- Investigation of the relationship between pollutant exposure and pediatric hospital visits for respiratory problems (i.e. asthma)
- Explore seasonal patterns

References

Austin, E., Coull, B. A., Zanobetti, A., & Koutrakis, P. (2013). A framework to spatially cluster air pollution monitoring sites in US based on the PM2. 5 composition. *Environment international*, *59*, 244-254.

Dominick, D., Juahir, H., Latif, M. T., Zain, S. M., & Aris, A. Z. (2012). Spatial assessment of air quality patterns in Malaysia using multivariate analysis. *Atmospheric Environment*, *60*, 172-181.

Dorman, M. (2014). *Learning R for geospatial analysis*. Packt Publishing Ltd.

Gramsch, E., Cereceda-Balic, F., Oyola, P., & Von Baer, D. (2006). Examination of pollution trends in Santiago de Chile with cluster analysis of PM10 and ozone data. *Atmospheric environment*, *40*(28), 5464-5475.

Iizuka, A., Shirato, S., Mizukoshi, A., Noguchi, M., Yamasaki, A., & Yanagisawa, Y. (2014). A cluster analysis of constant ambient air monitoring data from the Kanto Region of Japan. *International journal of environmental research and public health*, *11*(7), 6844-6855.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112). New York: springer.

Kassambara, A., (2017). *Practical Guide to Cluster Analysis in R.*

Kassambara, A., (2017). *Practical Guide to Principal Component Methods in R.*

Lu, W. Z., He, H. D., & Dong, L. Y. (2011). Performance assessment of air quality monitoring networks using principal component analysis and cluster analysis. *Building and Environment*, *46*(3), 577-583.

Pires, J. C. M., Sousa, S. I. V., Pereira, M. C., Alvim-Ferraz, M. C. M., & Martins, F. G. (2008). Management of air quality monitoring using principal component and cluster analysis—Part I: SO2 and PM10. *Atmospheric Environment*, *42*(6), 1249-1260.

Pires, J. C. M., Sousa, S. I. V., Pereira, M. C., Alvim-Ferraz, M. C. M., & Martins, F. G. (2008). Management of air quality monitoring using principal component and cluster analysis—Part II: CO, NO2 and O3. *Atmospheric Environment*, *42*(6), 1261-1274.

Pérez-Arribas, L. V., León-González, M. E., & Rosales-Conrado, N. (2017). Learning Principal Component Analysis by Using Data from Air Quality Networks. *Journal of Chemical Education*, *94*(4), 458-464.

Silva, C., & Quiroz, A. (2003). Optimization of the atmospheric pollution monitoring network at Santiago de Chile. *Atmospheric Environment*, *37*(17), 2337-2345.

Smith, L. I. (2002). *A tutorial on principal components analysis*.

http://www.nctcog.org/trans/air/ozone/formation.gifrial

https://science-edu.larc.nasa.gov/ozonegarden/images/page-graphics/pic1_sourceNOx.jpg