A data mining approach to identify the distinct multi-pollutant air quality profiles within the US

Project 3: Satish Chirra, Hengfang Deng, Zlatan Feric, Wenyi Kuang

1 Introduction

Air pollution monitoring is a program currently conducted by US EPA (United States Environmental Protection Agency) which provides publicly available temporal air quality data [1], it measures the concentration levels of certain pollutants in the outdoor air and forms a basis for long term assessment of the US air quality. With the availability of such data from the past two decades and the growing awareness of the respiratory diseases induced by various emission pollutants, there are now ample opportunities to apply unsupervised machine learning techniques to understand the heterogeneity in air quality composition and identify the most distinct multi-pollutant profiles across the United States.

There are currently six common air pollutants that are known as "criteria air pollutants" , EPA continuously monitor, assess and then review the Air Quality standards for such pollutants and collaborate with local governments of the areas where the concentration standard are exceeded to mitigate the harm on individual health and the environment. The six criteria pollutants are:

- Ground-level Ozone
- Particulate Matter
- Carbon Monoxide
- Lead
- Sulfur Dioxide
- Nitrogen Dioxide

In this project, we manually extract the historical data from the EPA Air Quality Data via Google Cloud Platform queries reporting criteria chemical pollutants readings and ambient weather conditions. The overall objective of this project is to develop a statistical framework that facilitates perceiving the mixture of air pollution and meteorological patterns, multiple cluster analysis will be implemented to spatially distinguish such patterns and the seasonality would also be taken into consideration by comparing results on different subsets. We hope to shed some lights on further investigation of pollutant mixtures and their impacts on individual health.

Since each observation has two dimensions including location and time, we first use both the whole data set with season as a categorical variable as well as four seasonal subsets to spatially cluster different monitor sites. This allows us to better distinguish the geographical aerial quality profiles induced by urban and industry phenomena with consideration of the impacts from seasonality. After the spatial clustering , four distinct monitor sites are selected and we then extract about 3 years of daily particulate and gaseous pollutants data for each site to find distinct groups of days. This would help us further associate the chemical characteristics with weather patterns across different seasons and months.

2 Data

Ambient concentrations of pollutants have been continuously measured at sites across the US, and currently EPA allows users to acquire daily data for a specific location and time period of a specific pollutants. In this project, we aim at developing a empirical framework to

spatially cluster monitor sites considering multiple pollutants composition jointly as well as the meteorological factors. The data acquisition was conducted as multi-step screening process on the Google Cloud Platform with big query. Our goal is to:

1, Extract the mean value of each pollutant on the same day at the same site.

2, Create a new column for each measured variable in the joint dataset with the most frequently used unit of measure to ensure each variable has only one standard unit.

3, Remove the rows/instances that have pollutants of which the sampling period is shorter than 8 hours since the time resolution for the final dataset is 1 day.

4, Finally, after the previous update the team has realized that there are still significant amount of duplicated rows existed in the final subset and this is largely due to that the meteorological factors are generally reported multiple times per day per site. We further calculated the daily average meteorological conditions for each site.

Based on the query, a dataset constaining 195,279 instances was generated while the final number of instances is 36,712 after eliminating the duplicates. The final dataset reports daily readings from 5,098 days from 55 distinct sites across the US including six criteria pollutants and three meteorological features. The details data features and names are:

- O3 (Ozone) in (Parts per million)
- PM 2.5 (Particulate Matter 2.5) in (Micrograms/cubic meter)
- PM 10 (Particulate Matter 10) in (Micrograms/cubic meter
- CO (Carbon Monoxide) in (Parts per million
- SO2 (Sulfur Dioxide) in (Parts per billion
- NO2 (Nitrogen Dioxide) in (Parts per billion
- TEMP (Temperature) in (Degrees Fahrenheit)
- RH (Relative Humidity) in (Percent Relative Humidity)
- WIND (Wind Speed) in (Knots)

3 METHODS

3.1 Clustering

We mainly focused on prototyped-based clsutering methods in this project: Both K-means and K-medoids have been implemented and compared. Medoid is one representative observation (instance) of one cluster whose average dissimilarity to all the observations in the cluster is the least [2]. Compared to K-means, K-medoids could be more robust to extreme values compared to k-means since the objective function is to minimize total pairwise dissimilarities instead of taking the sum of squared errors. [3]. PAM (Partition Around Medoid) as the most common algorithm for K-mediods have been chosen for demonstrative purpose. Furthermore, from the correlation result, we can see that there are some correlations exhibited between the variables which suggest that the spherical assumption might not stand, hence Hierarchical and Distribution-based are also the candidates for model selection.

For the Site-specific analysis, another variation of K-means - K-medians [4] has also been compared against other algorithms. This method is similar to K-medoids to some extent, while a medoid needs to be an actual observation for all features from the original data, while for the 'median' on a multivariate data this would only be required for one feature. In other words, The actual median is a combination of multiple instances with specific feature values..

3.2 Criteria for K

The two standard metrics for evaluating and determining the number of clusters are SSE (sum-of squares) and Average Silhouette Width (ASW). The SSE for the clustering result is defined as:

$$SSE = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} ||x_n - \mu_k||^2$$

For the sihouette width, it measures both cohesion and separation [5]. As the formula below shows, *a(i)* is the average distance of *i* with all other data within the same cluster while *b(i)* is the lowest average distance of *i* to all points in any other clusters.

$$Sihouette\ Width = \frac{b(i) - a(i)}{maxx(a(i), b(i)}$$

The third metric incorporated in our project is Gap Statistics [6]. Which was initially proposed in 2001 and has been proved to applicable to most clustering methods. The main essence of this new metric is to utilize a reference distribution of the whole which is obtained through a bootstrapping process. Compared to the tradition within clustersum-of-squares, the gap statistics computes the log of the within cluster sum-of-squares and compare it with the reference distribution:

$$Gap_n(k) = E_n^* log(W_k) - log(W_k)$$

4 Results

4.1 Statistical Analysis

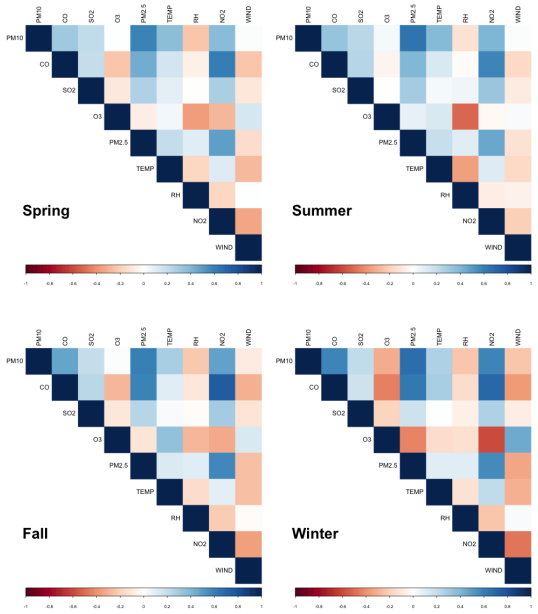


Fig. 1: Correlation Results

The next step is to perform a linear corelation analysis among all the variables. Since all readings are continuous random variables, the Pearson correlation coefficient (PCC) has been used to measure the linear correlation between two variables. With the hypothesis that pollution profile pattern would have discrepancies among different seasons, the seasonality has been taken into consideration. From Figure 1, we can see that the correlation results are similar for spring and fall. However, there are some variations between the summer and winter. We can see also that NO2 are positively correlated with the most pollutants especially in winter. In addition, PM2.5 and PM10 also demonstration strong association as expected. Interestingly, Ozone

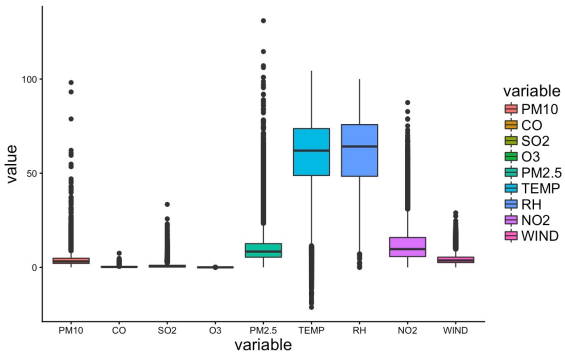


Fig. 2: Boxplot of the Criteria Pollutants

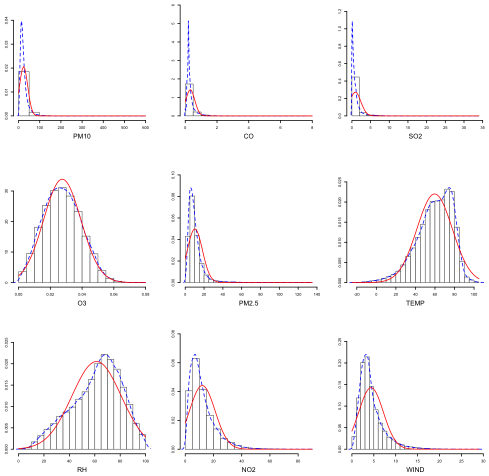


Fig. 3: Histogram of the Criteria Pollutants

Since the data comes from monitors at different sites, the sensors at each site would require proper calibrations to obtain accurate readings. There are about one thousand instances where negative concentrations were reported and these negative numbers are generally treated as calibration readings and hence removed. Boxplot is then used to to facilitate the outlier identification and the PM10 has

been scaled to fit in the figure. From Figure 2 it can be inferred that most of the data appear to have very sparse patterns, and the out-of-box observations don't appear to be random noises and instead just represent some extreme scenarios with high level of stagnation and pollution. Therefore, we decided to keep some data points to better capture all possible scenarios of pollutant profiles. Figure 3 also shows that the only Ozone, Relative Humidity and Temperature appear to have approximately normal distribution while others present significant long-tail distribution patterns.

4.2 Number of Clusters

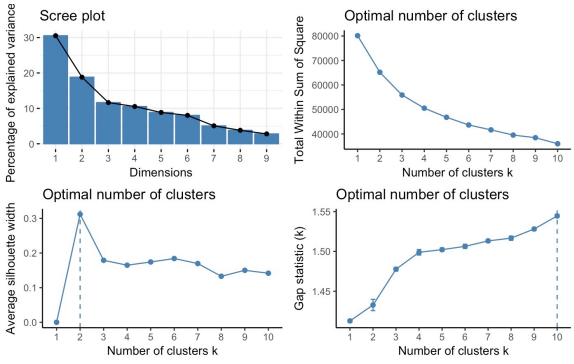


Fig. 4: Using K-means to find K on the Summer Subset

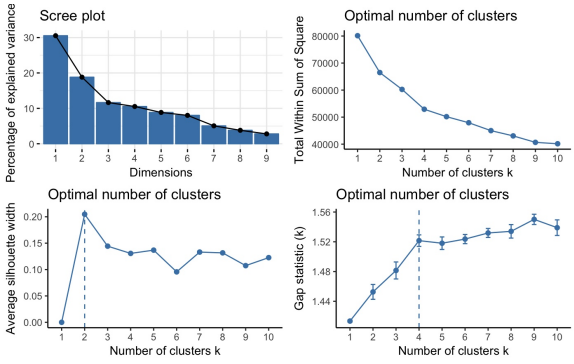


Fig. 5: Using PAM to find K on the Summer Subset

The first step in determining the optimal number of clusters is to undergo the PCA procedure that uses an orthogonal transformation to possibly decor-relate variables while reducing the dimension and then transform original features into a set of values called principal components. When finding the number of K, both Euclidean and Manhattan distances have been experimented and the Euclidean distance appear to have better performances on the Gap Statistics and hence we proceed the analysis with Euclidean distance. The Figure 4 below demonstrate that using principal components we are not able to successfully achieve the dimension reduction since the first two components explain approximately half of the total variance and other component are represent less than 10 percent. Using all three metrics on the K-means algorithm suggests the optimal number of clusters is 2 for ASW and 10 for Gap Statistics. Interestingly, Using PAM with Gap Statistics 4 stands out as the ideal number.

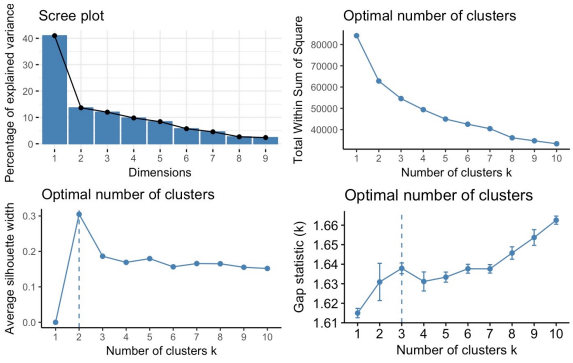


Fig. 6: Using K-means to find K on the Winter Subset

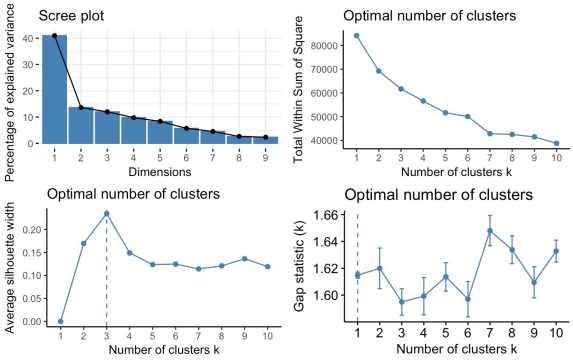


Fig. 7: Using PAM to find K on the Winter Subset

For the winter subset shown in Fig 6 and 7, again PCA could not help reduce the dimensionality effectively while K-means indicates that 3 is the optimal number of clusters.

TABLE 1: Centroid Information on Summer Subset with Kmeans

Cluster	1	2	3	4
PM10	22.82	26.2	14.59	55.52
CO	0.22	0.21	0.15	0.47
SO2	0.70	0.65	0.41	2.56
O3	0.03	0.04	0.03	0.03
PM2.5	10	9	5.89	20.81
TEMP	77.20	81.36	68.57	78.46
RH	71.57	37.32	66.11	61.91
NO2	8.10	8.69	4.71	19.91
WIND	2.84	3.92	6.31	3.75

TABLE 2: Medoid Information on Summer Subset with PAM

Cluster	1	2	3	4
PM10	25.00	43.00	11.50	30.00
CO	0.14	0.33	0.14	0.20
SO2	0.40	1.98	0.10	0.99
O3	0.03	0.04	0.03	0.05
PM2.5	7.25	16.05	3.30	9.30
TEMP	78.79	82.50	63.12	78.83
RH	65.75	70.00	61.54	30.58
NO2	2.84	13.07	5.80	11.25
WIND	2.56	2.38	4.49	3.45

4.3 Cluster Composition in Summer

After scaling all the variable to have a mean of 0 and a standard deviation of 1. We then set K=4 using both K-means and PAM. Table 1 and 2 describe the centroid/medoid information while Fig 8 and 9 show the locations and counts for the clustering result using both K-means and PAM. We can see that using both methods produces similar results for the summer subset, however, the prototypes from each method appear to have some variations. Both Cluster 2 from PAM and Cluster 4 rrepresent the sites that are heavily polluted but the individual pollutant values for the centroid are larger than the medoids . The size of the circles indicate how many observations from that site and the color indicates the cluster number.

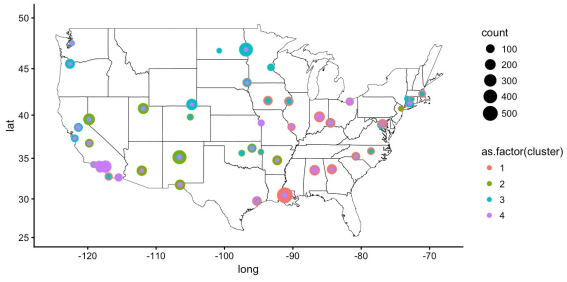


Fig. 8: Heat Map of Clustered Monitor Site with Kmeans in Summer(K=4)

For the K- means, Cluster 4 (Purple) represents heavily polluted air quality and dominates the Southern California region. Cluster 3 (Blue) demonstrates the best air quality and also characterized by high wind speed and relatively low temperature. Cluster 1 and 2 are both median quality clusters and the main difference is the dramatic relative humidity difference.

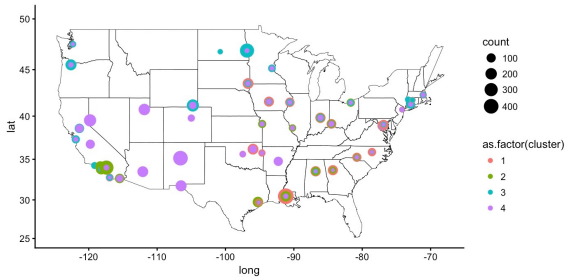


Fig. 9: Heat Map of Clustered Monitor Site with PAM in Summer(K=4)

We also implemented Gaussian Mixture Model which is a distribution based approach on the summer subset. We used the R package mclust to both identify the optimal G (number of mixture components) and to carry out expectation-maximization for fitting the model and extracting classifications. The type of mixture model that we use has an ellipsoidal distribution with variable shape, volume, and orientation as this model will adapt to the most diverse cluster shapes in our data. Similar to the K-means model, we are faced with finding the optimal number of mixture components., For this model we will use a measure known as BIC (Bayesian Information Criterion) [7] which a penalized log-likelihood function by

the number of components given by the formula:

$$BIC = ln(n)k - 2ln(L)$$

Based on the figure below, we decided to proceed with 4 components as the BIC increase was still significant and also to readily compare the clustering with our k- means implementation.

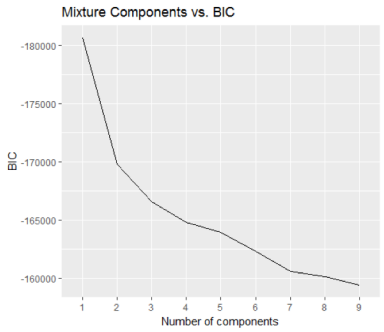


Fig. 10: Number of Components VS BIC

In applying the Gaussian Mixture Model we observed clusters of varying shape and orientation in two dimensional distributions where the model attempted to best segment the data by taking data variance and covariance into account. The GMM is beneficial in situations such as O3 vs. WIND and NO2 vs WIND as visualized by the diverse cluster shapes and sizes. However, the case for picking a GMM for our purposes was not entirely clear. For example, as we observed in the initial one dimensional distributions, some of our features were not normally distributed (i.e. PM10.) Furthermore, one certain situations such as SO2 vs (O3, PM2.5, TEMP, and RH) would be partitioned equally well by the k-means algorithm. We also note that no transformation was applied to the real data (i.e. box-cox transformation) which may have improved the performance of the Gaussian by normalizing long tail distributions observed in the PM10 histogram. Furthermore, in the next figure we can see that the overall conclusion and results of our hypothesis due not change much by applying the GMM which is why we only attempted this method for the most variable season of Summer.

From Fig 12 we can see that the GMM clustering results are also similar to K-means and

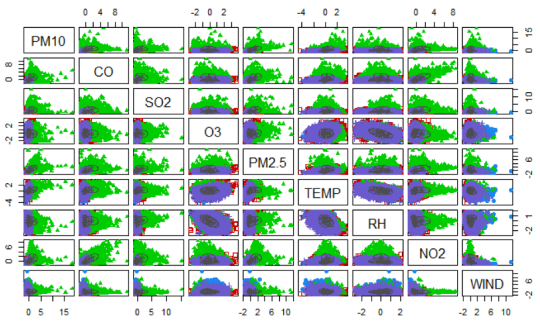


Fig. 11: Gaussian Clustering Visualization on 2D Distributions

PAM. However, it is also notable that the clustered sites which are shown as colored circles appear to have higher purity on the GMM plot. From the plots in K-means and PAM, we observed some overlapping circles yet for the GMM plot, we can visually see with 4 distributions the purity for the clustering result appears to be higher.

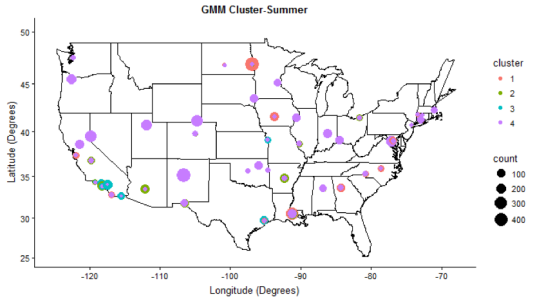


Fig. 12: Heat Map of Clustered Monitor Site with GMM in Summer

4.4 Cluster Composition in Winter

Fig 13 and 14 show the inter clustering distribution across the US using both K-means and PAM with k=3. We can see that the heterogeneity is less significant in winter compared to summer. More specifically, the separation between the east and west region has been replaced by them merging into one cluster. It can also be seen that the K-means result differ from the PAM results for the Northern regions and PAM has assigned the Northern region to the majority of the monitor sites in the US.

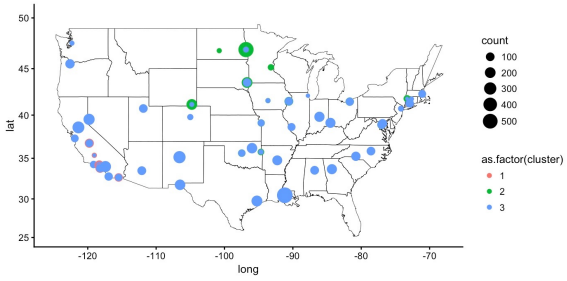


Fig. 13: Heat Map of Clustered Monitor Site with K-means in Winter(K=3)

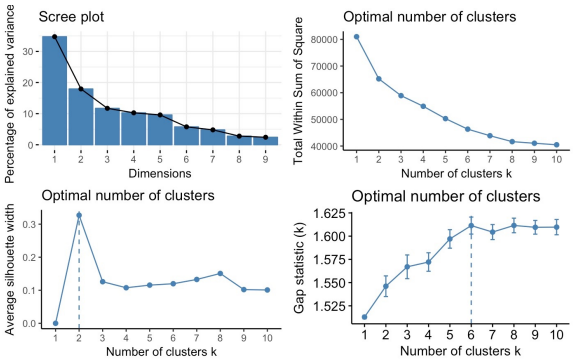


Fig. 15: Using PAM to find K on the Fall Subset

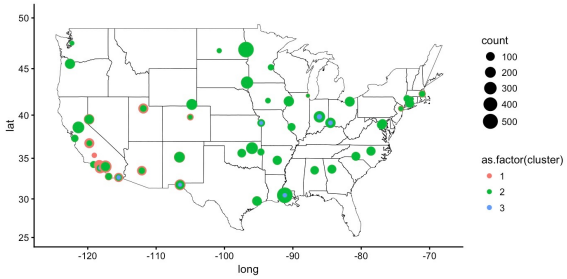


Fig. 14: Heat Map of Clustered Monitor Site with PAM in Winter(K=3)

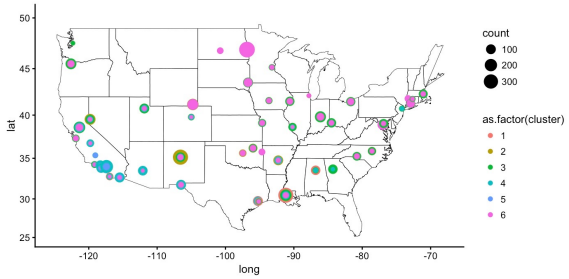


Fig. 16: Heat Map of Clustered Monitor Site with PAM in Fall(K=6)

4.5 Cluster Composition in Fall

With the similarities between Spring and Fall, the cluster distribution for fall has been only analyzed and shown below. Fig 15 indicates that when using PAM as the clustering method, the optimal number of clusters to maximize the Gap Statistics is 6. Fig 16 and 17 show the cluster distribution across the US in Fall using both K-means and PAM with k=6. We can see that K-means and PAM again produced different results: PAM has distinguished different groups with air quality patterns and generated similar results as we obtained from the summer subset while Kmeans produced one large central cluster with other small-sized clusters. Hence, for different season there might be different clustering methods that are more suitable for such data.

4.6 Site-specific Analysis

For the Site-specific analysis, we extract all the observations for a single site and determine for a single site what are the possible pollution scenarios (days). The models have been considered include K-means, K-median, and agglomerative clustering (hierarchical). Without the ground truth, we used SSB/SSW and Silhouette Width to compare each model's performances as well as with different number of clusters. SSB/SSW is the ratio of the Sum of Squared of distance Between clusters and Sum of Squared of distance Within cluster which is also known as a F-test [8]. The ratio generally increases as we increase the number of K and we use it as a relative indicator select the best algorithm.

There are several sites we have considered that have distinct chatacteristic based on the work: EL Paso, Texas; Washoe, Nevada; Sacramento, California; Washington, D.C; Honolulu, HI; East Baton Rouge, Louisiana. Here the re-

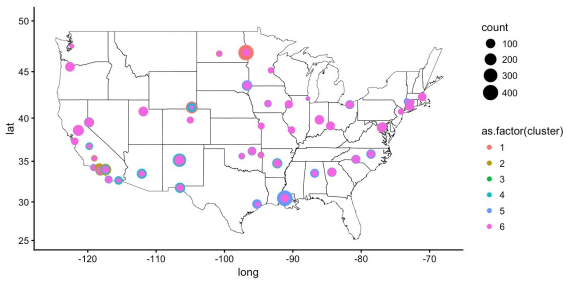


Fig. 17: Heat Map of Clustered Monitor Site with Kmeans in Fall(K=6)

sults from Sacramento, Honolulu and East Baton Rouge have been shown for demonstrative purpose.

Fig 18 depicts the Silhouette Width vs number of Clusters with different algorithms about for Sacramento and Fig 19 shows the SSB/SSW. As we increase k, the quality of clustering will decrease monotonically. And for all models, the K-means is the best one in all metrics for both sites.

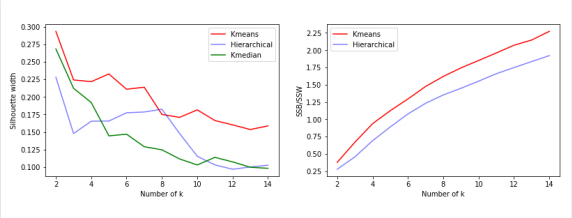
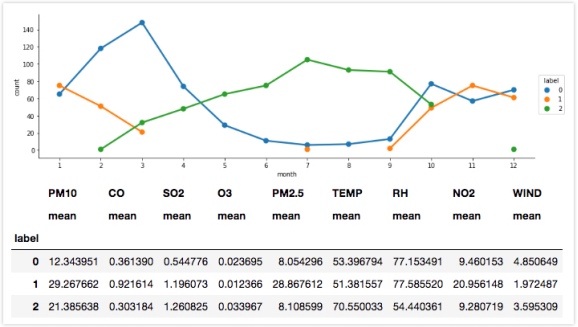


Fig. 18: Number of Clusters vs both Metrics in Sacrenmento

Fig 19 to 21 shows the trendlines of the occurences on month as well as the centroid information for all 3 sites. We can see that for Sacranmento three clusters represent good, fair and bad air qualities. We can also see that the blue cluster (good) has the peak in March and also very low frequency in summer.The orange cluster (bad) only occurs in winter months.

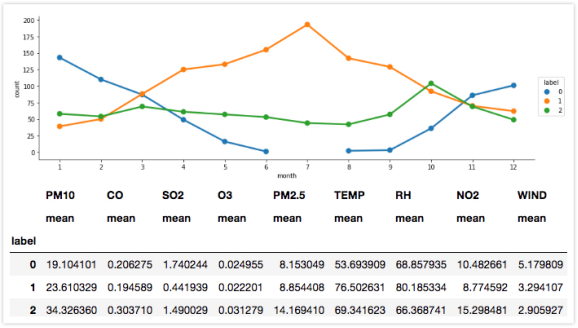
For East Baton Rouge in Louisiana, we can see the green cluster represents heavy pollution, and it is not sensitive to any seasonality, in other words, the monthly distribution of the 'bad quality' days is pretty even. It is also notable that most of good air quality (blue)



label	PM10 mean	CO mean	SO2 mean	O3 mean	PM2.5 mean	TEMP mean	RH mean	NO2 mean	WIND mean
0	12.343951	0.361390	0.544776	0.023695	8.054296	53.396794	77.153491	9.460153	4.850649
1	29.267662	0.921614	1.196073	0.012366	28.867612	51.381557	77.585520	20.956148	1.972487
2	21.385638	0.303184	1.260825	0.033967	8.108599	70.550033	54.440361	9.280719	3.595309

Fig. 19: Month Frequency Histograms for Cluster Composition in Sacramento

days are around February and March.



label	PM10 mean	CO mean	SO2 mean	O3 mean	PM2.5 mean	TEMP mean	RH mean	NO2 mean	WIND mean
0	19.104101	0.206275	1.740244	0.024955	8.153049	53.693909	68.857935	10.482661	5.179809
1	23.610329	0.194589	0.441939	0.022201	8.854408	76.502631	80.185334	8.774592	3.294107
2	34.326360	0.303710	1.490029	0.031279	14.169410	69.341623	66.368741	15.298481	2.905927

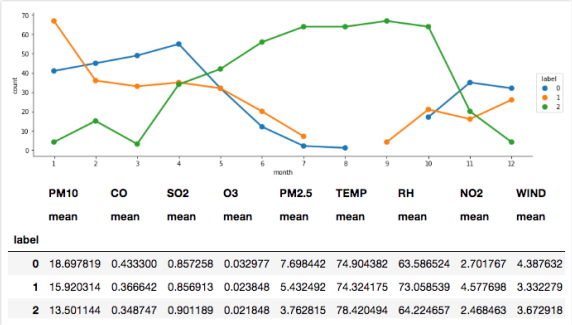
Fig. 20: Month Frequency Histograms for Cluster Composition in East Baton Rouge

Based on the historical air quality data from Honolulu, we can find a different pattern compared to the previous sites. In the summer days (May-Aug), the air quality is the best as indicated by the green line. Winter and Spring appear to have the most pollution. It is also notable that the pollution level in Hawaii is significantly lower than the previous sites which is due to the fact that tourism is the main industry for the state. Consequently, with less air pollution source, seasonality plays a more important role in the air-mixture profiles.

5 Conclusion

Based on the algorithms we have implemented and results we obtained so far, we can draw the following conclusions:

The statistical framework implemented can be used as a tool to indentify and further



label	PM10 mean	CO mean	SO2 mean	O3 mean	PM2.5 mean	TEMP mean	RH mean	NO2 mean	WIND mean
0	18.697819	0.433300	0.857258	0.032977	7.698442	74.904382	63.586524	2.701767	4.387632
1	15.920314	0.366642	0.856913	0.023848	5.432492	74.324175	73.058539	4.577698	3.332279
2	13.501144	0.348747	0.901189	0.021848	3.762815	78.420494	64.224657	2.468463	3.672918

Fig. 21: Month Frequency Histograms for Cluster Composition in Honolulu

classify air monitor sites based on their multi-pollutant composition. Using the tradition SSE generally generates a L-shape plot vs number of clusters which makes it difficult using the elbow method to determine the optimal K. Average Silhouette Width and Gap Statistics are proved to be more effective in picking the optimal number of K and especially Gap Statistics since it compares the changes of the Log scale of within-cluster errors with respect to a reference null distribution from a bootstrapping on all observations. However, it is also more computationally expensive, and reducing the bootstrapped sample size could potentially speed up the comparison process. Finally, the first two components for both summer and winter subsets explain about half of the total variance while the other components only stand for less than 10 percent of the total variance, hence orthogonal transformation could not help us effectively reduce the dimensionality if we want to keep very large portion of the original variance.

The heterogeneity among air monitor sites in the US is more significant during the summer months, and Northern US such as Minnesota represent the regions with the best air quality while the representative sites for the worst air quality are located in California mostly. Furthermore, K- means and K-medoids appear to have similar clustering results in summer but generate rather different results for other seasons even with the same K. During the fall season, K-medoids is able to extract distinct groups in the central and western US while K-means has merged those regions into one cluster which validates that medoid-based models can be more robust to noises and extreme values compared to centroid-based models.

Furthermore, we experimented using the entire dataset with season as a categorical variable, with the cost of computation, ASW was the only metric used to determine the optimal number of clusters. The optimal number of clusters is 6 for K-means and 6 for PAM using euclidean distance. The figures in the appendix demonstrate that for K-means and PAM using season as a variable to spatially cluster the sites the heterogeneity could not be accurately evaluated since from the cluster plots we mainly notice that the sites have been partitioned with two large groups that represent the east and west region. This separation is largely dominated by the influence from the climate and geographic regions instead of the pollution emission profiles.

In summer, GMM also presented very similar results as prototype-based methods and we also implemented DBSCAN on the summer subset, however, with multiple trials on the eps and number of neighbors we set them as 20 and 3, the model picked the regions with the best air as noises (See Appendix). Hence, we propose that careful tuning on such parameters is needed.

For the site specific analysis, we can see that for different sites, the influence from seasonality would vary. For a site with very low level of pollution sources such as Hawaii, season makes a difference on the air quality while for sites that are heavily industrialized or urbanized, the heavily polluted time span could occur all year long depending on the emission conditions.

REFERENCES

[1] Ambient Monitoring Technology Information Center (AMTIC) https://www.epa.gov/amtic

[2] Struyf, Anja; Hubert, Mia; Rousseeuw, Peter (1997), Clustering in an Object-Oriented Environment, Journal of Statistical Software, 1 (4): 130

[3] H.S. Park , C.H. Jun, A simple and fast algorithm for K-medoids clustering, Expert Systems with Applications, 36, (2) (2009), 33363341

[4] P. S. Bradley, O. L. Mangasarian, and W. N. Street, Clustering via Concave Minimization, Advances in Neural Information Processing Systems, vol. 9, M. C. Mozer, M. I. Jordan, and T. Petsche, Eds. Cambridge, MA: MIT Press, 1997, pp. 368374

[5] Peter J. Rousseeuw (1987), Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis, Computational and Applied Mathematics, 20: 5365

[6] Tibshirani, R., Walther, G. and Hastie, T.(2001). Estimating the number of clusters in a data set via the gap statistic, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 63: 411423. doi:10.1111/1467-9868.00293

[7] Bhat, H. S.; Kumar, N (2010). On the derivation of the Bayesian Information Criterion

[8] Hinkelmann and Kempthorne (2008), Approximating the randomization test, Design and Analysis of Experiments, Volume 1, Section 6.6.

6 APPENDIX

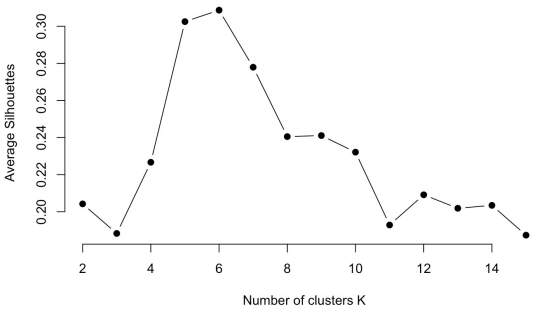


Fig. 22: ASW of K-means on Whole Dataset with Season as a binary variable

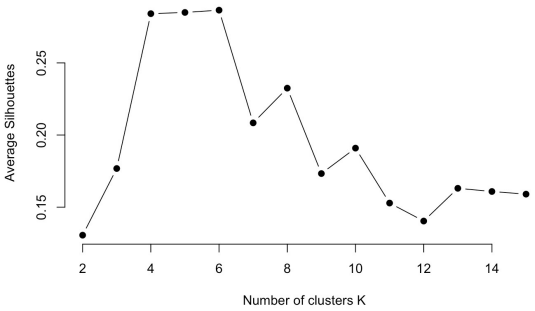


Fig. 23: ASW of PAM on Whole Dataset with Season as a binary variable

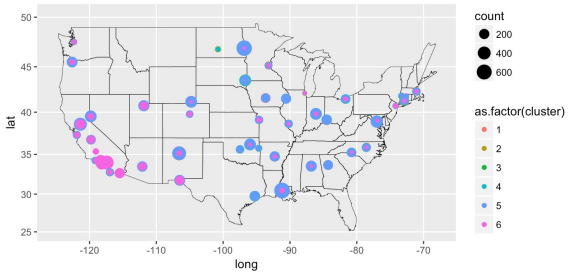


Fig. 24: K-means on Whole Dataset with K=6

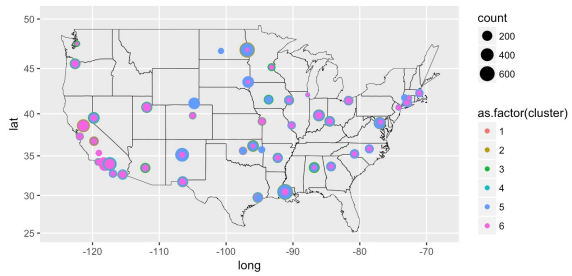


Fig. 25: PAM on Whole Dataset with K=6

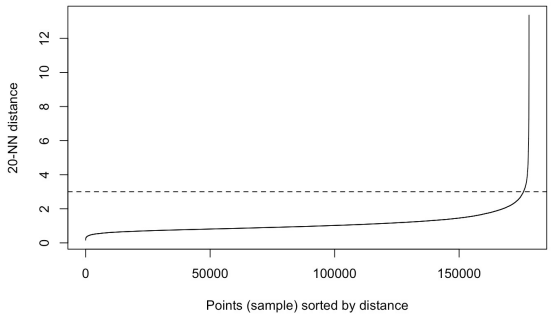


Fig. 26: DBSCAN on Summer Subset 20NN Plot

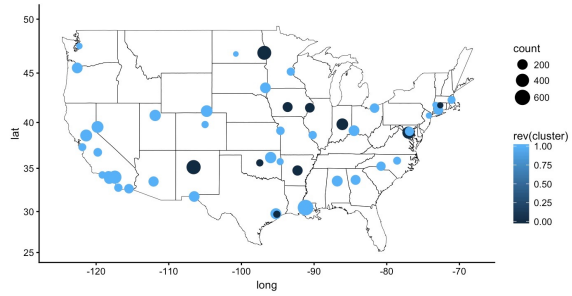


Fig. 27: DBSCAN on Summer Subset with eps =3 and MinPts = 20