**Clustering on Air Quality Data in the United States**Jiayang Xu

**Abstract**

The grouping of the temporal characteristics of air quality pattern is worth taking into account when investigating the air quality situation in an area. In this project, Euclidean k-means, Soft-DTW-k-means and Self-Organizing Map(SOM) are used and compared for the exploration of clustering on time series air quality data in the Continental United States. In the results, the concentration of PM2.5 and SO2 in 2019 did not show distinct spatial clusters while NO2 hadan apparent cluster in the California and O3 also had different seasonal fluctuation patterns in corresponding regions.

**I. Introduction**

Clustering of time series is an important problem arising in situations where a set of unlabelled time series needs to be split into classes according to an appropriate notion of similarity. It can be based on unsupervised machine learning algorithms and is employed in the areas such as financial time series analysis, Biodiversity analysis and identification of planets from luminosity readings[[1]](#endnote-1). Air quality data is a typical kind of time series data. Clustering of air quality data from different monitering sites in a certain area under a certain time span is of significance in terms of investigating the type of air quality, or even optimizing the monitoring networks in regional scale[[2]](#endnote-2). Some studies have practiced on the temporal patterns of air quality in the United States based on clustering. For example, Park et al.[[3]](#endnote-3) used k-means and rotated principal component analysis and produced six clusters of sites with distinct temporal patterns.

Here, this paper focused on finding the temporal patterns of pollutants in the Continental United States through time series clustering and visualizing the spatial patterns of the cluster groups on the map to confirm the rationality of the clusters.

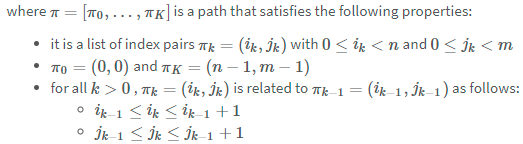
**II. Data**

The dataset is downloaded from the U.S. Environmental Protection Agency (EPA) Air Quality System’s Data Mart[[4]](#endnote-4). Four atmospheric components, O3, NO2, SO2 and PM2.5 were included in this project in the form of daily arithmetic mean of concentration data. These pollutants are among the six criteria pollutants identifified by EPA under the federal Clean Air Act[[5]](#endnote-5)(The other two criteria pollutants are carbon monoxide and lead). The dataset contains missing data. Monitoring sites that had values missing were excluded in this project.

**III. Methods**

Kmeans and Self-Organizing Map(SOM) are two major methods for clustering in this project. Kmeans clustering is a common method for unlabelled cluster analysis. The algorithm starts with initial estimates for the Κ centroids, assigns the data points to its nearest centroid and iterates until no data points change clusters. For TimeSeries Kmeans, python module tslearn[[6]](#endnote-6) is applied in the project, it provides machine learning tools for the analysis of time series. From this package, Euclidean k-means and Soft-DTW-k-means are used and compared in this project. Common [Euclidean distance](https://en.wikipedia.org/wiki/Euclidean_distance) metric is invariant to time shifts, ignoring the time dimension of the data, and thus [unsuitable](http://alexminnaar.com/2014/04/16/Time-Series-Classification-and-Clustering-with-Python.html) for time series[[7]](#endnote-7). Dynamic Time Warping(DTW) is an algorithms for measuring similarity between two temporal sequences. It is calculated as the squared root of the sum of squared distances between each element in X and its nearest point in Y:

(1)



DTW can compute the distance between the corresponding points by warping the time, while the normal Euclidean distance just computes the distance point to point. [Soft-DTW](https://arxiv.org/abs/1703.01541) is a differentiable variant of DTW that replaces the non-differentiable operation with a differentiable operation.

Self-Organizing Map(SOM) is a type of artificial neural network (ANN) that is trained using unsupervised learning for dimensionality reduction. It is also an excellent tool for clustering. In this project, Minisom[[8]](#endnote-8), a minimalistic and Numpy based implementation of the Self Organizing Maps (SOM) are used.

For the data preparation, the raw data was transformed into the correct format for training and was normalized. The visualization was mainly conducted on Basemap and matplotlib.pyplot.

The codes of this project can be obtained on the Github through the link https://github.com/jiangyoufeng/Project\_202012 (some repetitive codes and results are not shown in the Github).

**IV. Results**

**4.1 Determination of the number of clusters in kmeans**

Elbow method and silhouette score were tried to choose the best number of clusters in kmeans. Elbow method can plot the explained variation as a function of the number of clusters, so the elbow of the curve can be picked as the number of clusters to use in the clustering algorithm. Silhouette score computes the mean silhouette coefficient of all samples and silhouette coefficient is the distance between a sample and the nearest cluster that the sample is not a part of.

In this case, almost all the distortion lines did not present obvious elbows (shown in Figure 1, the example for PM2.5, 2019). Other figures were not shown here but they were in the similar situation as Figure 1. As for silhouette score, if the cosilhouette score is close to 1, the performance of the cluster is good. In the results, the silhouette scores of nearly all clustering for both Euclidean kmeans and Soft DTW kmeans were low(<0.3), which indicated that the data did not show distinct patterns. The silhouette scores for the clustering of SO2 was even close to 0 and for the clutering of NO2, O3 and PM2.5, the average of the highest scores were about 0.1-0.3. The number of the clusters of the highest cosilhouette scores were also changed when the random seeds were altered.



Figure 1. Distortion lines(PM2.5, 2019)

**4.2 Clustering patterns in the results of Euclidean kmeans**

The outputs resulted from 3 different seeds(0, 1, 2) were presented to show the effects of the random initialization. As shown in the results, kmeans is a method vulnerable to the randomly chosed initial center so the outcome patterns were not stable when the random seeds were changed. However, the patterns of times series could also be artifically identificated by comparing the time series plots of clusters. Different numbers of clusters were also tried to visually display the characteristic of the spatial patterns since there were no evident elbows observed in distortion lines. As the figures shown below, the number of clusters did not always make sense because some clusters which concentration were close to average and did not strongly fluctuate in a year fail to show obvious differences that can be easily recognized.

For SO2 and PM2.5, the outcome of spatial patterns were not apparent since different clusters overlapped when the random seeds were changed. The clusters are more scattered spatially compared to O3 and NO2(Figure 4&5).

As for SO2(Figure 2), in sites in the North Dakota(yellow circles), the SO2 concentration had a peak in the beginning of the year while in some sites scattered in the eastern United States(blue circles), SO2 concentration had peaks in winter and spring. For other sites, SO2 concentration was stable in the whole year. For PM2.5(Figure 3), there was not strong patterns in specific areas.

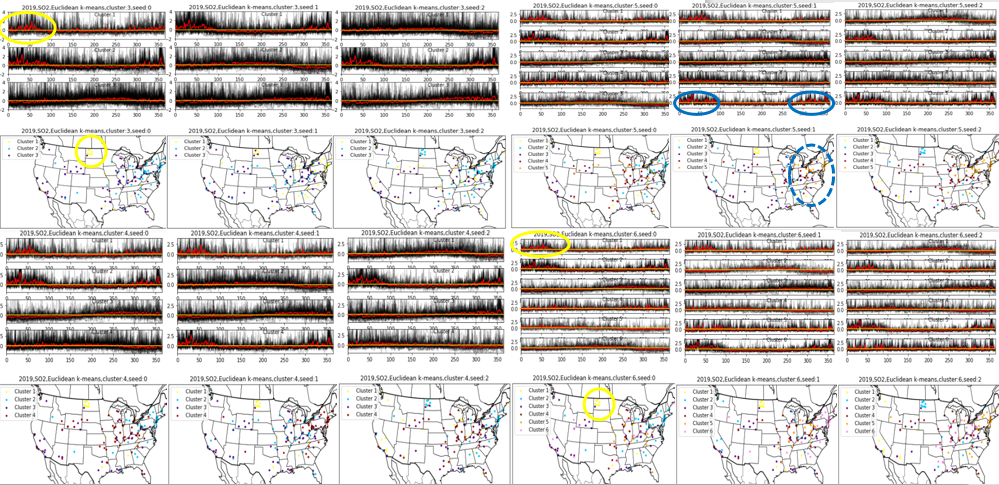


Figure 2. Clusters by Euclidean k-means (SO2, 2019)

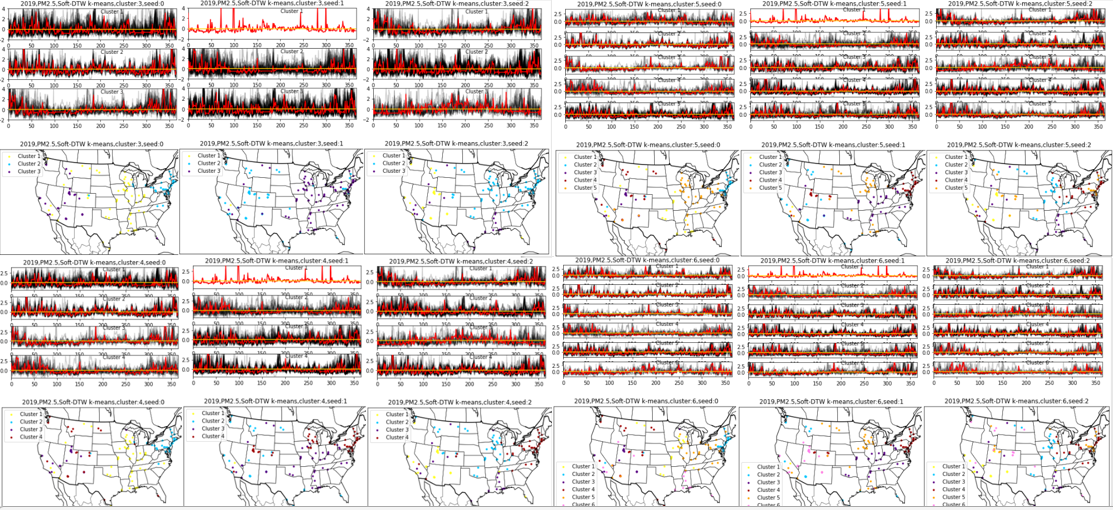


Figure 3. Clusters by Euclidean k-means (PM2.5, 2019)

In terms of O3 and NO2, in some regions, the spatial patents were relatively clearer and the differences of the cluster assignment derived from different initialized centers were smaller.

For O3 (Figure 4), the concentration of O3 in several states in the southwest of the United States(blue circles) was relatively higher from Feburary to September, and became lower in late autumn and winter. However, the west coast area(yellow circles), especially the sites near the San Francisco, showed a different pattern characterizing by two slow peaks -- the concentration of O3 dropped in June. The northern regions(green circles) tended to have a peak of O3 concentration in the spring.

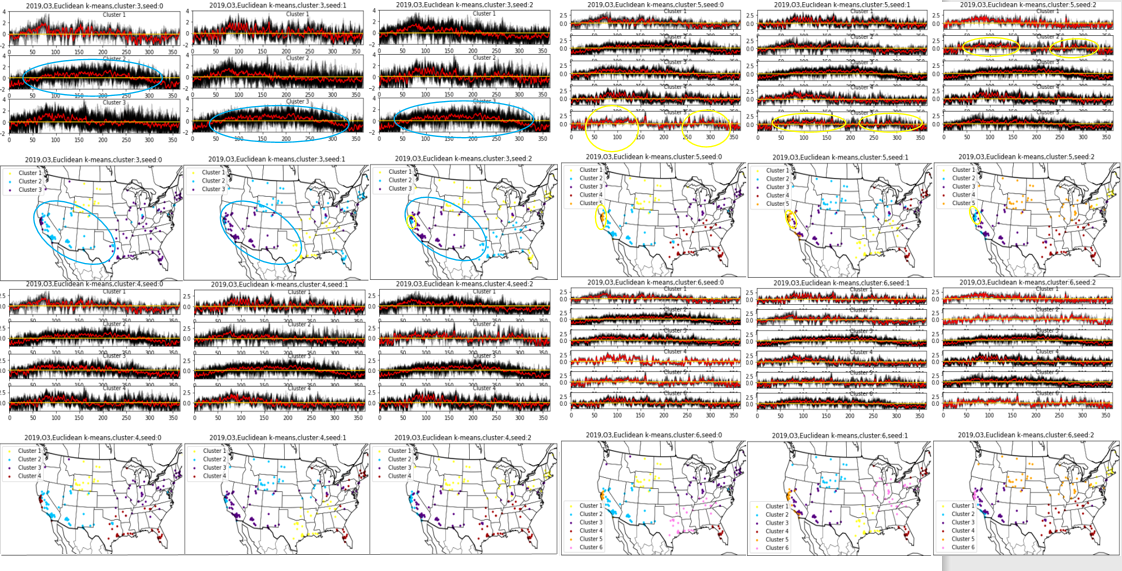


Figure 4. Clusters by Euclidean k-means (O3, 2019)

For NO2 (Figure 5), California(yellow circles) has a distinct pattern featured by an significant rise in October, which indicates that Euclides k-means clustering has the ability to discern the obvious local fluctuations. Except this cluster, other clusters show similar pattern—the concentration of NO2 changed greatly in spring and winter and was more consistent in the middle of the year.

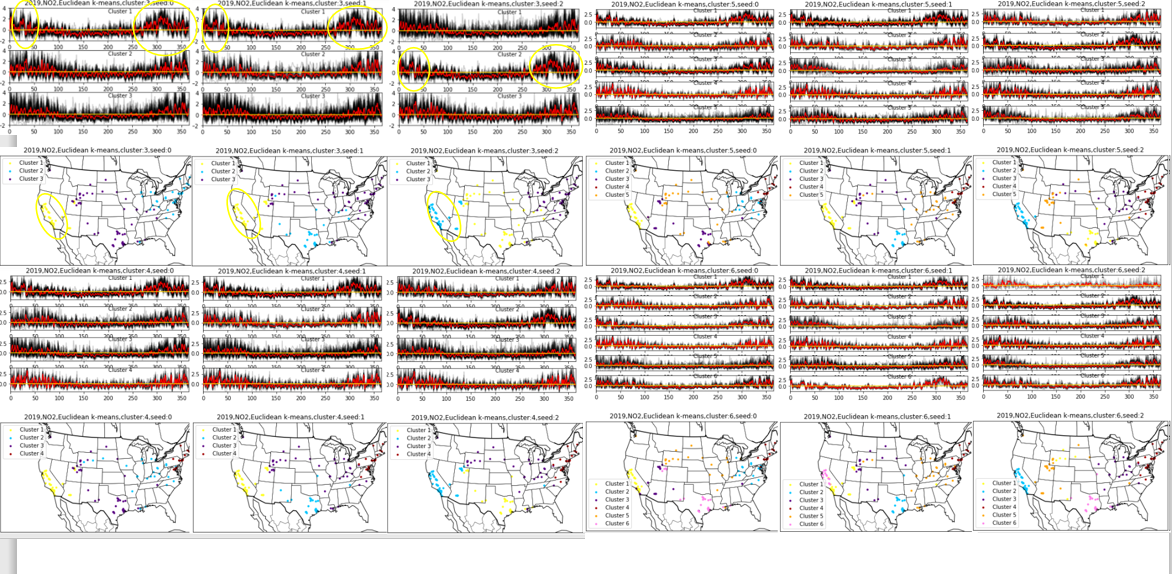


Figure 5. Clusters by Euclidean k-means (NO2, 2019)

In summary, roughly some spatial pattents in certain areas were consistant when the random seeds and the number of clusters varied and they showed a distinct characteristic of the fluctuation of air quality over time compared to other regions.

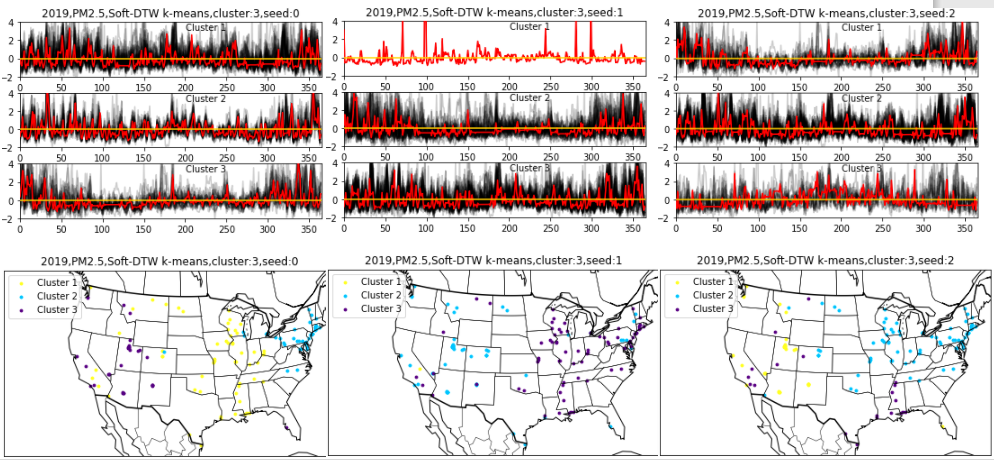
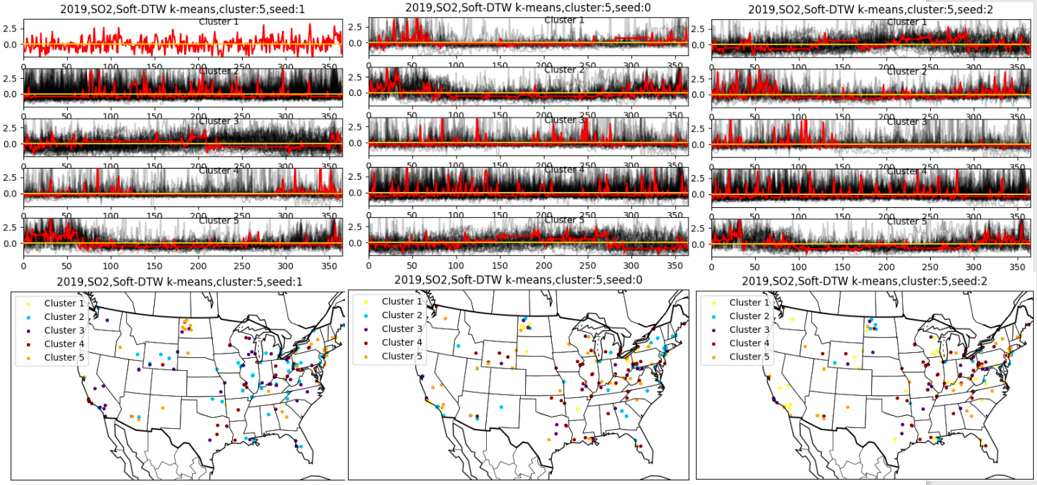
**4.3 Results in Soft DTW kmeans**

Running Soft-DTW-k-means algorithm took much longer time, generally spend about 40-50 minutes. Since the python program was not constructed on tensorflow, multiprocessing(3 processes) was tried to accelerate the processes.



The results of Soft DTW kmeans clustering did not display an obvious optimization compared to Euclidean k-means clustering. The information they obtained were quite similar(except SO2). Since clustering is an unsupervised learning method, it was also hard to use a concrete result-testing standard to compare between these two algorithums.

Some examples of the outcomes of Soft DTW kmeans clustering were shown in Figue 6. As mentioned in 4.2, the results of PM2.5 did not have obvious clusters for spatiotemporal distribution. Three different seasonal variation patterns could be detected for O3 and the west coast had a relatively different pattern for NO2. As for SO2, Soft DTW kmeans clustering showed different results. Like PM2.5, the results of SO2 showed no obvious clusters for spatiotemporal distribution. The special pattern of North Dakotadetected in Euclidean k-means was not displayed in the results of Soft-DTW-k-means(red dotted circle), which might indicate that the former pattern could be not that obvious.

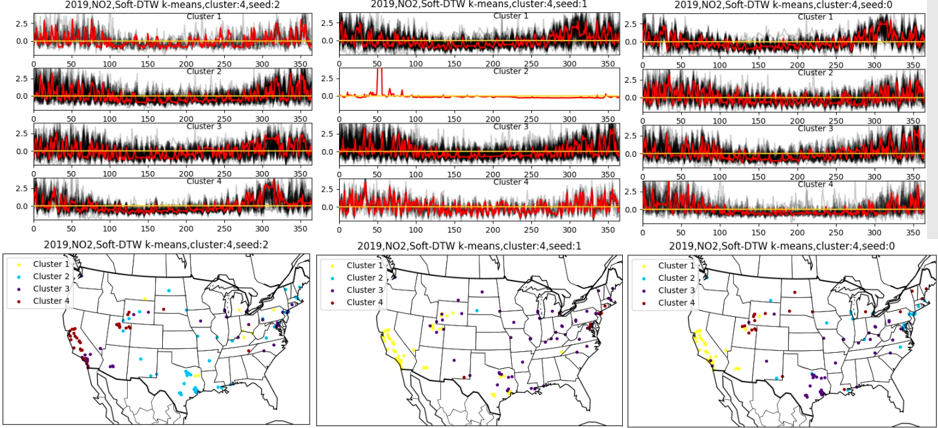
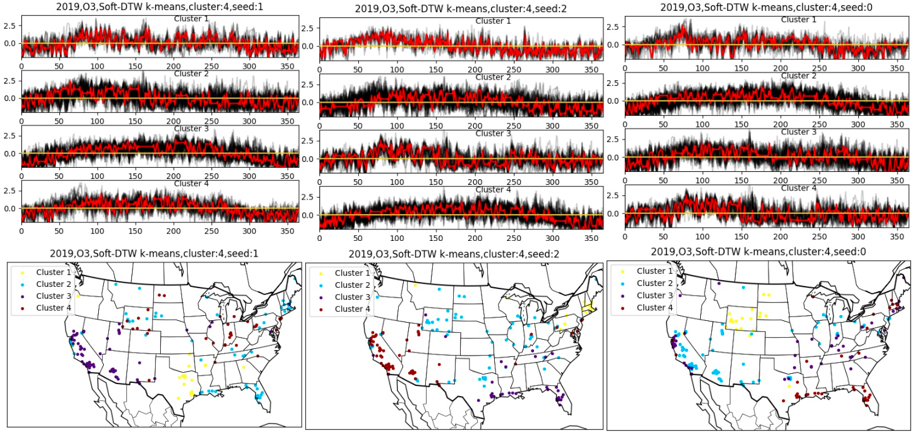


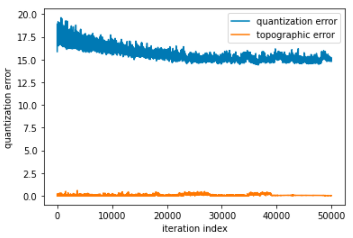
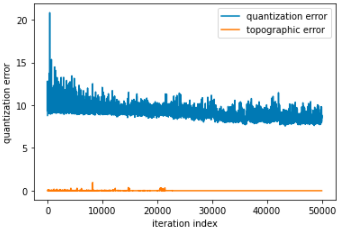
Figure 6. Clusters by Soft-DTW-k-means

(PM2.5, SO2, O3, NO2 from left to right, from top to bottom)

**4.4 Application of SOM**

SOM algorithm could be completed within one minute. Unlike kmeans, one of the advantages of SOM is that its results did not vary with random seeds.

In Minisom, quantization error can be used to show how much information lost. If the quantization error is 0, the weights of the network are exactly as the original data[[9]](#endnote-9). Learning curves were used to estimate the number of iterations. The maximum iteration was save to set 20000(Figure 7), but 50000 was also acceptable because SOM run quite fast. The quantization errors were still high(>8) when the curves reached pleateau.

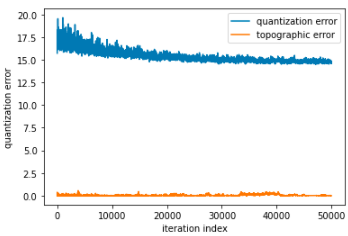
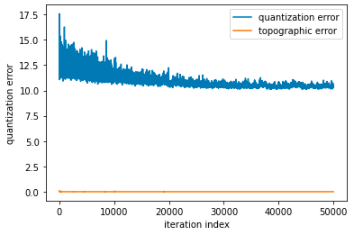
 

Figure 7. Learning curves (PM2.5, SO2, O3, NO2 from left to right, from top to bottom, (2,3))

When the parameter sigma in SOM was changed, the clustering outcome was changed and the quantization error was slightly changed. Take O3 for example, when the value of sigma were 0.1, 0.5, 0.8, 1, the quantization error were 15.5, 15.7 15.8, 16.4. The variation was not very large so sigma was set to be 0.5(0.5 is also the default).

As shown in Figure 8, the clustering results of SOM for PM2.5 could detect the sites where the concentrations were more volatile (cluster 1, 2). For SO2, some outliers(cluster 3, 4, 5, 6) could be detected. For O3, there were patterns a bit similar to the results of previous 2 methods(like cluster 2), but seasonal fluctuation trends of the mean of the clusters were smaller. For NO2, the feature previously observed in California (a rise in October) still exsisted but not that clear.

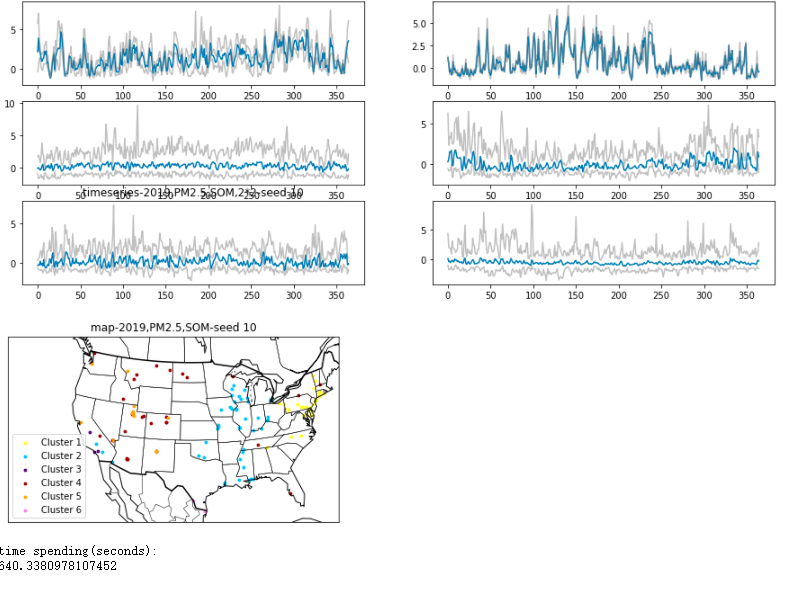
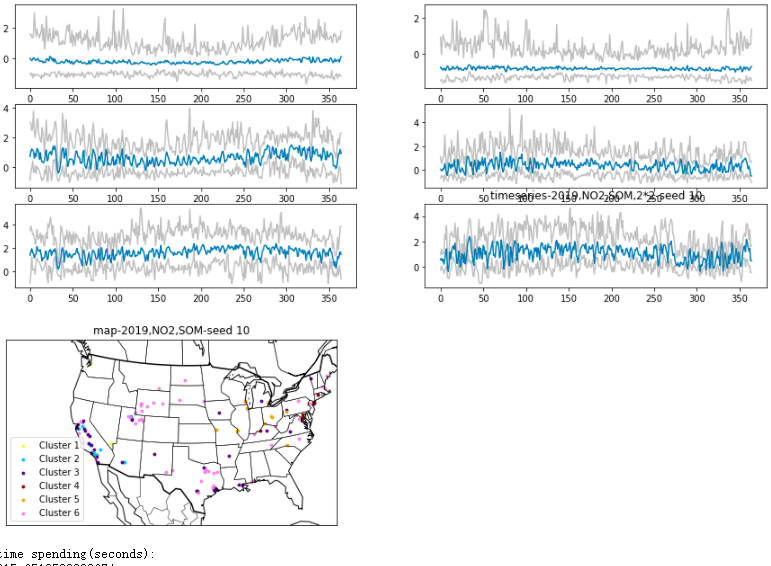
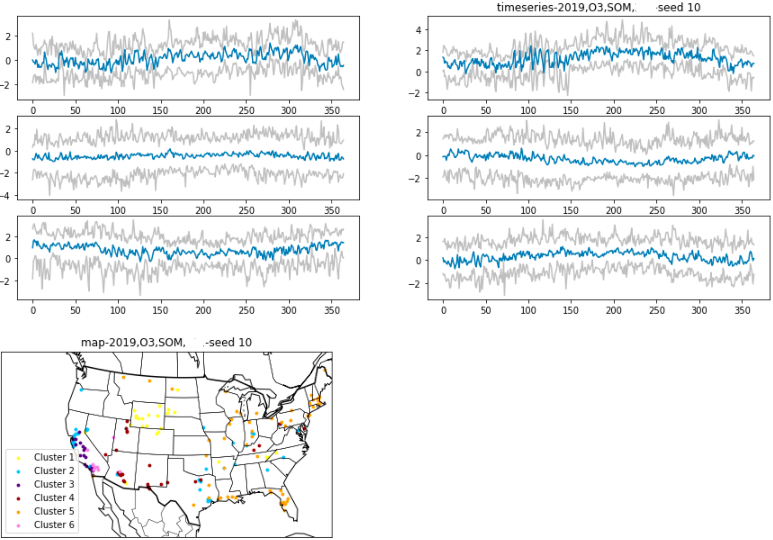
  

Figure 8. Clusters by SOM (PM2.5, SO2, O3, NO2 from left to right, from top to bottom)

**V. Conclution**

Overall, the information about spatiotemporal characteristics of air quality patterns obtained from Euclidean-k-means, Soft-DTW-k-means and SOM were generally consistent. The concentration of PM2.5 and SO2 in 2019 did not show distinct spatial clusters while NO2 hadan apparent cluster in the California and O3 also had some seasonal fluctuation patterns in corresponding regions.

In this case of air quality time-series data clustering, simple clustering methods like Euclidean k-means already has the ability to detect the distinct patterns of seasonal fluctuation. These clustering methods are useful to detect the regions of some characteristic air quality time series patterns. However, it was hard to use these clustering results to separate the details since most clusters were not distinct from each others, and the the resutls were not stable when the random seed were changed. These poor performances may be partly in consistent with the poor cosilhouette scores. It's also important to note that kmeans clustering results are sensible to the position of randomly choosed initial points, so runnning more than one random state and taking all their outcomes into account may be helpful for a more solid conclusion, especially when the clustering efficiency is relatively low.

Although the clustering efficiency between Euclidean k-means and Soft-DTW-k-means could not be easily compared due to the lack of labels and the poor cosilhouette scores for both methods, Soft-DTW-k-means did not seem to perform much better than Euclidean k-means through the direct observation of the results. The Soft-DTW-k-means was theoretically more reasonable but it was much too time-consuming.

**Reference:**

1. J. Vesanto and E. Alhoniemi. Clustering of the self-organizing map. IEEE Transactions on Neural Networks, 2000(11): 586-600, doi: 10.1109/72.846731. [↑](#endnote-ref-1)
2. P. Govender, V. Sivakumar. Application of k-means and hierarchical clustering techniques for analysis of air pollution: A review (1980–2019),Atmospheric Pollution Research, 2020(11): 40-56. https://doi.org/10.1016/j.apr.2019.09.009. [↑](#endnote-ref-2)
3. Sun-Kyoung Park, Amit Marmur, Armistead G. Russell. [Environmental Risk Assessment: Comparison of Receptor and Air Quality Models for Source Apportionment](https://www.tandfonline.com/doi/abs/10.1080/10807039.2012.730475). Human and Ecological Risk Assessment: An International Journal. 2013(19): 1385-1403. [↑](#endnote-ref-3)
4. US Environmental Protection Agency. Air Quality System Data Mart [internet database] available at <http://www.epa.gov/ttn/airs/aqsdatamart> (accessed 2020). [↑](#endnote-ref-4)
5. Clean Air Act, U.S. Environmental Protection Agency. available at <http://www.epa.gov/oar/caa/caa.txt> (accessed 2008). [↑](#endnote-ref-5)
6. Romain Tavenard, et al. Tslearn, A Machine Learning Toolkit for Time Series Data, Journal of Machine Learning Research, 2020(21):1-6. http://jmlr.org/papers/v21/20-091.html [↑](#endnote-ref-6)
7. Alexandra Amidon. How to Apply K-means Clustering to Time Series Data. https://towardsdatascience.com/how-to-apply-k-means-clustering-to-time-series-data-28d04a8f7da3#:~:text=Dynamic%20Time%20Warping%20%28DTW%29%20

   is%20used%20to%20collect,a%20group%20of%20time%20series%20in%20DTW%20space. [↑](#endnote-ref-7)
8. Giuseppe Vettigli. MiniSom: minimalistic and NumPy-based implementation of the Self Organizing Map, https://github.com/JustGlowing/minisom/ [↑](#endnote-ref-8)
9. Issues: quantization error (theoretical question) https://github.com/JustGlowing/minisom/issues/36 [↑](#endnote-ref-9)