Identification of air quality redundant stations through a clustering ensemble method

T Stolz¹, M Huertas² and A Mendoza²
¹ETH Zurich, Institute of Geophysics, Sonneggstrasse 5, 8092 Zurich, Switzerland
²Tecnologico de Monterrey, Escuela de Ingeniería y Ciencias, Av. Eugenio Garza Sada 2501, Monterrey, N.L., 64849, México

Corresponding author e-mail: mendoza.alberto@tec.mx

Abstract. The main purpose of this study is to assess the performance of three air quality monitoring networks of Mexico. Emphasis is placed on an ensemble method to combine the results of the different clustering techniques: Principle Component Analysis, Hierarchical Clustering and k-means. The specific objectives of this paper are: (i) finding similar and redundant stations using the ensemble method and (ii) giving a physical meaning to groups of similar stations by evaluating additional information like emission sources, meteorology and topography of the area of interest. The study was applied on time series data of particulates that have aerodynamic diameters less than or equal to 10 µm (PM₁₀) and ozone (O₃), acquired from the air pollutant monitoring systems in the metropolitan areas of Mexico City (MCMA), Monterrey (MMA) and Guadalajara (GMA). These three conurbations are characterized by diverse meteorological and geographical conditions. The findings show that the GMA has a well distributed air quality network with the fewest number of similar stations. The MMA presents the same clusters of stations for PM₁₀ and O₃, while in the MCMA a cluster of possible redundant stations is found. Results confirm that the clustering ensemble method is a confidence tool to identify similar stations.

Keyword: Cluster analysis, PCA, ozone, particulate matter, hierarchical clustering.

1. Introduction

The design of an AQMN requires to define the objectives of the network, which can depend on the required data usage, the area to be covered and the spatial variability of the pollutants [1]. These aspects define the distribution and the final number of air quality stations. Similar, but not redundant, air quality stations are expected to optimize and manage AQMN. An air quality station is defined as similar when its measurements are correlated with the data of another station, but still contributes with new information about the pollutant concentration. Whereas, a redundant station duplicates the same data of another air quality station.

Diverse approaches have been applied to identify redundant stations in AQMN: positive matrix factorization, artificial neural networks, multi-objective optimization, correlation analysis, Self-Organizing Maps, among others [2]. Principle Component Analysis (PCA) and hierarchical clustering are the most common approaches used to aggregate air quality stations [2]–[6]. However, there is not a general consent criteria to select the most proper clustering technique [7]. The literature suggests the use of a clustering ensemble method, which is a process of aggregation constructed from multiple clustering
techniques into a single robust result [8]. Normally, clustering aggregation provides a more reliable, accurate and consistent clustering result compared to the results from individual members of the ensemble [7].

In this study, a clustering ensemble method was implemented using three clustering models: PCA, hierarchical clustering, and k-means. The clustering aggregation was applied on the PM$_{10}$ and O$_3$ data from the air quality monitoring networks of the three largest metropolitan areas of Mexico: Mexico City (MCMA), Monterrey (MMA) and Guadalajara (GMA). The three conurbations are characterized by different geographic and meteorological conditions. The objective is to assess the spatial representativeness of the AQMN, identifying similar and possible redundant stations including a physical analysis of the results from the clustering ensemble method.

2. Material and methods

2.1. Study area
The MCMA is the most populated area in Mexico and the seventh largest megacity in the world [9] with 20.1 million inhabitants [10]. Its atmospheric monitoring network (SIMAT by its Spanish acronym) consists nowadays of more than 40 stations and was established in 1986. The MMA is the industrial center and the largest metropolitan area of northern Mexico with a population of 4.1 million inhabitants [10]. The air pollutants in the MMA are measured by the Integral Environmental Monitoring System (SIMA by its Spanish acronyms) which was established in 1992 and currently consists of 13 stations. The second largest city in Mexico is the GMA with a population of 4.4 million people [10]. The atmospheric monitoring network of the State of Jalisco (SIMAJ by its Spanish acronyms) monitors the pollutant concentration in the GMA since 1993 and includes 10 stations.

2.2. Database
Long-term hourly measurements of PM$_{10}$ and O$_3$ were extracted from records of the MCMA, MMA and GMA air quality networks. PM$_{10}$ and O$_3$ data quality is assured by the measurement methods and calibration procedures defined by the Mexican legislation. The air quality stations in each metropolitan area were selected depending on the PM$_{10}$ and O$_3$ data availability for the most recent years. The data sets were prepared by removing outliers, interpolating single missing data values and standardizing the time series. All processing was carried out with Excel 2018 and Python 3.6.

Table 1 shows the selected data set in each metropolitan area. For the MCMA data was taken from 2014 to 2016, for the MMA from 2012 to 2013 and for the GMA from 2014 to 2016.

| Metropolitan area | Name         | Symbol | Latitude | Longitude |
|-------------------|--------------|--------|----------|-----------|
| MCMA              | Atizapan     | ATI    | 19.576963| -99.254133|
|                   | Camarones    | CAM    | 19.468404| -99.169794|
|                   | Cuajimalpa   | CUA    | 19.365313| -99.291705|
|                   | FES Acatlán  | FAC    | 19.482473| -99.243524|
|                   | Hospital General | HGM | 19.411617| -99.152207|
|                   | Iztacalco    | IZT    | 19.384413| -99.117641|
|                   | Merced       | MER    | 19.424610| -99.119594|
|                   | Tláhuac      | TAH    | 19.246459| -99.010564|
|                   | Talnepantla  | TLA    | 19.529077| -99.204597|
|                   | Tultitlán    | TLI    | 19.602542| -99.177173|
2.3. Statistical methods

The clustering ensemble method is an aggregation of multiple clustering techniques. In this study, hierarchical clustering, PCA, and k-means were the clustering techniques selected. The clustering ensemble framework is shown in Figure 1, where $P_k^n$ describes the possible $k^{th}$ partition of the network with $k = 1, \ldots, K$ where $K$ are different individual results from the application of each clustering technique, allowing exactly $n$ clusters. The number of clusters ranges from two to the maximum number of principal components considered for each metropolitan area.

$$CR_k^n(\%) = 100 \left[ \frac{1}{3} P_{k, \text{PCA}}^n + \frac{1}{3} P_{k, \text{k-means}}^n + \frac{1}{3} \sum_{j=1}^{7} \frac{1}{7} P_{k, \text{HCl}}^n \right]$$

Figure 1. Clustering ensemble method framework applied for O$_3$ and PM$_{10}$ datasets in the MMA, MCMA and GMA.

The consensus result ($CR_k^n$) is a measure of similarity among the selected clustering methods. $CR_k^n$ describes how often $P_k^n$ was chosen by the different clustering methods and it is computed according to Equation 1. In this case, seven approaches of hierarchical clustering are possible ($j$).

| Location | Code | Latitude  | Longitude |
|----------|------|-----------|-----------|
| UAM Iztapalapa | UIZ | 19.360794 | -99.073880 |
| Villa de las Flores | VIF | 19.658223 | -99.096590 |
| Xalostoc | XAL | 19.525995 | -99.082400 |
| La Pastora | SE | 25.668000 | -100.249000 |
| San Nicolás | NE | 25.750000 | -100.255000 |
| Obispo | CE | 25.670000 | -100.338000 |
| San Bernabé | NW | 25.757000 | -100.366000 |
| Santa Catarina | SW | 25.676000 | -100.464000 |
| García | NW2 | 25.783000 | -100.586000 |
| Escobedo | N | 25.800000 | -100.344000 |
| Apodaca | NE2 | 25.777000 | -100.188000 |
| Atemajac | ATM | 20.719444 | -103.35527 |
| Oblatos | OBL | 20.700278 | -103.29638 |
| Loma Dorada | LOD | 20.629167 | -103.26388 |
| Tlaquepaque | TLA | 20.640833 | -103.14166 |
| Centro | CEN | 20.673611 | -103.33305 |
| Vallarta | VAL | 20.680000 | -103.39833 |
| Aguilas | AGU | 20.630833 | -103.41666 |
| Las Pintas | LPS | 20.576667 | -103.32638 |
| Santa Fe | SAN | 20.528888 | -103.376944 |
\( P_{k,n} \) is the partition from the clustering technique \( i \). \( P_{k,n} \) takes a value of one (1) if the partition is equal to at least one partition with exactly \( n \) clusters within other clustering method and zero (0) if the partition does not coincide with any other partition from a different clustering technique. The \( CR_{k,n} \) is computed iteratively from two clusters to the maximum number of principal components, and its maximum value is 100%. The final clustering solution (FCR) is the consensus result with the highest similarity (Equation 2).

\[
FCR (\%) = \max \{ CR_{k} \}
\]  

(2)

3. Results and discussion

3.1. Clustering ensemble result in the MCMA

The FCR for PM\(_{10}\) and O\(_{3}\) measurements in the MCMA are shown in Figure 2. The best agreement among the different methods for PM\(_{10}\) is found for seven clusters with a FCR of 100%. The PM\(_{10}\) clusters from the ensemble method consist of one, two and four air quality stations. Monitoring sites aggregate in the same cluster are direct neighbors and have a separation distance of around 10 km. The FCR for O\(_{3}\) consists of two groups which perfectly divide the metropolitan area in north and south. The basin circulation in the MCMA is mainly affected by the interaction of strong thermally induced local winds and synoptic conditions, which promotes convergence zones that crosses the MCMA as the day progresses [11]. This phenomenon explained the results found in the MCMA for O\(_{3}\).

![Figure 2](image-url)

**Figure 2.** Final clustering result from the ensemble method in MCMA for PM\(_{10}\) (left) and O\(_{3}\) (right).

3.2. Clustering ensemble result in the MCMA

The FCR for O\(_{3}\) and PM\(_{10}\) are shown in Figure 3. The air quality network is divided in five clusters for PM\(_{10}\) as well as O\(_{3}\). In both cases, the ensemble method shows a FCR of 100%, being the unique metropolitan area with the same number of clusters for both pollutants. Based on the available information is not possible to clearly deduce the reason of this match, maybe, this result can be influenced by the geographical and meteorological conditions in the MMA. This metropolitan area is highly affected by easterly anticyclonic air masses that come from the Gulf of Mexico [12] and it is surrounded by Sierra Madre Oriental mountains chain having high mountains (>1000 m a.s.l.), which act as natural barriers to the pollutant dispersion.
The stations which are group together are probably similar but not redundant taking into account the distance (>8 km) between SW and CE stations, and the mountain that divides the N, NW, and NE stations.

![Figure 3. Final clustering solution from the ensemble method in the MMA for PM$_{10}$ and O$_3$](image)

3.3. *Clustering ensemble result in the GMA*

The FCR suggests eight clusters for PM$_{10}$ and six clusters for O$_3$ (Figure 4), showing a similarity value of 67% and 90%, respectively. For PM$_{10}$ and O$_3$, a maximum of two air quality stations are grouped together, and they are probably similar but not redundant. This result can be explained because the AQMN in the GMA was designed to have stations with a spatial representativeness of 5 km for O$_3$ and 2 km for PM$_{10}$ [13], and the current stations are located in a distance larger than 5 km. The GMA was the region with the lowest FCR in both pollutants, which indicates that this air quality network is well distributed over the metropolitan area.

![Figure 4. Clustering result from ensemble method in the GMA for PM$_{10}$ (left) and O$_3$ (right).](image)

4. *Conclusions*

The clustering ensemble method was a robust technique to identify similar and probable redundant stations in the air quality monitoring networks of the MMA, MCMA and GMA. This is the first time where a clustering ensemble method is applied to evaluate similar air quality stations.
The final clustering results showed that the GMA was the region with the lowest percentage of similarity in both pollutants, which support the idea that this air quality network is well distributed over this urban area. Only in the MMA the same number of clusters were found for PM$_{10}$ and O$_3$ with an FCR of 100%. The clustering ensemble method clearly shows two areas of O$_3$ in the MCMA that are well explained by the thermal induced local wind. In the center of MCMA, for PM$_{10}$, a cluster was detected that probably includes redundant stations. However, further tools are required to fully proof redundancy of monitoring stations.

Acknowledgement
We thank to the Secretariat for the Environment of Mexico City, the Secretariat for Sustainable Development of the Nuevo León State, and the Secretariat for the Environment and Territorial Development of the Jalisco State for the public domain records.

References

[1] CCME, Ambient Air Monitoring Protocol for PM2.5 and Ozone. 2011. Canada-wide Standatds for Particulate Matter and Ozone.
[2] Zhao L., Xie Y., Wang J., and Xu X. 2015. A performance assessment and adjustment program for air quality monitoring networks in Shanghai, Atmos. Environ.
[3] Lu W., He H., and Dong L. 2011. Performance assessment of air quality monitoring networks using principal component analysis and cluster analysis, Build. Environ.
[4] Pires J., Sousa S., Pereira M., Alvim-Ferraz M., and Martins F. 2008. Management of air quality monitoring using principal component and cluster analysis-Part I: SO2 and PM10, Atmos. Environ.
[5] Pires M., Sousa S., Pereira M., Alvim-Ferraz M., and Martins F. 2008. Management of air quality monitoring using principal component and cluster analysis-Part II: CO, NO2 and O3, Atmos. Environ.
[6] D’Urso P., Massari R., Cappelli C., and De Giovanni L. 2017. Autoregressive metric-based trimmed fuzzy clustering with an application to PM10 time series. Chemom. Intell. Lab. Syst., pp. 1–20.
[7] Alqrashi T. and Wang T. 2018. Clustering ensemble method,” Int. J. Mach. Learn. Cybern., pp. 1–20.
[8] Wang W. 2008. Some fundamental issues in ensemble methods, in Proceedings of the International Joint Conference on Neural Networks.
[9] United Nations (UN). 2016. The World’s Cities in 2016: Data Booklet.
[10] INEGI. 2014. Las zonas metropolitanas en México, Censos Económicos 2014,
[11] de Foy B., Fast, J., Paech S., Phillips D., Walters J., Coulter R., Martin T., Pekour M., Shaw W., Kastendeuch P., Marley, Retama A., Molina,L., 2008. Basin-scale wind transport during the MILAGRO field campaign and comparison to climatology using cluster analysis. Atmos. Chem. Phys.
[12] Hernández-Paniagua I., Clemitshaw K., and A. Mendoza. 2017. Observed trends in ground-level O3 in Monterrey, Mexico, during 1993-2014: Comparison with Mexico City and Guadalajara, Atmospheric Chemistry and Physics.
[13] SEMARNAT. 2010. Programa para Mejorar la Calidad del Aire, Jalisco 2011-2020.