Analysis of air quality data in the city of Bogotá through clustering techniques

IOP Conf. Series: Materials Science and Engineering **872** (2020) 012027 doi:10.1088/1757-899X/872/1/012027

Analysis of air quality data in the city of Bogotá through clustering techniques

Jesús Silva[1], Luz Adriana Londoño[2], Noel Varela[3], Omar Bonerge Pineda Lezama[4], Liliana Patricia Lozano Ayarza[5], Julio Cesar Mojica Herazo[6]

[1]Universidad Peruana de Ciencias Aplicadas, Lima, Perú.
[2,3,5,6]Universidad de la Costa, Barranquilla, Atlántico, Colombia
[4]Universidad Tecnológica Centroamericana (UNITEC), San Pedro Sula, Honduras

[1]Email: jesussilvaUPC@gmail.com

Abstract. Climate change is one of the problems facing society today because of its impacts on the health of living beings. That is why the authorities need tools that provide them with the information necessary to make decisions that will reduce the impact of such change. This paper proposes a strategy to group the air quality data of the metropolitan area of the City of Bogotá from 2014 to 2018, in order to recognize the measurement patterns in the environmental contaminants that cause pre-contingencies and environmental contingencies in the area of the City of Bogotá.

1. Introduction

In the course of his existence, the human being has changed his environment to live comfortably and safely, proof of which are the great distances he has travelled through sky, sea, land and space. Technological advances have facilitated daily habits, business, the manufacture of large quantities of products, etc. However, these advances have led to environmental degradation that seriously threatens the current and future development of nations [1], [2].

Air pollution or atmospheric contamination is a problem that produces climate change throughout the world and affects the health of millions of people. It is for this reason that technological tools that contribute to the study of this pollution are of vital importance in the development of policies that eradicate pollution or mitigate its effects [3].

Currently, several organizations and governments have implemented mechanisms to measure air pollutants in order to know the air quality indices of the different regions of the planet. The air quality indices are numbers used by government agencies to determine air quality. In the city of Bogotá and in the Bogotá Valley area (ZVB) air pollution is measured by the Metropolitan Air Quality Index (MAQI). The MAQI is used to show the level of pollution and the level of risk it represents to human health in a given time in order to take protective measures [4][5].

In [6], the author proposes a Business Intelligence application to analyze climate change data for the southern zone of the Puebla Valley. The results of the Business Intelligence processes applied to the air quality data of the southern zone of the Puebla Valley, presented by the author, point to a very strong relationship between air quality and climate variables, and also show that air quality with respect to the concentration levels of atmospheric pollutants is determined by the presence of particles

smaller than 10 micrometers (PS10) and ozone chemicals (O3). In the city of Puebla there are 3 stations for measuring pollutants, but there is no model of its own that provides information on air quality. This study seeks to obtain similar conclusions to those obtained by the author of the previously described research, although the focus of its strategy only takes into account the air quality information for the Metropolitan Area of the City of Bogotá.

Another study related to the analysis of climate change data is that presented by [7], which proposes a genetic algorithm to group climate change data from the Bogotá Valley Zone. In the paper, the authors create the patterns from the measurements of several stations in the studied region and group them to determine the type of pollutants that are key for the activation of an environmental contingency, according to air quality standards. This grouping strategy resulted in the obtention of 10 clusters, in which climate change data can be grouped, concluding that the patterns that represented higher measurement levels of certain pollutants, such as (PS10) and ozone coincide with high MAQI measurements [8].

According to the strategy proposed in this paper, each known measurement of pollutants is taken as a pattern in which the attributes of the pattern are the values of each pollutant and thus their clustering will lead us to conclusions about the relationship between pollutant values and air quality. The literature offers several techniques for grouping data [9], however, a K-Means method has been used for grouping air quality data in this strategy.

2. Proposed clustering strategy

The proposed clustering strategy for climate change data is to use the K-means method. The instances are formed from the information of the pollutant's measurements of the Bogotá Valley Zone. Each instance consists of a vector that contains the measurements of six pollutants criteria for each hour of a certain year (from 2014 to 2018), resulting in a total of 11,254 instances per year. These measurements are grouped using the technique mentioned above and the groups generated from the silhouette of the resulting clusters are validated.

2.1.Data preparation

The first phase of the strategy is to prepare the data for clustering. The original data consists of a set of spreadsheets containing the measurements from various stations. However, many of the stations report negative values, which is impossible and indicates a failure in the station, which is why this study is based on only one station, which is the one that reports the least inconsistent values. The presence of these incorrect values was corrected by substituting those incongruent values with the arithmetic mean of the correct values.

2.2.K-means clustering

One of the most commonly used non-hierarchical clustering algorithms is the K-means algorithm which is used to find clusters of air quality data, due to its easy implementation and fast execution [10]. This is due to its easy implementation and fast execution [10]. This algorithm was introduced in the sixties [11] [12], and starts with a problem of m attributes, that is, each instance is moved to m-dimensional space. The centroid of the cluster describes each cluster and is a point in the m-dimensional space around which each instance is grouped. The most used distance from the instance to the centroid of the cluster is the Euclidean distance. The K-means algorithm consists of two main steps:

1. The assignment step consists of moving each instance to the nearest class.
2. The re-estimation step consists in recalculating the class centroids from the instances assigned to each class (cluster).

The two steps of the algorithm are repeated until the re-estimation step produces a minimal change in the centroids of the classes.

Once the data is corrected, the construction of the K-means algorithm can be used to group the data. The instances are formed by the measurements of the criteria pollutants. Criteria pollutants is a term

used internationally to describe air pollutants that have been regulated and are used as indicators of air quality. The criteria pollutants are: Ozone (O_3), Sulfur Dioxide (SO_2), Nitrogen Dioxide (NO_2), Carbon Monoxide (CO) and Suspended Particles (SP). With that information, the instances are formed as vectors as follows [13]:

$$X = (X1, X2, X3, X4, X5, X6) \quad (1)$$

Where:

X1 corresponds to the measurement of Carbon Monoxide (CO), X2 corresponds to the measurement of Nitrogen Dioxide (NO_2), X3 corresponds to the measurement of Nitrogen Oxide (NOX), X4 corresponds to the measurement of Ozone (O_3).

X5 corresponds to the measurement of suspended particles smaller than 10 micrometers (PM10).

X6 corresponds to the measurement of Sulphur Dioxide (SO_2).

It is necessary to mention that the measurement of nitrogen oxide (NOX) is added, since ozone is created by chemical reactions within this compound [14].

3. Experiments and results

As mentioned above, there are 11,254 instances per year, and this is the five years over which the data analysis reported in this paper was done. This information was obtained from the Ministry of Environment of Colombia [15]. The K-means clustering algorithm was applied to each annual data set, and to verify the number of clusters found, the validation technique was used with the silhouettes of the clusters. The result of the application of the K-means algorithm to the 5 data sets that represent the annual measurements of the climate change data, provide the information reported in Table 1.

Table 1. Optimal number of groups for air quality data.

Year	Optimal number of groups
2014	9
2015	9
2016	6
2017	6
2018	6

3.1. Validation with the silhouettes of the clusters

The optimal number of groups was calculated from the information obtained from the silhouette of each execution of the K-means method with different number of groups. Table 2 shows the tests that were made with a different number of groups and the error that the silhouette of each one of them shows, for the 5 years of measurements that are being studied in this research. The results with the least error in the silhouettes of each test are highlighted, thus justifying the optimum number of groups for each annual measurement. The results of the tests show groupings with a minimum of 6 groups and a maximum of 11 groups, since with a number of groups lower than 6 and higher than 11 the error increases, and for practical purposes it was decided to omit these results.

Once the air quality data for each year, from 2014 to 2018, are grouped together, Figure 1 shows the silhouettes of these clusters (except 2016).

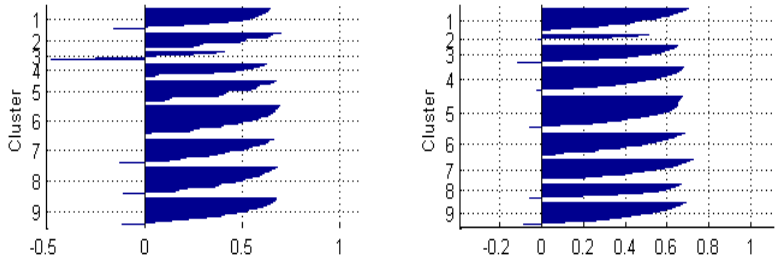

Figure. 1. Silhouettes of clusters of annual air quality data 2014 and 2015

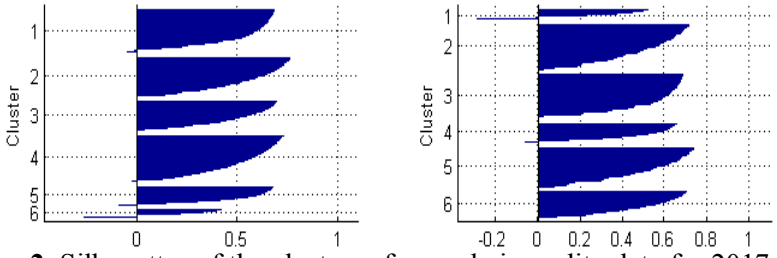

Figure. 2. Silhouettes of the clusters of annual air quality data for 2017 and 2018

Table 2. Errors thrown by the silhouette of each grouping.

Number of groups / Year	6	7	8	9	10	11
2014	0.5487	0.5368	0.5125	0.5012	0.5012	0.5230
2015	0.4962	0.3978	0.4201	0.3952	0.4236	0.4025
2016	0.3123	0.5012	0.4362	0.4420	0.4625	0.4231
2017	0.4652	0.4362	0.4752	0.4630	0.4750	0.4425
2018	0.4325	0.4785	0.4521	0.4850	0.5012	0.5365

With the groups obtained by the K-means method for each annual data set, the information of the pollutants associated to each data cluster is related. Tables 3, 4 and 5 concentrate the information on the maximum and minimum values for each pollutant criterion in each group. With this information, it can be noted that air quality is strongly influenced by suspended particles smaller than 10 micrometers (PM10).

Table 3. Maximum and minimum values for each cluster I.

Data from clustering of 2014 air quality data							
Pollutant / Group		CO	NO_2	NOX	O_3	PM10	SO_2
1	Maximum	16.5000	0.2500	0.4750	0.2720	133.0000	0.2500
	Minimum	0.1001	0.0033	0.0040	0.0010	92.0000	0.0010
2	Maximum	12.0000	0.2050	0.4450	0.2000	25.0000	0.1820
	Minimum	0.1000	0.0023	0.0033	0.0010	17.0000	0.0010
3	Maximum	13.5000	0.2923	0.4800	0.2824	788.0000	0.1524
	Minimum	0.2001	0.0110	0.0055	0.0034	135.0000	0.0011
4	Maximum	10.7000	0.1680	0.3440	0.1730	18.0000	0.1750
	Minimum	0.2000	0.0030	0.0030	0.0030	1.0004	0.0011
5	Maximum	7.4000	0.0850	0.2410	0.1820	33.0000	0.1550
	Minimum	0.1200	0.0020	0.0010	0.0010	26.0000	0.0014
6	Maximum	13.7000	0.1500	0.4813	0.2340	68.0000	0.1800
	Minimum	0.1200	0.0030	0.0034	0.0020	54.0000	0.0020
7	Maximum	8.3000	0.2390	0.3540	0.2160	42.0000	0.2050
	Minimum	0.2000	0.0020	0.0030	0.0010	34.0000	0.0010
8	Maximum	13.4000	0.1510	0.4330	0.2160	53.0000	0.1240
	Minimum	0.1010	0.0050	0.0040	0.0010	42.0000	0.0010
9	Maximum	14.1010	0.1510	0.3250	0.2450	92.0000	0.1300
	Minimum	0.2000	0.0023	0.0055	0.0012	70.0000	0.0014

Table 4. Maximum and minimum values for each cluster II.

Data from the clustering of air quality data for 2016							
Pollutant / Group		CO	NO_2	NOX	O_3	PM10	SO_2
1	Maximum	12.3040	0.1810	0.5000	0.2700	110.0000	0.2440
	Minimum	0.1000	0.0030	0.0033	0.0020	67.5000	0.0030
2	Maximum	8.4000	0.1270	0.4450	0.1820	44.0000	0.1600
	Minimum	0.1100	0.0020	0.0020	0.0030	30.0000	0.0020
3	Maximum	9.1000	0.1700	0.2630	0.1500	18.0000	0.0820
	Minimum	0.2000	0.0040	0.0020	0.0020	0.1000	0.0013
4	Maximum	10.6000	0.1330	0.3660	0.1662	28.0000	0.1950
	Minimum	0.1400	0.0040	0.0020	0.0031	17.0000	0.0010
5	Maximum	13.4100	0.2410	0.5030	0.2320	675.0000	0.1600
	Minimum	0.2070	0.0030	0.0030	0.0050	110.0000	0.0010
6	Maximum	9.8000	0.1820	0.4470	0.2201	69.0000	0.2030
	Minimum	0.2000	0.0047	0.0015	0.0033	44.0000	0.0020

Table 5.Maximum and minimum values for each cluster III.

Data from the clustering of air quality data for 2018							
Pollutant / Group		CO	NO_2	NOX	O_3	PM10	SO_2
1	Maximum	12.9000	0.1740	0.4020	0.1780	378.0000	0.2520
	Minimum	0.1700	0.0050	0.0200	0.0020	99.0000	0.0030
2	Maximum	7.8000	0.1330	0.3300	0.1450	38.0000	0.1920
	Minimum	0.1000	0.0070	0.0220	0.0010	21.0000	0.0010
3	Maximum	6.3000	0.1410	0.3000	0.1500	22.0000	0.2350
	Minimum	0.1040	0.0050	0.0040	0.0010	0.0000	0.0010
4	Maximum	12.2000	0.1420	0.3520	0.2010	94.0000	0.2750
	Minimum	0.2000	0.0070	0.0200	0.0020	70.0000	0.0010
5	Maximum	6.7000	0.1400	0.3130	0.1750	54.0000	0.2620
	Minimum	0.1030	0.0020	0.0040	0.0010	38.0000	0.0010
6	Maximum	9.0000	0.1120	0.3490	0.1870	709.0000	0.1910
	Minimum	0.1200	0.0050	0.0130	0.0030	55.0000	0.0020

4. Conclusions

The interest of this research was to answer the following questions: Is there a pattern in the records of each year, is only one pollutant triggered by measurement, what are the pollutants that are triggered most frequently? These questions cannot be answered by simply having the air quality record at a given time, but require analysis of the air quality measurements to see how the data behave, in order to obtain the conclusions. The clustering strategy presented in this paper provides a tool for the analysis of air quality data, specifically the testing of air quality information from the Bogotá Valley Zone. The study carried out demonstrated that there is an important variation in air quality data from one year to another, since the number of clusters varies from year to year, which with the help of experts, can be interpreted as a cause of climate change.

The exact interpretation of the clustering obtained in this study is not a trivial task, due to the lack of a model to identify the criteria pollutants that influence the increase of MAQI levels and consequently, the declaration of an environmental contingency. The strategy proposes a way to group these values and can be used for any other region that has stations that measure criterion pollutants.

References

[1] Jyothi, S. N., Kartha, K., Mohan, A., Pai, J., & Prasad, G. (2019, November). Analysis of Air Pollution in Three Cities of Kerala by Using Air Quality Index. In Journal of Physics: Conference Series (Vol. 1362, No. 1, p. 012110). IOP Publishing.

[2] Viloria, A., & Gaitan-Angulo, M. (2016). Statistical Adjustment Module Advanced Optimizer Planner and SAP Generated the Case of a Food Production Company. Indian Journal Of Science And Technology, 9(47). doi:10.17485/ijst/2016/v9i47/107371

[3] Yan, J., Li, X., Shi, Y., Sun, S., & Wang, H. (2019). The effect of intention analysis-based fraud detection systems in repeated supply Chain quality inspection: A context of learning and contract. Information & Management, 103177.

[4] Niu, X., Wang, X., Gao, J., & Wang, X. (2020). Has third-party monitoring improved environmental data quality? An analysis of air pollution data in China. Journal of environmental management, 253, 109698.

[5] Wahi, J. S., Thar, M. D., Garg, M., Goyal, C., & Rathi, M. (2019). Analysis of Air Quality and Impacts on Human Health. In Smart Healthcare Systems (pp. 109-123). Chapman and Hall/CRC.

[6] Yang, C. T., Chen, C. J., Tsan, Y. T., Liu, P. Y., Chan, Y. W., & Chan, W. C. (2019). An implementation of real-time air quality and influenza-like illness data storage and processing platform. Computers in Human Behavior, 100, 266-274.

[7] Niu, X., Wang, X., Gao, J., & Wang, X. (2020). Has third-party monitoring improved environmental data quality? An analysis of air pollution data in China. Journal of environmental management, 253, 109698.

[8] Yang, R., Zhang, X., Ye, X., Wang, C., & Li, X. (2020). Ventilation modes and greenhouse structures affect 222 Rn concentration in greenhouses in China. Journal of Radioanalytical and Nuclear Chemistry, 1-9.

[9] Rekhi, J. K., Nagrath, P., & Jain, R. (2020). Forecasting Air Quality of Delhi Using ARIMA Model. In Advances in Data Sciences, Security and Applications (pp. 315-325). Springer, Singapore.

[10] Zhang, B., Li, L., Yao, X., Gong, Y., Zhang, Y., Yang, H., ... & Jia, H. (2020). Analysis of Air Purification Methods in Operating Rooms of Chinese Hospitals. BioMed Research International, 2020.

[11] Rodríguez-Camargo, L. A., Sierra-Parada, R. J., & Blanco-Becerra, L. C. (2020). Análisis espacial de las concentraciones de PM2. 5 de acuerdo con los valores guía de calidad del aire de la OMS para enfermedades cardiopulmonares en Bogotá, DC, 2014-2015. Biomédica, 40(1).

[12] Sanchez, L., Vásquez, C., & Viloria, A. (2018, June). Conglomerates of Latin American countries and public policies for the sustainable development of the electric power generation sector. In International Conference on Data Mining and Big data (pp. 759-766). Springer, Cham.

[13] Guzmán, J. S., Hernandez, H. T., Mora, J. M., Piracoca, A., Vanegas, J. S., & Pachon, J. E. (2019, June). Análisis espacio temporal de concentraciones de material particulado en Bogotá: un año de operación de una red independiente. In Congreso Colombiano y Conferencia Internacional de Calidad del Aire y Salud Pública.

[14] Luna, G., Andrés, M., Gonzalez, V., Mario, J., Muñoz, T., Ceron, B., & Carlos, L. (2019, August). Evaluación espacial y temporal de PM 10 y PM 2.5 en Colombiautilizando información satelital (CAMS, MODIS-AOD) y mediciones de calidad del aire en superficie. In 2019 Congreso Colombiano y Conferencia Internacional de Calidad de Aire y Salud Pública (CASP) (pp. 1-5). IEEE.

[15] Espinosa, Mónica, and Juan F. Franco. "GESTIÓN DE LA CALIDAD DEL AIRE EN BOGOTÁ." (2019).