Study of the principal component analysis in air quality databases

IOP Conf. Series: Materials Science and Engineering **872** (2020) 012195 doi:10.1088/1757-899X/872/1/012195

Retraction: Study of the principal component analysis in air quality databases (*IOP Conf. Series: Materials Science and Engineering* **872** 012030)

Jesús Silva[1], Luz Adriana Londoño[2], Noel Varela[2] and Omar Bonerge Pineda Lezama[3]

[1] Universidad Peruana de Ciencias Aplicadas, Lima, Perú.

[2] Universidad de la Costa, Barranquilla, Atlántico, Colombia

[3] Universidad Tecnológica Centroamericana (UNITEC), San Pedro Sula, Honduras

Published 15 September 2020

This article, and others within this volume, has been retracted by IOP Publishing following clear evidence of plagiarism and citation manipulation.

This work was originally published in Spanish (1) and has been translated and published without permission or acknowledgement to the original authors. IOP Publishing Limited has discovered other papers within this volume that have been subjected to the same treatment. This is scientific misconduct.

Misconduct investigations are ongoing at the author's institutions. IOP Publishing Limited will update this notice if required once those investigations have concluded.

IOP Publishing Limited request any citations to this article be redirected to the original work (1).

Anyone with any information regarding these papers is requested to contact conferenceseries@ioppublishing.org.

(1) Sánchez López, A., Cruz-Gutiérrez, V., Posada-Zamora, M., Torrijos M., M., & Osorio Lama, M. (2016). Estudio del análisis de componentes principales en bases de datos de calidad del aire. Research In Computing Science, 120(1), 9-19. doi: 10.13053/rcs-120-1-1

IOP Conf. Series: Materials Science and Engineering **872** (2020) 012030 doi:10.1088/1757-899X/872/1/012030

Study of the principal component analysis in air quality databases

Jesús Silva[1], Luz Adriana Londoño[2], Noel Varela[3], Omar Bonerge Pineda Lezama[4]

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

[1]Email: jesussilvaUPC@gmail.com

Abstract. Technological development has facilitated daily habits, business, the manufacture of large quantities of products, among other types of industrial activities; however, these advances have caused environmental deterioration that seriously threatens the development of society. The increase of greenhouse gases in the atmosphere affects the health of millions of people and is the main factor that has modified the climate on planet Earth. Faced with this situation, it is necessary to carry out actions that allow to quickly adapt to this change and mitigate its effects. The present study proposes the analysis of main components in the data of the pollutant measurements in the city of Bogota, Colombia with the purpose of obtaining a more compact representation of these data, to later apply grouping techniques and obtain factors that allow the emission of an alert for pre-contingency and contingency.

1. Introduction

Currently, several organizations and governments have implemented schemes to measure pollutants and obtain the air quality indices (AQI) of the different regions of the planet [1]. In the city of Bogota, air pollution is measured with the Metropolitan Air Quality Index (MEAQI), which is used to show the level of pollution and the level of risk it represents to human health only in this region, in a determined time and thus to be able to take protective measures [2].
This paper proposes to apply the technique known as Principal Component Analysis (PCA) [3] to the pollutant measurement in order to establish a pattern. The attributes of each pattern are the values of each pollutant and, in this way, their grouping can be compared with the data to which the PCA has not been applied. In [4][5][6], basic clustering techniques commonly used such as the K-means and K-medoids method for clustering are described, but in this paper more complex algorithms such as Fuzzy c-Means, Possibilistic c-Means, Competitive Leaky Learning and Valey Seeking are explored. In Section 2, the development of the PCA is presented, then, in Section 3, the results and relevant comparisons of the algorithms are presented; finally, section 4 shows the conclusions.

2. Study proposal

Historical data were obtained for the criterion pollutants considered in the study, which are ozone (O3), nitrogen dioxide (NO2), nitrogen monoxide (NO), sulfur dioxide (SO2), carbon monoxide (CO), particles less than 10 microns (PM10), particles less than 2.5 microns (PM2.5) and coarse fraction particles or "coarse" (PMCO), with a database available since 1986 [7][8]. Because in some years the measurements and the number of criterion pollutants were not consistent, it was decided to start from the year 2000 until 2018, with the following considerations:

- From 1995 to 2003, the pollutants (CO), (NO2), (NO), (O3), (PM10) and (SO2) are taken into account.
- From 2004 to 2011, the pollutant criterion (PM2.5) is added.
- From 2014 onwards, it is mentioned in [9] that the measurement of (PMCO) started.

Once the data is cleaned, the PCA is carried out:

- **Analysis of correlation matrix:** A principal component analysis makes sense if there are high correlations between the variables (this indicates that there is redundant information and therefore few factors will explain a large part of the total variability [10].
- **Selection of factors:** It is done so that the first factor collects the greatest possible proportion of the original variability, the second factor must therefore collect the maximum variability not collected by the first one, etc. From the factors, those that collect a percentage of variability considered sufficient (main components) will be chosen [11].
- **Factorial matrix analysis:** Once the main components are selected, they are represented in matrix form. Each element therefore represents the factorial coefficients of the variables (the correlations between the variables and the main components) [12].
- **Interpretation of factors:** For a factor to be easily interpreted, it must exhibit the following characteristics [13]:

- Factorial coefficients should be close to 1
- A variable must have high coefficients with only one factor.
- There should be no factors with close coefficients.
- Calculation of factorial scores: These are the scores that have the main components in each case and which allows them to be graphed.

3. Results

When analyzing all the data sets, a total of 18,235 records per pollutant (corresponding to the hourly measurement) are observed, for each of the stations. It was decided to take the data from one monitoring station in order to create an initial model. The station to be chosen should measure all the pollutants since there are also stations that do not provide records of some particles. The station chosen was that of the Chapinero delegation. Figure 1 shows the results obtained.
With the results obtained from the previous process, a cluster study was carried out to see if the consistency of the data persisted [14]. For the analysis, the K-means and K-medoids methods were used with the raw data and later with the data obtained after the PCA, in order to make a comparative study of both and to determine if the PCA maintains the consistency of the information, and thus be able to test the hypothesis that, with the reduction of the size of each instance, the same result can be reached or present an acceptable approximation [15].

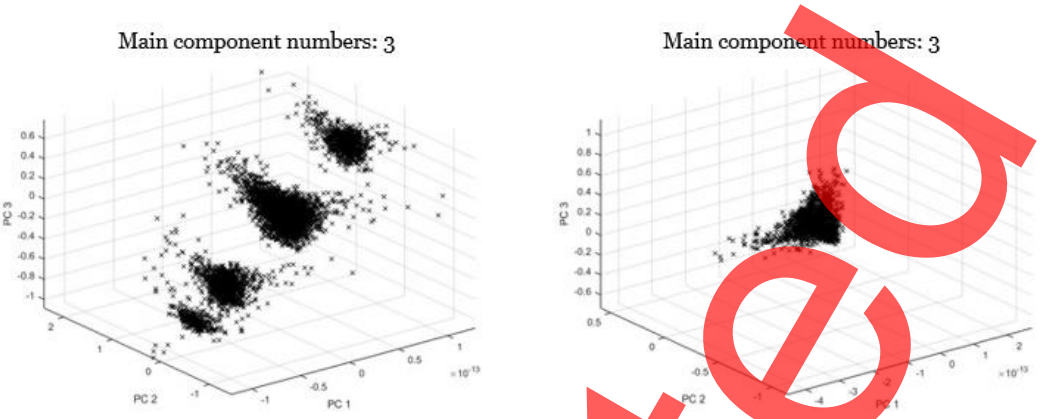

Figure 1. PCA results for Chapinero station in 2017 and 2018.

Due to the nature of the methods and given that the initialization of the centroids is random, the method was executed several times and the results were evaluated with the average of the values with the silhouette index in order to determine the optimal number of clusters [16]. In Table 1, the averages of the results of 20 runs with the K-means method from 2006 onwards are shown, since the previous data sets do not show a large change with the number of optimal clusters that this measurement yielded. The best values in each cluster, representing the optimal one, are highlighted. Table 2 shows the optimal number of each data set with the K-medoids method.
Cluster tests were again conducted with the data obtained from the PCA in order to determine the optimal number of partitions and to make a comparison. The tests were performed with the same limit iterations, executions, measurement mechanism and silhouette index; with this, the purpose was to determine if the results were maintained by performing the dimension reduction.

Table 1. Silhouette indices in K-means.

Clusters	4	5	6	7	8	9	10	11
Years								
2006	0.62414	0.58114	0.5206	0.67506	0.53564	0.60405	0.55538	0.5697
2007	0.51274	0.58134	0.5224	0.57556	0.48174	0.51315	0.45308	0.4822
2008	0.55364	0.60464	0.5283	0.56786	0.49334	0.53005	0.46868	0.488
2009	0.56784	0.57814	0.5699	0.59666	0.54484	0.52605	0.47428	0.4914
2010	0.52414	0.51944	0.4865	0.52676	0.46244	0.49825	0.44928	0.4739
2011	0.53054	0.51354	0.5049	0.54966	0.49164	0.52675	0.47438	0.4896
2012	0.58404	0.56224	0.5382	0.62036	0.51014	0.54775	0.47398	0.509
2013	0.55274	0.54644	0.5092	0.54856	0.48624	0.52035	0.45858	0.4711
2014	0.40104	0.38064	0.35	0.39926	0.32314	0.35975	0.29258	0.3235
2015	0.45394	0.43904	0.3796	0.43336	0.34354	0.35695	0.31748	0.3655
2017	0.40534	0.43494	0.4063	0.45396	0.34944	0.38395	0.31978	0.302
2018	0.46614	0.43524	0.4088	0.44706	0.35684	0.38435	0.32738	0.3464

Other algorithms with different behaviors were also used: Fuzzy c-Means (FcM), Possibilistic c-Means (PcM), Competitive Leaky Learning and Valey Seeking [11][13]. These algorithms have strengths and weaknesses that were confirmed with the tests performed on the contaminants data [7], using the PCA and the data without the use of dimensionality reduction. These algorithms are not optimal for large amounts of data, due to their iterative behavior and although they have an external shutdown condition, they are often delayed in execution time without performing the PCA, however, for creating a comparison, some tests will be performed.

Table 2. Final results of the clusters with K-medoids.

Years	Cluster	Index
2006	5	0.71254
2007	5	0.6962
2008	5	0.6952
2009	4	0.7012
2010	5	0.6932
2011	5	0.7012
2012	5	0.6333
2013	5	0.7002
2014	5	0.4952
2015	5	0.4214
2017	5	0.4536
2018	5	0.4124

For the Fuzzy c-Means (FcM) algorithm, the vector compatibility degree of a target function with a certain cluster is used; the algorithm is sensitive to outliers. It is also sensitive to the degree of defuzzification that the value must be in a given range of tests [14].
When using the PCA, it makes the separation similar to K-means and K-medoids, but these methods provide more separation. The number of clusters ranges from 5-10 groups, and the display shows the same separation with the simple K-medoids and K-means methods.
The Possibilistic c-Means (PcM) algorithm is ideal for revealing compact clusters, as FcM has a degree of belonging to each cluster that is defined, but is less sensitive to the exact number of clusters; this algorithm is iterative and has a computational cost not so high for its behavior (Figures 2 and 3).
The Leaky Learning (LLA) algorithm [17], is an appropriate algorithm to reveal compact clusters. The number of clusters is assumed, so tests must be done to find the right number of clusters. This algorithm uses the term density, which requires to know in which region the competition strategy is used. In the data without using the PCA, it is noted that the clusters overlap, which indicates that it is essential for pre-processing the data (Figures 4 and 5).
In the Valey Seeking Clustering (VS) algorithm, clusters are considered as peaks of data described by individuals, and these are separated by valleys.

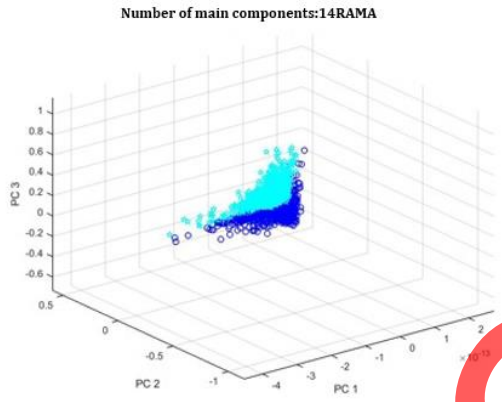

Figure 2. Two groups found, using the PcM algorithm for the year 2018.

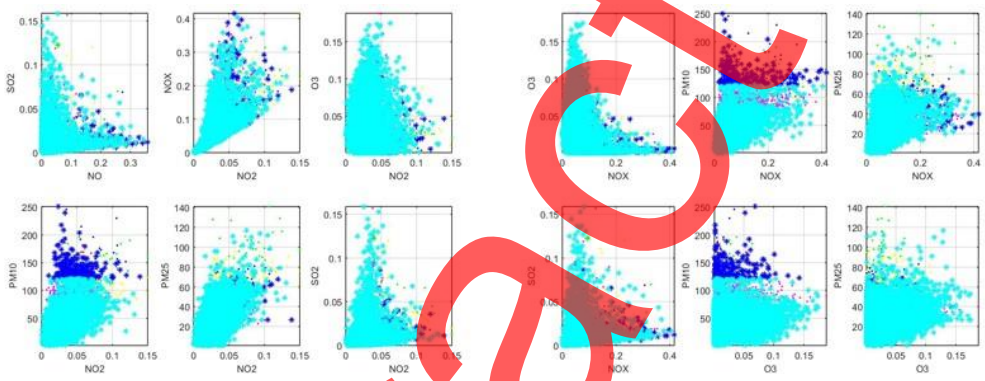

Figure 3. 2007 results without the PCA with the PcM algorithm.

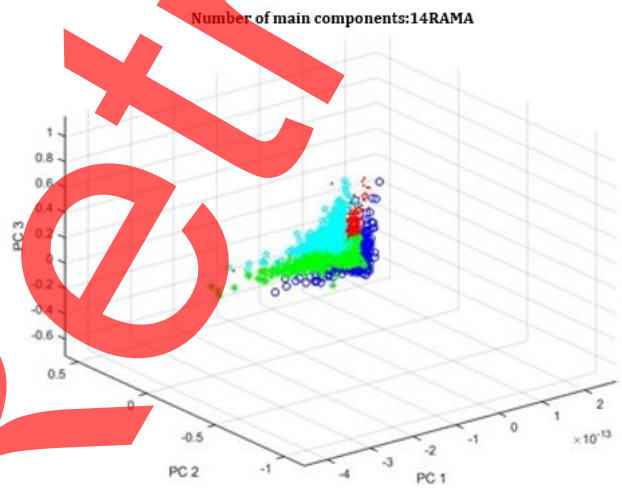

Figure 4. 2018 result without the ACP with the LLA algorithm.

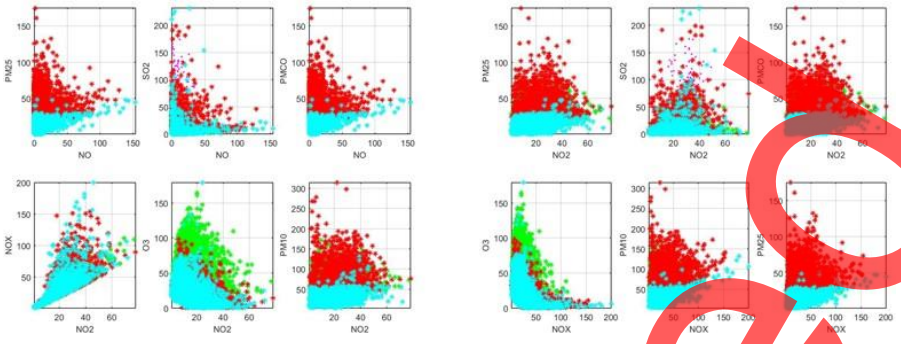

Figure 5. The groups found with the LLA algorithm for the year 2018.

Conclusions

The interest in carrying out this research is to answer the following questions: is there a pattern in the records of each year, is only one pollutant measured, 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 and to draw the appropriate conclusions. Cluster analysis is considered a good technique for uncovering many hidden patterns in the data.

References

[1] Dogruparmak, S. C., Keskin, G. A., Yaman, S., & Alkan, A. (2014). Using principal component analysis and fuzzy c–means clustering for the assessment of air quality monitoring. Atmospheric Pollution Research, 5(4), 656-663.

[2] 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.

[3] 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

[4] Lin, Y. C., Lee, S. J., Ouyang, C. S., & Wu, C. H. (2020). Air quality prediction by neuro-fuzzy modeling approach. Applied Soft Computing, 86, 105898.

[5] Ding, C., He, X.: K-means clustering via principal component analysis. In: Proceedings of the 20th International Conference on Machine Learning (2004)

[6] Zare, A., Young, N., Suen, D., Nabelek, T., Galusha, A., & Keller, J. (2017, November). Possibilistic fuzzy local information c-means for sonar image segmentation. In 2017 IEEE Symposium Series on Computational Intelligence (SSCI) (pp. 1-8). IEEE.

[7] Pholsena, K., & Pan, L. (2018, June). Traffic status evaluation based on possibilistic fuzzy c-means clustering algorithm. In 2018 IEEE Third International Conference on Data Science in Cyberspace (DSC) (pp. 175-180). IEEE.

[8] Stockwell, W. R., Saunders, E., Goliff, W. S., & Fitzgerald, R. M. (2020). A perspective on the development of gas-phase chemical mechanisms for Eulerian air quality models. Journal of the Air & Waste Management Association, 70(1), 44-70.
[9] Psiloglou, B. E., Kambezidis, H. D., Kaskaoutis, D. G., Karagiannis, D., & Polo, J. M. (2020). Comparison between MRM simulations, CAMS and PVGIS databases with measured solar radiation components at the Methoni station, Greece. Renewable energy, 146, 1372-1391.
[10] Johnson, T. (2002). A guide to selected algorithms, distributions, and databases used in exposure models developed by the office of air quality planning and standards. Research Triangle Park, NC, US Environmental Protection Agency, Office of Research and Development.
[11] Singh, K. P., Gupta, S., & Rai, P. (2013). Identifying pollution sources and predicting urban air quality using ensemble learning methods. Atmospheric Environment, 80, 426-437.
[12] Elbir, T. (2004). A GIS based decision support system for estimation, visualization and analysis of air pollution for large Turkish cities. Atmospheric Environment, 38(27), 4509-4517.
[13] Yatkin, S., Gerboles, M., Belis, C. A., Karagulian, F., Lagler, F., Barbiere, M., & Borowiak, A. (2020). Representativeness of an air quality monitoring station for PM2. 5 and source apportionment over a small urban domain. Atmospheric Pollution Research, 11(2), 225-233.
[14] Ganbold, G., & Chasia, S. (2017). Comparison between Possibilistic c-Means (PCM) and Artificial Neural Network (ANN) Classification Algorithms in Land use/Land cover Classification. International Journal of Knowledge Content Development & Technology, 7(1), 57.
[15] Grace, R. K., & Manju, S. (2019). A Comprehensive Review of Wireless Sensor Networks Based Air Pollution Monitoring Systems. Wireless Personal Communications, 108(4), 2499-2515.
[16] Rodrıguez-Camargo, L. A., Sierra-Parada, R. J., & Blanco-Becerra, L. C. (2020). Spatial analysis of PM2. 5 concentrations according to WHO air quality guideline values for cardiopulmonary diseases in Bogotá, DC, 2014-2015. Biomedical, 40(1).
[17] Casallas, A., Celis, N., Ferro, C., Barrera, E. L., Peña, C., Corredor, J., & Segura, M. B. (2020). Validation of PM 10 and PM 2.5 early alert in Bogotá, Colombia, through the modeling software WRF-CHEM. Environmental Science and Pollution Research, 1-11.