ISSN 1870-4069

Estudio del análisis de componentes principales en bases de datos de calidad del aire

Abraham Sánchez López[1], Viridiana Cruz-Gutiérrez[1], Mario Alberto Posada-Zamora[1], M. Teresa Torrijos M.[1], M. Auxilio Osorio Lama[2]

[1] Benemérita Universidad Autónoma de Puebla, Facultad de Ciencias de la Computación, México

[2] Benemérita Universidad Autónoma de Puebla, Facultad de Ingeniería Química, México

asanchez@cs.buap.mx,{viricruz, mariop}@rockruz.net, {tere.torrijos, mariauxosorio}@gmail.com

Resumen. El cambio climático es un problema de la sociedad actual que repercute en muchos aspectos de la vida cotidiana. Como es conocido, el aumento de los gases de efecto invernadero en la atmósfera afecta la salud de millones de personas. Es de suma importancia que las autoridades cuenten con herramientas que les permitan realizar una mejor toma de decisiones ante estos eventos. El presente trabajo propone el estudio del análisis de componentes principales en los datos de las mediciones de contaminantes de la ciudad de México con la finalidad de conseguir una representación más compacta de dichos datos, para posteriormente aplicar técnicas de agrupamiento y con ello obtener factores que permitan la emisión de una alerta de pre contingencia y contingencia.

Palabras clave: Componentes principales, cambio climático, algoritmos de clustering.

A Study of Principal Components Analysis of Air Quality Databases

Abstract. The climatic change is an escalated issue of the actual society that have repercussions in so many aspects of everyday life. As it is known, the increasing levels of the greenhouse effect gases have a negative impact on the overall health of millions of people. It is greatly important that the authorities count with tools that let them to prepare and execute a better decision making related to this type of events. The present work proposes the study of Principal Components Analysis in the air pollution measurements data of the Mexico City with the main goal to get a compacted representation of the data, to later be able to apply clustering techniques and with that obtain results that let the authorities

to emit precontingency and contingency alerts to the population.

Keywords: Principal components, climatic change, clustering algorithms.

1. Introducción

El desarrollo tecnológico ha facilitado los hábitos cotidianos, los negocios, la fabricación de grandes cantidades de productos, entre otro tipo de actividades industriales; sin embargo, estos avances han provocado un deterioro ambiental que amenaza seriamente al desarrollo de la sociedad.

El aumento de los gases de efecto invernadero en la atmósfera afecta la salud de millones de personas y es el principal factor que ha modificado el clima en el planeta Tierra. Ante esta situación, es necesario llevar a cabo acciones que nos permitan adaptarnos rápidamente a este cambio y mitigar los efectos que está produciendo. Si bien esta problemática se ha estudiado a nivel mundial por el grupo de expertos sobre cambio climático, corresponde ahora a los jefes de las naciones bajar el nivel de contaminación, establecer estrategias y coordinar acciones que contribuyan al conocimiento de esta problemática, ya que se presenta en diferentes formas, dependiendo de la situación geográfica y del nivel de desarrollo de cada país. En este sentido, se propone como alternativa viable, aplicar la tecnología de la inteligencia de negocios, como una estrategia de análisis que contribuya a llevar un control de la calidad del aire y de la variables climáticas asociadas; así como para proporcionar instrumentos que ayuden a establecer políticas públicas para realizar una mitigación en el cambio climático en la Zona Metropolitana del Valle de México (ZMVM) y en otras ciudades, siempre que se cuenten con estaciones de monitoreo para poder disponer de datos útiles.

En la actualidad, diversas organizaciones y gobiernos han implementado esquemas de medición de contaminantes y así obtener los índices de calidad de aire (ICA) de las diferentes regiones del planeta. En la Ciudad de México y en la ZMVM, la contaminación del aire se mide con el Índice Metropolitano de Calidad del aire (IMECA), el cual es usado para mostrar el nivel de contaminación y el nivel de riesgo que representa a la salud humana únicamente en esta región, en un tiempo determinado y así poder tomar medidas de protección.

En el presente trabajo se elabora una propuesta de aplicar la técnica conocida como análisis de componentes principales (ACP) en las mediciones de los contaminantes para establecer un patrón; los atributos de cada patrón, son los valores de cada contaminante y de esa forma su agrupamiento se podrá comparar con los datos a los que no se le haya aplicado el ACP. En [2,4] se describen técnicas de agrupamiento básicas utilizadas comúnmente como el método K-means y K-medoids para la agrupación, pero en este trabajo se exploran algoritmos más complejos como Fuzzy c-Means, Possibilistic c-Means, Competitive Leaky Learning y Valey Seeking. En la *Sección 2* se abordan definiciones conceptuales de clustering y de los aspectos generales del ACP, posteriormente, en la *Sección 3* se presenta el desarrollo del ACP, después en la *Sección 4* se dan a conocer

los resultados y las comparativas pertinentes de los algoritmos, finalmente, en la *Sección 5* se muestran las conclusiones y trabajo futuro de este trabajo.

2. Marco teórico

El clima es una descripción estadística de las condiciones de tiempo y sus variaciones, incluyendo condiciones promedio y extremas. Los gases de efecto invernadero juegan un rol importante en el deterioro del clima y provocan el cambio climático. Los gases de efecto invernadero incluyen vapor de agua, dióxido de carbono (CO_2), metano (CH_4), óxido nitroso (N_2O) y algunos gases industriales tales como cloro fluorocarbonos ($CFCs$). Estos gases actúan como una manta aislante, manteniendo la superficie de la tierra más caliente de lo que debería estar, esto es debido a que no se reflejan los rayos del sol al espacio. Una vez que estos gases son liberados a la atmósfera, muchos de ellos permanecerán ahí por un largo periodo de tiempo [1,3,4].

La primera fase de la estrategia consiste en preparar los datos para su agrupamiento, y dado que el conjunto inicial de datos tiene su origen en un grupo de hojas de cálculo, es necesario aplicar algunas técnicas como la reducción de dimensiones. Este proceso consiste en encontrar un espacio adecuado que es menor en la dimensión en el cual se representan los datos originales. Se espera que la representación de los datos ayude a

- explorar datos de gran dimensión con el objetivo de explorar estructuras o patrones que dirija a la formación de una hipótesis estadística,
- visualizar los datos con diagramas de dispersión, donde la dimensión es reducida a 2 o 3 dimensiones,
- analizar los datos en métodos estadísticos, como clustering, estimación de la densidad de probabilidad o de clasificación.

Un posible método para la reducción de dimensiones sería únicamente elegir subconjuntos de las variables para procesar y analizar estos grupos, sin embargo, en algunos casos, eso significaría deshacerse de mucha información útil. Una alternativa podría ser crear nuevas variables a partir de las variables originales en forma de combinaciones lineales [2,5].

2.1. Análisis de componentes principales

El propósito de análisis de componentes principales es reducir un espacio de dimensión p a un nuevo espacio de dimensión d, donde d es mucho menor que p, mientras que al mismo tiempo representa la variación de datos como sea posible. Con el ACP, se transforman los datos en un nuevo conjunto de coordenadas o variables que son una combinación lineal de las variables originales. Además, las observaciones en el nuevo espacio de componentes principales no están correlacionadas. Se espera obtener información y comprensión de los datos al analizar las observaciones en el nuevo espacio [2].

2.2. Técnicas de agrupamiento

Entre las técnicas de agrupamiento, se encuentra la de tipo no jerárquico [2], que consiste en dividir los datos en k particiones o grupos donde cada partición representa un grupo. El funcionamientos básicos de dicho método son

1. seleccionar K centroides iniciales, siendo K el número de grupos deseados,
2. asignar cada patrón al grupo que le sea más cercano,
3. reasignar o relocalizar cada observación a uno de los K grupos de acuerdo con algún criterio de paro,
4. termina si no hay reasignaciones de los puntos o si la reasignación satisface la regla de paro. En otro caso regresa al paso dos.

Entre estos métodos existen dos que son los más utilizados que es el K-means y K-PAM (Partitioning Around Medoids), estos métodos son de utilidad para encontrar grupos (clusters). Para poder unir variables o individuos es necesario tener algunas medidas numéricas que caractericen las relaciones entre las variables o los individuos. La medida de asociación puede ser una distancia o una similaridad; a continuación se describen brevemente dichas medidas:

- Cuando se elige una distancia como medida de asociación (por ejemplo la distancia euclideana) los grupos formados contendrán individuos parecidos de forma que la distancia entre ellos ha de ser pequeña.
- Cuando se elige una medida de similaridad (por ejemplo el coeficiente de correlación) los grupos formados contendrán individuos con una similaridad alta entre ellos.

Entre los algoritmos más utilizados en el análisis de clusters se encuentran: el algoritmo K-means y el algoritmo K-medoids.

2.3. Validación de clusters, índice de la silueta

El índice de silueta es una métrica para evaluar el buen funcionamiento de los algoritmos de aprendizaje no supervisado. El objetivo de este índice es identificar el número óptimo de agrupamientos. En los algoritmos de aprendizaje no supervisado, el número de clústeres puede ser un parámetro de entrada del algoritmo (K-means). La determinación del número óptimo de clústeres tiene que ser realizado mediante alguna medida externa al algoritmo. El índice silueta es indicador del número ideal de clústers. El valor de la silueta para cada punto es una medida de que tan similar es ese punto hacia los puntos que están en su propio clúster, cuando este es comparado con puntos en otros clusters.

El valor de la silueta para el i-ésimo punto, si está definido como

$$S_i = \frac{(b_i - a_i)}{max(a_i, b_i)}, \tag{1}$$

donde a_i es la distancia promedio desde i-ésimo punto hacia los otros puntos en el mismo cluster, y b_i es la distancia mínima promedio desde el i-ésimo punto

hacia los puntos en un clúster diferente. El valor de la silueta tiene rangos de -1 a +1. Un alto valor de silueta indica que i está bien asignado a su propio clúster, y pobremente mal asignado a clústers vecinos. Si muchos puntos tienen un alto valor de silueta, entonces la solución del clúster es apropiada. Por el contrario, si muchos puntos tienen un valor bajo o negativo, entonces la solución puede tener muchas o pocas particiones. El criterio de evaluación de clúster con siluetas puede ser utilizada con cualquier métrica de asociación.

2.4. Otros algoritmos de agrupamiento

Los algoritmos Fuzzy c-Means son algunos de los principales algoritmos utilizados en el agrupamiento difuso y pertenecen a una clase de algoritmos basados en funciones objetivo [7]. Definen un criterio de agrupamiento en la forma de una función objetivo que depende de la partición difusa.

Los algoritmos Possibilistic c-Means aparecen con el objetivo de resolver el mal comportamiento de los algoritmos Fuzzy c-Means, al ser utilizados en conjuntos de datos con mucho ruido [8]. Estos algoritmos se caracterizan por interpretar los valores como grados de compatibilidad con los grupos, en lugar de probabilidades de pertenencia. Para esto, se relaja la restricción de las particiones difusas que obliga a que la suma de los grados de pertenencia de un elemento hacia todos los grupos sea uno, exigiendo solamente que al menos uno de los grados de pertenencia sea positivo.

3. Propuesta del estudio

Dada la insuficiencia de datos de calidad del aire, disponibles de la ciudad de Puebla, se optó por utilizar los datos para la Ciudad de México. Se obtuvieron los datos históricos de los contaminantes criterio considerados en el estudio, los cuales son ozono (O_3), dióxido de nitrógeno (NO_2), monóxido de nitrógeno (NO), dióxido de azufre (SO_2), monóxido de carbono (CO), partículas menores a 10 micrómetros (PM_{10}), partículas menores a 2.5 micrómetros ($PM_{2,5}$) y partículas fracción gruesa o "coarse"(PM_{CO}), se cuenta con una base de datos desde 1986. Debido a que en algunos años las medidas y el número de contaminantes criterio no eran consistentes, se decidió comenzar desde el año 1995 hasta 2014, con las siguientes consideraciones:

- De 1995 a 2003 se toman en cuenta los contaminantes (CO), (NO_2), (NO), (O_3), (PM_{10}) y (SO_2).
- De 2004 a 2011 se agrega el contaminante criterio ($PM_{2,5}$).
- De 2013 en adelante, se menciona en [6] que comenzó la medición de (PM_{CO}).

Concluida la limpieza de los datos, se lleva a cabo el ACP:

- **Análisis de la matriz de correlaciones:** Un análisis de componentes principales adquiere todo su sentido, si existen altas correlaciones entre las variables (esto indica que existe información redundante y por lo tanto, pocos factores explicarán una gran parte de la variabilidad total).

- **Selección de los factores:** Se realiza de forma que el primer factor recolecte la mayor proporción posible de la variabilidad original, el segundo factor debe por lo tanto recolectar la máxima variabilidad no recolectada por el primero, etc. De los factores se elegirán aquellos que recolecten un porcentaje de variabilidad considerado como suficiente (componentes principales).
- **Análisis de la matriz factorial:** Una vez que se han seleccionado los componentes principales, estos se representan en forma matricial. Cada elemento representa por lo tanto los coeficientes factoriales de las variables (las correlaciones entre las variables y los componentes principales).
- **Interpretación de los factores:** Para que un factor sea interpretado con facilidad, este tiene que exhibir las siguientes características:
  - Los coeficientes factoriales deben ser próximos a 1.
  - Una variable debe tener coeficientes elevados sólo con un factor.
  - No debe haber factores con coeficientes cercanos.
- **Cálculo de las puntuaciones factoriales:** Son las puntuaciones que tienen las componentes principales en cada caso y esto permite graficarlos.

4. Resultados

Al analizar todos los conjuntos de datos, se observa un total de 8760 registros por contaminante considerado (lo que corresponde a la medición por hora), por cada una de las estaciones. Se decidió tomar los datos de una estación de monitoreo para poder así crear un modelo inicial. La estación a ser elegida deberá medir todos los contaminantes ya que también existen estaciones que no proporcionan registros de algunas partículas. La estación elegida fue la de la delegación Tlalpan. En la Figura 1, se muestran los resultados obtenidos.

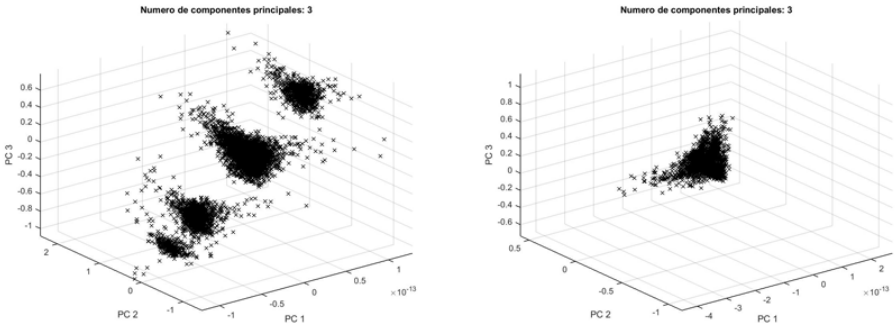

Fig. 1. Resultados del ACP para la estación Tlalpan en los años 2013 y 2014.

Al final del análisis se esperaba que cuando se realizaban las mediciones de $PM10$, $PM2,5$ y PM_{CO} existiera un agrupamiento, al observar los datos se puede apreciar que ya no hay una estabilidad de los contaminantes, así que se debe estudiar cuales son factores que hacen que la distribución se dispare.

También, se esperaba que, al incrementar las variables con datos dispersos, se obtuvieran más componentes, lo cual resultó cierto; pero efectuando un análisis visual, se observan algunos conglomerados de grupos con los que es posible hacer un análisis y agruparlos como se menciona en [3] para poder llegar a realizar los experimentos y ver qué resultados arrojan.

Con los resultados obtenidos del proceso anterior se prosiguió a realizar un estudio de clústers para observar si la consistencia de los datos persiste. Para el análisis se ocuparon los métodos K-means y K-medoids con los datos sin procesar y posteriormente con los datos obtenidos después del ACP, para poder hacer el estudio comparativo de ambas y así determinar si el ACP mantiene la consistencia de la información, y con ello poder probar la hipótesis que con la reducción de la dimensión de cada instancia se pueda llegar al mismo resultado o tener una aproximación aceptable.

Por la naturaleza de los métodos y dado que la inicialización de los centroides es de manera aleatoria se realizó la ejecución del método varias veces y después se evaluaron los resultados con la media de los valores con el índice de silueta para poder determinar el número óptimo de clústers. En la Tabla 1, se muestran las medias de los resultados de 20 ejecuciones con el método K-means del año 2003 en adelante, ya que en los conjuntos de datos anteriores no se muestra un gran cambio con el número de clúster óptimo que arrojó esta medición; se destacan los mejores valores en cada clúster, que representan el óptimo. En la Tabla 2 se muestra el número óptimo de cada conjunto de datos con el método K-medoids.

Se realizaron nuevamente las pruebas de clústers, con los datos obtenidos del ACP, con la finalidad de determinar el número óptimo de particiones y efectuar una comparativa. Las pruebas se realizaron con las mismas iteraciones límite, ejecuciones, mecanismo de medición e índice de silueta; con ello se trata de determinar si se mantuvieron los resultados realizando la reducción de dimensión.

Tabla 1. Índices de siluetas en K-means.

Clústers / **Año**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
2003	**0.6496**	0.5596	0.5329	0.6327	0.5682	0.5917	0.5975	0.5842
2004	0.5382	**0.5598**	0.5347	0.5332	0.5143	0.5008	0.4952	0.4967
2005	0.5791	**0.5831**	0.5406	0.5255	0.5259	0.5177	0.5108	0.5025
2006	**0.5933**	0.5566	0.5822	0.5543	0.5774	0.5137	0.5164	0.5059
2007	**0.5496**	0.4979	0.4988	0.4844	0.4950	0.4859	0.4914	0.4884
2008	**0.5560**	0.4920	0.5172	0.5073	0.5242	0.5144	0.5165	0.5041
2009	**0.6095**	0.5407	0.5505	0.5780	0.5427	0.5354	0.5161	0.5235
2010	**0.5782**	0.5249	0.5215	0.5062	0.5188	0.5080	0.5007	0.4856
2011	**0.4265**	0.3591	0.3623	0.3569	0.3557	0.3474	0.3347	0.3380
2012	**0.4794**	0.4175	0.3919	0.3910	0.3761	0.3446	0.3596	0.3800
2013	**0.4308**	0.4134	0.4186	0.4116	0.3820	0.3716	0.3619	0.3165
2014	**0.4916**	0.4137	0.4211	0.4047	0.3894	0.3720	0.3695	0.3609

También se utilizaron otros algoritmos con distintos comportamientos: Fuzzy c-Means (FCM), Possibilistic c-Means (PCM), Competitive Leaky Learning y Valey Seeking. Estos algoritmos tienen fortalezas y debilidades que se confirmaron con las pruebas ejecutadas sobre los datos de los contaminantes, utilizando el ACP y los datos sin el uso de la reducción de la dimensionalidad. Estos algoritmos no son óptimos para grandes cantidades de datos, debido a su comportamiento iterativo y aunque tengan una condición de paro externa, estos suelen ser tardados en tiempo de ejecución sin realizar el ACP, sin embargo, para crear un comparativo, se realizarán algunas pruebas.

Tabla 2. Resultados finales de los clústers con K-medoids.

Año	Clúster	Índice
2003	4	0.7321
2004	4	0.6872
2005	4	0.6870
2006	5	0.6951
2007	4	0.6849
2008	4	0.6936
2009	4	0.6999
2010	4	0.7066
2011	4	0.4563
2012	4	0.4711
2013	4	0.4553
2014	4	0.4475

Para el algoritmo Fuzzy c-Means (FcM), se usa el grado de compatibilidad del vector de una función objetivo con cierto clúster (pertenencia), el algoritmo es sensible a los 'outliers' o datos fuera de rango. También es sensible al grado de 'defuzzificación' que el valor debe de estar en un rango dado de pruebas. El algoritmo considera poder separar los clústers según un grado difuso de pertenencia, se hizo antes una prueba para poder buscar el número de clústers óptimo, con los métodos ya existentes, el cual con ciertas medidas arroja un número adecuado de división en las que se podría representar los datos.

Cuando se usa el ACP hace la separación casi igual a K-means y a K-medoids pero estos métodos hacen más separación. El número de clusters oscila entre 5-10 grupos, y en la visualización se nota la misma separación con los métodos simples de K-medoids y K-means.

El algoritmo Possibilistic c-Means (PcM), es ideal para revelar clústers compactos, igual que FcM se tiene un grado de pertenencia a cada clúster que se defina, pero es menos sensible al número exacto de clústers, este algoritmo es iterativo y tiene un costo computacional no tan grande por su comportamiento (Figuras 2 y 3).

El algoritmo Leaky Learning (LLA), es un algoritmo apropiado para revelar clústers compactos. Se asume el número de clústers, por lo cual hay que hacer

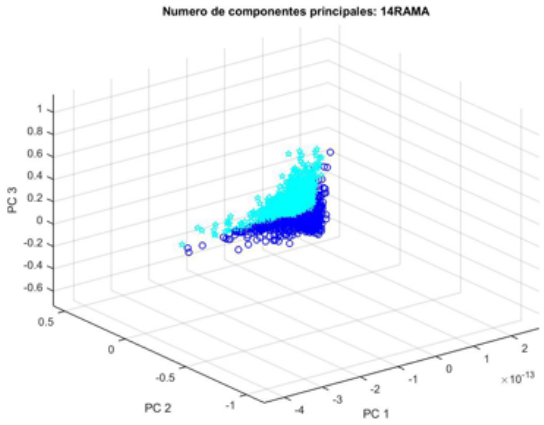

Fig. 2. Dos grupos encontrados, utilizando el algoritmo PcM para el año 2014.

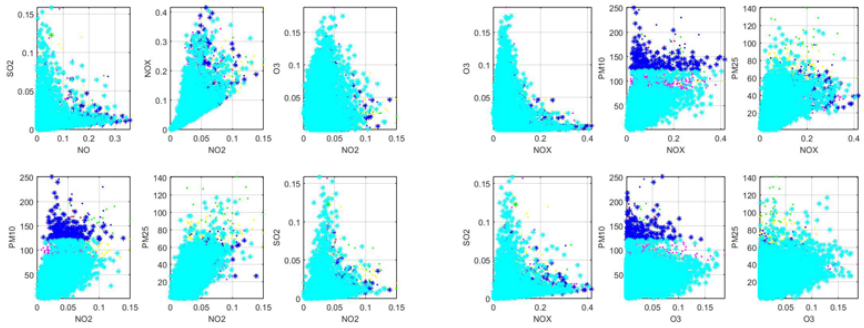

Fig. 3. Resultados del año 2007 sin el ACP con el algoritmo PcM.

pruebas para encontrar el número adecuado de clústers. Este algoritmo usa el término *densidad*, el cual requiere conocer en qué región se usa la estrategia de competición. Tiene parámetros de paro, los cuales deben de estar en un rango. El comportamiento del algoritmo muestra que con un número de 4 clústers, llega a ser obvia la separación de los datos y poder así interpretarlos. Este algoritmo es similar al algoritmo K-means, pero mucho más estable. Otro factor a considerar es que es mucho más rápido y las condiciones de paro son más claras. En los datos sin usar el ACP, se nota que se traslapan los clústers, esto indica que es indispensable hacer un tratamiento previo de los datos (Figuras 4 y 5).

En el algoritmo Valey Seeking Clustering (VS), los clusters son considerados picos de datos descritos por los individuos, y estos son separados por valles. El algoritmo es muy sensible con las variables con las que se inicializa, ya que se podría efectuar un refinamiento erróneo, esto se notó en todos los años, excepto en el año 2012, ya que se tiene un parámetro que se compara con las distancias, se tiene que encontrar el ideal, dado que, en estos datos con el ACP, las distancias entre los puntos es muy pequeña y si se excede comienza a absorber

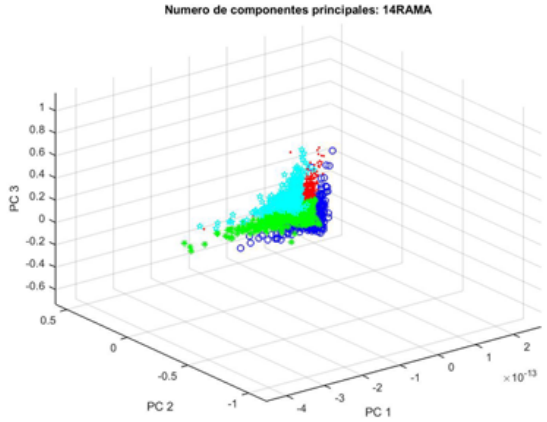

Fig. 4. Resultado del año 2014 sin el ACP con el algoritmo LLA.

a los demás grupos. Para poder tener resultados óptimos con este algoritmo hay que encontrar el paso anterior de la asignación para así se pueda hacer un refinamiento, y el valor de *desimilaridad* correcto para el que no desaparezca los grupos. El proceso de pruebas de VS puede ser muy tardado, al encontrar el valor de *desimilaridad*, y hacer bastantes pruebas.

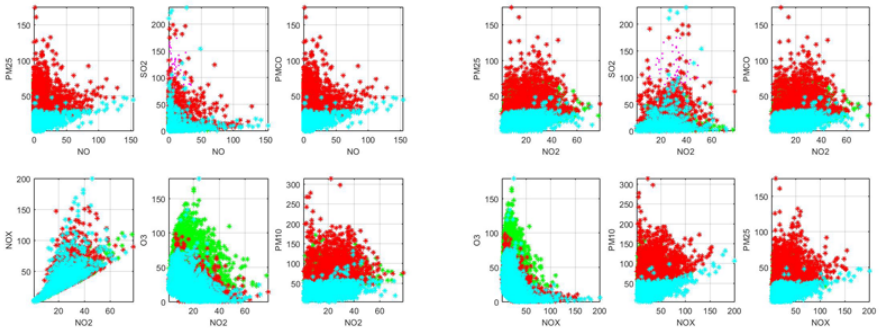

Fig. 5. Los grupos encontrados con el algoritmo LLA para el año 2014.

5. Conclusiones y trabajo futuro

El interés por realizar la presente investigación es para contestar las siguientes preguntas: ¿existe un patrón en los registros de cada año?, ¿sólo un contaminante criterio se dispara por medición?, ¿cuáles son los contaminantes que se disparan con mayor frecuencia?. Estas preguntas no pueden ser contestadas con solo tener el registro de la calidad del aire en un momento determinado, si no que se

requiere del análisis de las mediciones de calidad del aire para observar cual es el comportamiento de los datos y obtener las conclusiones pertinentes. Se considera que el análisis de clúster es una buena técnica para descubrir muchos patrones ocultos en los datos. Es claro que el análisis de componentes principales aporta una disminución importante en la dimensión. Con este tipo de estudios se pueden conocer cuáles son los contaminantes que están agrupados en el momento de declarar una contingencia, por ejemplo.

Como trabajo futuro inmediato queda por analizar si se obtienen los mismos resultados que en los experimentos propuestos por [1] y [3] y hacer un comparativo con el enfoque presentado, para hacer una destacada aproximación de un modelo para la toma de decisiones en la mejora de la "calidad de aire". También se puede hacer un análisis de la "descomposición de valores singulares" como otro método lineal de reducción de dimensión o aplicar un método no lineal y hacer experimentos comparativos.

Referencias

1. Camarillo Ramírez, P., Sánchez López, A., Calva Rosales, L. J., Pérez Vásquez, I.: Análisis de datos de calidad del aire de la Zona Metropolitana del Valle de México mediante técnicas de agrupamiento. Journal Research in Computing Science 72, pp. 137–150 (2014)
2. Theodoridis, S., Koutroumbas, K.: An Introduction to Pattern Recognition: A MATLAB Approach. Elsevier Inc., USA (2010)
3. Sanchez, A., Reyes, J.: Analysis of air quality data in Mexico city with clustering techniques based on genetic algorithms. Electronics, Communications and Computing CONIELECOMP, pp. 27–31 (2013)
4. Everitt, B. S., Landau, S., Leese, M., Stahl, D.: Cluster analysis. Wiley 5th Edition, England (2011)
5. Ding, C., He, X.: K-means clustering via principal component analysis. In: Proceedings of the 20th International Conference on Machine Learning (2004)
6. Secretaría del Medio Ambiente, Ciudad de México, http://www.aire.cdmx.gob.mx/
7. Yang, M.S.: A Survey of Fuzzy Clustering. Mathl. Comput. Modelling 18 (11), pp. 1–16 (1993)
8. Krishnapuram, R., Keller, J.M.: A Possibilistic Approach to Clustering. IEEE Transactions on Fuzzy Systems 1, pp. 98–110 (1993)