Análisis de la calidad del aire en el área metropolitana de Monterrey (AMM)

Miguel Ángel Chávez Robles A01620402

***Resumen*—La contaminación del aire es un problema que impacta directamente en la calidad de vida de las personas. Sin dudas la contaminación del aire es un serio problema, particularmente en la área metropolitana de Monterrey (AMM). El objetivo del análisis es comprender la interacción que tienen los contaminantes y los indicadores meteorológicos entre si, dando una perspectiva sobre el impacto de la actividad industrial en San Nicolás de los Garza sobre el ambiente y la población del AMM.**

***Index Terms*—Conglomerados de variables, contaminantes, análisis multivariado.**

I. Problematización

Nuevo León es uno de los estados mas industriales de México, en este se tiene una alta actividad económica por la variedad de actividades que se desempeñan dentro del mismo. Entre los artículos que se elaboran en el estado se encuentran: vidrio, cemento, acero, cerveza, componentes electrónicos, diversos electrodomésticos, medidores para la industria aeroespacial, entre muchos otros productos [Lara, 2018] (todo esto sin mencionar la gran actividad agrícola y ganadera del estado, la cual también es bastante alta). Si bien esto es excelente para la economía del estado, su impacto en el ambiente es un problema innegable y creciente; en 2019 el senador Panista Víctor Fuentes solicitó la intervención del gobierno federal en el asunto por medio de una carta enviada a la SEMARNAT, entre sus demandas se encontraba el estudio de las aportaciones de contaminantes de vehículos, transporte público, industrias y construcción, así como implementar acciones para combatir el problema [Ayala, 2019].

No es para menos, pues la contaminación del aire es un problema que impacta directamente en la calidad de vida de las personas, según la Organización Mundial de la Salud es un determinante directo de la salud pública que además tiene impactos socio económicos tales como absentismo laboral y escolar, así como una relación directa con la incidencia de enfermedades respiratorias en la población [Becerra et al., 2021].

Por si fuera poco, la contaminación del aire también influye directamente en las condiciones climáticas. El cambio climático tiene como principal causa el aumento de la concentración de gases de invernadero [Research Starters, 2021]. Entre estos se encuentran el **dióxido de carbono, el metano y el óxido de nitrógeno**.

Miguel Ángel Chávez Robles pertenece al Instituto Tecnológico y de Estudios superiores de Monterrey. Monterrey, N.L. C.P. , 64849, México

II. Enfoque

Sin dudas la contaminación del aire es un serio problema, particularmente en la área metropolitana de Monterrey (AMM), conformada por este y 12 municipios más, entre ellos San Nicolás de los Garza. Es este último donde se centrará el análisis. San Nicolás de los Garza tiene una actividad industrial, dedicándose principalmente en la producción de cableado, laminados y acero , logrando una actividad económica de 2.89 mil millones de dólares en exportaciones en 2020 [DataMéxico, 2020] por lo que es de particular interés saber como afectan los contaminantes a esta región, y sobre todo como estos interactúan entre si con así como con los indicadores meteorológicos para poder así comprender los efectos que la actividad industrial tiene sobre los mismos y en consecuencia sobre la población del AMM.

En concreto son de interés las medidas correspondientes a:

- Monóxido de carbono (CO)
- Dióxido de carbono (CO_2)
- Dióxido de nitrógeno (NO_2)
- Óxido de nitrógeno (NO_x)
- Ozono (O_3) (lectura a nivel del suelo)
- Material particulado PM_{10} (material particulado como cenizas o polvo con un diámetro menor a 10 μm
- Material particulado $PM_{2.5}$ (material particulado como cenizas o polvo con un diámetro menor a 2.5 μm
- Dióxido de azufre (SO_2)
- Plomo (Pb)
- Hidrocarburos contaminantes tales como:
  Metano (CH_4)
  Benceno (C_6H_6)

pues estos son los contaminantes que han sido reconocidos como causantes de contaminación ambiental.

El análisis se centra en un periodo de tiempo obtenido de una base de datos proporcionada por el Sistema Integral de Monitoreo Ambiental (por sus siglas SIMA) perteneciente al Gobierno del Estado de Nuevo León. Esta base de datos cuenta con lecturas hechas cada hora sobre el nivel de contaminantes, así como de indicadores meteorológicos de una estación ubicada en San Nicolás de los Garza, así como una estampa de tiempo para cada registro. Los datos disponibles corresponden a lecturas que se realizaron cada hora desde 2017 el primero de enero de 2017 hasta el 31 de diciembre de 2020.

III. Propósito

Como se mencionó en la sección II, el objetivo del análisis es comprender la interacción que tienen los contaminantes y los indicadores meteorológicos entre si, dando una perspectiva

sobre el impacto de la actividad industrial en San Nicolás de los Garza sobre el ambiente y la población del AMM. La técnica a utilizar se conoce como conglomerados de variables, se trata de un método jerárquico que calcula la similitud que existe entre las variables, agrupando las que guardan correlación entre si [Johnson, 2016].

A diferencia de otras técnicas existentes, el análisis de conglomerados no parte de suposiciones previas sobre los datos (tales como el número de grupos a considerar). Se busca un agrupamiento natural de los datos a partir de las medidas de similitud encontradas, por lo que es una técnica bastante adecuada para realizar análisis exploratorio [Johnson, 2016].

Para realizar un análisis de conglomerados es necesario conocer las medidas de similitud o distancias entre las variables. En el caso de conglomerados de variables, se requieren coeficientes de asociación como lo es la correlación, para posteriormente obtener la distancia entre las variables a partir de este mismo , cabe mencionar que esta última puede variar dependiendo de si se calcula usando la correlación o la correlación absoluta [Minitab, ND].

Este análisis responde las preguntas:

- ¿Que variables están asociadas entre si?
- ¿Cual es la medida de similitud entre las mismas?
- ¿Existe relación entre la variable x_i y la x_j?
- ¿Que agrupación de variables aporta más información en la menor cantidad de grupos posible?

IV. Información

IV-A. Vista general de el análisis de grupos de variables

La técnica en cuestión es útil para responder a los cuestionamientos planteados en la sección III, pues como se menciona en la sección II la cantidad de datos a analizar es bastante grande, y estos se desglosan en un total de 15 variables continuas y 1 categórica (la estampa de tiempo), por lo que buscar representar información con alta similitud en un menor número de variables sería algo natural, y por medio de los grupos de variables se comprende la relación que estas tienen favoreciendo este propósito.

IV-B. Métodos de enlace

Son adecuados para agrupar tanto variables como observaciones. En el caso de los métodos jerárquicos, inicialmente se tienen tantos grupos como observaciones (en este caso variables), y conforme avanza el proceso se agrupan entre si por similitud, terminando con un solo gran grupo en el último paso. Esto no es muy útil; tener tantos grupos como variables no aporta información, pues solo se están observando por separado, por el contrario tener un solo grupo ignora la distancia que existe entre los mismos y por ende es como observar a todas las variables como si solo fuese una. Es aquí donde entra el concepto de un dendograma, esta figura ilustra todos los posibles agrupamientos que se pueden hacer dadas n variables. Esto lo hace mostrando la distancia entre los grupos en el eje y y el objeto en cuestión en el eje x. Un ejemplo de este tipo de figura se puede observar en la figura 1. Un buen criterio para escoger un número adecuado de grupos es leer el dendograma, observar donde sucede el cambio más pronunciado en distancia, y tomar el número de grupos previo a este cambio (pues es allí donde crece más la similitud entre los grupos). A continuación se explican los métodos de enlace entre los grupos en base a los que se calcula la distancia usados en el análisis.

IV-B1. De enlace simple: También se conoce como de vecino más cercano. Los grupos se forman uniendo las entidades mas cercanas con su vecino más cercano donde el vecino más cercano es el que tiene la menor distancia. Este método de enlace define la distancia entre un grupo U y un grupo V de cualquier otro grupo W como:

$$d_{(UV)W} = min\{5d_{UW}, d_{VW}\} \tag{1}$$

[Johnson, 2016]

IV-B2. De enlace completo: Funciona de una manera muy similar a el enlace simple, pero a diferencia de este entre cada paso se computa la distancia entre los elementos más distantes de un grupo, es decir toman el vecino más lejano. Entonces, la ecuación 1 se convierte en:

$$d_{(UV)W} = max\{5d_{UW}, d_{VW}\} \tag{2}$$

[Johnson, 2016]

IV-B3. De enlace promedio: Este método de enlace trata la distancia entre grupos como la distancia promedio entre pares de elementos de los grupos en cuestión. Toma el elemento más cercano y lo agrega a su grupo. Su fórmula de distancia está dada por:

$$d_{(UV)W} = \frac{\sum_i \sum_k d_{ik}}{N_{(UV)} N_W} \tag{3}$$

Dónde $N_{(UV)}$ y N_W son el número de elementos en el grupo $N_{(UV)}$ y N_W respectivamente [Johnson, 2016].

IV-C. Enlace Ward

Este método de enlace se basa en minimizar la pérdida de información al unir 2 grupos. Dicho de otra forma, busca que los grupos representen la mayor cantidad de información haciendo la menor cantidad de uniones posible. Se basa en la suma de errores cuadráticos. En cada paso, se crea el grupo que crea la suma de errores cuadráticos más chica [Johnson, 2016].

IV-D. Periodo temporal usado

Como se menciona en la sección II los datos cubren desde el primero de enero de 2017 hasta el 31 de diciembre de 2020, siendo un total de 35064 registros. Sin embargo, existen periodos con una pérdida de información considerable, tanto dentro de las observaciones como de las propias variables, de hecho no existe una sola variable (a excepción de fecha) que cuente con todos sus registros. Existen meses enteros con lecturas de ciertas variables faltantes. En concreto las variables NO_2 y NO_x cuentan con un número tan pequeño de observaciones que han sido excluidas del análisis. Las estadísticas descriptivas del periodo 2017-2019 se pueden observar en la figura 2.

Debido a que este análisis se beneficia de tener la mayor cantidad de información posible, no es viable tomar la base de datos entera ya que se requieren registros donde estén presentes todas las variables a analizar, por lo que se delimita un subconjunto para trabajar. Tras realizar un análisis exploratorio se determina que el periodo Agosto 2018 a Julio 2019 es adecuado. Tomar este subconjunto permite tener datos más consistentes. Cabe aclarar que la variable NEWD (North East Wind Direction) se elimina casi por completo, quedando solo con 10 registros. Pese a esto, sigue siendo el periodo más estable encontrado para realizar el análisis. Las estadísticas descriptivas del periodo 01/08/2018-31/07/2019 se pueden observar en la figura 3.

V. Razonamiento

Tras definir el conjunto de datos para el análisis, se realiza el mismo con cada uno de los métodos de enlace descritos en las secciones IV-B1, IV-B2, IV-B3 y IV-C. Se aplicó la siguiente metodología en cada uno de los métodos:

1. Obtener el dendograma completo para el método, así como la matriz de semejanzas y distancias
2. Observar en que paso sucede el salto de distancia más grande
3. Tomar el número de grupos previo a este salto
4. Generar el dendograma con el nuevo número de grupos

Los resultados obtenidos usando esta metodología se pueden observar en las figuras 4, 5, 6 y 7 para el método de enlace simple, 8, 9, 10 y 11 para el método de enlace completo, 12, 13, 14 y 15 para el método de enlace por promedio y 16, 17, 18 y 19 para el método de enlace Ward.

VI. Conclusiones y Recomendaciones

En la sección IV-A se detalla que se espera obtener de estos agrupamientos, por lo que el método que se pueda considerar más significativo es el que haya obtenido una combinación de: diferencia menor de distancia entre los grupos, mayor similitud y menor número de conglomerados tras haber fijado el número de grupos. En este caso los análisis más significativos son tanto el de enlace simple como el de enlace completo, esto es esperable pues la manera en la que funcionan es bastante similar. Si bien es cierto que los resultados muestran que el nivel de semejanza es ligeramente mayor en el caso del método de enlace simple, la mejora es tan baja que es despreciable, ambos análisis son igualmente significativos, aunque personalmente prefiero el método de enlace simple. Ambos métodos agruparon las variables de la misma manera; los conglomerados finales llevan a concluir que:

1. PM_{10} y $PM_{2.5}$ presentan similitud. Ambas hacen referencia a material particulado, aunque es interesante que ambos se presenten en cantidades correlacionadas pues dependiendo de la actividad en la región este podría no ser el caso
2. O_3, Radiación solar, NO y la temperatura presentan similitudes. Si bien esto es de esperarse de la radiación solar y la temperatura, el hecho de que gases contaminantes también lo estén sugiere que los niveles de actividad industrial afectan a el AMM
3. La velocidad del viento, el monóxido de carbono y la lluvia no presentan similitud con ninguna otra variable
4. La humedad relativa y la presión atmosférica presentan similitudes, esto tiene sentido pues la humedad relativa aumenta conforme el aire se enfría, mientras que la presión atmosférica hace referencia a la fuerza que ejerce el aire

Haciendo énfasis en el segundo hallazgo, se sugiere investigar si la similitud encontrada entre esas variables es un fenómeno de causa efecto, y si es así que se tomen las medidas pertinentes, pues tener niveles altos de radiación solar y temperatura en consecuencia de la actividad industrial del AMM perjudicaría a la población.

Referencias

[Ayala, 2019] Ayala, V. (2019). Exige ante contaminación intervención de semarnat. *El Norte (Monterrey, Nuevo León, Mexico)*, page 3.

[Becerra et al., 2021] Becerra, D., Ramírez, L. F., Niño, M. V., Oviedo, C. H., and Plaza, L. F. (2021). Relationship between air quality and the incidence of respiratory diseases in the municipality of san josé de cúcuta, norte de santander. *Ingeniería y Competitividad*, 23(2):1 – 13.

[DataMéxico, 2020] DataMéxico (2020). San nicolás de los garza: EconomÍa, empleo, equidad, calidad de vida, educación, salud y seguridad pública.

[Johnson, 2016] Johnson, Richard A. Wichern, D. W. (2016). *Applied Multivariate Statistical Analysis.* Pearson.

[Lara, 2018] Lara, I. (2018). ¿qué se produce desde nuevo león?

[Minitab, ND] Minitab (N.D). Medidas de distancia para conglomerados de variables.

[Research Starters, 2021] Research Starters (2021). Climate change: Research starters topics. *Climate Change: Research Starters Topics*.

VII. Apéndice

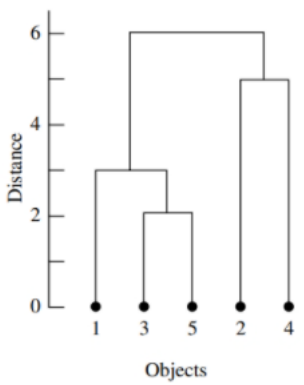

Figura 1. Ejemplo de dendograma

Estadísticas

Variable	N	N*	Media	Error estándar de la media	Desv.Est.	Varianza	Mínimo	Q1	Mediana
PM10	33863	1201	60,083	0,223	40,944	1676,440	2,000	34,000	51,000
PM2.5	19271	15793	20,527	0,130	18,057	326,046	2,020	11,200	16,570
O3	21165	13899	21,710	0,108	15,727	247,333	1,000	10,000	18,000
SO2	25657	9407	8,583	0,0385	6,169	38,059	0,500	2,500	10,300
CO	21746	13318	1,7962	0,00775	1,1430	1,3065	0,0500	0,9700	1,6600
NO	8470	26594	5,442	0,273	25,109	630,452	0,500	0,500	0,600
NO2	2101	32963	2,5531	0,0467	2,1413	4,5853	0,5000	1,3000	1,8000
NOx	2093	32971	22,51	1,05	48,09	2312,88	2,90	5,30	6,30
TOUT	32815	2249	22,737	0,0393	7,114	50,605	-1,860	18,030	23,410
RH	32443	2621	66,594	0,116	20,931	438,102	0,000	52,000	70,000
SR	33902	1162	0,14064	0,00130	0,23878	0,05702	0,00000	0,00000	0,00000
RAINF	33307	1757	0,8292	0,0206	3,7635	14,1639	0,0000	0,0000	0,0000
PRS	32773	2291	719,36	0,0222	4,01	16,09	682,50	716,70	718,90
WS	30577	4487	7,130	0,0243	4,253	18,091	0,300	3,900	6,500
NEWD	15981	19083	132,60	0,804	101,61	10324,19	1,00	56,00	103,00

Variable	Q3	Máximo
PM10	75,000	731,000
PM2.5	24,690	660,500
O3	31,000	120,000
SO2	12,900	151,600
CO	2,7200	9,7700
NO	0,800	500,000
NO2	3,0000	26,3000
NOx	13,20	500,00
TOUT	27,740	41,570
RH	84,000	96,000
SR	0,19100	0,94500
RAINF	0,0000	17,9900
PRS	721,60	736,00
WS	9,900	127,500
NEWD	197,00	360,00

Figura 2. Estadísticas descriptivas de la base de datos entera

Estadísticas

Variable	N	N*	Media	Error estándar de la media	Desv.Est.	Varianza	Mínimo	Q1	Mediana
PM10	8630	130	57,694	0,461	42,836	1834,916	2,000	31,000	49,000
PM2.5	7834	926	20,766	0,253	22,427	502,991	2,020	10,650	16,155
O3	5103	3657	20,748	0,206	14,716	216,572	2,000	10,000	17,000
SO2	7884	876	4,7068	0,0550	4,8828	23,8419	0,5000	1,6000	2,7000
CO	2170	6590	2,1377	0,0318	1,4823	2,1972	0,0500	1,0750	1,9700
NO	598	8162	0,56806	0,00309	0,07547	0,00570	0,50000	0,50000	0,60000
TOUT	8721	39	22,349	0,0797	7,440	55,355	2,850	16,790	23,340
RH	8465	295	69,341	0,219	20,147	405,893	2,000	56,000	73,000
SR	8743	17	0,15741	0,00259	0,24248	0,05880	0,00000	0,00000	0,00300
RAINF	8723	37	0,00296	0,000388	0,03627	0,00132	0,000000	0,000000	0,000000
PRS	8723	37	719,06	0,0422	3,95	15,57	707,80	716,50	718,60
WS	8718	42	7,1246	0,0417	3,8921	15,1482	0,7000	3,9000	6,5000
NEWD	10	8750	1,0000	0,000000	0,000000	0,000000	1,0000	1,0000	1,0000

Variable	Q3	Máximo
PM10	73,000	731,000
PM2.5	24,590	660,500
O3	29,000	111,000
SO2	4,7000	35,1000
CO	3,3000	9,7700
NO	0,60000	0,70000
TOUT	27,660	41,570
RH	86,000	96,000
SR	0,24500	0,91600
RAINF	0,000000	1,60000
PRS	721,40	733,10
WS	9,9000	22,0000
NEWD	1,0000	1,0000

Figura 3. Estadísticas descriptivas del subconjunto seleccionado

Pasos de amalgamación

Paso	Número de conglomerados	Nivel de semejanza	Nivel de distancia	Conglomerados incorporados		Nuevo conglomerado	Número de obs. en el conglomerado nuevo
1	11	92.2639	0.154723	6	7	6	2
2	10	86.3481	0.273039	3	9	3	2
3	9	85.3523	0.292953	3	6	3	4
4	8	81.0041	0.379919	1	2	1	2
5	7	73.2469	0.535061	8	11	8	2
6	6	66.0575	0.678851	3	12	3	5
7	5	63.6579	0.726841	1	4	1	3
8	4	63.1899	0.736202	1	3	1	8
9	3	62.3843	0.752313	1	5	1	9
10	2	60.0895	0.798210	1	8	1	11
11	1	55.7218	0.885564	1	10	1	12

Figura 4. Matriz con enlace simple

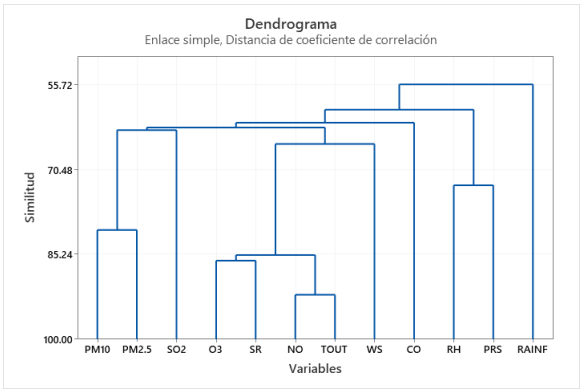

Figura 5. Dendograma preliminar usando enlace simple

Partición final

	Variables
Conglomerado 1	PM10 PM2.5
Conglomerado 2	O3 NO TOUT SR
Conglomerado 3	SO2
Conglomerado 4	CO
Conglomerado 5	RH PRS
Conglomerado 6	RAINF
Conglomerado 7	WS

Figura 6. Matriz de grupos final con enlace simple

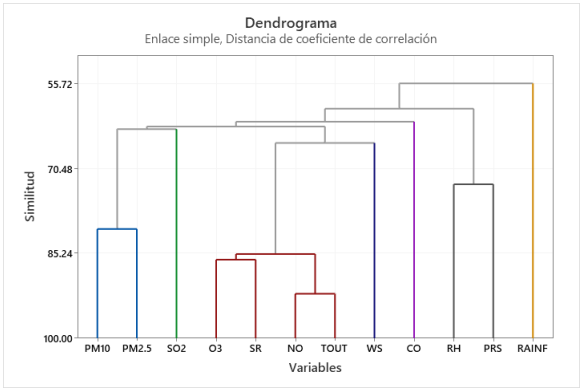

Figura 7. Dendograma usando enlace simple

Pasos de amalgamación

Paso	Número de conglomerados	Nivel de semejanza	Nivel de distancia	Conglomerados incorporados		Nuevo conglomerado	Número de obs. en el conglomerado nuevo
1	11	92.2639	0.15472	6	7	6	2
2	10	86.3481	0.27304	3	9	3	2
3	9	81.0041	0.37992	1	2	1	2
4	8	77.8343	0.44331	3	6	3	4
5	7	73.2469	0.53506	8	11	8	2
6	6	61.8098	0.76380	1	4	1	3
7	5	57.7711	0.84458	5	12	5	2
8	4	50.4666	0.99067	8	10	8	3
9	3	47.5854	1.04829	3	5	3	6
10	2	41.1264	1.17747	1	3	1	9
11	1	2.7884	1.94423	1	8	1	12

Figura 8. Matriz con enlace completo

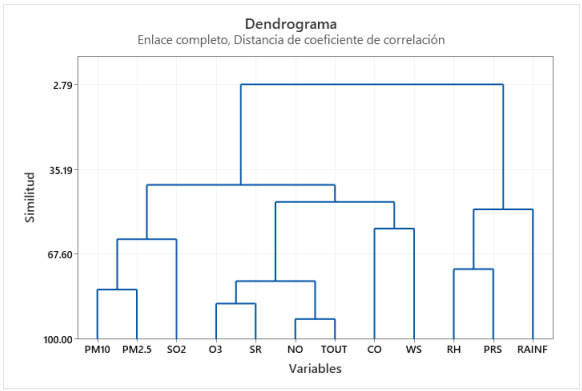

Figura 9. Dendograma preliminar usando enlace completo

Partición final

	Variables
Conglomerado 1	PM10 PM2.5
Conglomerado 2	O3 NO TOUT SR
Conglomerado 3	SO2
Conglomerado 4	CO
Conglomerado 5	RH PRS
Conglomerado 6	RAINF
Conglomerado 7	WS

Figura 10. Matriz de grupos final con enlace completo

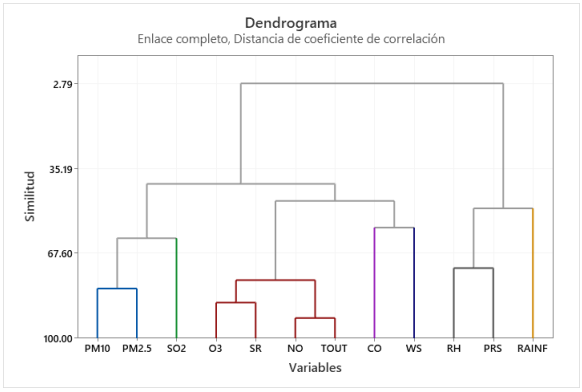

Figura 11. Dendograma usando enlace completo

Pasos de amalgamación

Paso	Número de conglomerados	Nivel de semejanza	Nivel de distancia	Conglomerados incorporados		Nuevo conglomerado	Número de obs. en el conglomerado nuevo
1	11	92.2639	0.15472	6	7	6	2
2	10	86.3481	0.27304	3	9	3	2
3	9	81.9672	0.36066	3	6	3	4
4	8	81.0041	0.37992	1	2	1	2
5	7	73.2469	0.53506	8	11	8	2
6	6	62.7339	0.74532	1	4	1	3
7	5	58.6177	0.82765	1	5	1	4
8	4	57.4577	0.85085	3	12	3	5
9	3	51.8494	0.96301	1	3	1	9
10	2	51.6447	0.96711	8	10	8	3
11	1	40.0337	1.19933	1	8	1	12

Figura 12. Matriz con enlace de promedios

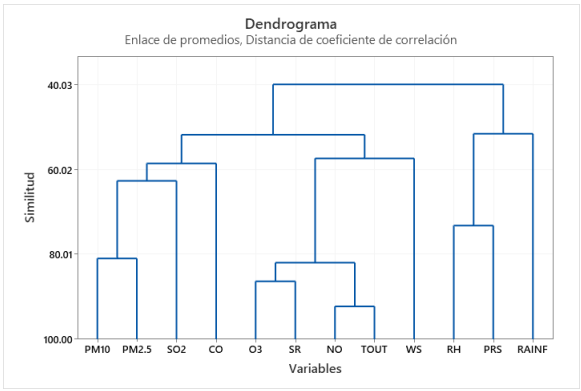

Figura 13. Dendograma preliminar usando enlace de promedios

Partición final

	Variables
Conglomerado 1	PM10 PM2.5 SO2
Conglomerado 2	O3 NO TOUT SR
Conglomerado 3	CO
Conglomerado 4	RH PRS
Conglomerado 5	RAINF
Conglomerado 6	WS

Figura 14. Matriz de grupos final con enlace de promedios

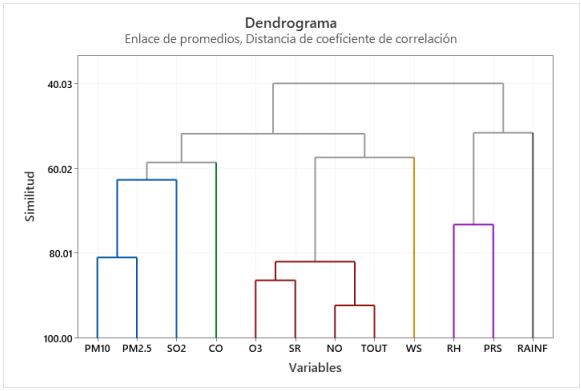

Figura 15. Dendograma usando enlace de promedios

Pasos de amalgamación

Paso	Número de conglomerados	Nivel de semejanza	Nivel de distancia	Conglomerados incorporados		Nuevo conglomerado	Número de obs. en el conglomerado nuevo
1	11	92.2639	0.15472	6	7	6	2
2	10	86.3481	0.27304	3	9	3	2
3	9	81.0041	0.37992	1	2	1	2
4	8	74.6285	0.50743	3	6	3	4
5	7	73.2469	0.53506	8	11	8	2
6	6	57.7711	0.84458	5	12	5	2
7	5	56.6438	0.86712	1	4	1	3
8	4	45.9008	1.08198	5	10	5	3
9	3	39.9869	1.20026	1	5	1	6
10	2	18.8489	1.62302	1	8	1	8
11	1	-57.7977	3.15595	1	3	1	12

Figura 16. Matriz con enlace Ward

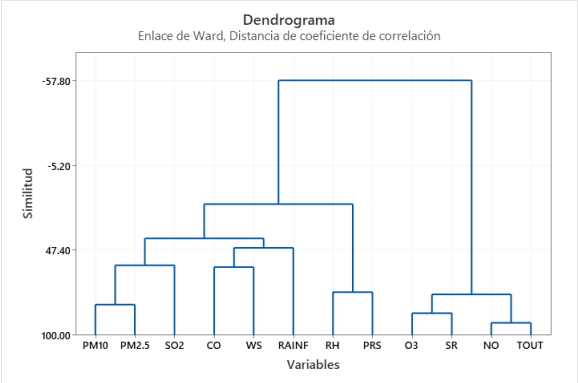

Figura 17. Dendograma preliminar usando enlace Ward

Partición final

	Variables
Conglomerado 1	PM10 PM2.5
Conglomerado 2	O3 NO TOUT SR
Conglomerado 3	SO2
Conglomerado 4	CO WS
Conglomerado 5	RH PRS
Conglomerado 6	RAINF

Figura 18. Matriz de grupos final con enlace Ward

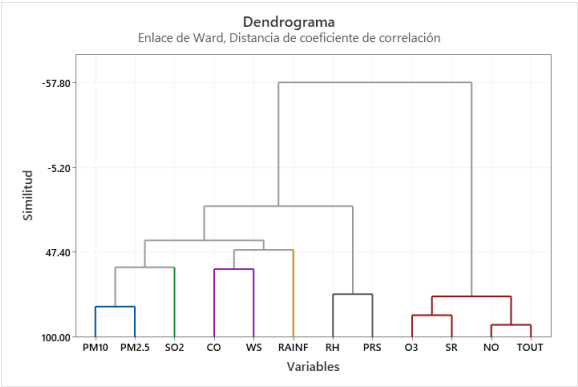

Figura 19. Dendograma usando enlace Ward