Análisis de contaminantes de San Nicolás de los Garza

Pedro Alan González Arámbula - A01625308, Miguel Ángel Chávez Robles - A01620402,
Francisco Leonid Gálvez Flores - A01174385
Instituto Tecnológico y de Estudios Superiores de Monterrey. Monterrey, Nuevo León

Introducción

El análisis de componentes principales es una técnica usada cuando se tienen datos con un alto número de dimensiones y se desea encontrar un número menor de variables no correlacionadas, denominadas “componentes principales” para representar la misma información. Estas son combinaciones lineales de las variables originales [2].

Base de datos

Se analizarón los datos de contaminantes e indicadores meteoreológicos proporcionados por el “Sistema de Monitoreo Ambiental” de la Dirección De Gestión Integral de la Calidad del Aire del Gobierno del Estado de Nuevo León. La estación analizada fue la estación noreste, ubicada en San Nicolás de los Garza, Nuevo León. El periodo de tiempo usado para el análisis va del primero de agosto de 2018 al 31 de julio del 2019. Este periodo fue escogido debido a que en el anásis exploratorio este demostró ser el más completo, pues tenía el menor número de datos faltantes.

Metodología

Análisis de valores propios

Al realizar un análisis de componentes principales se busca encontrar k variables que expliquen de buena manera las p variables originales de manera que $k < p$. Este puede realizarse de 2 maneras: por matriz de correlación y de covarianza; en nuestro caso se realizó el análisis por matriz de correlación pues las escalas de las variables no son iguales. Los valores propios, proporción de varianza explicada y varianza explicada acumulada por componente se muestran a continuación.

Análisis de los valores y vectores propios de la matriz de correlación

Valor propio	3,1752	1,8848	1,1960	0,9350	0,8109	0,4002	0,2639	0,1867	0,1473
Proporción	0,353	0,209	0,133	0,104	0,090	0,044	0,029	0,021	0,016
Acumulada	0,353	0,562	0,695	0,799	0,889	0,934	0,963	0,984	1,000

4450 casos utilizados, 4310 casos contienen valores faltantes

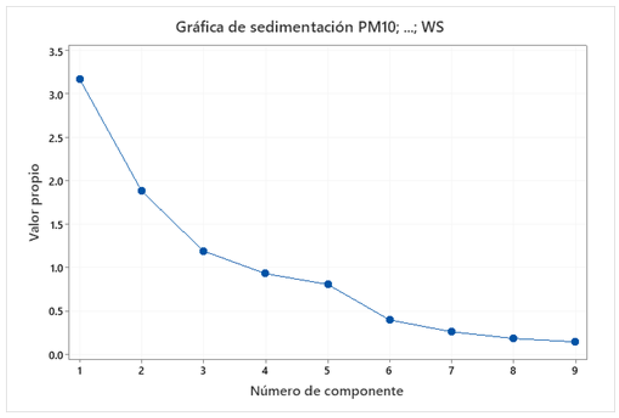
Al momento de escoger cuáles k componentes existen 2 técnicas bastante utilizadas, el “Criterio de Kaiser-Guttman”, el cual sugiere escoger componentes principales cuyo valor propio $\lambda > 1$ [1], alternativamente existe el criterio de varianza acumulada, el cual establece que deben tomarse k componentes tales que la varianza acumulada de estos sea mayor o igual al 80 % de la varianza total. Esta se obtiene con la fórmula mostrada a continuación.

$$\frac{\lambda_1 + \lambda_2 + ... + \lambda_k}{\lambda_1 + \lambda_2 + ... + \lambda_p}$$ [3]

En el análisis realizado se optó por el criterio de la varianza acumulada, quedando con 5 componentes principales.

Selección de componentes principales

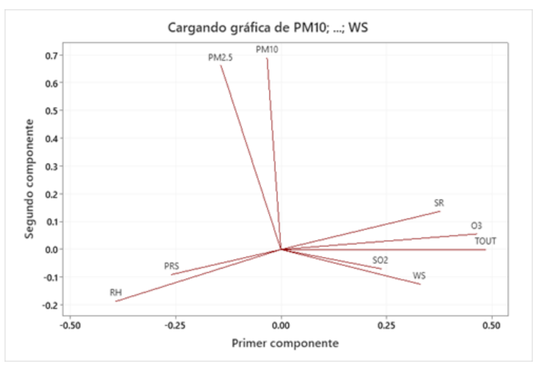
La gráfica de sedimentación representa la varianza acumulada de la matriz anterior.

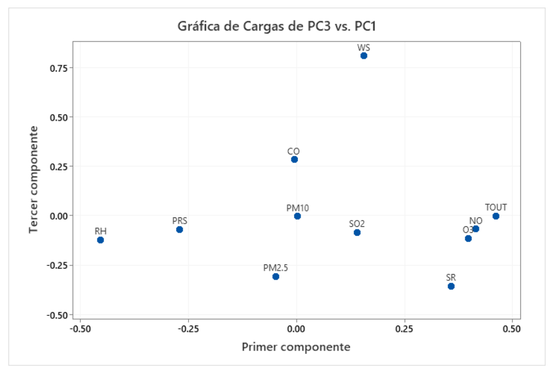


Tras haber decidido usar 5 componentes principales, se obtiene la matriz de eigenvectores. Esta se interpreta como la carga que tiene cada variable sobre el componente. Esta carga puede ser tanto positiva como negativa. A su derecha se muestra la interpretación gráfica de las 2 primeras columnas

Vectores propios

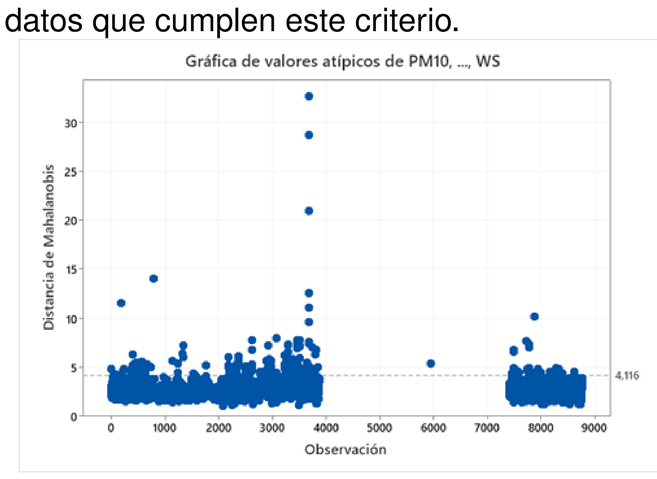
Variable	PC1	PC2	PC3	PC4	PC5
PM10	-0,033	0,690	0,065	0,066	0,123
PM2.5	-0,142	0,662	0,073	-0,006	0,114
O3	0,464	0,056	-0,301	-0,034	-0,205
SO2	0,237	-0,069	0,356	-0,579	0,637
TOUT	0,485	-0,001	0,318	0,000	-0,097
RH	-0,392	-0,187	0,352	-0,230	-0,154
SR	0,377	0,137	-0,326	-0,471	-0,239
PRS	-0,260	-0,090	-0,665	-0,175	0,451
WS	0,330	-0,125	0,005	0,594	0,480





Valores atípicos

Usando la distancia de Mahalanobis se determinó que los valores atípicos son todos los que se encuentren arriba de 4.116 de la misma. Existe un número considerable de datos que cumplen este criterio.



Conclusiones

Se escogió usar 5 componentes para describir de manera completa los datos originales, pues estos explican el 88.9 % de la varianza. No todos los componentes cumplen con el criterio de Kaiser-Guttman, lo cual puede esperarse de los datos, ya que estos tienen un comportamiento errático, son relativamente estacionales, además existen grandes periodos sin lecturas y una cantidad considerable de valores atípicos. A futuro sería valioso realizar el mismo análisis utilizando variables transformadas. Adicionalmente se sugiere estudiar los valores atípicos encontrados para confirmar si estos se tratan de errores de captura, fenómenos naturales ocasionales, o si realmente el clima en San Nicolás de los Garza tiene un comportamiento anormal.

Tecnológico de Monterrey

DIRECCIÓN DE GESTION INTEGRAL DE CALIDAD DEL AIRE

Referencias

[1] D. A. Jackson. Stopping rules in principal components analysis: A comparison of heuristical and statistical approaches. *Ecology*, 74(8):2204–2214, 1993.
[2] R. A. Johnson and D. W. Wichern. *Applied Multivariate Statistical Analysis.* Pearson, 2016.
[3] Minitab. Metodos y formulas para analisis de componentes principales.