CLI-MAX: Calidad del aire en Apodaca

Andrea Bravo[1], Renata Vargas[1], Alberto Lozano[1], Mariana Rincón[1], Leonardo Laureles[1], Carlos Mateos[1]; [*]equal contribution

[1]Tecnológico de Monterrey, México;

El reto consiste en diseñar una propuesta para la detección de las relaciones que se establecen entre los contaminantes del aire y su relación con el clima en las distintas estaciones meteorológicas de la Ciudad de Monterrey.

Introducción

La ciencia de datos es una rama muy moderna que a la vez incluye otras técnicas como Machine Learning, Deep Learning, Big Data, Artificial Intelligence, entre muchas otras, que dan oportunidad a las empresas de analizar sus datos de una manera más precisa y confiable al ser respaldadas por fundamentos matemáticos, estadísticos y de programación. La empresa socio formadora, SIMA y la dirección de gestión de calidad del aire, busca un armado de propuesta para el análisis de la relación entre los contaminantes del aire y las distintas estaciones metereológicas de la Ciudad de Monterrey.

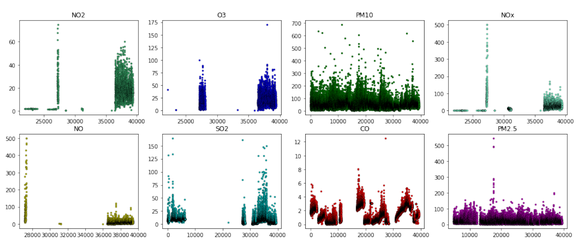
¿Cuáles son los factores que más se relacionan entre los contaminantes del aire y las distintas estaciones meteorológicas?

Justificación

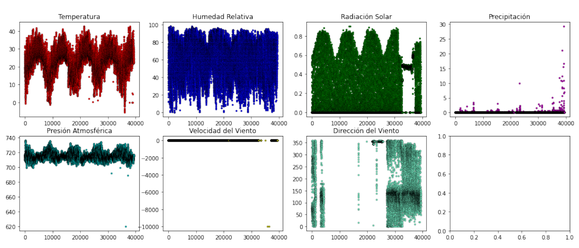
Al contar con varias variables en las bases de datos, se hace posible realizar un análisis de relaciones de interdependencia entre las variables, para ello se pueden utilizar distintos métodos en los que destacan el análisis de factores y el análisis de conglomerados, cuyos resultados arrojan información de valor para ver patrones o tendencias que comparten las variables entre sí; por lo que resulta factible hacer uso de estos métodos para encontrar una relación entre los contaminantes del aire y las variables meteorológicas de la estación de Apodaca, Nuevo León.

Información-Estación

La estación a analizar se encuentra ubicada en el municipio Apodaca, Nuevo León; la cuál la podemos encontrar en la base de datos como NE2, los contaminantes que consideramos importantes para el análisis fueron, PM10, PM2.5 Y SO2. Las variables metereológicas son TOUT (Temperatura), RH (Humedad relativa), SR (Radiación solar), RAINF (Precipitación), PRS (Presión atmsférica), WS (Velocidad del viento), WD (Dirección del viento).



Dispersión de contaminantes en Apodaca, Nuevo León Gráficas de dispersión para ver la distribución de datos en el tiempo



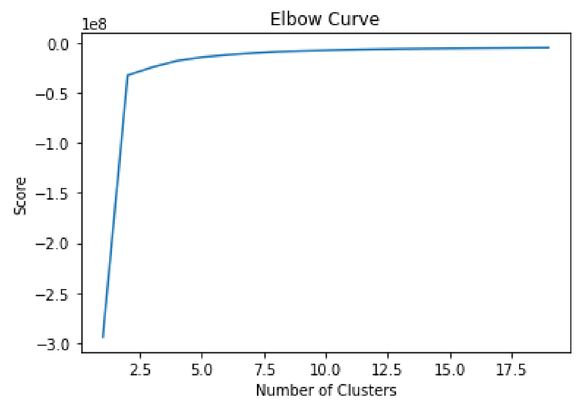
Dispersión de variables meteorológicas en Apodaca, Nuevo León Gráficas de dispersión para ver la distribución de datos en el tiempo

Modelación

Para la modelación de los datos se planteó un análisis dependiente e interdependiente, las pruebas realizadas arrojaron que los datos no tienen una distribución normal multivariada, por lo que se optó por utilizar 2 técnicas de análisis interdependiente: análisis de conglomerados y análisis factorial.

Análisis de conglomerados

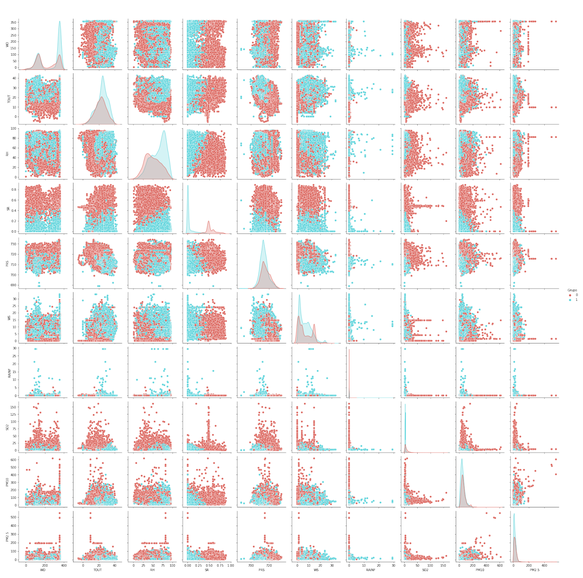
Se realizó un análisis de conglomerados por medio del método KMeans, para realizar dicho método, primero se necesita saber con cuántos clusters se va a trabajar, esto es una de las desventajas, ya que no se conoce con certeza, por lo que se optó por realizar un gráfico de codo.



Gráfica de codo Se opta por trabajar con 2 clusters debido a que se observa el punto de codo donde la tasa de ascenso se afila en ese punto de la gráfica.

Se ejecuta el algoritmo KMeans para 2 clusters

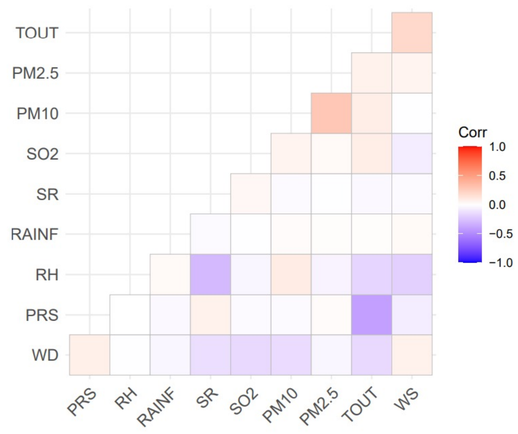
Cluster plot con R Se opta por descartar este método y realizar el análisis factorial



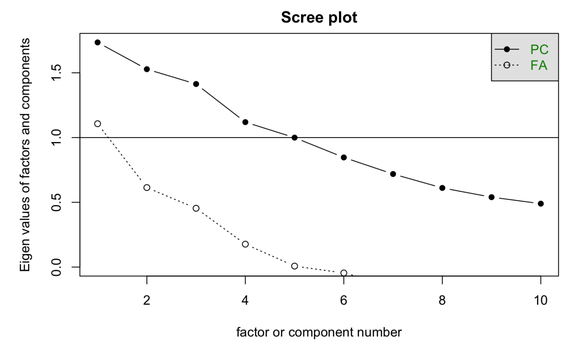
Pairplot con grupos por K-mediasGráfica de K-means Solamente se utilizan los quadrantes

Análisis de Factores utilizando variables binarias

En razón de que la base de datos cuenta con variables cuantitativas, que son numéricas continuas, pero ninguna de ellas se distribuye de manera normal, se decidió crear una variable dependiente de los contaminantes a partir de las Normas Oficiales Mexicanas (NOM) de Calidad del Aire Ambiente.



Matriz de correlación con variables binarias Se opta por descartar este método y realizar el análisis factorial



Gráfica de scree plot Encontrar el número de factores

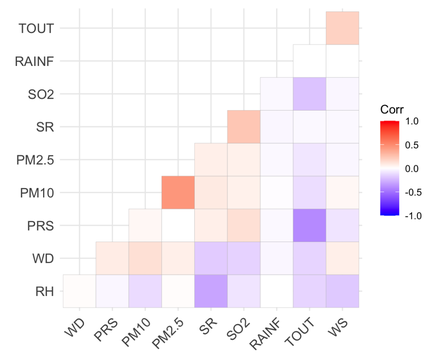
Una vez contando con el test de Bartlett, que indica que si hay suficiente significancia, se realizó un análisis de factores con el método de componentes principales, en el cual se pueden observar un factor y sus pesos con relación a cada variable

	V	PA1	h2	u2	com
	<S3: AsIs>	<S3: AsIs>	<dbl>	<dbl>	<dbl>
TOUT	2	0.87	0.7497014893	0.2502985	1
PRS	5	-0.43	0.1877200058	0.8122800	1
WD	1	-0.22	0.0505322013	0.9494678	1
WS	6	0.18	0.0330312634	0.9669687	1
PM10	9	0.18	0.0327509175	0.9672491	1
PM2.5	10	0.14	0.0197033233	0.9802967	1
SO2	8	0.13	0.0174450883	0.9825549	1
RH	3	-0.13	0.0164720862	0.9835279	1
RAINF	7	0.04	0.0014333148	0.9985667	1
SR	4	0.02	0.0004308373	0.9995692	1

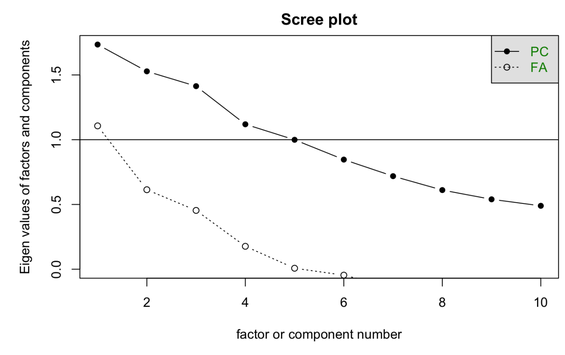
Tabla Análisis Factorial Se observan los pesos de cada variable

Análisis de Factores utilizando variables continuas

Se hizo un segundo análisis de factores, pero esta vez sin usar variables binarias para hacer un análisis interdependiente. Como primera instancia se generó una matriz de correlación para poder observar de una manera gráfica la existencia de correlación en las variables.



Gráfica de K-means Se opta por descartar este método y realizar el análisis factorial



Gráfica de K-means Se opta por descartar este método y realizar el análisis factorial

	V	PA1	h2	u2	com
	<S3: AsIs>	<S3: AsIs>	<dbl>	<dbl>	<dbl>
TOUT	2	0.87	0.7497014893	0.2502985	1
PRS	5	-0.43	0.1877200058	0.8122800	1
WD	1	-0.22	0.0505322013	0.9494678	1
WS	6	0.18	0.0330312634	0.9669687	1
PM10	9	0.18	0.0327509175	0.9672491	1
PM2.5	10	0.14	0.0197033233	0.9802967	1
SO2	8	0.13	0.0174450883	0.9825549	1
RH	3	-0.13	0.0164720862	0.9835279	1
RAINF	7	0.04	0.0014333148	0.9985667	1
SR	4	0.02	0.0004308373	0.9995692	1

Gráfica de K-means Se observan los pesos de cada variable

Discusión y conclusiones

En la matriz de correlación del análisis utilizando variables binarias se presentan 3 relaciones fuertes (PM10-PM2.5, SR-SO2, TOUT-WS), mientras que en la del análisis utilizando variables continuas se observan 2 (PM10-PM2.5, TOUT-WS), a partir de esto se puede decir que existe una relación entre las temperaturas y la dirección de viento con los contaminantes PM10 y PM2.5. Con estas relaciones, se puede definir las causas de las concentraciones de contaminantes en la región de Apodaca.