Evidencia final:
Reporte individual.

Equipo 5

Pablo Pérez Sandoval A01710355

7 de septiembre del 2024

Aplicación de métodos multivariados en ciencia de datos (Grupo 101)

Link a los archivos

Introducción

Debido a la libertad que se nos otorgó en este reto mi equipo y yo iniciamos con una lluvia de ideas para encontrar un tema en específico en el que pudiéramos trabajar y que fuera de valor. La primera idea que decidimos indagar fue el buscar una relación entre la inversión extranjera en la ciudad de Monterrey y la cantidad de contaminantes. El principal problema con esta idea fue la extracción de la base de datos y cómo implementar las técnicas vistas en la clase para obtener resultados de valor. Al seguir desarrollando esta idea observamos que no encontrábamos ninguna correlación significativa entre los contaminantes contra la inversión extranjera, principalmente por el tiempo en el que se veían reflejados esos montos de dinero invertidos en la ciudad de Monterrey.

Nuestra segunda idea iba de la mano con investigar las causas de los contaminantes por medio del análisis factorial exploratorio. Lo que hacía interesante a este objetivo era el observar si las causas de los contaminantes cambiaban por cada zona de Monterrey, en este análisis esperábamos encontrar grandes diferencias entre las zonas pero nunca nos imaginamos que el análisis factorial exploratorio nos llevaría tanto tiempo para cada una de las zonas.

Con estos aprendizajes en mente es como llegamos al proyecto que presentamos. Este consiste en aplicar un análisis factorial exploratorio a una de las estaciones de monitoreo de la zona central de Monterrey. El objetivo de este análisis era encontrar similitudes y diferencias entre las distintas épocas del año, y cómo es que las temporadas anuales afectan a la producción y comportamiento de los contaminantes presentes en la base de datos.

Con esto en mente procedemos a explicar la metodología que seguimos.

Metodología.

Después de leer las bases de datos y almacenarlas en nuestro programa procedimos a juntarlas en un solo dataset. A continuación observamos un poco los datos y eliminamos las filas y columnas que no utilizaremos en nuestra base de datos, en este caso las únicas columnas que fueron útiles para la persecución de nuestro objetivo fueron las columnas que cuantificaban los contaminantes presentes en las bases de datos. Con esto en mente seguimos observando nuestra base de datos y decidimos reducirla a un total de 5 zonas, ya que consideramos que estas zonas son una representación acertada de la totalidad de la zona metropolitana, además de ser zonas que contaban la menor cantidad de datos nulos (menor al 5%). Con nuestras filas y columnas seleccionadas nos encontramos con el reto de llenar los datos vacíos, que después de analizarlo llegamos a la conclusión que la mejor manera de sustituirlos era mediante el método del promedio móvil, ya que se trataba de una serie de tiempo.

Con la creación y limpieza de los datos nulos procedemos a realizar un análisis exploratorio de cómo es que se comportaba cada uno de los contaminantes en cada una de las zonas, estas observaciones se hicieron mediante las medidas de tendencia central y las medidas dispersión. Para poder realizar estas observaciones de manera más gráfica nos apoyamos de dos gráficos, uno enfocado en las medidas de tendencia central (histograma) y otro enfocado en las medidas de dispersión (Boxplot). Con estas observaciones llegamos a la conclusión que los valores extremos afectan de manera significativa a nuestros datos, por lo que decidimos removerlos de la base de datos.

Después de realizar estas modificaciones a las bases proporcionadas por el SIMA pasamos a tener los siguientes valores para nuestra base de datos.

Previo a las modificaciones		Posterior a las modificaciones	
Cantidad de columnas.	240	Cantidad de columnas.	42
Cantidad de registros.	19,726	Cantidad de registros.	19,479
Zonas del área metropolitana.	15	Zonas del área metropolitana.	5

Con esta base de datos limpios encontramos que la mejor manera de proseguir era mediante la graficación de la serie de tiempo de cada una de las 5 estaciones. Después de un cercano análisis de las series de tiempo observamos que la estación de Obispado (zona centro de Monterrey) tiene un gran componente de estacionalidad en su gráfica, por lo que la seleccionamos para seguir con nuestro posterior análisis.

Con nuestra base de datos completamente limpia y ordenada procedimos a separar en 4 base de datos dependiendo de las estaciones del año en la que se tomaron esas mediciones.

Ya con las bases de datos de las estaciones del año procedimos a aplicar las validaciones para nuestro modelo. Hicimos pruebas de normalidad, homocedasticidad y la prueba de Prueba de Kaiser-Meyer-Olkin (KMO). En este caso la única prueba totalmente necesaria para realizar el análisis exploratorio era la prueba de KMO ya que para el análisis exploratorio podemos omitir la normalidad de los datos.

Encontramos los siguientes resultados para la prueba de la prueba KMO:

- Datos Primavera: 0.61
- Datos Verano: 0.66
- Datos Otoño: 0.76
- Datos Invierno: 0.76

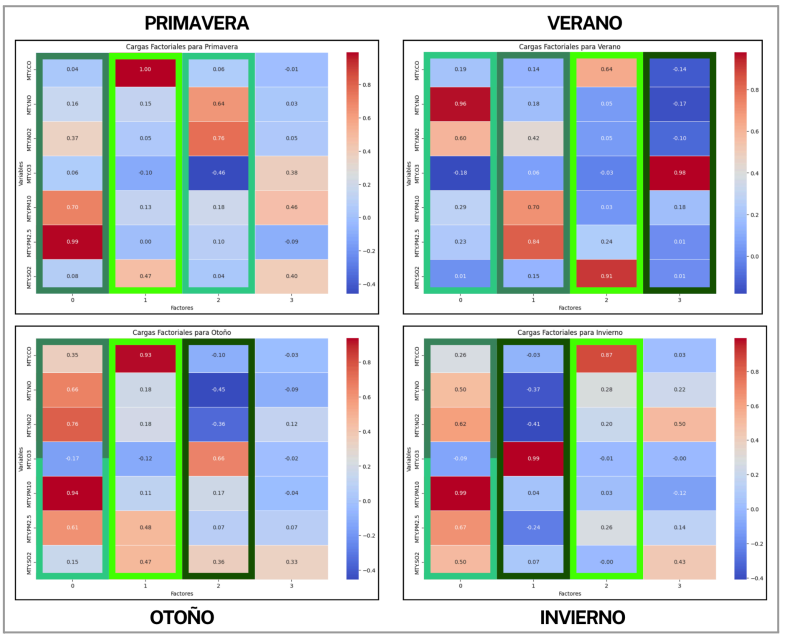
Con esto vemos que nuestro análisis factorial es plausible y que nos proporciona insights importantes para la posterior interpretación de los datos.

Resultados

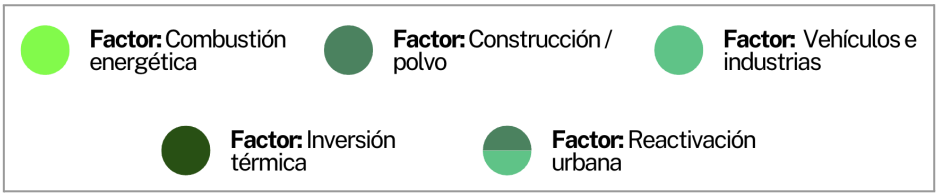
Empleamos el método de máxima verosimilitud con rotación ortogonal "Quartimax" para identificar las correlaciones más fuertes entre las variables de forma independiente. La selección del número adecuado de factores se basó en la varianza acumulada explicada por cada factor en cada estación. Con ayuda de un Scree Plot pudimos encontrar que la mejor cantidad de factores era de 4 para todas las estaciones.

Con el número óptimo de factores identificado y las hipótesis del equipo definidas, se generaron representaciones gráficas de los análisis factoriales, que permitieron evaluar en

detalle el comportamiento de los contaminantes. A continuación, se presentan las ilustraciones resultantes de dichos análisis.



De esta gráfica pudimos identificar que ciertos factores se repetían sin importar la época del año. Los indicamos de la siguiente manera.



En el análisis se identificaron cinco factores clave que explican las fuentes y la dispersión de contaminantes. El primer factor está relacionado con la "combustión energética", vinculada a contaminantes como el monóxido de carbono y el dióxido de azufre. El segundo factor, "construcción/polvo", agrupa partículas finas y gruesas asociadas a actividades industriales y de construcción. El tercer factor, "vehículos e industrias", se relaciona con contaminantes generados por el tráfico vehicular y las actividades industriales. Un cuarto factor, llamado "inversión térmica", está relacionado principalmente con el ozono, y un quinto factor, "reactivación urbana", resulta de la combinación de factores previos debido al aumento en la actividad económica y vehicular.

Conclusiones.

Los resultados del análisis muestran cómo estos factores predominan de manera diferente según las condiciones climáticas y las estaciones del año, observándose una mayor dispersión de contaminantes en estaciones más cálidas debido a la radiación solar, mientras que en las estaciones más frías la inversión térmica juega un rol importante en la acumulación de contaminantes. En general, se concluye que las condiciones meteorológicas, como la temperatura y la inversión térmica, junto con las actividades humanas, son determinantes en la concentración y dispersión de contaminantes en diferentes épocas del año.