Reflexión Individual: Análisis de Datos Ambientales y Métodos Estadísticos

Instituto Tecnológico de Estudios Superiores de Monterrey
MA2003B.102 Application of Multivariate Methods in Data Science

Introducción

El desarrollo del proyecto presentó una experiencia para aplicar conocimientos matemáticos, tecnológicos y especialmente estadísticos. La situación que revisamos fue la contaminación atmosférica en la Zona Metropolitana de Monterrey. Este trabajo consistió en tomar datos reales, complejos y con un nivel significativo de inconsistencias para transformarlos en información estructurada y analizable. Esto permitió comprender cómo la estadística aplicada y la programación pueden convertirse en herramientas utiles para apoyar a la toma de decisiones públicas y privadas.

Este reto también incluyo aprender a comunicar resultados científicos de manera clara, para poder crear explicaciones que sean entendidas por personas que no están tan familiarizadas con el campo y están más interesadas en ver resultados que procesos.

Base de datos

La información que se uso provino de la Red de Monitoreo Ambiental de Nuevo León (SIMA), la cual integra series temporales con registros horarios de contaminantes (PM10, PM2.5, NO_2, NO, CO, O_3, SO_2, NOx) y variables meteorológicas como la temperatura, humedad, radiación solr, precipitación, presión y viento.

Los datos originales tienen aproximadamente 779,000 observaciones. La fase de preparación de datos fue un proceso largo y esencial. El procesamiento incluyó los siguientes pasos:

1. Unificación de cinco estructuras de datos diferentes.
2. Homologación de nombres de variables y códigos de estaciones.
3. Eliminación de duplicados.
4. Filtrado de outliers por umbrales.

A partir de estas actividades, se construyó un dataframe maestro con 17 columnas estandarizadas, lo cual aseguro trazabilidad y replicabilidad. Me

atrevo a decir que esta etapa fue una de las que más aportaciones tuvo de mi parte y estoy satisfecho con la entrega de la misma.

Métodos estadísticos y tecnológicos

Se implementaron 2 bloques principales en el proyecto: imputación jerárquica de datos y análisis exploratorio de datos.

Imputación jerárquica de datos faltantes

Se diseñó un enfoque de tres niveles:

1. Interpolación temporal para brechas cortas (menos de 6 horas).
2. Borrowing espacial con estaciones vecinas (entre 6 y 48 horas)
3. Eliminar huecos extensos para evitar sesgos.

Este esquema se comparó con tres métodos específicos:

- MTB (Mean Top-Bottom): Promedio de valores adyacentes,
- Hearest Neighbour: relleno con observaciones más cercano en el tiempo.
- Iterative (MICE-like): Imputación multivariada.

La validación mediante MAE y RMSE mostró que el método MTB ofrecía mejor balance, sin embargo, en casos particulares, rellenaba datos con una media igual a la anterior, lo cual, viéndolo en un gráfico, se mostraba como una linea recta. Finalmente, se utilizó una mezcla entre los 3 metodos, dependiendo de la longitud de la brecha.

Análisis exploratorio

El análisis exploratorio inicio con estadísticas descriptivas, box plots, y series temporales. Los promedios móviles de 7 horas en contaminantes como el Ozono (O_3) y CO fueron útiles para cumplir con criterios normativos y detectar puntos críticos. También se realizaron comparaciones interanuales para identificar variaciones significativas en periodos clave: inicio y fin de pandemia (2020-2021), reactivación económica y situación actual.

Este proceso requirió integrar conocimientos de programación en Python, utilizando librerías estadísticas y visualización de datos.

Resultados

Durante la pandemia (2020), las reducciones más visibles, se observaron en contaminantes asociados a tránsito vehicular (NO_2, NOx, CO), confirmando el efecto inmediato de la reducción en movilidad. El PM2.5 también disminuyó aunque con picos mas aislados, presuntamente por factores meteorológicos.

En contraste, durante la reactivación económica en 2024 se alcanzaron niveles críticos de contaminación: el PM2.5 superó con mayor frecuencia umbrales de 50 $\mu g/m^3$, mientras que NOx y ozono presentaron episodios extremos. Finalmente, en 2025 se observó una mejora parcial, pero sin regresar a los niveles mínimos registrados durante el confinamiento.

Conclusión

Este proyecto representó una oportunidad real para unir la teoría y la práctica. Aprendí que los métodos estadísticos, más allá de generar números, permiten construir evidencia sólida para interpretar situaciones reales.

Particularmente estoy muy orgulloso de los procesos de limpieza de los datos y creo que más allá del impacto científico que tuvieron para el proyecto, creo que son de suma importancia para que este tipo de experimentos sean reproducibles.