Etapas 1. Conociendo del negocio

Integrantes

Ana Lidia Hernández Diaz — A00838643
Ana Paula García Valverde — A01174572
René Cumplido Feregrino — A01735183
Jorge Eduardo González Cantú — A00838643

Aplicación de métodos multivariados en ciencia de datos
MA2003B G 301

Blanca Rosa Ruiz Hernández
Raúl Gómez Muñoz

20 de noviembre de 2025

1. Introducción

La calidad del aire es importante para la ciudadanía porque esta puede presentar riesgos para la salud cuando esta tiene concentraciones altas de contaminantes. En estos casos la ciudadanía tendría que suspender las actividades al aire libre. En algunos casos más extremos se puede presentar una contingencia y se suspenden actividades académicas y laborales.

La medición de la calidad del aire es un pilar fundamental para la salud pública, la cual se determina cuantificando la concentración de contaminantes criterio (como PM2.5, PM10, ozono, etc.) a través de redes de monitoreo; estas mediciones se traducen luego en un índice (como el IMECA o AQI) que comunica el nivel de riesgo al público.

El análisis de la calidad del aire evalúa la concentración de contaminantes (como CO, NO2, PM2.5), sus fuentes de emisión (ej. vehículos) y factores meteorológicos y topográficos (ej. inversiones térmicas). Internacionalmente, se usa el Índice de Calidad del Aire (AQI), una escala de colores (Verde a Marrón) que estandariza el riesgo. El contaminante más dañino es el PM2.5, pues por su tamaño microscópico penetra el torrente sanguíneo, causando daños sistémicos. El CO es peligroso porque desplaza al oxígeno en la sangre (asfixiante), y el NO2 es un gas irritante que inflama las vías respiratorias y agrava el asma.

Los aspectos que se consideran cuando se analiza la calidad del aire son los siguientes:

- Presencia de gases contaminantes o partículas en la atmósfera
- Nivel de contaminantes que no se deben superar
- Zonas de evaluación
- Medidas de mejora es caso en el que se superen los valores máximos
- Información para la población en caso de superar los límites máximos

En el caso específico del Área Metropolitana de Monterrey, esta tarea es realizada por el SIMA (Sistema Integral de Monitoreo Ambiental), la red operada por el gobierno estatal encargada de recolectar, validar y difundir estos datos. Sin embargo, este proceso enfrenta dificultades significativas, como el alto costo de mantenimiento y calibración de los equipos, y la necesidad de una cobertura geográfica amplia para capturar la variabilidad espacial del aire.

A nivel regulatorio, la protección de la salud se basa en las Normas Oficiales Mexicanas (NOMs), que establecen los límites máximos permisibles de contaminantes. La problemática central aquí es que estas NOMs son considerablemente más laxas que las directrices de la Organización Mundial de la Salud (OMS), creando una falsa percepción de seguridad en días que, bajo estándares internacionales, serían considerados dañinos.

Por lo tanto, la mejora de la calidad del aire es un desafío complejo que depende de la intervención sobre variables clave, como la reducción de emisiones vehiculares e industriales, y también de factores meteorológicos no controlables, como el viento y la lluvia, que ayudan a la dispersión.

Específicamente, la Zona Metropolitana de Monterrey (ZMM) enfrenta un desafío crítico y persistente de contaminación atmosférica, donde las emisiones de fuentes móviles son señaladas como uno de los principales contribuyentes. Este trabajo se centra en los contaminantes criterio CO, NO2 y PM2.5, directamente ligados a la combustión vehicular.

El objetivo fundamental del proyecto es trascender la presunción y establecer, mediante un análisis estadístico robusto, la existencia de una relación directa y significativa entre el flujo vehicular y las concentraciones de estos contaminantes, probando que dicha correlación es representativa del comportamiento de la calidad del aire en la metrópoli.

Actualmente, es un área de investigación que se ha abordado en varias ocasiones con distintos acercamientos. En 2019 se realizó un análisis de la contaminación por PM2.5 (Partículas finas primarias y secundarias) en la ciudad de Monterrey, Nuevo León (Centro Mario Molina, 2019). El propósito de este análisis consistía en determinar posibles causas de estas emisiones y futuras formas de prevenirlas; sin embargo, la metodología consistió simplemente en observar el comportamiento de los datos y tratar de establecer algún patrón que los explicara, como el tráfico vehicular, por lo que la investigación no resultó concluyente. Además, no se realizó ningún tipo de análisis estadístico, lo cual es una importante limitación.

Además de esto, la relevancia de este estudio se fundamenta en el severo impacto a la salud pública que estos contaminantes conllevan, como lo documentan la COFEPRIS (s.f.-a, s.f.-b) y estudios epidemiológicos (Lee et al., 2017). Si bien análisis locales previos, como el del IMA-CCA (2019) sobre PM2.5, han diagnosticado la gravedad del problema en la ZMM, y la base de datos MWTP (Aguilar-Díaz et al., 2017) ha hecho disponibles los datos integrados de tráfico, clima y polución, aún existe una brecha en la cuantificación concluyente de esta relación.

La literatura internacional ofrece metodologías probadas para cerrar esta brecha. Investigaciones en otros contextos urbanos (Oh et al., 2016; Su et al., 2007) han demostrado la eficacia de modelos de regresión y análisis geoespaciales que incorporan variables meteorológicas (como el viento) para predecir con precisión el impacto del tráfico. Este proyecto se justifica al aplicar, por primera vez de forma integrada para estos tres contaminantes, dichos modelos estadísticos avanzados a los datos específicos de Monterrey, pasando de la descripción del problema a la prueba de su causalidad vehicular.

Por todo lo anterior, el objetivo del presente trabajo recae en la necesidad de establecer una relación entre las emisiones de gases por combustión automotriz en la zona metropolitana de Nuevo León (Monterrey) y la presencia de los contaminantes CO, NO2 y PM2.5 en el aire. Se desea probar, más allá de cualquier duda razonable, que dicha relación existe y es representativa del comportamiento de los datos en la vida real.

Para lograr el objetivo general anteriormente mencionado, se plantean tres objetivos específicos: 1) Cuantificar el grado de correlación estadística entre el aforo vehicular y los niveles de CO, NO2 y PM2.5. 2) Desarrollar un modelo de regresión múltiple que determine el impacto predictivo del tráfico sobre los contaminantes, aislando el efecto de covariables meteorológicas. 3) Mapear la fuerza de esta relación en la ZMM para identificar las zonas de mayor impacto. La consecución de estos objetivos aportará la evidencia científica ("más allá de duda razonable") necesaria para sustentar futuras políticas públicas de movilidad y gestión de la calidad del aire.

2. Preparación de datos

Para realizar el EDA, se utilizaron todas las bases de datos, excepto la del año 2020. De estas bases, se extrajeron las columnas relativas a condiciones climatológicas y las correspondientes a los siguientes contaminantes: CO, NO2 y PM2.5. Las estaciones seleccionadas fueron únicamente: CE, NTE y SE.

Estos fueron los datos elegidos debido a que son los más relevantes para la investigación. En el caso de las estaciones, son las que están localizadas en las zonas que registran mayor tráfico en horas pico de Monterrey (Tomtom, 2025).

Figura 1: Nivel de congestión promedio en Monterrey durante todos los días. Fuente: TomTom 2025.

En el caso de los contaminantes, se tomó la decisión de enfocar la investigación en las emisiones de CO, NO2 y PM2.5 porque son los principales componentes de la combustión de gasolina de automóviles, según diversos estudios realizados. Aunado a esto, se determinó que para realizar una investigación más profunda, es importante considerar diversos factores como la precipitación, dirección y velocidad del viento, pues todos ello influyen en la manera en que se esparcen los contaminantes a través del aire. De ser posible, se planea incluirlos en el modelo de predicción, a fin de realizar un modelo mucho más robusto y representativo del comportamiento real de los datos.

En cuanto a la limpieza de datos, se siguieron los siguientes pasos:

2.1. 1. Procesamiento de Archivos (2020 y 2022)

Se define una función llamada `limpieza1` que hace lo siguiente:

- Lee los archivos Excel especificados (los de 2020 y 2022).
- Para cada archivo, lee las hojas de cálculo específicas ('NORTE', 'SUR', 'CENTRO').
- Añade una nueva columna llamada 'Zona' a los datos de cada hoja, usando el nombre de la hoja como valor (p.ej., todos los datos de la hoja 'NORTE' obtienen 'NORTE' en la columna 'Zona').
- Combina (concatena) los datos de las tres hojas (Norte, Sur, Centro) de un mismo archivo en un solo DataFrame.
- El resultado es un DataFrame limpio para 2020 (`datos_limpios_20`) y otro para 2022 (`datos_limpios_22`).

2.2. 2. Procesamiento de Archivos (2023, 2024 y 2025)

Se realizó un procesamiento similar para los archivos restantes (2023, 2024 y 2025). El archivo de 2023 (`..._ITESM.xlsx`) tenía una estructura diferente, por lo que requirió pasos especiales para filtrar y reestructurar los datos para que coincidan con el formato de los demás. Los archivos de 2024 y 2025 se procesan con una función similar, llamada `limpieza2()`.

2.3. 3. Consolidación Final y Conversión de Tipos

Todos los DataFrames limpios (2020, 2022, 2023, 2024, 2025) se agruparon en una sola lista, convirtieron todas las columnas de datos (excepto date y Zona) a tipo numérico. Si un valor no se puede convertir (p.ej., un texto), se transforma en NaN. Finalmente, se concatenaron todos los DataFrames individuales en uno solo (`data_combinada_final`), ordenados por Zona y por Date.

2.4. 4. Verificación de Duplicados

Como último paso, se comprobó la integridad del dataset final. Se buscaron duplicados en filas que tengan la misma fecha y la misma zona. Se encontraron 31,685 de estos duplicados, lo que indica que puede haber registros superpuestos o múltiples mediciones para la misma hora y zona en los archivos fuente.

En cuanto a las filas completamente idénticas, se buscaron filas donde todos los valores (contaminantes, temperatura, zona, fecha, etc.) fueran exactamente iguales. El resultado es 0, lo que significa que no hay duplicados perfectos.

Figura 2: Evolución Diaria de CO por Zona (Mayo 2025)

Figura 3: Ciclo Diurno Promedio de CO por Zona (Mayo 2025)

3. Imputación de datos

El método que se usó para imputar los datos fue `interpolate`. Las ventajas de este método es que es bueno para series de tiempo, como los datos tienen una estructura secuencial clara (aunque con ruido por tener varias variables involucradas), se desempeña bien por este lado, sin embargo, este método no toma en cuenta otras variables, por lo que para métodos multivariados no es tan usado, quisimos usar este método inicialmente por su simpleza, sin embargo se planean probar con otros métodos como son:

- Multivariate Imputation by Chained Equations(MICE)
- Modelos autoregresivos (ARIMA, VAR)
- Redes neuronales (RNN, LSTM)

4. Base de datos resultantes

El dataset final de 176,183 filas.

Las mediciones de PM2.5 y WDR (Dirección del Viento) presentan la mayor cantidad de datos faltantes (25.75 % y 28.47 %, respectivamente). Esto deberá tenerse en cuenta durante el análisis, ya sea mediante técnicas de imputación o eliminando filas con datos nulos, dependiendo de la estrategia del modelo. Los contaminantes gaseosos (CO, NO, NO2, O3, SO2) y variables meteorológicas (RH, NOX) tienen un porcentaje de datos faltantes similar, entre 15.9 % y 18.9 %.

5. Link del repositorio

`https://github.com/anahedi/RETO-Multivariados`