Instituto Tecnológico y de Estudios Superiores de Monterrey

Aplicación de métodos multivariados en ciencia de datos

MA2003B, Grupo 101

Docentes:
Blanca Rosa Ruiz Hernandez
Monica Guadalupe Elizondo Amaya

Conocimiento de la naturaleza de contaminantes que influyen en la calidad del aire y sus interrelaciones con el medio

Equipo 5:

Daniela Márquez Campos	A00833345
David Fernando Armendáriz Torres	A01570813
Julio Eugenio Guevara Galván	A01704733
Karla Andrea Palma Villanueva	A01754270

9 de septiembre de 2023

Resumen

Actualmente vivimos en un mundo modernizado, en el cual la industria y la urbanización han prosperado, dicho progreso ha desencadenado un impacto perjudicial en la pureza del aire. Una de las principales consecuencias a raíz de dicha problemática, son las repercusiones en la salud, puesto que la esperanza de vida de las personas disminuye rotundamente.

Solucionar la problemática de una baja calidad del aire en la atmósfera es de suma importancia, por ende la organización SIMA tiene como función principal el tener un control y análisis de los contaminantes detectados para implementar medidas preventivas o correctivas.

La institución registra en una base de datos de 15 estaciones, algunos de los contaminantes y factores meteorológicos presentes por zona, fecha y hora, en donde se anotan a escala, basadas en las normas oficiales mexicanas para la calidad del aire. Al recabar el dataset se realizó una exploración de datos, en la cual se pudo identificar la existencia de valores erróneos, atípicos y faltantes.

Para realizar el análisis estadístico deseado se tenía que hacer una limpieza de datos, para la cual se optó por eliminar los datos con registros contrarios a su naturaleza, además de transformar los datos categóricos a cuantitativos y eliminar los datos nulos.

Es importante mencionar que los objetivos a lograr son descubrir si existe una relación entre los contaminantes y las condiciones meteorológicas, así como entre las propios contaminantes y si existe un patrón con respecto al tiempo, por lo cual, primeramente se desarrolló un análisis de correlación en donde se valoraba dicha relación entre las variables. Posteriormente, se hizo una valoración de la presencia de todos los contaminantes por hora y por día, a lo cual, se decidió seleccionar los contaminantes con mayor variación, siendo PM10, PM2.5 y O3.

Al ya contar con la información necesaria, se llevó a cabo la selección de variables para la implementación de dos modelos: Análisis Discriminante y Regresión Logística, aquí se utilizaron como variables predictoras las referentes a las condiciones meteorológicas con la finalidad clasificar los días dentro y fuera de la norma con respecto al contaminante en cuestión.

Ahora bien, el contaminante a predecir se tuvo que transformar a una variable binaria, de tal forma que fuera más fácil la clasificación; dicho proceso se basó en la normatividad que tienen registrada. El primer análisis implementado fue de manera general, es decir, abarcando todas las estaciones; esto no resultó no muy efectivo, puesto que la sensibilidad de ambos modelos era bajo. Por consiguiente, se optó por realizarse individualmente por estación, a lo que se detectó que el modelo cambiaba, debido a que existe mucha variabilidad en las zonas.

Finalmente, se llegó a la conclusión de que el modelo de regresión logística por estación posee un mejor desempeño en comparación de su contraparte, el modelo de análisis discriminante, con base en la tasa de error y la sensibilidad. No obstante, se fomenta la implementación de medidas correctivas ante los supuestos del modelo para generar mejores resultados. Asimismo, se encontraron interrelaciones entre la concentración de $PM10$ y los factores meteorológicos, destacando que las variables con mayor influencia fueron

precipitación ($RAINF$), presión atmosférica (PRS) y radiación solar (SR). Es importante mencionar que el análisis global y específico de las estaciones aportó diferentes perspectivas y hallazgos debido a la variación de sus registros. Finalmente, también fue posible reconocer diversos patrones y tendencias en el comportamiento de la concentración de los contaminantes de interés con respecto al tiempo.

Conociendo el negocio

1. SIMA y la dirección de gestión de calidad del aire

El Sistema Integral de Monitoreo Ambiental se creó con la finalidad de hacer una evaluación de los niveles de contaminación ambiental en el Área Metropolitana de Monterrey, a través de la obtención de información continua de los diversos impactos que suceden en el aire, procesamiento de datos y extracción de reportes, para lograr así incrementar la resiliencia al cambio climático.

2. Medición de la calidad del aire:

Para realizar el cálculo de la calidad del aire existe el ICA (Índice de la Calidad del Aire), cuyos valores se miden en escala de 0 y mayor a 500, estableciendo así seis categorías de peligrosidad que señala la gravedad de la calidad del aire. Para valorar las categorías se establecen los siguientes factores:

- Buena: Color verde (ICA de 0 a 50)
- Moderada: Color amarillo (ICA de 51 a 100)
- Dañina a la salud para grupos sensibles: Color naranja (ICA de 101 a 150)
- Dañina a la salud: Color rojo (ICA 151 a 200)
- Muy dañina a la salud: Color morado (ICA 201 a 300)
- Peligrosa: Color marrón (ICA superior a 300)

3. Introducción

Como dijo el físico William Thomson Kelvin, “lo que no se puede medir, no se puede mejorar”. Naturalmente, el cuidado de la calidad del aire -y por lo tanto su análisis-, debe ser un objetivo general, dada la magnitud que tienen la salud, ya que de acuerdo con la OMS, los efectos de la contaminación del aire se asocian a 6,7 millones de muertes prematuras anuales. Esto es tan solo un indicador de la influencia que tiene lo que respiramos en nuestra salud. Un factor tan determinante en la calidad de vida como lo es la salud debe ser siempre prioridad.

Ahora bien, para llevar a cabo este paso preliminar e indispensable -la medición- se consideran criterios específicos para el análisis de la calidad del aire, que incluyen las partes por billón de elementos y compuestos químicos presentes en el aire, como el ozono, dióxido

de azufre, dióxido de nitrógeno, monóxido de carbono, monóxido de nitrógeno y óxidos de nitrógeno, además de la cantidad de material particulado menor a 2.5 y 10 micrómetros por metro cúbico, y parámetros meteorológicos, como temperatura, humedad relativa, radiación solar, precipitación, presión atmosférica, dirección y velocidad del viento. Naturalmente, las variables que intervendrían en una mejora de la calidad del aire se constituyen por el nivel de presencia de los contaminantes, por lo que de igual manera, las fuentes de liberación de estos materiales al aire son la influencia que determina la calidad del aire.

Como el complejo proceso que es, la medición de la calidad del aire enfrenta una diversidad de dificultades, mismas que recaen principalmente en la cantidad de variables (sucesos y circunstancias) con la capacidad de afectar la recolección de información, sesgar los datos y en general obstaculizar el proceso de captura es más que considerable. Adicionalmente, eso provoca que se necesite un entendimiento muy completo de los datos para interpretar correctamente las fluctuaciones que se dan por estas circunstancias inesperadas, identificar las fuentes de estas y hacer el análisis correspondiente.

En la ciudad de Monterrey, esta ardua labor la realiza SIMA, el Sistema Integral de Monitoreo Ambiental. Esta entidad empresarial tiene como principal misión la reducción de la contaminación atmosférica en la región de Monterrey. Para lograr este propósito se toman mediciones y recopila información de la calidad del aire de tal forma que se analicen las alertas detectadas para así regular a las empresas y difundir a la población la condición ambiental en las que se encuentran.

Esta empresa se apoya en un sistema creado por la NASA que contribuye con el pronóstico de la calidad del aire. Además dispone de 15 estaciones de monitoreo ubicadas de manera estratégica, siendo respaldadas por centros de control y cámaras instaladas para realizar un análisis de la situación ambiental con mayor precisión. En su base de datos brindan información de los diversos contaminantes con mayor impacto, reflejando también la contaminación que existe en cada zona del área de Monterrey. La información extraída se almacena en una base de datos en SQL para posteriormente pre validarla; un factor importante en los datos es el tiempo, indicando su mayor grado de importancia a los de 3 horas más recientes.

En su funcionamiento, SIMA se rige por las Normas Oficiales Mexicanas para la Calidad del Aire, en donde se establecen los niveles permitidos de concentración y los lapsos de exposición seguros para los contaminantes. En situaciones donde llegara a existir una alerta en que los contaminantes afecten de forma significativa a la salud, se debe entonces tomar una regulación de fuentes de contaminación del aire, en especial las relacionadas con la

industria, puesto que es una de las principales que afectan lo que respiramos. Por tal motivo, cuentan con 24 horas para controlar sus niveles.

De acuerdo con el Gobierno de México, las clasificaciones de los contaminantes del aire ambiente se dan bajo tres categorías: criterio, tóxicos y biológicos. Los primeros son normados, es decir, se les han establecido un límite máximo permisible de concentración en el aire ambiente (ozono, dióxido de azufre, monóxido de carbono, dióxido de nitrógeno, las partículas en suspensión y plomo). Los tóxicos son compuestos en forma de gas o partículas que se encuentran en el aire en concentraciones bajas pero con características de toxicidad o persistencia, son altamente nocivos a la salud humana. Ejemplos: benceno, tolueno, xileno (estos forman parte de los compuestos orgánicos), amoníaco, cloro (no orgánicos), plomo, cromo y cadmio. Y por último, los biológicos son emitidos a partir de material vivo o en descomposición, por ejemplo, moho, esporas, partes de insectos, restos de piel humana o animal y plagas.

Es importante mencionar que los contaminantes específicos más dañinos o preocupantes para la salud pública según la Organización Mundial de la Salud (*WHO* por sus siglas en inglés) en 2019 son las partículas en suspensión, el ozono (O_3), el monóxido de carbono (CO), el dióxido de azufre (SO_2) y el dióxido de nitrógeno (NO_2), puesto que la contaminación del aire, tanto en interior como en exterior, es la causa de diversas enfermedades respiratorias y de otras índoles, identificadas como uno de los principales motivos de morbimortalidad.

Existen regulaciones para cada uno de los contaminantes mencionados previamente, en las cuales se establece el límite máximo permisible en las unidades correspondientes y distintos tiempos de exposición pertinentes para cada contaminante en particular. La vasta mayoría de estas normas se han modificado recientemente, en 2021. Se planea que se sigan modificando con reducciones paulatinas de los límites con el objetivo de reducir la presencia de estos contaminantes.

4. Descripción del problema

El problema específico sobre el cual se redacta en este texto comienza con el aumento de la cantidad de personas que viven en zonas urbanas, lo cual ha llevado a la conglomeración de actividades económicas y de producción en distintos sectores. En condiciones específicas, esto desemboca en empeorar ciertas situaciones, como los problemas vinculados con la baja calidad del aire (Compendio de Estadísticas Ambientales y de los

indicadores del Conjunto Básico, Clave y de Crecimiento Verde, 2018). A lo largo de los años la problemática relativa a la calidad del aire ha tomado relevancia y sigue latente hasta el día de hoy (Organización Panamericana de la Salud, s. f.). Aunado a ello, se plantea estudiar y analizar esta situación, para así dar respuesta a las siguientes interrogantes con el objetivo de aportar en su estudio estadístico y mejoría general:

1. ¿Qué tipo de relación existe entre los contaminantes presentes en el aire y las condiciones meteorológicas de cada región?
2. ¿Qué tipo de interrelaciones existen entre los contaminantes del aire recopilados en las mediciones?
3. ¿Existe algún patrón en las mediciones de los contaminantes con respecto al paso del tiempo?

4.1. Objetivos

Basados en la resolución de las preguntas de investigación, se tiene como objetivo una serie de puntos que trabajan en función de facilitar información útil para la toma de decisiones en aras de un avance en el cuidado del aire (específicamente indicando un direccionamiento adecuado de esfuerzos), principalmente a través de la política pública referente al cuidado de la calidad del aire pero no limitado a la misma. Los puntos son los siguientes:

- Modelar las interrelaciones de los contaminantes que influyen en el aire, a través de análisis de correlación, análisis de factorial e incluso, modelos de predicción.
- Analizar los factores ambientales que tienen influencia en la concentración de los contaminantes, para así detectar condiciones más desfavorables.
- Analizar y modelar las condiciones que están fuera y dentro de la normas oficiales establecidas como normas de calidad de salud.

4.2. Justificación de los objetivos

Anteriormente, gracias a la revisión de bibliografía relacionada al tema, fue posible reafirmar que la calidad del aire repercute directamente en la calidad y esperanza de vida de la comunidad, por lo que analizar el cómo se ve afectada por los diversos factores presentes en el ambiente es un tema de relevancia actual.

5. Descripción de las fuentes de información

Para el desarrollo de este proyecto se dispone de una base de datos brindada por SIMA, en la cual se recopila información de los diversos agentes contaminantes presentes en cada zona y subzona de la región de Monterrey. La disposición de dicha información se encuentra ordenada por hora iniciando desde el 1 de enero de 2022 y abarcando hasta el 17 de agosto del 2023, aquí se encuentran registrados los contaminantes expresados en niveles de presencia en la atmósfera.

6. Impacto social principal

Finalmente, hay que recalcar que la calidad del aire es un factor de enorme relevancia en la salud de la población, en vista de la severidad e incidencia de los perjuicios de salud ocasionados por esta problemática. En la proporción en la que se tenga un mayor entendimiento de las relaciones entre los elementos que componen la situación y las mismas consecuencias, será posible tomar las medidas necesarias para beneficiar a la sociedad en materia de salud y por lo tanto, calidad de vida.

Comprensión y preparación de los datos

1. Comprensión de los datos del negocio.

1.1. Dimensión del dataset: El dataset contiene 17 columnas, es decir variables, y un total de 191550 registros. Este se generó al inicialmente concatenar todos los registros de las diversas estaciones en una sola base de datos.

1.2. Descripción de las variables:

Nombre	Descripción	Tipo	Valores posibles	Valores nulos
date	Fecha y hora del registro	Categórica	14255	0
CO	Contaminante: Monóxido de Carbono *(ppm)*	Numérica	$\mathbb{R}$	7046
NO	Contaminante: Monóxido de Nitrógeno *(ppb)*	Numérica	$\mathbb{R}^+$	6598
NO2	Contaminante: Dióxido de Nitrógeno *(ppb)*	Numérica	$\mathbb{R}^+$	5179
NOX	Contaminante: Óxidos de Nitrógeno *(ppb)*	Numérica	$\mathbb{R}^+$	5152
O3	Contaminante: Ozono *(ppb)*	Numérica	$\mathbb{R}^+$	6744
PM10	Contaminante: Material Particulado menor a 10 micrómetros *($\mu g/m^3$)*	Numérica	$\mathbb{R}^+$	6157
PM2.5	Contaminante: Material Particulado menor a 2.5 micrómetros *($\mu g/m^3$)*	Numérica	$\mathbb{R}^+$	36785
PRS	Descripción de Parámetros Meteorológicos: Presión Atmosférica *(mm/Hg)*	Numérica	$\mathbb{R}^+$	4232
RAINF	Descripción de Parámetros Meteorológicos: Precipitación *(mm/Hr)*	Numérica	$\mathbb{R}^+$	2901

RH	Descripción de Parámetros Meteorológicos: Humedad Relativa *(%)*	Numérica	$\mathbb{R}$	13257
SO2	Contaminante: Dióxido de Azufre *(ppb)*	Numérica	$\mathbb{R}^+$	13256
SR	Descripción de Parámetros Meteorológicos: Radiación Solar *(kW/m^2)*	Numérica	$\mathbb{R}$	6800
TOUT	Descripción de Parámetros Meteorológicos: Temperatura *(°C)*	Numérica	$\mathbb{R}$	3389
WSR	Descripción de Parámetros Meteorológicos: Velocidad del Viento *(Km/hr)*	Numérica	$\mathbb{R}^+$	6289
WDR	Descripción de Parámetros Meteorológicos: Dirección del Viento *(°)*	Numérica	$\mathbb{R}$	7495
ubi	Ubicación de la estación meteorológica	Categórica	15	0

1.3. Exploración de los Datos:

1.3.1. Medidas Estadísticas:

1. Variables Cuantitativas:

1.3.1.1.1. Medidas de Tendencia Central y Dispersión:

Variable	Media	Mediana	Moda	Min	Max	Var	Std
CO	1.340	1.270	1.41	-0.130	7.460	0.4380453	0.6618499
NO	*11.09*	5.00	2.7	0.30	380.80	349.6533	18.69902
NO2	15.23	11.80	6.2	0.00	167.80	140.3209	11.84571
NOX	26.16	17.60	9.3	0.50	410.30	709.6271	26.63883
O3	26.25	23.00	16	0.70	171.00	318.2035	17.83826
PM10	61.73	52.00	43	2.00	1001.00	1827.069	42.74422
PM2.5	20.83	17.21	12	0.00	406.52	227.1291	15.0708

PRS	*715.7*	722.1	729	655.4	747.6	91.66277	9.574068
RAINF	0.0051	0.0000	0	0.0000	67.8000	0.06414696	0.2532725
RH	55.01	56.00	72	-9999.00	190.00	1591.408	39.89245
SO2	4.824	3.900	3.3	0.500	222.900	18.48793	4.29976
SR	-0.490	0.007	0	-9999.000	7.655	6493.846	80.5844
TOUT	23.42	24.27	24.9	-4.00	112.39	60.93	7.805767
WSR	8.981	8.200	1.4	0.100	8.981	31.45548	5.608518
WDR	128	107	21	-9999	360	9990.828	99.95413

2. **Variables Cualitativas:**

1.3.1.2.1. Moda:

date: 2022-12-01 01:00:00

ubi: SURESTE

1.3.1.2.2. Tablas de Frecuencia:

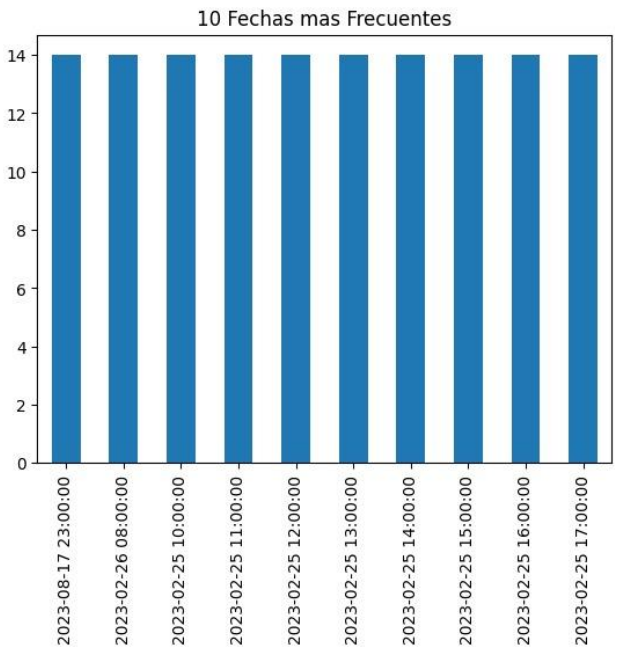

Figura 1: Frecuencias con respecto a las Fechas.

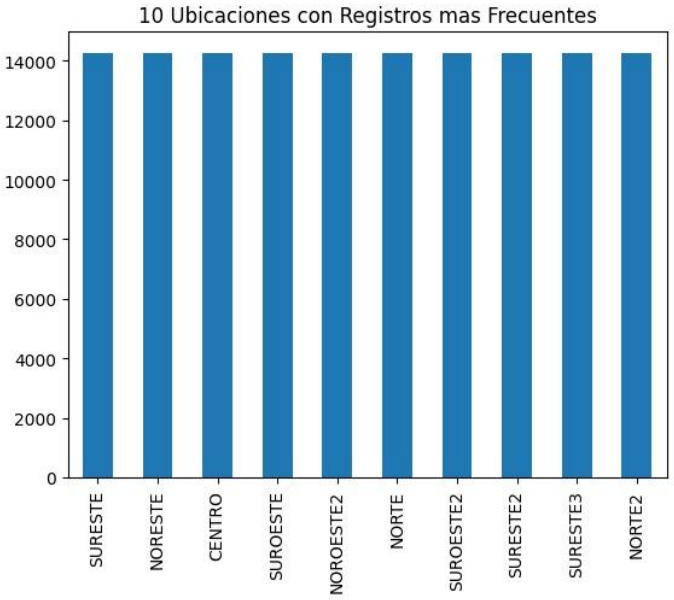

Figura 2: Frecuencias con respecto a las Zonas

1.3.2. Herramientas de Visualización:

1. **Variables Cuantitativas:**

1.3.2.1.1. Boxplots: A continuación se presentan los gráficos *boxplot* realizados de las variables más significativas y difíciles de analizar según SIMA. Los gráficos de las variables restantes se adjuntan en el código realizado.

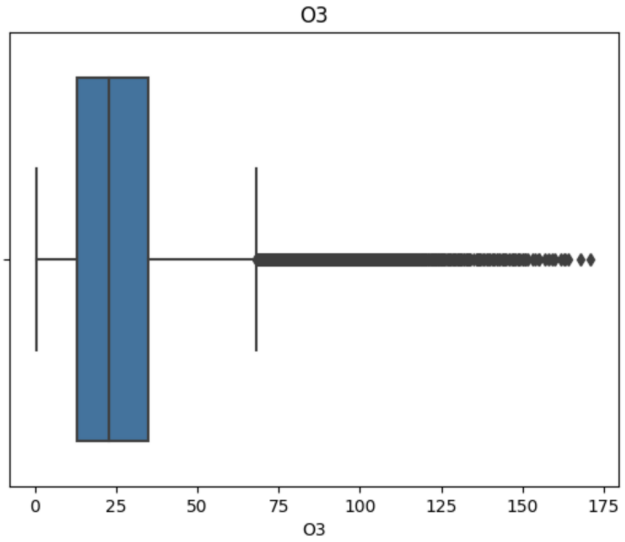

Figura 3: Diagrama de caja y bigotes de la variable Ozono (O_3)

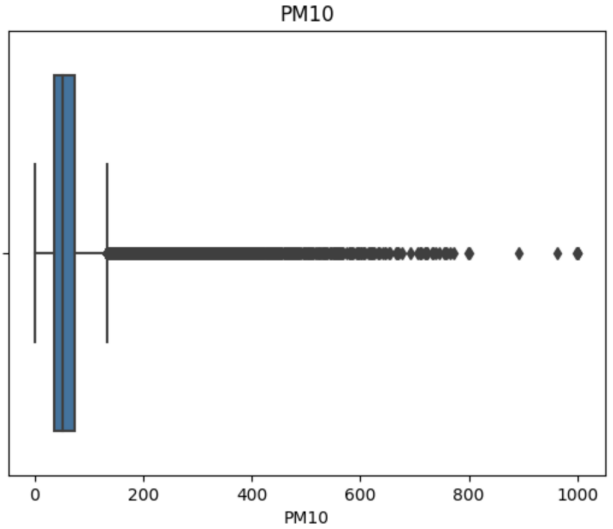

Figura 4: Diagrama de caja y bigotes de la variable Partículas suspendidas de menos de 10 micrómetros (PM_{10})

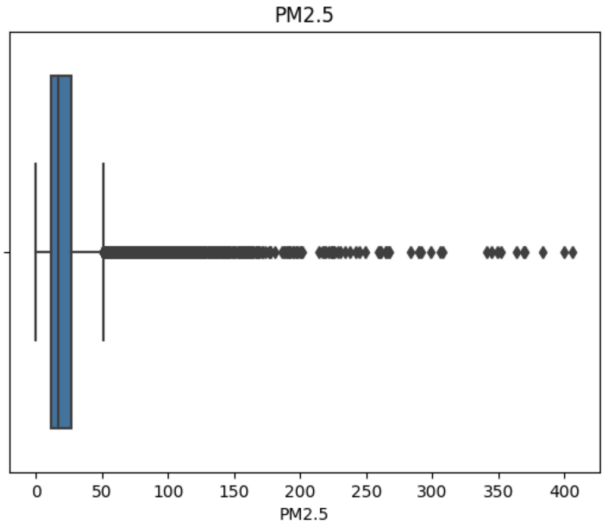

Figura 5: Diagrama de caja y bigotes de la variable Partículas suspendidas de menos de 2.5 micrómetro ($PM_{2.5}$)

1.3.2.1.2. **Histogramas**:

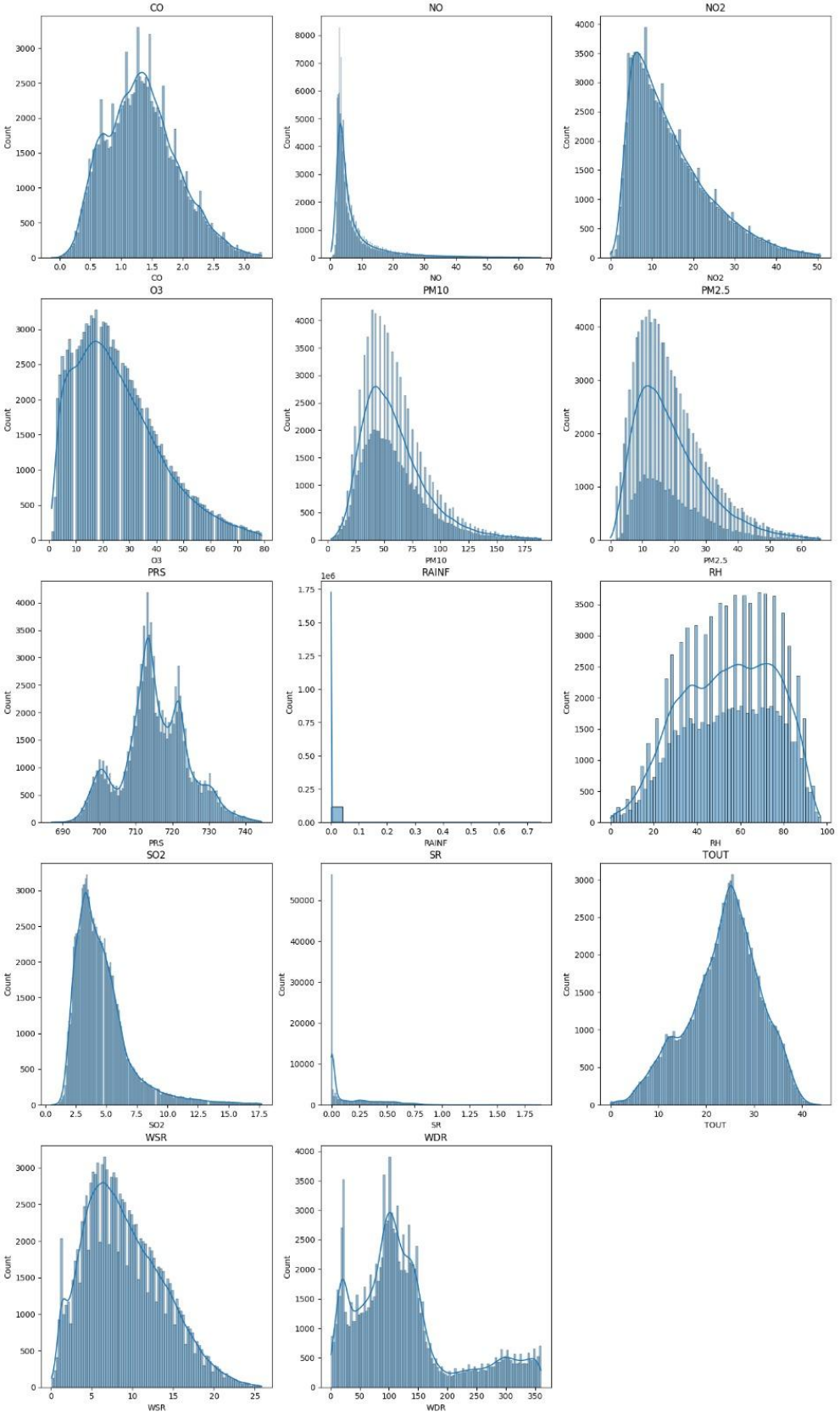
CO
Count
CO
NO
Count
NO
NO2
Count
NO2
O3
Count
O3
PM10
Count
PM10
PM2.5
Count
PM2.5
PRS
Count
PRS
RAINF
1e6
Count
RAINF
RH
Count
RH
SO2
Count
SO2
SR
Count
SR
TOUT
Count
TOUT
WSR
Count
WSR
WDR
Count
WDR

Figura 6. Histogramas

Todas las variables son numéricas, a excepción de dos de ellas, por lo que en su gran mayoría, desde un inicio se puede observar que el histograma no es una herramienta de visualización adecuada para esta base de datos. Ahora bien, dado que las dos excepciones son las variables de fecha y ubicación, tampoco es de utilidad realizar los histogramas, dado que la información que estos ofrecen no son más que parámetros de la operación de la recolección de información que no enriquecen el análisis.

1.3.2.1.3. Mapas de Calor:

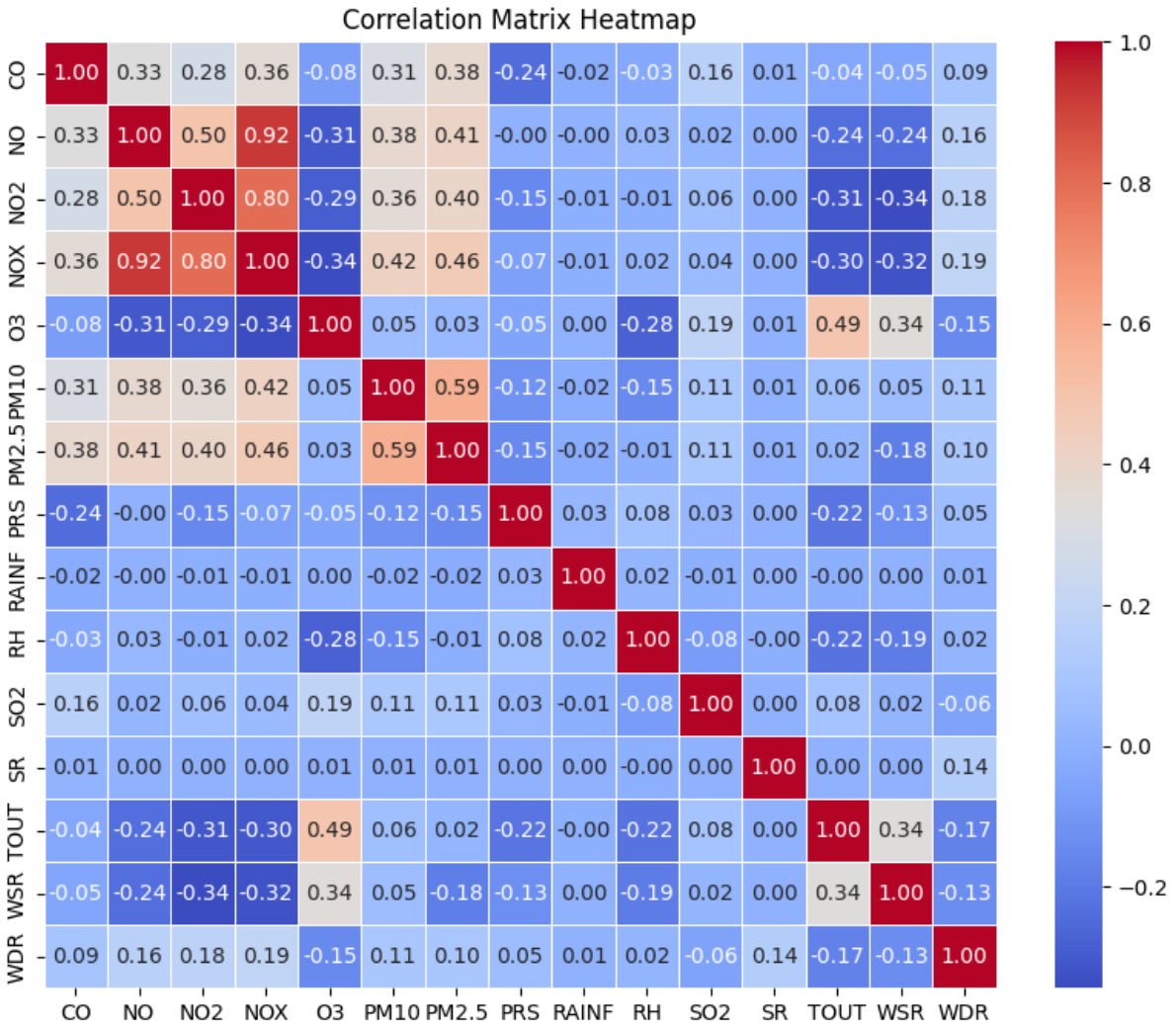

Figura 7. Matriz de correlaciones con indicadores por color

2. Variables categóricas:

1.3.2.2.1. Diagrama de Barras

Se obviaron los diagramas de barras de las únicas dos variables categóricas, fecha y ubicación (estación de registro) dada su naturaleza, ya que está estipulado que se genere una cantidad uniforme de registros para ambas. Es decir, es protocolo generar la misma cantidad de registros cada hora, así como también desde cada ubicación, por lo que la visualización de las mismas no ofrecería mayor información de las variables que la que ya se tiene por plena estructura.

Por el otro lado, se generaron gráficas de barras de las variables numéricas que representan la presencia de contaminantes con dos enfoques distintos. El primero consistió en obtener el promedio de cada variable agrupando por hora del día, así visualizando la fluctuación de cada indicador a lo largo de las 24 horas del día, en promedio. Para el segundo se obtuvo el promedio diario de cada indicador y así se visualizó la fluctuación al pasar de las semanas y meses.

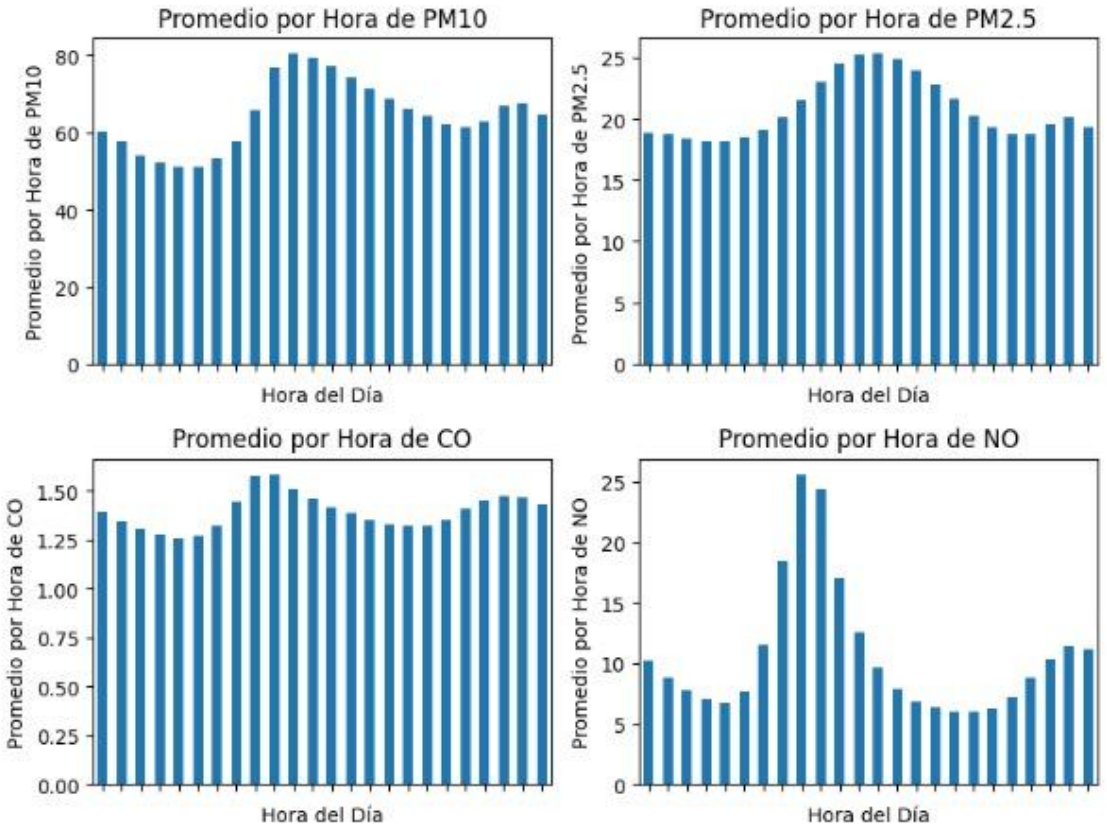

Figura 8. Promedios por hora de los indicadores de contaminantes

Gracias a estos visuales se puede observar que en promedio, al medio día se dieron las mayores concentraciones de partículas de 2.5 y 10 micrómetros.

Además, el óxido de nitrógeno demostró un aumento considerable alrededor de las 9 am.

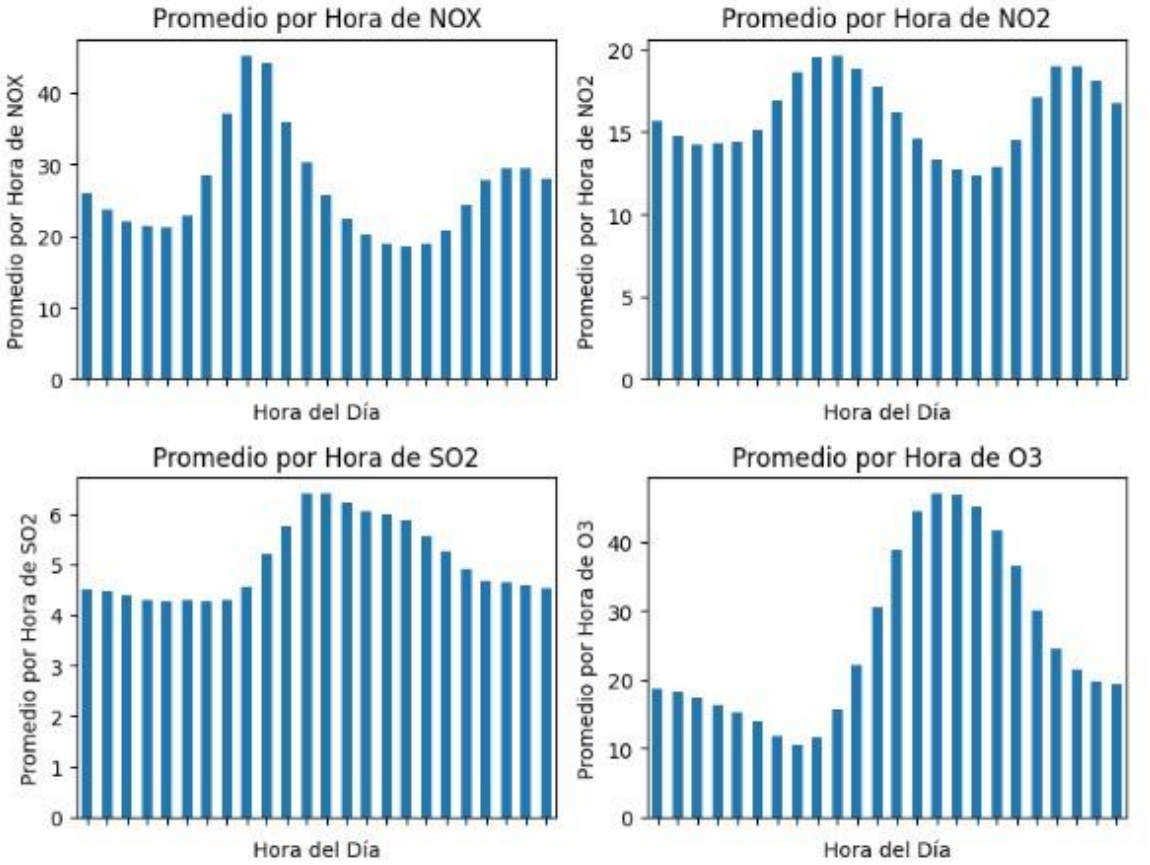

Figura 9. Promedios por hora de los indicadores de contaminantes

Esta figura indicó que los óxidos de nitrógeno también se dispararon alrededor de las 9 am, mientras que el ozono vio un aumento alrededor de las 4 de la tarde.

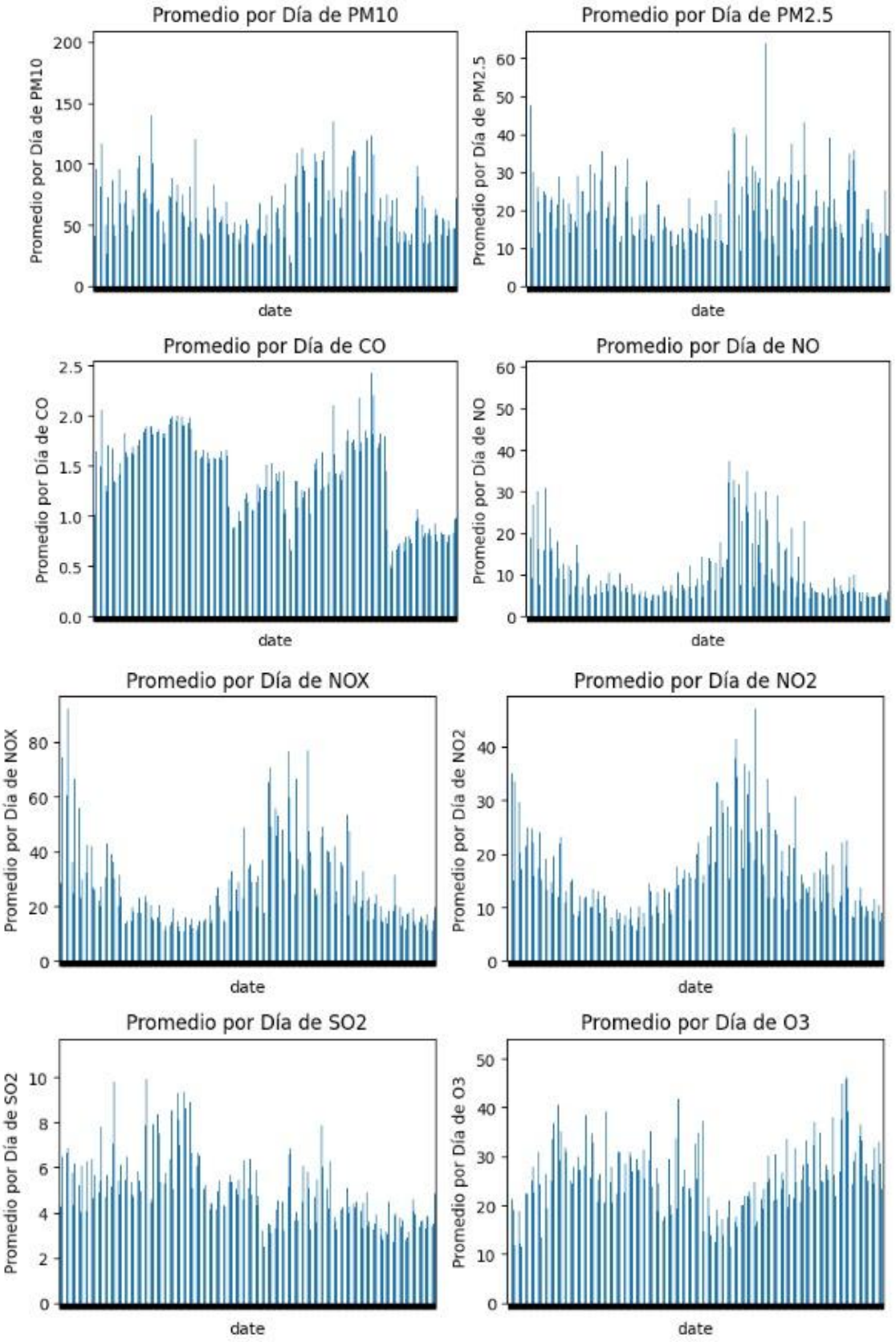

Figura 10. Promedios por día de los indicadores de contaminantes

Con las dos figuras más recientes, se confirmó que no se ha registrado un aumento constante de ninguno de los contaminantes con el paso del tiempo ni existe uno actual en progreso. Si bien existen indicadores que han tenido periodos de tiempo en los que se ha visto un aumento, estos son temporales y han regresado a los niveles previos a su incremento. Finalmente, se interpreta que las concentraciones de CO disminuyeron abruptamente y se mantienen en ese nivel hasta el final de los registros.

1.4. Verificación de la calidad de los datos: valores faltantes, valores de los datos, ortografía: El set de datos, al ser principalmente numérico, carecía de errores ortográficos, no obstante contaban con algunos errores numéricos. Obviando los valores faltantes o nulos, cuyo manejo se tratará más adelante, los datos trabajados gozan en su mayoría de una buena calidad.

2. Preparación de los Datos:

2.1. Conjunto de datos a utilizar: El set de datos a utilizar consiste en lo planteado anteriormente, un dataset que contiene los registros concatenados de las 15 estaciones de monitoreo. Dicho dataset contiene las 17 variables y 191550 registros. Dado el objetivo a trabajar se conservaron del dataset únicamente las variables objetivo y aquellas pertinentes como predictoras, es decir aquellas referentes a los factores ambientales (PRS, RH, $TOUT$, WSR, WDR, $RAINF$, SR).

2.1.1. Identificación de la columna objetivo: Se seleccionaron 3 variables objetivo sobre las cuales se realizarán los mismos análisis de manera individual. Estas variables, dado que son aquellas que SIMA indica generan mayores problemas, son: PM10, PM2.5 y Ozono

2.2. Limpieza de datos:

2.2.1. Valores duplicados: El dataset no cuenta con valores duplicados por lo cual ninguna acción fue necesaria.

2.2.2. Valores erróneos: El dataset cuenta con algunos valores que van en contra de la naturaleza de los mismos, aunado a mayor especificidad, algunas variables contaban con registros negativos de mediciones las cuales deberían poseer datos exclusivamente positivos. Las filas con dicho tipo de registros en las columnas pertinentes (“CO”, “RH”, “SR”, “WDR”) fueron eliminadas.

2.2.3. Manejo de datos categóricos: Los registros de la variable “ubi” se reemplazaron por contrapartes numéricas almacenadas en un diccionario.

A su vez, la columna “date” fue separada en 3 diferentes variables:

1. date: contiene el día del año [1 - 365] del registro.
2. hour: contiene la hora [0 - 23] del registro
3. year: contiene el año del registro [2022, 2023]

2.2.4. Manejo de valores atípicos: Los valores atípicos se identificaron como aquellos que se encontraban a una distancia de más de 3 desviaciones estándar de la media. Dichos valores se reemplazaron por registros nulos para ser eliminados.

2.2.5. Manejo de valores nulos/faltantes: Los registros con valores nulos o faltantes en cualquier columna se eliminaron del dataset.

2.3. Transformación del dataset:

2.3.1. Banderas: Con la intención de realizar un análisis de discriminante en los datos, lo cual se trata más adelante, se generó un nuevo dataset “flags.csv” el cual dadas las normas oficiales mexicanas para la calidad del aire clasifica los registros de $PM10$, $PM2.5$ y $O3$ como 1 en caso de que superen la concentración máxima aceptable y un 0 de lo contrario. Dicho dataset fungirá como las variables objetivo del análisis de discriminante.

1. $mean(PM10_{day,\,year}) \geq 70\ \mu g/m^3 \Rightarrow 1$
2. $mean(PM2.5_{day,\,year}) \geq 41\ \mu g/m^3 \Rightarrow 1$
3. $O3_{hour} \geq 90\ ppb \Rightarrow 1$

Adecuación y/o validación de modelos

3. Análisis de la estructura de los datos

Identificar las estrategias estadísticas que serían de utilidad para realizar un análisis significativo de la estructura de los datos fue todo un proceso. Se llevaron a cabo diversos modelos: inicialmente, se desarrollaron un modelos de regresión lineal múltiple en el cual se designaron como las variables dependientes a los tres contaminantes de interés respectivamente uno en cada modelo y como las variables independientes a los factores meteorológicos, buscando modelar e identificar las interrelaciones entre contaminantes y factores ambientales; no obstante, debido a que las variables no contaban con un comportamiento normal, los supuestos de validación no fueron solventados. Como solución a ello, se realizaron transformaciones de normalidad en las variables a través del método de Yeo Johnson con el objetivo de aproximar lo mayor posible el comportamiento de las variables a una distribución normal. Sin embargo, ni con las transformaciones realizadas en las variables los supuestos de validación en las regresiones fueron cumplidos. Posterior a ello, se planteó la realización de dos métodos: análisis de componentes principales y análisis factorial, con el objeto de, a través de la creación de dichos componentes o factores, determinar relaciones significativas entre contaminantes y factores meteorológicos u obtener componentes/factores que fungieran de variables dependientes o independientes para su posterior utilización en modelos de regresión. Aunque se generaron estos modelos, no se identificaron relaciones significativas por lo que se optó por abordar esta identificación desde otra perspectiva. Finalmente, se decidió implementar modelos de aprendizaje supervisado, en específico, modelos de clasificación que se detallan a continuación:

3.1. Aplicación de Análisis Discriminante:

3.1.1. Modelo Global: Se desarrolló un modelo de análisis discriminante con el propósito de clasificar los días dentro y fuera de la norma, de acuerdo a las concentraciones de contaminantes límite establecidas en los documentos de normatividad brindados por SIMA. Se plantearon tres modelos diferentes situando como la variable *target* cada una de las tres variables de interés previamente establecidas: $PM10$, $PM2.5$ y $O3$, respectivamente. Sin embargo, debido al gran desbalance en la proporción de registros de las variables definidas como *target* ($> 1\%$, $< 99\%$), en en el caso de $O3$ y $PM2.5$, el

modelo planteado no fue lo suficientemente eficiente ni significativo para la investigación en curso.

Por otro lado, un mayor balance en la proporción de los registros de las etiquetas del modelo, en el caso del contaminante $PM10$, este permitió el desarrollo adecuado de los clasificadores discriminantes utilizando todos los factores meteorológicos existentes en la base de datos como variables independientes.

Se generaron los *sets* de entrenamiento y prueba del modelo a raíz de la base de datos de estaciones concatenadas con una proporción de 0.7 y 0.3, respectivamente.

Para el caso de la variable $PM10$, algoritmo generó una función discriminante de la forma:

$$LD1 = -0.0267 * PRS - 0.0478 * RH - 0.0210 * TOUT - 0.0933 * WSR + 0.0049 * WDR - 0.9107 * RAINF - 0.5478 * SR$$

En esta ecuación se observa que la variable independiente que mejor "discrimina" en el modelo es la correspondiente a RAINF, correspondiente a la precipitación, puesto que, aunque de valor negativo, tiene un mayor valor absoluto de peso en la función.

Como se menciona al inicio de esta sección, este modelo es de clasificación, sin embargo, teniendo ya entrenado el modelo de la manera adecuada, puede utilizarse para realizar predicciones con respecto a si el día estará dentro o fuera de la normatividad, ingresando previamente los datos de las variables independientes implicadas en el modelo.

3.1.2. Modelo por Estación: Análogo al modelo global, se realizaron modelos de análisis por discriminante con cada una de las estaciones presentes en el dataset. El proceso, los datos empleados y las proporciones de entrenamiento y prueba se mantienen constantes. A continuación se presentan los coeficientes generados:

Estación	PRS	RH	TOUT	WSR	WDR	RAINF	SR
Sureste	-0.2288	-0.0431	-0.0977	-0.0774	0.0024	**-0.3116**	0.1729
Noreste	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Centro	**-0.2492**	-0.0519	-0.0960	-0.1308	0.0019	—	0.0246
Noroeste	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Suroeste	-0.1510	-0.0491	-0.0604	-0.0877	0.0040	**1.2026**	0.0367
Noroeste2	-0.1914	-0.0347	-0.0630	-0.0459	0.0050	—	**2.7960**
Norte	**-0.1929**	-0.0499	-0.1011	-0.0517	0.0059	—	0.1583
Suroeste2	-0.4835	-0.0496	-0.0845	-0.0620	0.0048	**-0.5540**	0.0698
Sureste2	-0.3142	-0.0466	-0.0566	-0.0691	0.0049	—	**-1.0216**
Sureste3	**-0.2350**	-0.0431	-0.1199	-0.0877	0.0033	—	-0.1582
Sur	-0.2435	-0.0532	-0.1208	-0.1046	0.0012	—	**-0.2449**
Norte2	-0.2251	-0.0486	-0.1063	-0.1355	0.0023	**-0.5491**	-0.0002
Noreste2	-0.2071	-0.0422	-0.1117	-0.0216	0.0033	—	**-0.4766**
Noreste3	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Noroeste3	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Tabla 1. Tabla de los coeficientes generados por estación del Modelo de Análisis Discriminante

Adicionalmente, aquellas estaciones sin registros corresponden a aquellos a las cuales no se les pudo aplicar el modelo dada la ausencia dado que contaban con clases unarias. La ausencia de algún coeficiente en la variable $RAINF$ de algunas estaciones se debe al comportamiento constante y nulo de dicha variable para dichas estaciones que imposibilita la implementación del modelo. Finalmente, de color azul se pueden observar los coeficientes de mayor peso en la función discriminante de cada estación, es decir, la variable con mayor influencia discriminante.

3.2. Modelo de regresión logística:

3.2.1. Modelo Global:

Se implementó un clasificador con regresión logística múltiple donde la variable dependiente y es una variable binaria que representa si el día correspondiente a dicho registro está fuera de la norma con respecto al valor de la concentración de material particulado menor a 10 micrómetros ($PM10$), mientras que las variables independientes o predictoras son los factores meteorológicos correspondientes a cada día: presión atmosférica

(PRS), humedad relativa (RH), temperatura ($TOUT$), velocidad del viento (WSR), dirección del viento (WDR), precipitación ($RAINF$) y radiación solar (SR). Con esto, se planteó predecir, con base en las variables que describen condiciones climatológicas, el exceso de concentración de material particulado por día, lo cual podría definir relaciones de interdependencia entre los factores meteorológicos y el contaminante mencionado, indicando además, posibles escenarios desfavorables conforme a las mediciones climatológicas.

El modelo resultante tiene la siguiente forma:

$$P(Logit(PM10)) = 18.5731 - 0.0231 * PRS - 0.0392 * RH - 0.0142 * TOUT - 0.0815 * WSR + 0.0040 * WDR - 46.5088 * RAINF - 0.4337 * SR$$

En esta ecuación de probabilidad demuestra que la variable independiente con mayor peso, así como en el modelo de Análisis Discriminante, es la correspondiente a RAINF, es decir a la precipitación, puesto que, aunque nuevamente cuenta con un valor negativo, su valor absoluto es mayor que el correspondiente al de cualquiera de las otras variables predictoras.

De la misma manera que en el modelo anterior, este también es un modelo de clasificación, al cual, si se le ingresan previamente los datos de las variables independientes implicadas en el modelo, siendo que éste ya está entrenado, es posible generar una predicción con respecto a si el día estará dentro o fuera de la normatividad con respecto a la concentración de $PM10$.

3.2.2. Modelo por Estación: De igual manera al modelo discriminante, el mismo proceso se lleva a cabo segmentando los datos para generar modelos de regresión logística por estación. Los coeficientes de la función logística se presentan a continuación:

Estación	PRS	RH	TOUT	WSR	WDR	RAINF	SR	β_0
Sureste	-0.2918	-0.0488	-0.1214	-0.0847	0.0025	**-53.7098**	0.3630	216.2778
Noreste	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Centro	**-0.3229**	-0.0665	-0.1169	-0.1726	0.0023	—	0.0720	236.1482
Noroeste	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Suroeste	-0.1918	-0.0613	-0.0656	-0.1158	0.0050	**-4.1896**	0.0951	138.6614
Noroeste2	-0.2502	-0.0460	-0.0803	0.0541	0.0061	—	**4.1895**	177.5734
Norte	**-0.2005**	-0.0487	-0.1007	-0.0472	0.0052	—	0.1326	147.6920
Suroeste2	-0.6292	-0.0610	-0.1025	-0.0601	0.0061	**-38.6461**	0.0535	453.8123
Sureste2	-0.3515	-0.0374	-0.0572	-0.0456	0.0037	—	**-0.4292**	256.2459
Sureste3	-0.2698	-0.0448	-0.1316	-0.0799	0.0034	**-222.4784**	-0.0235	202.5050
Sur	**-0.2877**	-0.0634	-0.1395	-0.1290	0.0011	—	-0.2501	211.6474
Norte2	-0.2693	-0.0530	-0.1174	-0.1380	0.0027	**-80.8594**	-0.1613	197.6771
Noreste2	-0.1954	-0.0387	-0.1007	-0.0167	0.0031	—	**-0.5532**	143.2977
Noreste3	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Noroeste3	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Tabla 2. Tabla de los coeficientes generados por estación del Modelo de Regresión Logística

De la misma manera que en la Tabla 1, se señala que las estaciones plasmadas sin registro son aquellas con las cuales no se pudo implementar el respectivo modelo debido a la falta de registros de clase. Por otro lado, los valores nulos en algunas estaciones respectivos al coeficiente de la variable *RAINF* se deben a la inexistencia de datos en los registros de la variable para dichas estaciones, por lo cual incluirla imposibilita el desarrollo del modelo. Finalmente, en color verde se resaltan los coeficientes de mayor peso en cada función.

4. Validación de la eficiencia de los modelos obtenidos

4.1. Modelo de análisis discriminante PM10:

Estación	~Tasa de Error	~Sensibilidad
Sureste	24.75%	51.02%
Centro*	22.91%	53.48%
Suroeste	21.78%	56.65%
Noroeste2*	22.94%	57.80
Norte*	22.33%	37.68%
Suroeste2	22.69%	56.41

Sureste2*	22.27%	29.24%
Sureste3*	25.45%	47.11%
Sur	23.86%	53.64%
Norte2*	24.09%	50.15%
Noreste2*	45.43%	50.24
Global	27.70%	40.10%

4.2. Modelo de regresión logística PM10:

Estación	~Tasa de Error	~Sensibilidad
Sureste	25.87%	60.73%
Centro*	23.96%	69.54%
Suroeste	22.55%	68.62%
Noroeste2*	22.22%	72.86%
Norte*	27.97%	70.92%
Suroeste2	22.93%	65.18%
Sureste2*	30.68%	68.76%
Sureste3*	28.30%	67.15%
Sur	26.38%	68.58%
Norte2*	27.53%	70.00%
Noreste2*	27.49%	62.96%
Global	31.08%	60.40%

5. Justificación de la elección de los métodos estadísticos empleados

5.1. Regresión Logística Múltiple: Con la intención de evidenciar la relación entre los factores ambientales y la presencia de contaminantes en el aire, específicamente el correspondiente a $PM10$, así como de, adicionalmente, predecir si el día correspondiente a dicho registro está fuera de la norma con respecto al valor de la concentración de material particulado menor a 10 micrómetros $(PM10)$, se optó por

un modelo de predicción. Además, dado el alcance del curso del cual forma parte este reto, se optó por un modelo de regresión logística múltiple.

5.2. Análisis de Discriminante: Dados los objetivos planteados previamente, se ideó el objetivo específico de desarrollar un modelo de aprendizaje supervisado con la capacidad de indicar si, dada información respecto a diversos factores ambientales a cierta hora, las mediciones de los contaminantes objetivo sobrepasarán o no el límite máximo aceptable. Con esto en mente, se optó por un análisis de discriminante ya que se ajusta a la necesidad planteada de clasificación binaria y no cuenta con supuestos lo cual se ajusta a las limitaciones del set de datos utilizado.

6. Validación de los supuestos requeridos para la aplicación de las técnicas estadísticas

6.1. Análisis de Discriminante: Debido a que el análisis de discriminante no cuenta con ningún supuesto estadístico, salvo a la presencia de una variable categórica que funja como etiqueta de clase y respectivas variables predictoras, ningún tipo de validación fue necesario.

6.2. Regresión Logística:

6.2.1. Variable de respuesta binaria: El documento "flags.csv" que cuenta con las etiquetas de clase de las variables objetivo es bivariado.

6.2.2. Independencia de las observaciones: Graficando los residuos del modelo contra el tiempo, no se observa ningún patrón lo que denota independencia.

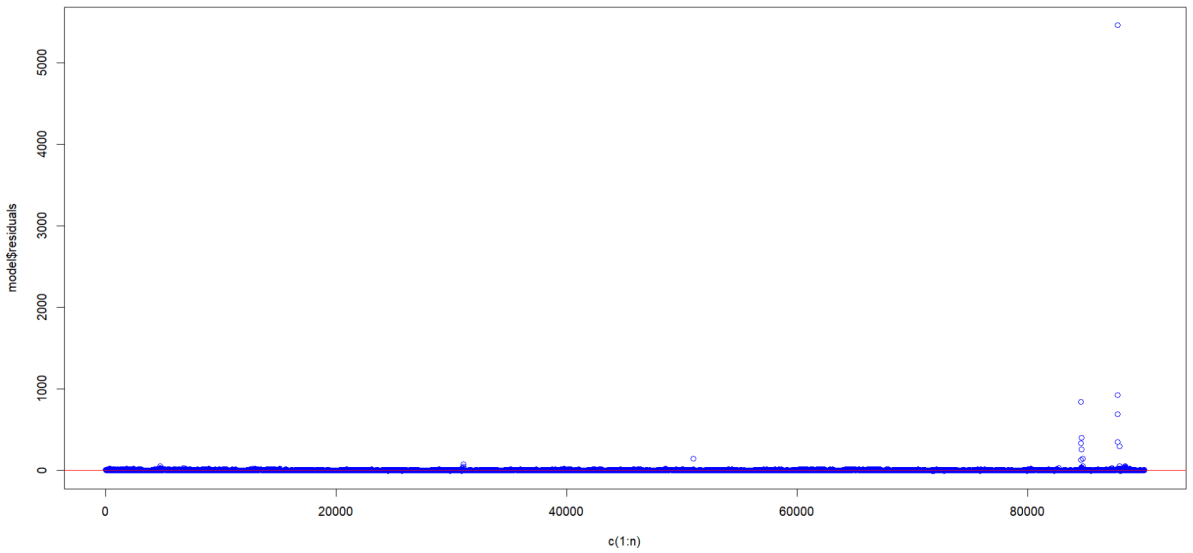

Figura 11. Gráfica de residuos del modelo de regresión logística vs tiempo.

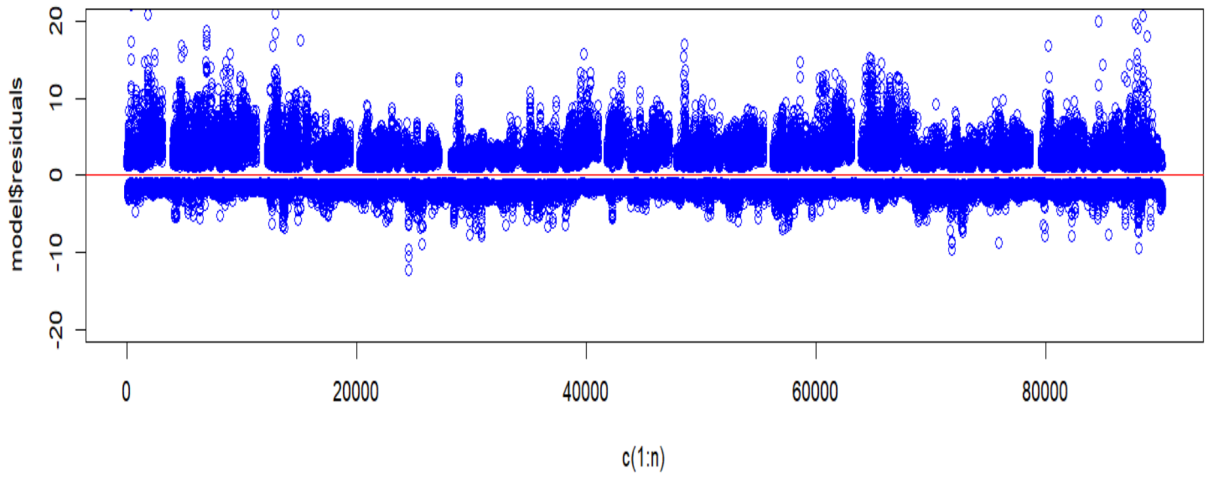

Figura 12. Gráfica de residuos del modelo de regresión logística vs tiempo | rango vertical = [-20,20].

6.2.3. Ausencia de Multicolinealidad: Un factor de inflación de varianza arrojó los siguientes resultados que indican una correlación moderada.

- PRS: 1.061170
- RH: 1.392263
- TOUT: 1.428280
- WSR: 1.349760
- WDR: 1.068301
- RAINF: 1.002533
- SR: 1.197195

6.2.4. Ausencia de Valores atípicos extremos: En la exploración de datos se identificaron valores atípicos en múltiples variables, sin embargo dada la naturaleza de los datos esto era esperado y vital para la representación adecuada de la información, por ende no se eliminaron o modificaron.

6.2.5. Relación lineal entre las variables explicativas y el logit de la variable de respuesta: Para ejemplificar este supuesto se realizó una regresión lineal múltiple entre las variables explicativas y los residuos del logit de la variable respuesta. A través del resumen de resultados de esta regresión se observó que algunos coeficientes, específicamente RH, $TOUT$, WSR y WDR, resultaron ser significativos, lo cual indica que la relación entre dicha variable independiente tiene es de carácter no lineal con el logit de la variable de respuesta.

Discusión y Conclusiones

Analizando los resultados de cada modelo, se puede observar que la variable $RAINF$, es decir la precipitación, es aquella que tiene mayor peso sobre tanto la función discriminante como el logit de los modelos globales. En cuanto a los modelos por estación, dicha variable con mayor influencia variaba entre las variables PRS, $RAINF$ y SR, correspondientes a presión atmosférica, precipitación y radiación solar respectivamente, las cuales, en el modelo de regresión logística, son aquellas que presentan una relación lineal con respecto al logit de la variable dependiente, hecho que es bastante lógico.

Por otro lado, en cuanto a la validación de la eficiencia de los modelos, se optó por utilizar como métricas de evaluación la tasa de clasificación errónea como métrica general y la sensibilidad como métrica específica; esto debido a que es particularmente es de interés reducir los falsos negativos, es decir, los días clasificados como dentro de la norma, cuando en realidad están fuera de ella. Conforme a ello, el modelo global de Análisis Discriminante tiene mejor valor en cuanto a la tasa de error, pero el de Regresión Logística supera el anterior modelo con respecto a la sensibilidad. Además, para analizar las métricas de los modelos por estación, se calculó un promedio de cada una para cada modelo, con respecto a lo cual el modelo de regresión logística genera mejores resultados tanto en la tasa de error como en la sensibilidad.

Asimismo, cabe mencionar que el modelo de Análisis Discriminante no cuenta con supuestos a validar como tal para el desarrollo del mismo, sin embargo, la Regresión Logística sí cuenta múltiples supuestos, los cuales no todos se cumplen, específicamente se cuenta con los supuestos de variable de respuesta bivariada e independencia de las observaciones; la ausencia de multicolinealidad se cumple parcialmente, mientras que la ausencia de valores atípicos y la relación lineal entre las variables explicativas y el logit de la variable de respuesta no se solventan de manera adecuada.

De acuerdo a lo planteado anteriormente, se pueden concluir las siguientes aseveraciones:

- El modelo de Regresión Logística es más eficiente que el de Análisis Discriminante conforme a la problemática establecida y de acuerdo a las métricas de evaluación seleccionadas.
- Ambos modelos de clasificación evidencian y demuestran las interrelaciones entre el contaminante objetivo, en este caso, $PM10$, y los factores meteorológicos a través de los coeficientes de sus ecuaciones. Asimismo, el peso de dichas variables refleja la

influencia de las mismas en la concentración del contaminante, definiendo proporciones en las que las características del ambiente podrían favorecer o viceversa la existencia y prolongación del mismo.

- Analizar de forma global y específica las estaciones a través de los modelos implementados proporcionó diferentes perspectivas, puesto que la variabilidad de los datos e incluso, la existencia de variables específicas, como es el caso de $RAINF$, para algunas estaciones es bastante diferente y aporta hallazgos particulares.
- El análisis exploratorio de la evolución de la concentración de los contaminantes conforme al tiempo permitió la identificación de patrones, así como de horas y estaciones características en las que la concentración de los contaminantes se dispara o disminuye, lo cual inherentemente va de la mano con los factores meteorológicos y es de utilidad para tomar las medidas preventivas necesarias para tratar la problemática abordada.

Adicionalmente se plantean las siguientes recomendaciones:

- La Regresión Logística no cumple con todos los supuestos de validación necesarios para su implementación, por lo que el tomar medidas correctivas para apegarse a los supuestos planteados podría mejorar el desempeño del modelo.
- Parece ser evidente la diferencia entre el comportamiento climatológico de las múltiples estaciones, se recomienda corroborar dicha afirmación con un análisis ANOVA que verifique la significancia de la ubicación por estación con respecto a la concentración de los contaminantes.

CÓDIGO

La etapa 1 y 2, comprensión y preparación de los datos, se realizó en Python. La etapa 3, la modelación y validación, se llevó a cabo en R.

Liga de acceso:

https://drive.google.com/drive/folders/1lMQ756oOcOLN9K3wOzUx9D6tSy_sv7iv?usp=drive_link

Referencias y fuentes bibliográficas

La Protección Contra Riesgos Sanitarios, C. F. P. (s. f.). *Clasificación de los contaminantes del aire ambiente*. gob.mx. https://www.gob.mx/cofepris/acciones-y-programas/2-clasificacion-de-los-contaminantes-del-aire-ambiente

Spurio, S. (s. f.). *Lo que no se mide no puede mejorarse*. https://www.cab.cnea.gov.ar/ieds/index.php/34-institucional/noticias/374-lo-que-no-se-mide-no-puede-mejorarse

Sistema Integral De Monitoreo Ambiental Nuevo León. (s. f.). Normatividad. http://aire.nl.gob.mx/nor_normatividad.html

World Health Organization: WHO. (2022). Contaminación del aire ambiente (exterior). *www.who.int*. https://www.who.int/es/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health#:~:text=Mediante%20la%20disminuci%C3%B3n%20de%20los,agudas%2C%20entre%20ellas%20el%20asma

World Health Organization: WHO. (2019). Contaminación atmosférica. *www.who.int*. https://www.who.int/es/health-topics/air-pollution#tab=tab_1

Compendio de Estadísticas Ambientales y de los indicadores del Conjunto Básico, Clave y de Crecimiento Verde. (2018). *Informe del Medio Ambiente*. Gobierno de México. https://apps1.semarnat.gob.mx:8443/dgeia/informe18/tema/cap5.html

Organización Panamericana de la Salud (OPS). (s. f.). *Calidad del aire*. OPS/OMS | Organización Panamericana de la Salud. https://www.paho.org/es/temas/calidad-aire