Análisis de Métodos Multivariados en Ciencia de Datos

Alteraciones en los contaminantes y factores por la pandemia de COVID-19

Miranda Isabel Rada Chau A01285243

Fedra Fernanda Mandujano López A00836747

Erick Isaac Lascano Otañez A00836571

Kevin Jesús Martínez Trinidad A00834493

17 de noviembre de 2024

1. Resumen

La calidad del aire ha ido empeorando con el paso del tiempo. Si la calidad del aire es baja, tiene consecuencias fuertes en la calidad de vida de las personas. En Monterrey, específicamente, este problema se ha vuelto más relevante en los últimos años. Este proyecto consiste en realizar un análisis estadístico multivariable sobre los niveles de contaminación en diferentes áreas de Monterrey, tomando en cuenta el periodo de la pandemia y los tiempos externos a la misma. Realizamos este análisis con el fin de identificar si la pandemia tuvo un impacto significativo en el nivel de contaminación en el área metropolitana de Monterrey. Conocer estos resultados permitirá analizar si el cambio en actividad humana durante la pandemia mejoró la calidad del aire.

Se utilizó la estadística multivariable y el análisis factorial para analizar si había relaciones entre los contaminantes y diferentes áreas de la zona metropolitana, para después modelar estos contaminantes y su relación con la pandemia. Para poder formular conclusiones apropiadas se utilizó la estadística inferencial y simulaciones de Montecarlo. A partir del análisis se permite identificar si la eliminación de algunas fuentes de contaminación mejora la calidad del aire.

Los resultados muestran que durante la pandemia hubo un cambio en el nivel de la mayoría de los factores de los contaminantes, lo que indica que las restricciones en la actividad humana tuvieron un impacto en la calidad del aire. Sin embargo, es necesario considerar que hubo cambios diferentes en cada estación de la ciudad. Por lo tanto, en una investigación futura sería interesante profundizar más en lo que generó cada cambio en cada estación y si esto fue afectado por la densidad de población del área.

Esta solución impacta a todos, debido a que una mala calidad del aire afecta directamente a nuestra salud, por esto mismo es importante concientizar a la población sobre la importancia de disminuir la emisión de contaminantes.

2. Introducción

2.1. Contexto general

Para la ciudadanía, la calidad del aire es un tema de gran importancia, ya que si se tienen niveles bajos de contaminación en lo que se respira, mejor será la salud cardiovascular y respiratoria de la población en general, tanto a largo como a corto plazo. (Figueroa, s.f.)

La diversidad de fuentes contaminantes hace que mejorar la calidad del aire sea un objetivo con muchas aristas. No obstante, la identificación precisa de las zonas, días y horarios con mayor contaminación es el primer paso para tomar medidas y realizar estrategias para controlar las acciones que causan o agravan las emisiones contaminantes. La medición constante de la calidad del aire no solo permite detectar problemas, sino también evaluar el impacto de las políticas ambientales y garantizar una mejora continua en la calidad del aire que se respira.

Si bien, es cierto que es posible percibir cuando el aire que respiramos está contaminado, ya sea porque se observa smog en el cielo, por un malestar a la hora de respirar o simplemente por la sensación del aire, estas observaciones no son precisas y muchas veces se pueden deber a factores externos. Una manera de medir el aire cuantitativamente es gracias a sensores diseñados para detectar contaminantes específicos, principalmente mediante el uso de láseres o imágenes satelitales. Con estos equipos es posible conocer la concentración de distintos contaminantes en la muestra de aire obtenida, pudiendo así generar una clasificación de la calidad dicho aire. (UNEP, 2022).

Ciertas dificultades pueden surgir a la hora de medir la contaminación del aire. Una de las más comunes es la de obtener muestras muy poco representativas. Esto es, que la evaluación obtenida solo representa el aire que está a poca distancia de la estación, el cual muchas veces puede estar influenciado por factores externos como chimeneas, asadores, humo de cigarros, entre otros. Por ello, se vuelve

esencial seleccionar un lugar adecuado para los equipos de medición (Hugo Barrera, 2021).

Monterrey, Nuevo León cuenta con el Sistema Integral de Monitoreo Ambiental (SIMA), el cual, junto a la Secretaría del Medio Ambiente se encargan de medir, monitorear y transmitir la información de la calidad del aire de manera eficiente y ordenada con el objetivo de crear estrategias y aplicar medidas para que la calidad del aire sea lo más limpio posible. Gracias a esto, actualmente, existen las Normas Oficiales Mexicanas para la Calidad del Aire, que establecen no solo los niveles máximos permitidos en el aire de las sustancias y partículas más peligrosas para la población sino también los métodos y requisitos de monitoreo de medición y el tipo de equipos para medir las concentraciones de dichas sustancias. Las normas más importantes son desplegadas en el Cuadro 11 que se encuentra en el anexo 8.2.

2.2. Delimitación del objeto de estudio

Para mejorar la calidad del aire, los principales factores a considerar son todas las fuentes de los principales contaminantes en el aire (Ver Sección 3). Al controlar o minimizar el uso de estas fuentes, también se estaría reduciendo la cantidad de contaminantes liberados a la atmósfera y, con ello, mejorar la calidad del aire (INECC, s.f.).

3. Marco Teórico

Como sabemos, en el aire no solamente se encuentra oxígeno, sino que también existen en él una gran variedad de compuestos y elementos que respiramos constantemente. De todas estas partículas, existen algunas que son especialmente perjudiciales para nuestra salud.

Según la OMS, 2022, los contaminantes con mayor evidencia de afectación a la salud humana son los siguientes:

- Partículas Menores a 2.5μg (PM2.5) y Partículas Menores a 10μg (PM10): Se refiere a partículas inhalables compuestas principalmente por sulfatos, nitratos, amonios, carbón negro, polvo mineral, entre otros componentes. Estas partículas son capaces de penetrar profundamente en nuestros pulmones y en nuestro torrente sanguíneo, causando afectaciones cardiovasculares, cerebrovasculares y respiratorias, sobre todo las PM2.5 (OMS, 2022).
- Dióxido de Nitrógeno (NO_2): Compuesto resultante de la combustión de combustibles, este es un importante precursor del ozono, relacionado al asma y a otras condiciones respiratorias. (OMS, 2022)
- Ozono (O_3): Emitido principalmente por vehículos y la industria, es uno de los principales compuestos del smog, provocando problemas para respirar, asma y afectaciones en los pulmones. (OMS, 2022)
- Monóxido de Carbono (CO): Emitido principalmente por los motores vehiculares, es un compuesto que viaja a lo largo de los pulmones, dificultando la distribución de oxígeno, lo que a su vez provoca dificultades para respirar, mareos, etc. (OMS, 2022).
- Dióxido de Azufre (SO_2): Producido principalmente durante la combustión de combustibles fósiles. Se asocia principalmente a ataques de asma. (OMS, 2022)

4. Problemática y Objetivos

4.1. Planteamiento del Problema

Como se mencionó con anterioridad, la calidad del aire es un problema que ha ido empeorando con el paso de los años en todo el mundo, aumentando considerablemente los problemas de salud relacionados a la respiración. Por otro lado, las pandemias son eventos con cierto grado de impredecibilidad, de las cuales se tiene relativamente poca información acerca de su efecto en la calidad del aire. En la última pandemia que sufrió el mundo, al ser una de índole

respiratorio, aquellas personas con problemas de este tipo se vieron mayormente afectadas, el mundo tenía tiempo sin ver una pandemia de esta escala y en muchos sentidos nos tomó desprevenidos.

4.2. Justificación

Comprender los efectos directos y subyacentes que tiene una pandemia mundial en la calidad del aire nos permitirá estar mejor preparados en caso de que ocurra otra, pues podremos entonces tener una mejor capacidad de predecir las consecuencias que esta tendrá en la concentración de contaminantes, permitiéndonos crear protocolos para contrarrestar los efectos negativos y/o alentar los positivos, de forma que se puedan salvar vidas si una nueva pandemia llegara.

4.3. Objetivos

El objetivo es realizar un análisis multivariado para determinar si existió un cambio en las variables y factores durante la pandemia de COVID-19 y en momentos previos y posteriores a esta, evaluando si durante dichos períodos, existieron cambios en la concentración de contaminantes y valores de los factores en distintas estaciones.

4.4. Hipótesis

Por lo tanto, la hipótesis con la que se trabaja se define de la siguiente manera: **existió una alteración entre las variables y factores durante la pandemia de COVID-19 y durante períodos previos y posteriores a esta**. Dicha hipótesis es racional porque durante la pandemia la actividad industrial y humana se vieron altamente modificadas, siendo estas factores de peso en la calidad del aire.

5. Metodología

5.1. Preparación de los Datos

Se decidió hacer uso de los datos otorgados por el Socio Formador de los años 2020 a julio de 2024, debido a que consideramos que para efectos de nuestra investigación, la cual tendrá un enfoque en los tiempos de pandemia, estos 4 años nos permiten comparar eficientemente varios de los factores que se relacionan con la calidad del aire. Por otro lado, se decidió hacer uso de todas las variables suministradas, pues en los métodos subsecuentes los factores meteorológicos serán relevantes.

5.2. Limpieza de los Datos

Después de la selección de los conjuntos que se iban a utilizar, fue necesario comenzar con la limpieza de los datos. Esta limpieza se conforma de varios procesos, incluyendo la eliminación de duplicados, el tratamiento de los valores nulos, análisis de valores erróneos, entre otros.

En general, este proceso no contó con grandes dificultades, debido a que en nuestras bases de datos no estuvimos manejando variables categóricas. Esto implica que se pueden analizar las columnas de manera directa sin necesidad de crear variables dummies.

Como primer paso, 719 datos erróneos fueron transformados a nulos (eliminados) considerando los siguientes supuestos:

1. Ningún componente del aire puede tener una concentración menor o igual a 0. (253 datos)
2. La presión atmosférica (PRS) no puede tener un valor menor o igual a 0. (1 dato)
3. La precipitación (RAINF) no puede tener valores menores a 0. (0 datos)
4. La humedad relativa (RH) debe de tener un valor de 0 a 100. (274 datos)

5. La radiación solar (SR) no puede tener valores menores a 0. (189 datos)

6. La dirección del viento (WDR) debe tener valores entre 1 y 360. (2 datos)

Asimismo, fue necesario eliminar también las filas duplicadas, las cuales creemos se debían a errores computacionales o humanos. En este caso, solo se eliminaron las filas que eran completamente iguales (sin contar la variable fecha). Estas filas correspondían a 0.9 % de los datos.

En la base de datos se pudo identificar que había una gran cantidad de datos faltantes. Esto se debía a algunas fallas que hubo en los analizadores entre los años 2019 y 2020. Estos problemas generaron una necesidad de rehabilitar todos los equipos causando que los datos no se pudieran recopilar apropiadamente en dicho lapso de tiempo. Considerando la enorme cantidad de datos faltantes, se decidió eliminar aquellas filas con más de seis variables vacías, representando 8.33 % de las filas. Esta decisión se tomó con el objetivo de mejorar la calidad de los datos y evitar sesgos en el análisis, al eliminar observaciones incompletas que podrían afectar significativamente los resultados.

Finalmente, para poder poner en práctica algunas de las técnicas de modelación utilizadas posteriormente, fue necesario utilizar únicamente filas carentes de datos faltantes, por lo cual 45.7 % de los datos fueron eliminados.

5.3. Técnicas de ingeniería empleadas

Durante la obtención de los resultados se hizo uso de herramientas de programación en R, técnicas estadísticas y modelado de los contaminantes, todo esto con una rigurosidad científica y basado en literatura válida previa.

5.4. Infraestructura y recursos utilizados

La infraestructura de soporte que puede beneficiar otros proyectos de análisis multivariante de datos a largo plazo incluye sistemas de información para la gestión de datos y equipos tecnológicos con capacidad para procesar las técnicas estadísticas avanzadas empleadas en el análisis para grandes volúmenes de datos. Además, es fundamental contar con espacios de trabajo, como oficinas, para el personal que desarrolla modelos, técnicas matemáticas y estadísticas. También se requiere del equipo necesario para la recolección y almacenamiento de datos del Sistema Integral de Monitoreo Ambiental (SIMA), que fue nuestra fuente principal de datos para el presente proyecto, pues nos suministró los datos históricos del clima de Monterrey desde el 2020 hasta el año actual. Para recaudar esta información, SIMA utilizó equipo de medición y plataformas de gestión para limpiar y depurar dichos datos.

Por otro lado, para implementar los modelos estadísticos no fue necesario adquirir ninguna licencia, pues únicamente se hizo uso del lenguaje de programación *R* en el software *R Studio* en los siguientes equipos:

- **Para Modelos Mixtos:** Windows 10 Pro, Intel(R) Core(TM) i3-6006U CPU @ 2.00GHz, 16GB de RAM.
- **Para Modelos Mixtos:** MacOS Sonoma 14.5, Apple M2, CPU de 8 núcleos, GPU de 10 núcleos, 8GB de RAM.
- **Para Análisis Factorial:** MacOS Sequoia 15, Apple M2 Max, CPU de 12 núcleos, GPU de 38 núcleos, 32GB de RAM.
- **Para Simulación de Montecarlo y test de permutaciones:** MacOS Sonoma 14.3, Apple M1, CPU 8 núcleos, GPU 8 núcleos, 8GB de RAM

5.5. Modelación y Validación

Durante el proceso de la modelación, se probaron varios tipos de análisis con la finalidad de encontrar el más apropiado que nos permitiría generar conclusiones en cuanto al impacto de la pandemia en la contaminación y por ende, en la calidad del aire. Algunos de los modelos que

probamos fueron: ANOVA, análisis discriminante lineal, análisis discriminante cuadrático y la prueba de Kruskal-Wallis.

Después de realizar estos análisis, nos dimos cuenta de que los resultados dados por estos no son válidos debido a que los datos que estamos tratando no muestran distribución normal, homoscedasticidad, ni independencia. Conociendo esto, podemos ver que los datos no cumplen con algunos de los 3 supuestos más comunes en cuanto a las pruebas estadísticas.

A raíz de que no pudimos utilizar estos modelos, decidimos crear modelos mixtos y analizarlos para ver si estos nos podrían dar resultados válidos. Este proceso se explica a continuación.

5.5.1. Modelos Mixtos

Modelos Mixtos Generales

Después de probar las técnicas anteriores, las cuales no cumplieron con los supuestos, se aplicó el método de Modelos Mixtos que nos permite identificar el peso de las variables en cada contaminante. Para esto se determinaron los factores de efecto aleatorio y fijo.

Para elegir cuáles de las variables se iban a utilizar como efectos fijos, se calculó la matriz de correlación y se eligió mantener a todas las variables que tuvieran correlaciones mayores a 0.3. La cantidad de variables varía por contaminante, había algunos con 4 o 5 correlaciones mayores a 0.3, mientras otros tenían 1. Además de las variables con las características mencionadas, se mantuvo a las estaciones del año como efecto fijo para analizar su relación con cada uno de los contaminantes.

En cuanto los factores aleatorios, se establecieron a las variables relacionadas con el viento, y la columna binaria "pandemia" que es la que nos interesa, puesto que queremos determinar si los niveles de los contaminantes fueron afectados por la pandemia.

La varianza resultante del modelo es determinante para precisar a los contaminantes influenciados por la pandemia. A continuación se pueden ver los resultados de tanto el análisis de correlaciones realizado para el modelo correspondiente a PM10, como los resultados del modelo. Cabe mencionar que se eligió mostrar al PM10 en su totalidad debido a que es el contaminante con la mayor varianza, indicando que es el contaminante con mayor impacto en la variable pandemia.

Contaminante	Correlaciones
PM10	PM2.5: 0.6142959 NOX: 0.4489002 NO: 0.3926326 NO2: 0.3889203

Cuadro 1: PM10 y los contaminantes correlacionados

En el caso del PM10, solamente PM2.5, NOX, NO y NO2 tuvieron correlaciones mayores a 0.3. Por esto mismo, solo se modelaron con estas 4 variables fijas, además de la estación del año que se utilizó en todos los modelos. La varianza y los coeficientes resultantes de este modelo se muestran en el cuadro 2.

Contaminante	Varianza	Factores fijos
PM10	2.385	PM2.5: 1.338207 NO2: 1.286414 NO: 0.964164 NOX: -0.691697 season: 0.061565

Cuadro 2: PM10 y sus factores fijos

En el cuadro 2 se pueden ver los factores fijos y su coeficiente correspondiente, así como la varianza relacionada con la variable aleatoria de la pandemia. Para este modelo, también se tomaron en cuenta las variables aleatorias relacionadas con el viento y con la pandemia. Asimismo, se puede ver que la varianza de la velocidad del viento es 433, un valor muy alto, el cual indica que la velocidad del viento tiene una gran influencia en los niveles de PM10.

La gráfica correspondiente a este modelo es:

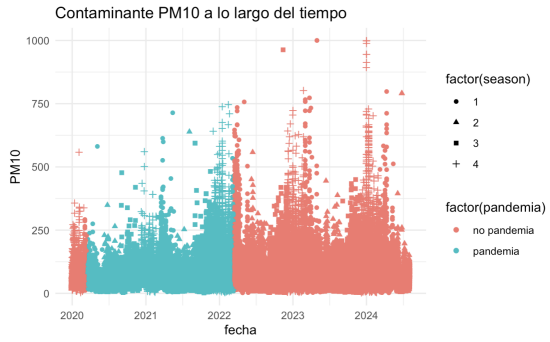

Figura 1: PM10 y la pandemia

En este gráfico se puede ver que en general los niveles de contaminación durante la pandemia no son tan diferentes a los niveles de contaminación en otros periodos del tiempo. Además, se puede ver que el nivel del contaminante muestra un comportamiento cíclico. Esto puede significar que durante el invierno, los niveles de contaminación son más altos. Esto se puede apreciar en el gráfico 1. Las varianzas y los coeficientes correspondientes a otros 2 contaminantes con varianzas relativamente altas se muestran en los Cuadros 3 y 4. Los gráficos de estas variables se pueden ver en el anexo 8.3.1.

Contaminante	Varianza	Factores fijos
PM2.5	0.6353	CO: 2.6331982 NO2: -0.8054022 NOX: 0.9452399 PM10: 0.1972244 NO: -0.8612525 season: -1.0339223

Cuadro 3: PM2.5 y sus factores fijos

Contaminante	Varianza	Factores fijos
SO2	0.2223	O3: 0.0473815 season: 0.2233605

Cuadro 4: S02 y sus factores

Después de crear los modelos, se aplicó la validación para los que tuvieron varianzas mayores. Analizando los resultados del proceso de validación que se puede ver en el anexo 8.4.1, podemos ver que no podemos crear predicciones confiables a partir de este modelo. El hecho de que no se cumplan estos supuestos se puede deber a la cantidad de valores nulos presentes, la manera en que se recopilaron los datos, entre otros.

Modelos Mixtos por Estación

Después de observar los resultados del modelo mixto general y de conocer un poco más el contexto de los contaminantes en Monterrey, nos dimos cuenta de que había una posibilidad que al dividir los datos por estación pudiéramos lograr obtener un mejor comportamiento en los modelos, puesto que las estaciones tienen una aportación diferente de contaminación, por lo que al generalizarlo todo en un solo modelo la complejidad de los datos no podía ser representada correctamente.

Se realizaron sub-sets por estación al dataset completo (sin eliminación de todos los datos nulos), pero ya con las columnas: periodo, pandemia y estación del año, y se calcularon los datos faltantes por estación, eliminando aquellas con un porcentaje mayor del promedio (19 %).

Se procuró que las estaciones utilizadas fueran representativas de la zona metropolitana, considerando que los niveles de los contaminantes pueden variar mucho entre estaciones debido a la presencia de factores como la refinería en Cadereyta o la cantidad de automóviles que transitan en Escobedo, esto puede tener un gran impacto en el nivel y la clase de contaminantes predominantes en cada área. Cabe mencionar que se consideró que las condiciones climatológicas pueden ser muy diferentes debido a que las estaciones analizadas se encuentran en diferentes zonas del área metropolitana de Monterrey. Para elegir cuáles de las variables se iban a utilizar como efectos fijos se realizó lo mismo que en el modelo general, calculamos las correlaciones, discriminamos aquellas con menos de 0.3 de correlación, manteniendo a su vez a las estaciones del año como efecto fijo y estableciendo las variables relacionadas con el viento y a la pandemia como efectos aleatorios. De igual manera nos dimos cuenta de que la cantidad de variables variaba mucho no solo por contaminante, sino ahora también por estación (lugar).

La varianza resultante del modelo en el factor aleatorio de la variable pandemia es determinante para precisar el

nivel de influencia que tuvo la pandemia en los valores registrados de los contaminantes.

En el Cuadro 5, se pueden apreciar los resultados de las siete estaciones elegidas con el contaminante con mayor varianza para cada una de estas. Ahí se puede ver que el contaminante con la mayor influencia (varianza) en el factor aleatorio pandemia es el PM10 en la estación Escobedo.

En el anexo 8.4.2 se puede ver un ejemplo del método de validación que se le aplicó a los modelos mixtos como parte de este análisis. Los resultados de esta validación indican que el modelo actual no es adecuado para realizar predicciones confiables debido a la violación de los supuestos fundamentales. Estos problemas podrían estar relacionados con la calidad de los datos, como la presencia de valores faltantes o errores en la recolección de datos.

5.5.2. Análisis Factorial

El propósito del análisis factorial es describir, si es posible, las relaciones de covarianza entre muchas variables en términos de unas pocas cantidades aleatorias subyacentes, pero no observables, llamadas factores (Johnson y Wichern, 1998).

Contemplando el objetivo de esta investigación, el modelo factorial está motivado por el argumento de que las variables se pueden agrupar según sus correlaciones. Es decir, todos los contaminantes dentro de un grupo particular están altamente correlacionados entre sí, pero tienen correlaciones relativamente pequeñas con contaminantes de un grupo diferente. Entonces, es concebible que cada grupo de contaminantes represente un único constructo o factor subyacente que sea responsable de las correlaciones observadas, en este caso, se puede relacionar con similares fuentes de contaminación para los contaminantes.

¿Es factible su uso?

Considerando la prueba de Kaiser-Meyer-Olkin (KMO), la cual es una medida de la idoneidad de los datos para el análisis factorial, con un valor de 0.5378448, es factible realizar el análisis factorial, ya que esto nos indica que existe suficiente correlación entre las variables, dado que es mayor a 0.5. Sin embargo, aunque es posible realizar el análisis factorial, las correlaciones entre las variables no

Escobedo		
Contaminante	Varianza	Influencias fijas
PM10	45.95	NO2:-6.39913 NOX:6.88193 NO:-6.64091 PM2.5:1.68691 season:2.94547
García		
Contaminante	Varianza	Influencias fijas
O3	23.18	NOX:-0.072710 RH:-0.379587 TOUT:0.322075 season:-1.021743
Santa Catarina		
Contaminante	Varianza	Influencias fijas
O3	7.654	NOX:-0.04913 RH:-0.249266 SR:14.222509 TOUT:0.600753 season:0.189662
Obispado		
Contaminante	Varianza	Influencias fijas
PM10	1.681	NO2:1.22584 NOX:-0.60218 NO:0.73675 PM2.5:1.31288 SO2:0.95250 season:1.51630
Cadereyta		
Contaminante	Varianza	Influencias fijas
PM10	15.707	CO:2.67270 NO2:5.95631 NOX:-4.24895 NO:4.63677 PM2.5:1.60516 season:-1.33301
Juarez		
Contaminante	Varianza	Influencias fijas
PM10	15.524	CO:1.86024 NO:-1.19156 NO2:-0.43877 NOX:1.44063 PM2.5:1.06543 season:-1.20277
San Pedro		
Contaminante	Varianza	Influencias fijas
PM10	4.909	NO2:0.60141 NOX:0.06758 PM2.5:1.81864 season:-0.90454

Cuadro 5: Contaminantes, factores influyentes y pandemia

son lo suficientemente fuertes como para asegurar factores claramente definidos o interpretables.

A partir de la Figura 2, donde se visualizan los valores propios de la matriz de correlación R y del análisis factorial, se puede observar que un número adecuado de factores a retener está entre 3 y 4. Debido a eso, se realizaron pruebas con 3 y 4 factores, y tras realizar un trabajo de interpretación, se detectó que con 4 factores se daría la mejor interpretación. Por tanto, en este análisis se procederá con la extracción de 4 factores, además que una mayor cantidad de factores puede equilibrar el hecho de que los datos tienen un bajo nivel de factibilidad para el análisis factorial.

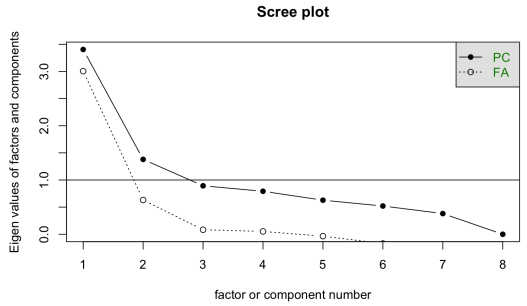

Figura 2: Scree Test

El método utilizado para crear el modelo, fue el de **Solución de Factores Principales**, que consiste en extraer factores subyacentes de una matriz de correlación, ajustando las comunalidades iniciales y maximizando la varianza explicada por los factores comunes, mientras se minimiza la influencia de las varianzas específicas (Johnson y Wichern, 1998). Además, se aplicó la rotación "varimax", que busca maximizar la dispersión de los cuadrados de las cargas factoriales, seleccionando una transformación ortogonal que "extiende"los valores de las cargas factoriales en cada factor (Johnson y Wichern, 1998), facilitando la identificación de patrones en las variables, ya que se cargan más a ciertos factores que a otros.

A continuación, en el Cuadro 6 se muestran las comunalidades para cada variable en el modelo generado. La comunalidad indica la proporción de la varianza total de cada variable que es explicada por los factores extraídos.

Variable	**M**
NO2	1.0111583
NOX	1.0014552
NO	0.9895817
PM2.5	0.6978573
PM10	0.5492104
O3	0.3673687
CO	0.2262845
SO2	0.1779734

Cuadro 6: Comunalidades de las variables en el modelo M.

Existen variables como "O3", "CO" y "SO2"que tienen una comunalidad menor a 0.5, lo que indica que los factores extraídos no explican adecuadamente la varianza de estas variables. Esto se relaciona con el hecho de que la prueba (KMO), había indicado que un análisis por factores no iba a ser del todo favorable para todas las variables.

A continuación, en el Cuadro 7 se presentan las cargas por cada factor.

Variable	**Factor 1**	**Factor 2**	**Factor 3**	**Factor 4**
CO	0.176	0.354	0.109	-
NO	0.928	0.339	0.113	-
NO2	0.301	0.308	0.909	-
NOX	0.805	0.376	0.461	-
O3	-0.315	-	-0.187	0.299
PM10	0.142	0.710	0.136	-
PM2.5	0.114	0.818	0.107	-
SO2	-	0.102	-	0.596

Cuadro 7: Cargas por Factor

Sabiendo que cada puntuación de los factores se obtiene como una combinación lineal de las cargas de los factores y los valores de las variables, entonces, sea $\mathbf{F}_j$ el j-ésimo factor, se tiene que:

$$\mathbf{F}_j = \sum_{i=1}^{n} \mathbf{L}_{ij} \mathbf{X}_i$$

Donde:

$\mathbf{L}_{ij}$ es la carga de la i-ésima variable en el j-ésimo factor.

$\mathbf{X}_i$ son los registros de valores de la i-ésima variable.

n es el número total de variables.

Una vez calculadas las puntuaciones de los 4 factores, para facilitar su interpretación, se las puede visualizar en la Figura 3, en donde cada factor está asociado con una combinación de contaminantes, considerando las cargas con el mayor valor.

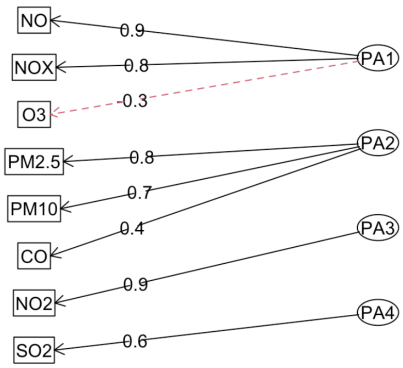

Figura 3: Diagrama de Factores por Contaminante

Como parte de nuestra metodología, a continuación se explica qué razones podrían estar detrás del agrupamiento de las variables en los 4 factores, considerando los limitados conocimientos que el equipo posee en química, física y dinámica de partículas. Se cree que la principal razón de estos factores es su origen, como se da a conocer a continuación.

- **Factor 1 (Industria y Vehículos):** Creemos que este factor se relaciona con los automóviles, centrales eléctricas y otras fuentes industriales, principales responsables de la creación de óxidos de nitrógeno en general. Por otro lado, el ozono se encuentra en este factor porque una de las principales formas en la que se crea es precisamente a partir del óxidos de nitrógeno.
- **Factor 2 (Actividad No Industrial):** Parece ser que aquí se encuentran compuestos generados por actividades no industriales, como la construcción. Ya que que estos tienden a generar una gran cantidad de partículas microscópicas de diversos tamaños, así como CO, la cual se genera en procesos industriales sin combustión o en el sector agropecuario.
- **Factor 3 (Doméstico):** NO2 se encuentra en su propio factor debido posiblemente a que su fuente de origen se relaciona ampliamente con la actividad humana dentro del hogar, como las estufas, la soladura, el humo de tabaco, queroseno, etcétera.
- **Factor 4 (Combustión de combustibles fósiles):** Finalmente, el factor 4 contiene únicamente a SO2, ya que su origen se relaciona principalmente con la combustión de combustibles fósiles que contienen azufre, lo que lo diferencía del resto de contaminantes.

5.5.3. Test de Permutaciones

Considerando que los factores son ortogonales, no correlacionados, sin normalidad ni homocedasticidad, no se cumplen los supuestos para realizar pruebas de comparación de medias usando, por ejemplo, la prueba paramétrica t de dos muestras o la no paramétrica U de MannWhitney.

Como alternativa se usa un test de permutación, para el estudio de la diferencia entre dos grupos, calculando la distribución de un estadístico. En este caso, el estadístico es la diferencia media de las observaciones factoriales entre grupos Δ_{obs}, para todas las posibles reorganizaciones de las observaciones en los distintos grupos, generando una distribución de diferencia media permutada Δ_{perm} (Whitlock y Schluter, 2014), con el objetivo de determinar si existieron cambios significativos en los factores de contaminantes entre los grupos de pandemia y no pandemia, por estación.

Debido a que en el test de permutación, el valor p obtenido es exacto, ya que se calculan todas las posibles permutaciones de las observaciones, resulta computacionalmente imposible realizarlo dado que el tamaño muestral es considerable, entonces se emplea una simulación de Monte Carlo, haciendo que el test se tenga que realizar únicamente sobre una muestra aleatoria de todas las posibles permutaciones. El valor p se calcula como la pro-

porción de permutaciones en las que la diferencia absoluta es mayor o igual que la diferencia observada:

$$\text{p-valor} = \frac{r}{k}$$

Donde:

r es el número de permutaciones en las que:

$$|\Delta_{\text{perm}}| \geq |\Delta_{\text{obs}}|$$

k es el número total de permutaciones empleadas en la simulación.

Evaluándose sobre las siguientes hipótesis:

- H_0: No hay diferencia significativa en los factores de contaminantes entre los grupos de pandemia y no pandemia.
- H_1: Hay una diferencia significativa en los factores de contaminantes entre los grupos de pandemia y no pandemia.

6. Resultados

6.1. Propuesta Metodológica a Utilizar

Se optó por ejecutar tantos la metodología de modelos mixtos como la del test de permutaciones, el primero con la finalidad de analizar la alteración de los compuestos contaminantes y el segundo con la finalidad de analizar la alteración de los factores obtenidos con el análisis factorial.

6.2. Modelos Mixtos

6.2.1. Modelos Mixtos Generales

El contaminante con mayor varianza influenciada por la pandemia en los Modelos Mixtos Generales fue PM10. Las medidas correspondientes a este modelo se mencionaron anteriormente en el Cuadro 2.

En este modelo, la velocidad del viento mostró una varianza de 433, indicando una fuerte influencia sobre los niveles de PM10. No obstante, los supuestos de normalidad, homoscedasticidad e independencia no se cumplieron, lo que afecta la capacidad predictiva del modelo. Los resultados de estos supuestos se pueden ver en las Figuras 11, 13, 14 y 12.

6.2.2. Modelos Mixtos por Estación

Al aplicar los Modelos Mixtos por Estación, se observó un aumento considerable en la varianza del contaminante PM10 en la estación de Escobedo, lo que sugiere una mejor representación de la relación entre la pandemia y los contaminantes en modelos segmentados por estación, indicando a su vez, que las zonas tienen diferentes niveles de contaminación.

Estación	Contaminante	Varianza
Escobedo	PM10	45.95

Cuadro 8: Varianza de PM10 en la Estación de Escobedo

Los factores fijos para este modelo fueron:

Factor Fijo	Coeficiente
NO2	-6.39913
NOX	6.88193
NO	-6.64091
PM2.5	1.68691
Estación	2.94547

Cuadro 9: Factores Fijos en la Estación de Escobedo (PM10)

El modelo mixto por estación, específicamente para PM10 en Escobedo, presentó una mayor varianza (45.95 vs. 2.385 en el modelo general). Sin embargo, al igual que en el modelo general, no se cumplieron los supuestos de normalidad, homocedasticidad e independencia.

Esto se puede visualizar en las Figuras 15 y en 17.

6.3. Análisis Factorial

Utilizando los 4 factores obtenidos durante la etapa de modelación, se realizó un análisis a nivel general sobre el comportamiento de estos durante el tiempo de pandemia (Figura 4) y fuera de él (Figura 5), esto con la finalidad

de observar qué factores sufrieron más cambios y la naturaleza de estos.

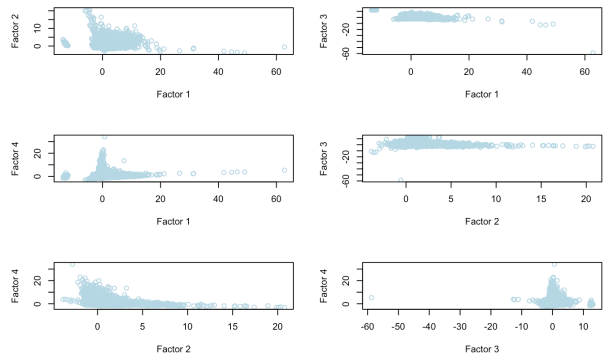

Figura 4: Comparación de factores durante la pandemia

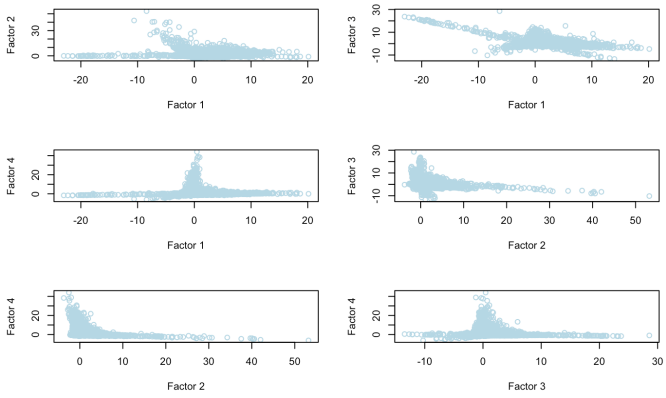

Figura 5: Comparación de factores fuera de la pandemia

De las gráficas anteriores es posible obtener las siguientes conclusiones:

- El factor 1 tomó valores positivos en muchas más ocasiones durante la pandemia que fuera de ella. De hecho, en el período sin pandemia la distribución de este factor parece estar centrada alrededor del 0, mientras que durante esta la mayor parte de los valores que tiene son positivos.
- Respecto al factor 2, este tuvo una varianza considerablemente menor durante la pandemia, siendo su valor máximo igual a 20.75, mientras que en el otro periodo de tiempo su máximo fue de 50.19.
- El factor 3 también presentó una varianza menor durante la pandemia, así como una mayor cantidad de datos atípicos.
- El factor 4 no parece haber cambiado de forma sustancial durante la pandemia, pues sus valores siguen una distribución muy parecida en los dos períodos.

6.4. Test de Permutaciones

El test de permutaciones con simulación de Monte Carlo fue aplicada para los 4 factores en las 7 estaciones utilizadas en los modelos mixtos, se obtuvieron los resultados observables en la Tabla 10.

Estación	PM1	PM2	PM3	PM4
Cadereyta	7×10^{-4}	0,0033	1×10^{-4}	1×10^{-4}
Obispado	1×10^{-4}	0,4217	1×10^{-4}	1×10^{-4}
Santa Catarina	1×10^{-4}	1×10^{-4}	1×10^{-4}	1×10^{-4}
García	1×10^{-4}	1×10^{-4}	1×10^{-4}	1×10^{-4}
Escobedo	0,2001	1×10^{-4}	1×10^{-4}	1×10^{-4}
San Pedro	5×10^{-4}	1×10^{-4}	1×10^{-4}	0,0027
Juárez	1×10^{-4}	1×10^{-4}	1×10^{-4}	1×10^{-4}

Cuadro 10: Resultados de Test de Permutaciones.

Analizando estos resultados, podemos notar que en la gran mayoría de los tests se obtuvieron valores muy pequeños, indicándonos que la hipótesis nula es rechazada, lo que nos otorga suficiente evidencia estadística para decir que existió un cambio significativo en los factores 3 y 4 en todas las estaciones durante la pandemia. Asimismo, estas pruebas nos permiten afirmar que el factor 1 se alteró en todas las estaciones a excepción de la de Escobedo, mientras que el factor 2 se alteró en todas las estaciones a excepción de la de Obispado. A grandes rasgos y a juzgar por los valores p obtenidos, se puede afirmar que el factor que experimentó menos cambios fue el segundo, mientras que aquel que experimentó más fue el primero.

7. Discusión y Conclusiones

Al comparar los resultados de los Modelos Mixtos Generales y los Modelos Mixtos por Estación, se observa que los modelos por estación permiten captar mejor la influencia de la pandemia en los niveles de contaminantes, especialmente para PM10 en Escobedo. Sin embargo, la falta

de cumplimiento en los supuestos estadísticos limita la capacidad predictiva de ambos enfoques y no se puede generar una conclusión concreta debido a que los modelos no se consideran válidos.

Sin embargo, analizando los resultados del Test de Permutaciones, se puede ver que estos tienen evidencia estadística significativa para rechazar la hipótesis nula de igualdad en todas las estaciones de monitoreo. Por consiguiente, dado que se observa un cambio significativo en los cuatro factores, podemos inferir que durante la pandemia se sufrieron cambios en las emisiones de contaminantes provenientes de las cuatro fuentes correspondientes: industria y vehículos, actividad no industrial, doméstica y combustión de combustibles fósiles.

A futuro, si tenemos la oportunidad de extender este proyecto, nos gustaría probar diferentes transformaciones en las variables con la intención de mejorar los modelos y que estos sean significativamente válidos. También sería interesante probar un tratamiento diferente de los modelos nulos, ya sea a través de métodos de interpolación u otras técnicas. Consideramos que esto podría ayudarnos a evitar las dificultades que surgieron a partir de la cantidad de valores nulos que estábamos manejando en esta base de datos.

Esto se nos hace interesante, ya que consideramos que si logramos que los modelos mixtos pasen las pruebas de validación, podríamos comparar los resultados de estos modelos con los del Test de Permutaciones. Esto nos ayudará a complementar la conclusión actual del proyecto al corroborar o refutar los resultados del Test de Permutaciones. Por ende, podríamos dar una respuesta concreta en cuanto a si la pandemia generó un impacto real en el nivel de contaminación. Consideramos que esta conclusión sería muy importante, ya que sí se corrobora la conclusión del Test de Permutaciones, podríamos utilizar este estudio como evidencia de que reduciendo ciertas actividades cotidianas como sucedió en la pandemia, podríamos mejorar la situación actual de contaminación y calidad del aire en nuestra zona.

La nueva pregunta que surge posterior al finalizar este análisis es: ¿Las zonas de alta densidad poblacional experimentan mayores concentraciones de ciertos contaminantes después de la pandemia?, ya que una vez reactivada la movilidad y la actividad económica, las áreas densamente pobladas pudieron haber experimentado un rebote en la contaminación en comparación con zonas menos pobladas.

Referencias

de Salud, S. (2021). *Criterios para la ejecución de actividades de promoción de la salud en materia alimentaria*. https://www.dof.gob.mx Recuperado 11/08/2024.

Figueroa, J. V. H. P. (s.f.). *Estado de la Calidad del Aire en México*. https://www.gob.mx/inecc/articulos/estado-de-la-calidad-del-aire-en-mexico?idiom=es#:~:text=Cuanto%20m%C3%A1s%20bajo%20sean%20los,largo%20como%20a%20corto%20plazo. Recuperado 11/08/2024.

Hugo Barrera, A. B. y. M. G. (2021). *Estudio para el rediseño de la red de monitoreo de la calidad del aire de Monterrey*. https://experiencia21.tec.mx/courses/520587/pages/referencias-reto-recomendadas-por-socio-formador?module_item_id=30641080 Recuperado 11/08/2024.

INECC. (s.f.). *Type of pollutants*. https://sinaica.inecc.gob.mx/scica/ Recuperado 11/08/2024.

Johnson, R. A., & Wichern, D. W. (1998, enero). *Applied Multivariate Statistical analysis*. Pearson.

OMS. (2022). *¿Cómo se mide la calidad del aire?* https://www.who.int/teams/environment-climate-change-and-health/air-quality-and-health/health-impacts/types-of-pollutants Recuperado 11/08/2024.

UNEP. (2022). *Type of pollutants*. https://www.unep.org/es/noticias-y-reportajes/reportajes/como-se-mide-la-calidad-del-aire Recuperado 11/08/2024.

Whitlock, M., & Schluter, D. (2014). Analysis of Biological data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *11*(6), 1272. https://doi.org/10.1109/tcbb.2014.2374979

8. Anexos

8.1. Carpeta en drive con los documentos de trabajo

- Liga que dirige a documentos de trabajo:

https://drive.google.com/drive/folders/1-ydpSGiXGwwtFaN_5U-XToSAWHZa97dC?usp=sharing

8.2. Normas oficiales de calidad del aire

Normatividad			
Contaminante	Norma	Concentración	Tiempo de Exposición(horas)
Monóxido de Carbono (CO)	NOM-021-SSA1-2021	26.0 ppm	1
Monóxido de Carbono (CO)	NOM-021-SSA1-2021	9.0 ppm	8
Bióxido de Azufre (SO2)	NOM-022-SSA1-2019	0.075 ppm	1
Bióxido de Azufre (SO2)	NOM-022-SSA1-2019	0.04 ppm	24
Ozono (O3)	NOM-020-SSA1-2021	0.090 ppm	1
Ozono (O3)	NOM-020-SSA1-2021	0.060 ppm	8
Bióxido de Nitrógeno (NO2)	NOM-023-SSA1-2021	0.106 ppm	1
Bióxido de Nitrógeno (NO2)	NOM-023-SSA1-2021	0.021 ppm	Promedio Anual
Partículas Menores a 10 Micras (PM10)	NOM-025-SSA1-2021	60 $\mu g/m^3$	24 (Promedio Anual)
Partículas Menores a 10 Micras (PM10)	NOM-025-SSA1-2021	28 $\mu g/m^3$	24 (Promedio Anual)
Partículas Menores a 2.5 Micras (PM2.5)	NOM-025-SSA1-2021	33 $\mu g/m^3$	24 (Promedio Anual)
Partículas Menores a 2.5 Micras (PM2.5)	NOM-025-SSAI-2021	10 $\mu g/m^3$	Promedio Anual

Cuadro 11: Normas Oficiales Mexicanas (de Salud, 2021)

8.3. Modelos Mixtos Generales

8.3.1. Resultados correspondientes a otros modelos

Contaminante	Varianza	Factores fijos
O3	0.01252	RH: -0.295252 TOUT: 0.341883 SR: 23.057821 NOX: -0.123755 season: 0.227076

Cuadro 12: O3 y sus factores fijos

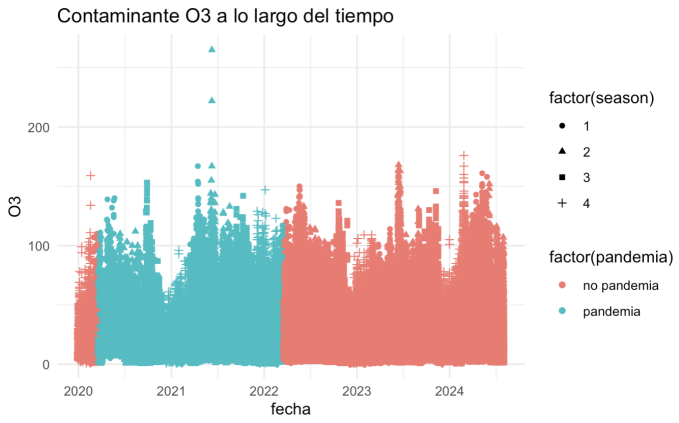

Figura 6: O3 y la pandemia

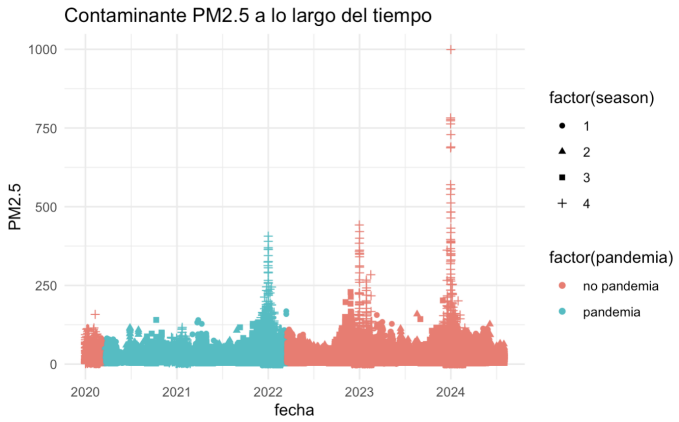

Figura 7: PM2.5 y la pandemia

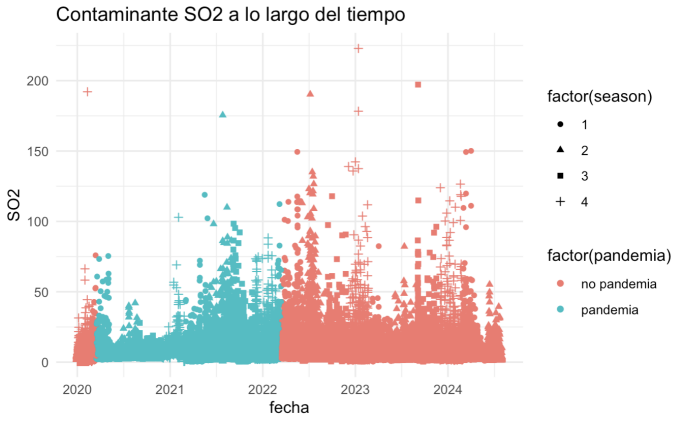

Figura 8: SO2 y la pandemia

8.3.2. Modelos Mixtos por Estación

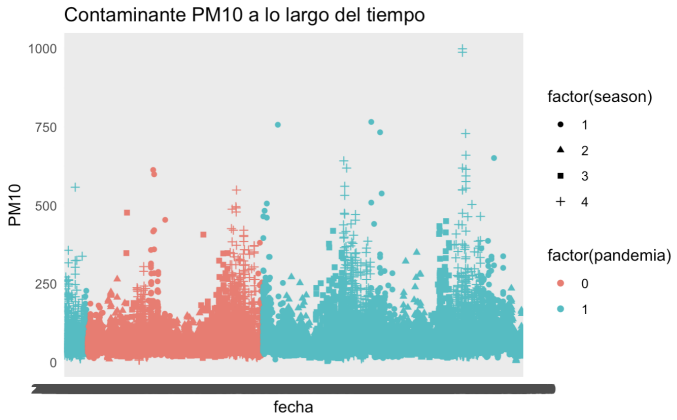

Figura 9: PM10-Cadereyta y la pandemia

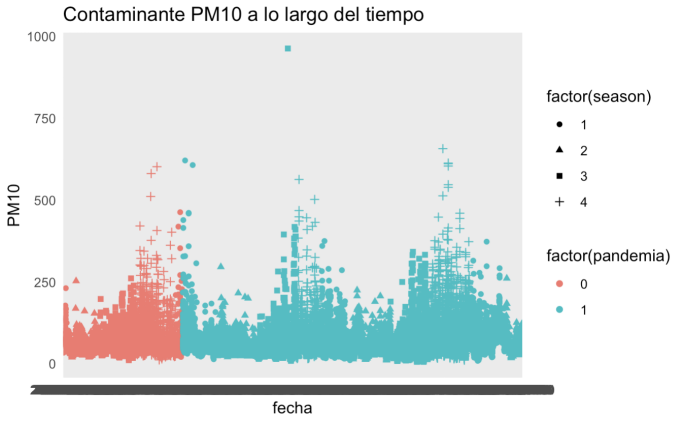

Figura 10: PM10-Juarez y la pandemia

8.4. Validación de los modelos mixtos

8.4.1. Modelos mixtos generales

En este caso, analizamos los supuestos de los 3 modelos con mayor varianza con respecto a la pandemia, debido a que estos son los que tuvieron mejores resultados, tomando en cuenta nuestro objetivo y nuestra pregunta de investigación. Para validar el modelo, probamos la normalidad, la homoscedasticidad y la independencia de los residuos. Esta validación se muestra a continuación.

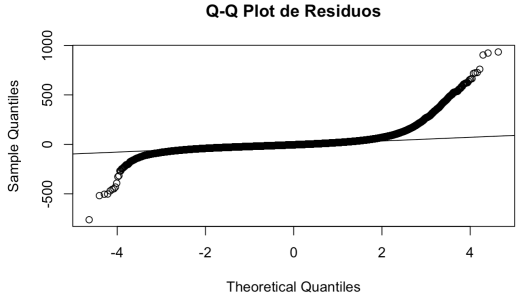

Figura 11: Q-Q plot de la normalidad de los residuos

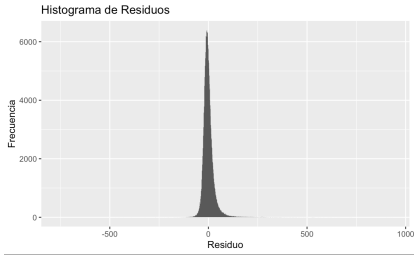

Figura 12: Histograma de la normalidad de los residuos

Analizando tanto el QQ-plot como el histograma 12, se puede notar que los residuos no tienen una distribución normal. En la distribución de estos residuos se puede identificar un sesgo muy marcado. Esto implica que el modelo no pasa el supuesto de normalidad de los residuos.

Como parte de este análisis de supuestos, hicimos la prueba de Durbin-Watson para revisar si los datos muestran homoscedasticidad. Esta prueba resultó en un p-valor muy pequeño, menor a 0.05, lo que nos indica que no se cuenta con esta característica. Esta conclusión se puede corroborar con la Figura 13.

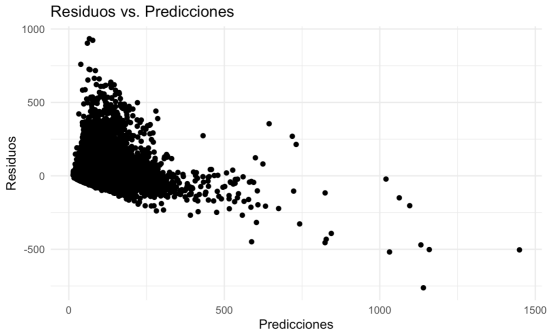

Figura 13: Gráfica para revisar homoscedasticidad

En este caso, se puede identificar un patrón muy marcado en la dispersión de los puntos. Esto corrobora la conclusión anterior que indica que no hay varianzas constantes. Por lo tanto, este modelo no pasa el supuesto de homoscedasticidad.

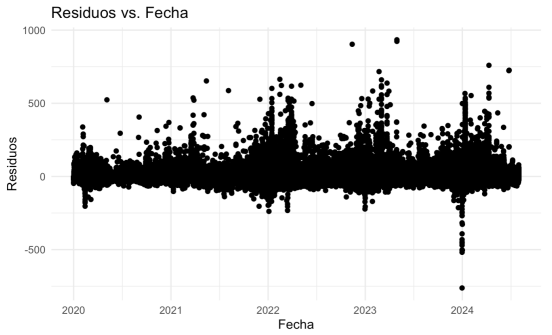

Figura 14: Gráfica para revisar independencia

En el caso de este gráfico se puede ver que sí se aprecia un patrón ligero. Esto parece indicar que los residuos no son independientes. Por lo tanto, este modelo no pasa los supuestos de la independencia.

8.4.2. Modelos mixtos por estaciones

A continuación se presentará el método de validación del modelo aplicado al contaminante con la varianza más alta (45.95), siendo este el PM10 en la estación de Escobedo.

Tanto el QQ-plot de los residuos (Figura 15) como el histograma (Figura 16), muestran una evidente desviación de la normalidad, evidenciando un sesgo pronunciado, esto se puede ver sobre todo en el QQ-plot cuando las colas se despegan completamente de la línea de tendencia.

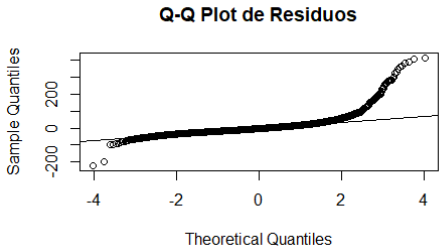

Figura 15: QQ-plot de la normalidad de los residuos (Escobedo-PM10)

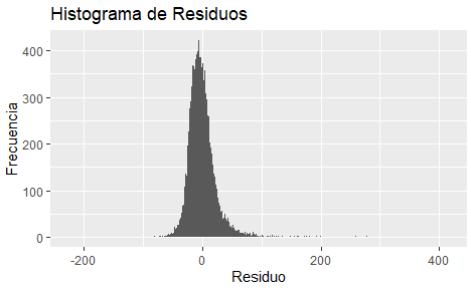

Figura 16: Histograma de la normalidad de los residuos (Escobedo-PM10)

La prueba de Durbin-Watson calculó un valor inferior a 0.05, indicando la presencia de autocorrelación en los residuos y, por tanto, el rechazo de la hipótesis de homoscedasticidad, conclusión que se corrobora visualmente con el siguiente gráfico.

El gráfico de dispersión de los residuos (Figura 17) muestra un patrón no aleatorio, evidenciando heteroscedasticidad. Esta observación confirma los resultados de la prueba de Durbin-Watson.

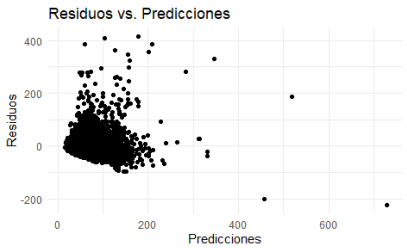

Figura 17: Gráfica para revisar homoscedasticidad (Escobedo-PM10)

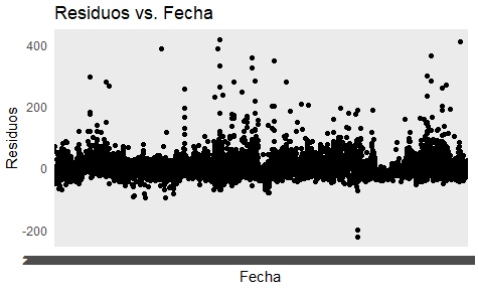

Figura 18: Gráfica para revisar independencia (Escobedo-PM10)

El gráfico de residuos por fecha (Figura 18) revela un patrón discernible, lo que sugiere una violación del supuesto de independencia de los errores.