Instituto Tecnológico y de Estudios Superiores Monterrey

Escuela de Ingeniería y Ciencias

Ingeniería en Ciencia de Datos y Matemáticas

Campus Monterrey

Aplicación de métodos multivariados en ciencia de datos

Grupo 301

Blanca Rosa Ruiz Hernández Mónica Guadalupe Elizondo Amaya

REPORTE FINAL

Equipo 2

Ricardo Vargas Garduño	A01026909
Esteban Ángel Pérez Muraira	A00832329
Juan David Rivera	A01026193
Alejandro Medrano Torres	A00831829
Rogelio Lizárraga Escobar	A01742161

03 de diciembre de 2023, Monterrey, N.L.

1. RESUMEN

En este proyecto de análisis de la calidad del aire en la zona metropolitana de Monterrey, se destacó la importancia de monitorear y abordar la presencia de contaminantes atmosféricos. Se identificaron los contaminantes más perjudiciales, como PM10, PM2.5, CO, NO2, SO2 y O3. Se llevó a cabo un análisis detallado de la base de datos, identificando y limpiando valores nulos, datos errados y selección del conjunto de datos adecuado.

Los objetivos iniciales de relacionar la contaminación con enfermedades respiratorias se ajustaron a nuevos enfoques, como la clasificación eficiente de la calidad del aire y la identificación de las partículas más influyentes. Se implementaron modelos de regresión logística multinomial y series de tiempo con medianas mensuales, aunque se encontraron limitaciones en la predicción.

Finalmente, con la prueba de ANOVA se revelaron diferencias significativas en las concentraciones de contaminantes entre zonas, destacando la importancia del PM10, PM2.5 y SO2 y la similitud del SO2 en las tres zonas estudiadas. Se concluyó que en los contaminantes si se presentan diferencias en concentraciones, resaltando en una variabilidad geográfica en la calidad del aire.

2. INTRODUCCIÓN Y JUSTIFICACIÓN

a. INTRODUCCIÓN

Es evidente que, en una macrociudad, el análisis de la calidad del aire es crucial para la ciudadanía por diversas razones. En primer lugar, impacta directamente en la salud pública al identificar y abordar contaminantes que pueden causar problemas respiratorios y enfermedades. Además, promueve la conciencia ambiental, incentivando prácticas sostenibles y medidas contra la contaminación. Contribuye a una mejor calidad de vida al crear entornos más saludables, afecta positivamente al medio ambiente al preservar la biodiversidad y garantiza el cumplimiento de normativas ambientales. Por otro lado, la información resultante es vital para la planificación urbana y tiene implicaciones económicas al afectar sectores como el turismo y la agricultura. En definitiva, el análisis de la calidad del aire es esencial para salvaguardar la salud, promover la sostenibilidad y mejorar la vida en las comunidades.

Luego de remarcar la importancia de analizar la calidad del aire, es relevante también centrarnos en aquellos aspectos más relevantes para realizar dicho análisis. Esto abarca la medición de la concentración de contaminantes, como partículas en suspensión y gases, expresados en microgramos por metro cúbico o partes por millón. La naturaleza de los contaminantes, su comparación con normativas y estándares ambientales, y la evaluación de zonas geográficas específicas también son cruciales. Las estaciones de monitoreo distribuidas, los factores meteorológicos y la identificación de fuentes de contaminación, como el tráfico vehicular o la industria, se consideran en conjunto. Además, se evalúan los efectos en la salud, se analizan tendencias temporales y se comunica esta información a la población para promover medidas protectoras. Este enfoque integral permite comprender y abordar eficazmente la calidad del aire en diferentes contextos.

Ahora bien, las normativas para clasificar a los contaminantes más relevantes son las Normas Oficiales Mexicanas para la Calidad del Aire; cabe destacar que estas se basan en el impacto en la salud humana y el medio ambiente. Entre los contaminantes más destacados se encuentran el ozono (O3), dióxido de azufre (SO2), monóxido de carbono (CO), dióxido de nitrógeno (NO2), y las partículas en suspensión (PM10, PM2.5), así como el plomo (Pb). Luego, al centrarnos en la gravedad de estos contaminantes, podemos recalcar que las partículas finas (PM2.5) son consideradas los contaminantes más perjudiciales, capaces de penetrar en los pulmones y el torrente sanguíneo,

causando problemas respiratorios y cardiovasculares. El ozono (O3) también impacta negativamente en la salud respiratoria, especialmente en áreas urbanas, mientras que el dióxido de nitrógeno (NO2), dióxido de azufre (SO2), y monóxido de carbono (CO) son igualmente perjudiciales para la salud humana. Esta información resalta la importancia de monitorear y abordar específicamente estos contaminantes clave para salvaguardar la salud pública.

Una vez identificados los contaminantes más perjudiciales para el ser humano, es de suma importancia prestar atención a cómo es que estos pueden ser identificados. La medición de la calidad del aire implica la detección y análisis de contaminantes atmosféricos mediante sistemas de monitoreo equipados con sensores específicos, abarcando contaminantes como ozono (O3), dióxido de nitrógeno (NO2), dióxido de azufre (SO2), monóxido de carbono (CO), partículas en suspensión (PM10, PM2.5) y plomo (Pb). Estas mediciones se realizan cada minuto, generando promedios ponderados cada hora para evaluar la concentración de contaminantes.

b. JUSTIFICACIÓN

Consideramos que es pertinente para la ciudadanía de la ciudad de monterrey identificar la manera en que la presencia y la distribución de contaminantes afecta a las personas de bajos recursos; es por ello que se ha optado por realizar un análisis estadístico robusto para identificar estas relaciones y de esta manera ayudar a las autoridades pertinentes a tomar decisiones más informadas que logren minimizar los efectos negativos de la contaminación en las comunidades más vulnerables de la zona metropolitana.

Tomando en consideración el trabajo de investigación “Determinación del Nivel de Vida por Municipio en el Estado de Nuevo León, México.” realizado por Cantú Martinez, P. C., y Gómez Guzmán, L. G., dentro de la zona metropolitana se trabajará con el municipio de de San Nicolás de los Garza, por tener el mejor índice de nivel de vida en Nuevo León, con Cadereyta Jimenez debido a ser el municipio con menor índice de nivel de vida dentro del alcance de las estaciones de monitoreo, y finalmente con el municipio de Garcia, que cuenta con un nivel de vida entre los anteriores. Contando entonces con 3 municipios, uno con un Grado de nivel de vida muy alto, otro medio-alto y uno medio-bajo.

3. PROBLEMÁTICA Y OBJETIVO

a. PROBLEMÁTICA

La medición precisa de la calidad del aire enfrenta desafíos como la selectividad y sensibilidad de los sensores, que pueden no ser suficientes para detectar contaminantes específicos sin pruebas en condiciones reales. Las condiciones ambientales extremas, junto con la variabilidad espacial y temporal de la calidad del aire, afectan la precisión de las mediciones, exigiendo la calibración adecuada de los sensores. Además, la amenaza de partículas ultrafinas, como el PM2.5, representa un desafío adicional debido a su tamaño diminuto. En resumen, se requieren tecnologías avanzadas y una comprensión profunda de los factores atmosféricos para lograr mediciones precisas de la calidad del aire.

Una de las organizaciones dedicadas a esta labor en el estado de Nuevo León es el Sistema Integral de Monitoreo Ambiental de Nuevo León (SIMA). Es una organización gubernamental que se encarga de medir y analizar la calidad del aire en la Zona Metropolitana de Monterrey. El monitoreo ayuda a identificar áreas con altos niveles de contaminación y tomar medidas para proteger a la población. SIMA es responsable de brindar información obtenida del Sistema de Monitoreo Atmosférico, mediante la medición de los parámetros meteorológicos e indicadores de calidad del aire denominados como contaminantes criterio. La información sobre la calidad del aire es valiosa para la planificación urbana y permite tomar decisiones informadas sobre el desarrollo de infraestructuras y el diseño de ciudades más sostenibles y saludables.

Es pertinente recalcar que la mejora de la calidad del aire implica la reducción de emisiones contaminantes, abordando sectores clave como transporte, industria, agricultura y generación de energía. Además, se han propuesto medidas como el uso de transportes más eficientes, tecnologías industriales más limpias, prácticas agrícolas sostenibles y el fomento de fuentes de energía renovable para lograr una mejora significativa en la calidad del aire.

b. OBJETIVO GENERAL

El objetivo de este proyecto es identificar la manera en que la contaminación afecta a las comunidades de bajos recursos en la zona metropolitana de Monterrey.

c. OBJETIVOS PARTICULARES

Objetivos anteriores:

- Identificar los contaminantes más perjudiciales para la salud del ser humano
- Obtener información sobre reportes de padecimientos de enfermedades respiratorias en personas de bajos recursos
- Identificar si hay alguna correlación entre la exposición a contaminantes y el padecimiento de enfermedades respiratorias en personas de bajos recursos
- Identificar zonas de alta contaminación

Objetivos nuevos:

- Clasificar de manera eficiente la calidad del aire en las categorías buena, aceptable y mala.
- Identificar qué partículas influyen más sobre la clasificación de la calidad del aire, con el fin de que se puedan enfocar en las fuentes que producen estas partículas.
- Hacer un pronóstico de series de tiempo de las concentraciones de los contaminantes seleccionados
- Identificar si hay diferencia significativa entre las concentraciones de contaminantes en alguna de las tres estaciones de monitoreo.

4. PREPARACIÓN DE LA BASE DE DATOS

4. COMPRENSIÓN DE LOS DATOS

a. Dimensiones del dataset

El archivo se encuentra en formato de Excel, y contiene las siguientes pestañas correspondientes a las diferentes zonas que abarca cada una de las estaciones meteorológicas.

- Sureste
- Noreste
- Centro
- Noroeste
- Suroeste
- Noroeste 2
- Norte
- Suroeste 2
- Sureste 2
- Sureste 3
- Sur
- Norte 2
- Noreste 2
- Noreste 3
- Noroeste 3

- **Columnas:** 17 (16 para NORESTE Y NOROESTE 3 , pues PM2.5 tiene celdas vacías).
- **Registros:** 14,255 (14254 para SUR, NORESTE y 6237 para NOROESTE 3).

b. Descripción de las variables

Nombre	Descripción	Tipo	Valores posibles	Valores nulos
date	Fecha	Numérico	yyyy/mm/dd	no tiene
CO	Monóxido de carbono	Numérico	[0, ∞]	nan
NO	Monóxido de Nitrógeno	Numérico	[0, ∞]	nan
NO2	Dióxido de Nitrógeno	Numérico	[0, ∞]	nan
NOX	NO + NO2	Numérico	[0, ∞]	nan
O3	Ozono	Numérico	[0, ∞]	nan
PM10	Material particulado menor a 10 micrómetros	Numérico	[0, ∞]	nan
PM2.5	Material particulado menor a 2.5 micrómetros.	Numérico	[0, ∞]	nan
PRS	Presión atmosférica	Numérico	[0, ∞]	nan
RAINF	Precipitación	Numérico	[0, ∞]	nan
RH	Humedad relativa	Numérico	[0, 100]	nan
SO2	Dióxido de azufre	Numérico	[0, ∞]	nan
SR	Radiación solar	Numérico	[0, ∞]	nan
TOUT	Temperatura	Numérico	[0, ∞]	nan
WSR	Velocidad del viento	Numérico	[0, ∞]	nan
WDR	Dirección del viento	Numérico	[0-359]	nan

c. Calidad de los datos

Luego de tener un primer análisis exploratorio de los datos se pudo observar presencia de algunos indicadores de mala calidad:

- Presencia de valores nulos: los dataset tenían bastantes datos **NULL** (o **nan** en Python), lo cual fue debido a falta de información de esas instancias.
- Presencia de datos atípicos en algunas variables: algunas variables tenían *outliers* o valores muy altos. Esto es debido al comportamiento del aire y la producción dentro de los procesos de industrialización, sectores agrícolas o contaminaciones cotidianas cercanas, como lo puede ser una *carne asada*, *cigarros*, entre otros.

d. Selección del conjunto de datos a utilizar

Una vez explorados los datos se decidió utilizar únicamente los datos históricos del 2022 al 2023, ya que tenían una mejor calidad y menor presencia de valores nulos. Posteriormente, de acuerdo con la investigación realizada acerca de los contaminantes más relevantes, se decidió trabajar únicamente con las siguientes variables. Cabe resaltar que se mantuvieron todos los registros:

- date
- CO
- NO
- NO2
- O3
- PM10
- PM2.5
- SO2

Luego de la etapa de preprocesamiento, se terminó con distintas observaciones para cada estación. Al final, como se busca analizar por indicadores socioeconómicos, nos quedamos con las siguientes estaciones:

- Noreste: 14 063 registros.
- Noroeste 2: 14 137 registros.
- Sureste 3: 13 957 registros.

PREPARACIÓN DE LOS DATOS

- Identificar las columnas objetivo: Para las columnas objetivo, escogimos aquellas que influyan como contaminantes, las cuales son principalmente: CO, NO, NO2, O3, PM10, PM2.5 y SO2.
- Limpieza de datos: Lo que se hizo fue eliminar todas aquellas filas que tuviesen sus columnas llenas de nan (una fila de puros nan).
- Manejar valores espurios o erróneos: Afortunadamente, en nuestro dataset no había datos erróneos distintos a nan, por lo que solo nos tocó eliminar estos.
- Maneja valores faltantes: Se reemplazó el valor de los nan por el promedio del valor previo válido y valor posterior válido (por válido nos referimos a que sea distinto a nan).
- Manejar los valores atípicos (outliers) que encuentres en el dataset: Los outliers son muy importantes en nuestro análisis, pues nos permite entender el comportamiento de la contaminación del aire, por lo que es importante conservarlos y analizar el porqué de su comportamiento.

TRANSFORMACIÓN DE LOS DATOS

Dado el contexto del problema seleccionado, no se optó aplicar un escalamiento de datos hasta la creación del modelo, pues se busca comprender los distintos comportamientos.

Por otro lado, se buscó transformar los datos por medio de Box-Cox y Yeo-Johnson, aunque los resultados no fueron exitosos, por lo que se trabajó con los datos originales.

Finalmente, debido a que estamos trabajando con datos que pueden llegar a tener outliers que son importantes e influyentes para la distribución de estos, hacer una discretización nos haría perder mucha información.

Nos dimos cuenta que todas las variables que se están tomando en cuenta se tienen en ppb, a excepción de PM10, PM2.5 y CO, por lo que tenemos que dividirlos entre 1000 para que coincida con los datos proporcionados por la OSF, los cuales están en ppm.

MANEJO DE DATOS ERRÓNEOS

El mínimo de PM2.5 no puede ser mayor al mínimo de PM10, lo cual no siempre se cumple. Algunos valores de los contaminantes son negativos, lo cual no es posible. Para la limpieza de estos datos, se modificaron los valores negativos al cambiarlos tomando el valor promedio de sus vecinos. Para el caso de los valores PM2.5 superiores a PM10, se eliminaron aquellas filas erradas.

5. MODELACIÓN, VALIDACIÓN y RESULTADOS OBTENIDOS

Observando los datos y después de analizarlos, llegamos a la conclusión de que los objetivos iniciales que planteamos no eran posibles de lograr debido a la falta de información sobre las enfermedades respiratorias, otro de los impedimentos fue el tiempo con el que se contaba.

Los nuevos objetivos buscan clasificar de manera eficiente la calidad del aire, al igual que identificar qué partículas tienen una mayor influencia a la hora de esta clasificación. Para la realización de esto se empleó una regresión logística multinomial, donde se logra identificar qué partículas tienen un mayor peso, al igual que se crea un clasificador de la calidad del aire. Para la regresión logística multinomial se decidió eliminar el NO como variable predictora, debido a que no se contaba con información para clasificar niveles de óxido de nitrógeno establecidos por normas nacionales o internacionales. Posteriormente, se optó por hacer una predicción de la calidad del aire al hacer un análisis de series de tiempo con la mediana de los datos mensuales. Para realizar las predicciones se intentó aplicar el modelo ARIMA.

I. REGRESIÓN LOGÍSTICA MULTINOMIAL

Con la base de datos ya limpia, llevamos a cabo el análisis previamente mencionado. Inicialmente, dividimos nuestro conjunto de datos en train y test, asignando el 80% para train y el 20% para test. Sin embargo, esta división se realizó de manera aleatoria, dado que no hay un balanceo de datos para cada clase si tomamos los conjuntos cronológicamente. Por ende, se seleccionaron los conjuntos aleatoriamente de forma totalmente aleatoria.

Posteriormente se creó el modelo de regresión logística, donde las clases a predecir son “buena”, “aceptable” y “mala” (en mala se incluyó mala, muy mala y extremadamente mala). Se entrenó el modelo con el train set y por último, se probó con el test set. Observando el summary del modelo concluimos que la partícula que tiene un mayor peso es el PM10, por lo que la recomendación sería enfocarse en aquellas fuentes emisoras de esta partícula.

Cabe recalcar que las estaciones se nombran de distintas maneras, pero representan exactamente la misma estación:

- Estación 2 = Noreste = San Nicolás.
- Estación 6 = Noroeste 2 = García.
- Estación 10 = Sureste 3 = Cadereyta.

Supuestos (para mayor detalle revisar el código de R)

Se cumple la significancia de los coeficientes, balanceo de clases e independencia para la estación 2. No se cumple la multicolinealidad.

Se cumple la significancia de los coeficientes, multicolinealidad, balanceo de clases e independencia para la estación 6.

Se cumple la significancia de los coeficientes, balanceo de clases e independencia para la estación 10. No se cumple la multicolinealidad.

Resultados de la estaciones (para mayor detalle revisar el código de R)

- Tiene una mayor influencia PM10, seguido de PM2.5 en la estación de San Nicolás.
- Tiene una mayor influencia PM10, seguido de PM2.5 en la estación de García.
- Tiene una mayor influencia PM10, seguido de SO2 en la estación de Cadereyta.
- Los tres modelos hicieron una clasificación correcta para más del 90% de los datos del conjunto de prueba.

Se pueden observar los resultados de las predicciones gráficamente en las figuras 4.25, 4.26 y 4.27.

II. PREDICCIÓN DE SERIES DE TIEMPO DE VARIABLES

Para esta tarea, lo primero que se realizó fue el cálculo de la mediana de los datos recopilados para cada una de nuestras variables en 30 días. Posteriormente, graficamos estas mediciones para cada una de las tres estaciones, obteniendo los siguientes resultados.
Al analizar detenidamente cada uno de estos gráficos de la figura 5.1 podemos obtener lo siguiente:

1. **CO (Monóxido de Carbono):** Las líneas parecen fluctuar a lo largo del tiempo. No se puede determinar una tendencia clara; además, no parecen existir semejanzas significativas entre las diferentes estaciones.

2. **NO2 (Dióxido de Nitrógeno):** Las mediciones de las tres estaciones son sumamente similares, además, parecen mostrar un patrón que se repite alrededor de cada año. Además, podemos notar que entre los meses 5 y 10 de cada año parece haber menores concentraciones de este contaminante.

3. **O3 (Ozono):** Se observan semejanzas en las mediciones de las tres estaciones; sin embargo, la estación sureste 3 parece ser la que tiene los más altos niveles. Nuevamente, parece existir un patrón que se repite cada año, pero en esta ocasiones los niveles mínimos se encuentran entre el último y el primer mes del año.

4. **PM10 (Partículas Menores a 10 Micrómetros):** Se observa un patrón claro en las mediciones, y los valores más bajos parecen encontrarse entre los meses 8 y 10 del año. Además, las mediciones entre las estaciones parecen ser significativamente distantes. La estación Noreste 2 es la que mayor reporta estas partículas.

5. **PM2.5 (Partículas Menores a 2.5 Micrómetros):** Nuevamente se observa una especie de patrón en estas partículas y las mediciones entre las estaciones muestran semejanzas respecto a sus patrones.

6. **SO2 (Dióxido de Azufre):** No se observa un patrón claro que se repita a lo largo del año; sin embargo, las mediciones parecen ser semejantes e indicar las mismas subidas y bajadas de esta partícula.

Posteriormente de este primer análisis exploratorio, procedimos a realizar el Modelo Autorregresivo Integrado de Media Móvil (ARIMA); sin embargo, los resultados no fueron tan buenos. Consideramos que esto se debe a que al calcular las medianas de los datos por mes, se terminó con un total de 20 datos, uno por cada mes desde 2022 hasta agosto de 2023, y en ese lapso de tiempo no se generaron los suficientes patrones como para que el modelo aprenda. Se intentaron distintas técnicas pero el mejor resultado obtenido, y aun así insatisfactorio, fue el de la figura 5.2, el cual corresponde a la predicción de la variable que más claro parece seguir un patrón, el dióxido de nitrógeno (NO2).

III. PRUEBA ANOVA ENTRE CONTAMINANTES POR ZONA

Finalmente, con el objetivo de analizar si había una diferencia significativa entre los niveles de contaminantes por zona a lo largo de todo el periodo de tiempo considerado, se optó por realizar una prueba ANOVA.

El primer procedimiento consistió en la creación de un nuevo conjunto de datos que agrupara las mediciones de las tres zonas en un solo data frame con las siguientes columnas:

- Una columna para la fecha y hora
- Una columna para cada uno de los seis contaminantes

- Una columna indicando la zona a la que pertenecen las mediciones

Con esto, se obtuvo un conjunto de datos de **42,206** renglones y **8** columnas.

Luego, se realizó un gráfico de cajas y bigotes para cada uno de los seis contaminantes, agrupándolos por cada una de las tres zonas, además, se aplicó la prueba ANOVA para cada contaminante, comparando las concentraciones del mismo en las tres zonas y con el objetivo de identificar si alguna zona es estadísticamente más contaminada que el resto. El resultado fue el siguiente.

Como se puede observar en la figura 6.1, a pesar de que, a primera vista, pareciera que las concentraciones de cada contaminante son iguales en las tres zonas, la prueba ANOVA nos indica que existe, para cada contaminante, al menos una zona que tiene una concentración estadísticamente diferente del resto. Para identificar esa zona, se aplicó la prueba de Tukey, observando los intervalos de confianza de las diferencias entre zonas.

Como se puede observar en la figura 6.2, en este primer análisis con los datos sin suavizar, se observa que todas las zonas son estadísticamente diferentes para casi todos los contaminantes, con excepción del PM2.5.

Luego, se realizaron los supuestos del modelo para cada uno de los contaminantes, analizando normalidad de residuos, homocedasticidad de residuos e independencia de residuos. **Ninguno de los supuestos de cada modelo fue validado**. Por ello consideremos un suavizamiento de los datos: medianas móviles mensuales.

Posteriormente, aplicando la técnica de suavizamiento por medianas móviles mensuales, para cada agrupación de los contaminantes por cada zona. El resultado fue que, agrupando los datos, se observaron diferencias entre las distintas zonas. Por ello, aplicamos la prueba ANOVA para cada una de los contaminantes, comparándolos en las tres zonas. En la siguiente imagen observamos los gráficos de cajas para cada contaminante, así como el resultado de la prueba ANOVA y el valor p obtenido (se redondeó a tres decimales para la visualización).

Estos primeros resultados (Figura 6.3) nos indican que, para todos los contaminantes, con excepción del dióxido de azufre (SO2), hay al menos una zona que difiere estadísticamente del resto. Ahora bien, para identificar esa zona, realizamos la prueba de Tukey, con el objetivo de identificar el par de zonas cuyas concentraciones del contaminante son estadísticamente diferentes entre sí con un 95 % de confianza, observable en figura 6.4.

En la figura 6.5 de los anexos podemos notar dos cosas significativas. Por un lado, el SO2 es estadísticamente similar en las tres zonas. En segundo lugar, las zonas Noroeste2 y Sureste3 son las que presentan diferencias significativas en la mayoría de los contaminantes analizados. Posteriormente, procedimos a realizar la validación de los supuestos de cada modelo ANOVA creado para cada contaminante. Se realizó el análisis de supuestos para cada uno de los seis modelos creados.

Normalidad de residuos

En primer lugar realizamos pruebas de hipótesis y gráficos para observar la normalidad de los residuos. Los resultados fueron satisfactorios, pues la mayoría de los valores p fueron superiores al nivel de significancia alfa = 0.05, por lo que este supuesto se validó para la mayoría de los casos.

Contaminante	Valor p	Validez del supuesto
CO	0.9236	SÍ
NO2	0.04616	NO
O3	0.1472	SÍ
PM10	0.7262	SÍ
PM2.5	0.7074	SÍ
SO2	0.182	SÍ

Homocedasticidad de residuos

Para la validación de este supuesto, se realizó la prueba de Breusch-Pagan, y los resultados fueron satisfactorios para la mayoría de los modelos. A continuación se reportan los valores p obtenidos y los gráficos de los residuos.

Contaminante	Valor p	Validez del supuesto
CO	0.001178	NO
NO2	0.3513	SÍ
O3	0.8368	SÍ
PM10	0.9075	SÍ
PM2.5	0.1295	SÍ
SO2	0.814	SÍ

Independencia de residuos

Para la validación de este supuesto, se realizó la prueba de Durbin-Watson, y los resultados en esta ocasión no fueron favorables. A continuación se reportan los valores p obtenidos y los gráficos de los residuos.

Contaminante	Valor p	Validez del supuesto
CO	4.134e-07	NO
NO2	3.702e-12	NO
O3	0.0002335	NO
PM10	6.506e-05	NO
PM2.5	0.0683	SÍ
SO2	2.641e-06	NO

Como se puede apreciar en los valores p y en los gráficos de dispersión de los errores, se puede concluir que estos están autocorrelacionados, y era de esperarse, pues se trata de una serie de tiempo y los datos no son independientes entre sí.

En conclusión, puesto que se validaron los supuestos, salvo el de independencia, podemos decir que los modelos de ANOVA, utilizando medianas móviles mensuales, son válidos para uno de los contaminantes con excepción del CO y NO2, pues sus modelos no fueron validados.

6. CONCLUSIONES

1. Impacto del PM10: El PM10 se destacó como el contaminante más influyente en todos los modelos de clasificación estudiados. Su presencia significativa destaca su relevancia en la calidad del aire y su potencial impacto en la salud pública. Esta alta influencia resalta la necesidad de medidas específicas para controlar y reducir las emisiones de partículas suspendidas en el aire.

2. Importancia del dióxido de azufre (SO2): Aunque el dióxido de azufre (SO2) no alcanzó la influencia del PM10, su constante presencia y significancia estadística en las tres zonas evaluadas lo posicionan como un factor relevante para la calidad del aire. Este hallazgo resalta la importancia de monitorear y regular las emisiones de SO2 para mitigar sus efectos adversos en la salud y el medio ambiente.

3. Similaridad del dióxido de azufre (SO2) entre zonas: A pesar de su importancia, el dióxido de azufre (SO2) mostró una similitud estadística en las tres zonas estudiadas. Esta consistencia sugiere una distribución homogénea del SO2 en estas áreas, lo que enfatiza la necesidad de estrategias integrales y coordinadas para abordar este contaminante a nivel regional.

4. Diferencias en concentraciones entre zonas: Las zonas de García y Cadereyta emergieron como las áreas con concentraciones diferentes en la mayoría de los contaminantes evaluados. Esta disparidad destaca la variabilidad geográfica en la calidad del aire, subrayando la importancia de considerar factores locales y específicos al implementar políticas de control de la contaminación.

7. DATOS Y CÓDIGOS EMPLEADOS:

Liga a los datos y códigos empleados: https://drive.google.com/drive/folders/171AoWUOwLUGrm0GTkkUtSP5eO49fJzDB?usp=sharing

8. BIBLIOGRAFÍA:

La Protección Contra Riesgos Sanitarios, C. F. P. (n.d.). Normas Oficiales Mexicanas (NOM) de Calidad del Aire Ambiente. gob.mx. https://www.gob.mx/cofepris/acciones-y-programas/4-normas-oficiales-mexicanas-nom-de-calidad-del-aire-ambiente

United Nations Environment Programme. (n.d.). ¿Cómo se mide la calidad del aire? UNEP. https://www.unep.org/es/noticias-y-reportajes/reportajes/como-se-mide-la-calidad-del-aire

De La Megalópolis, C. A. (n.d.). Contaminantes en el aire que afectan nuestra salud. gob.mx. https://www.gob.mx/comisionambiental/es/articulos/contaminantes-en-el-aire-que-afectan-nuestra-salud?idiom=es

Roldán, L. F. (2020, January 30). Qué es y cómo se mide la calidad del aire. ecologiaverde.com. https://www.ecologiaverde.com/que-es-y-como-se-mide-la-calidad-del-aire-2423.html

Editor rdu. (2019, May 20). Calidad del aire y monitoreo atmosférico - RDU UNAM. RDU UNAM. https://www.revista.unam.mx/2019v20n3/calidad-del-aire-y-monitoreo-atmosferico/

Ecozap. (2023, May 6). Identificando el contaminante más dañino para la salud. https://ecozap.es/moda-etica-sostenible/identificando-el-contaminante-mas-danino-para-la-salud/

Digital, M. (2022, March 31). ¡Entérate! Así se mide la calidad del aire. Grupo Milenio. https://www.milenio.com/estados/calidad-del-aire-como-se-mide-y-que-es

Cantú Martinez, P. C., & Gómez Guzmán, L. G. (s. f.). Determinación del Nivel de Vida por Municipio en el Estado de Nuevo León, México. Coordinación General de Investigación, Facultad de Salud Pública y Nutrición (UANL). https://respyn.uanl.mx/index.php/respyn/article/download/34/34/67

ANEXOS:

Figuras:

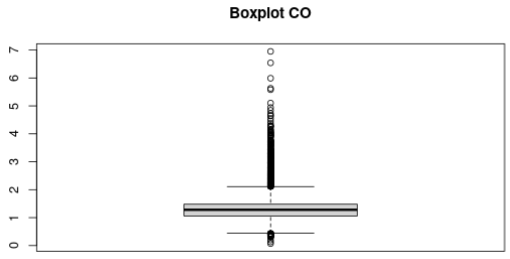


Figura 1.1: Gráfico de caja y bigotes para CO estación noreste

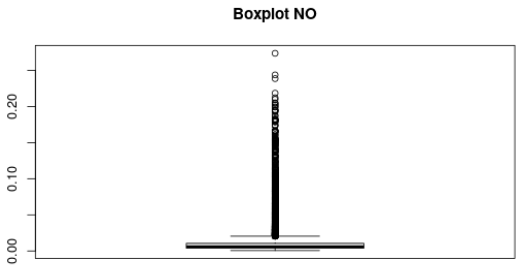


Figura 1.2: Gráfico de caja y bigotes para NO estación noreste

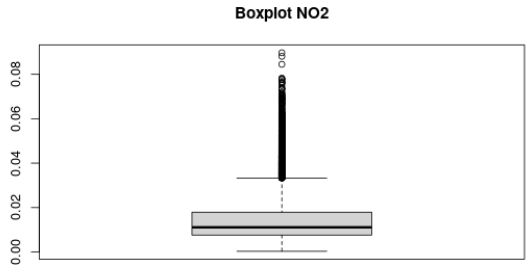


Figura 1.3: Gráfico de caja y bigotes para NO2 estación noreste

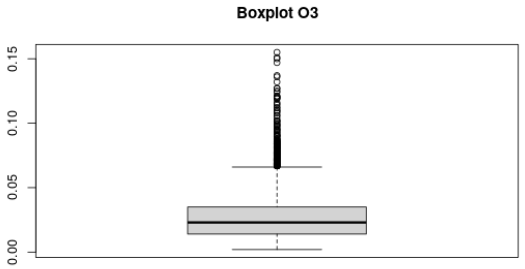
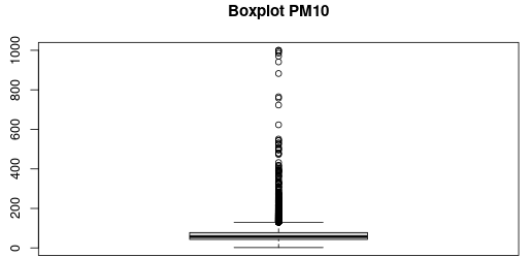


Figura 1.4: Gráfico de caja y bigotes para O3 estación noreste



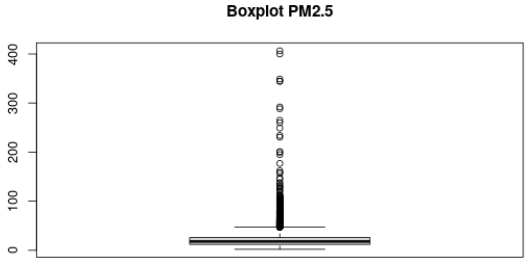


Figura 1.5: Gráfico de caja y bigotes para PM10 estación noreste

Figura 1.6: Gráfico de caja y bigotes para PM2.5 estación noreste

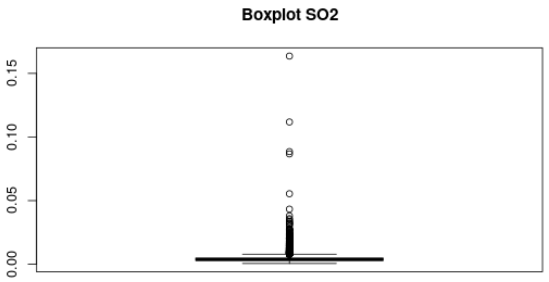
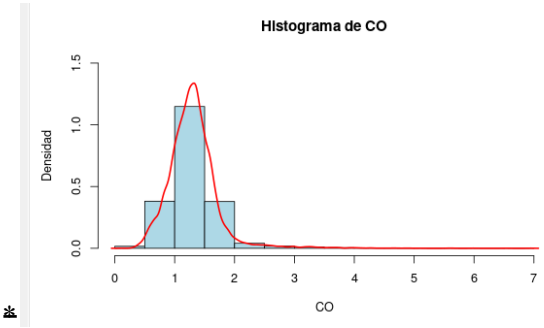


Figura 1.7: Gráfico de caja y bigotes para SO2 estación noreste



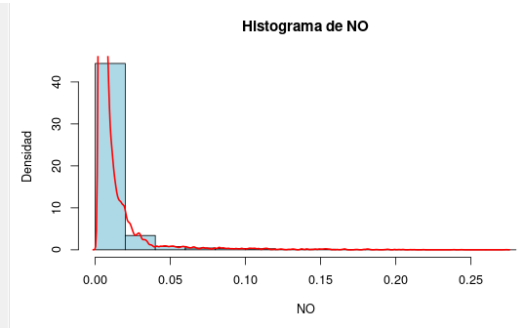


Figura 1.8: Histograma datos de CO estación noreste

Figura 1.9: Histograma datos de NO estación noreste

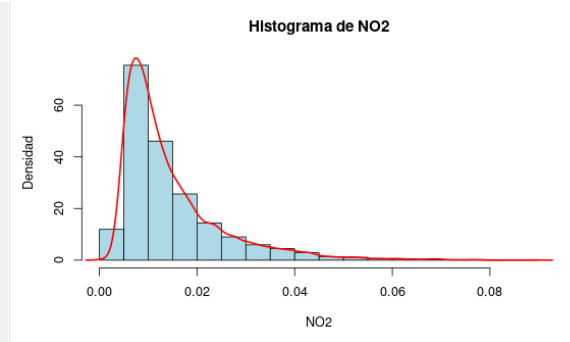


Figura 1.10: Histograma datos de NO2 estación noreste

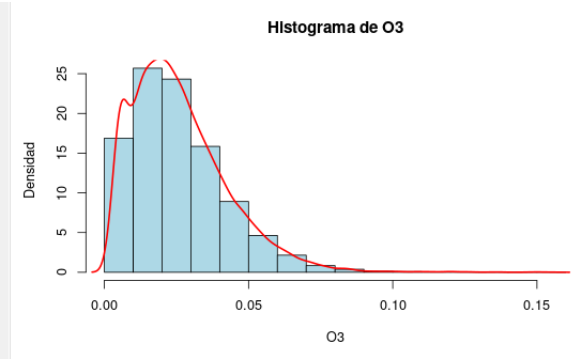


Figura 1.11: Histograma datos de O3 estación noreste

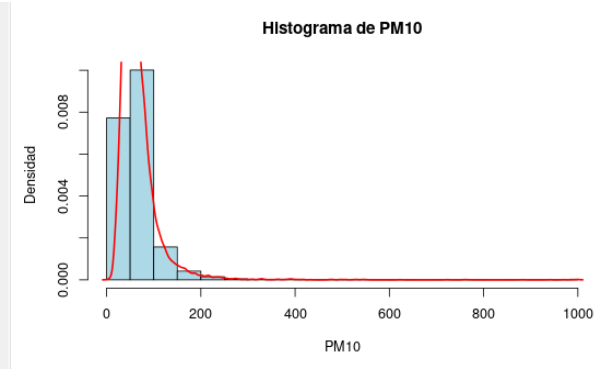


Figura 1.12: Histograma datos de PM10 estación noreste

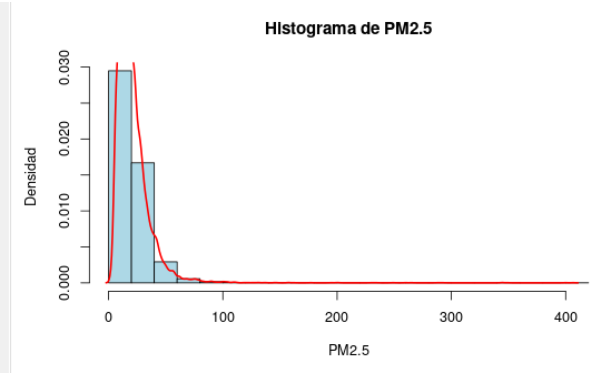


Figura 1.13: Histograma datos de PM2.5 estación noreste

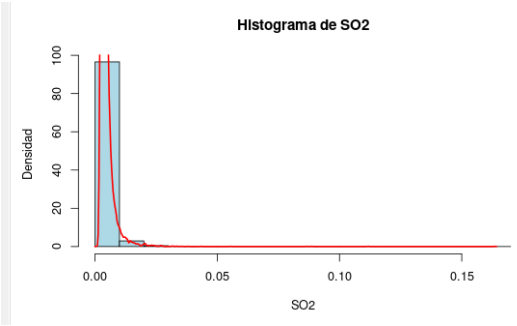


Figura 1.14: Histograma datos de SO2 estación noreste

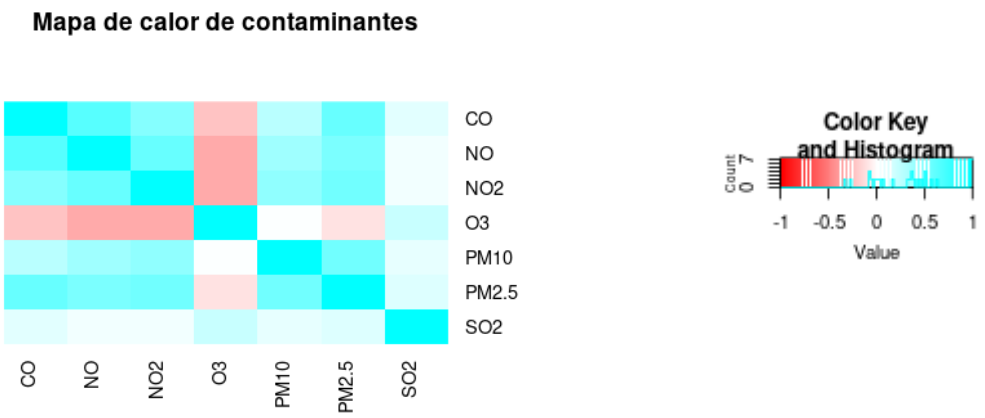


Figura 1.15: Matriz de correlación de los contaminantes estación noreste

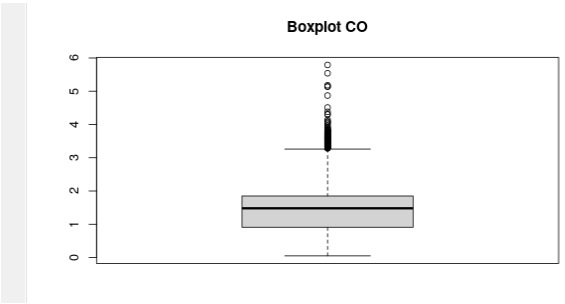


Figura 2.1: Gráfico de caja y bigotes para CO estación noroeste

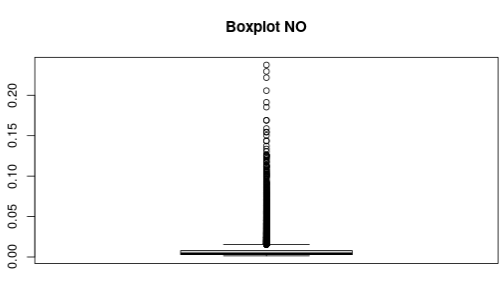


Figura 2.2: Gráfico de caja y bigotes para NO estación noroeste

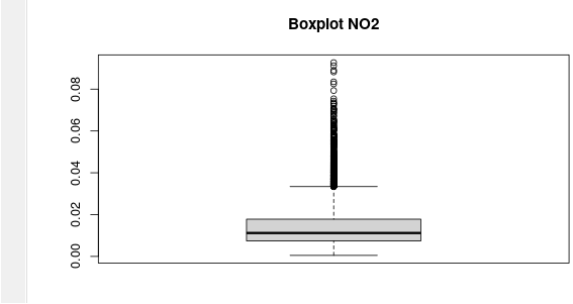


Figura 2.3: Gráfico de caja y bigotes para NO2 estación noroeste

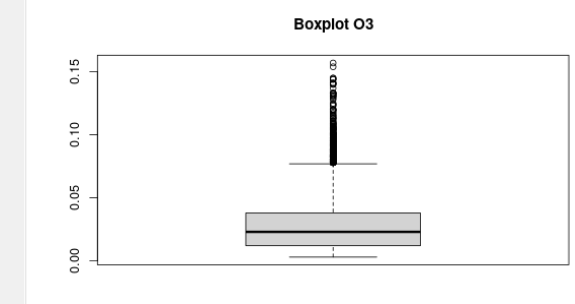


Figura 2.4: Gráfico de caja y bigotes para O3 estación noroeste

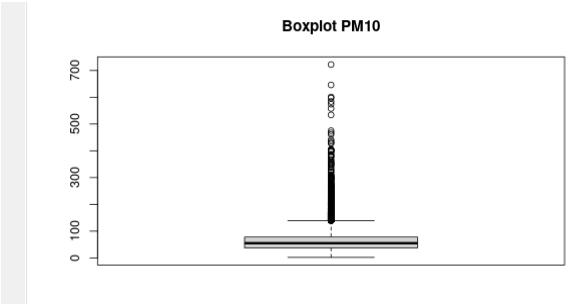


Figura 2.5: Gráfico de caja y bigotes para PM10 estación noroeste

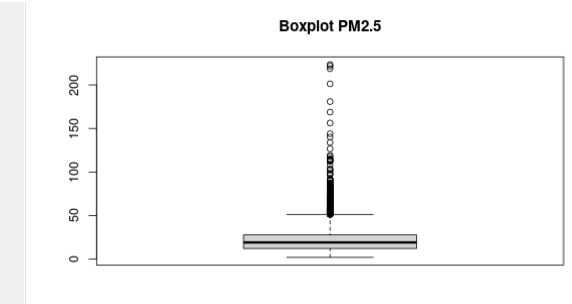


Figura 2.6: Gráfico de caja y bigotes para PM2.5 estación noroeste

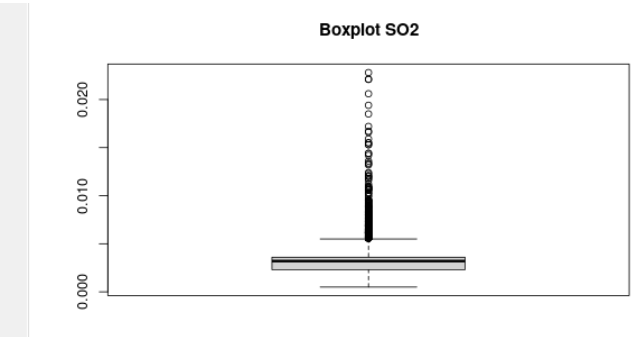


Figura 2.7: Gráfico de caja y bigotes para SO2 estación noroeste

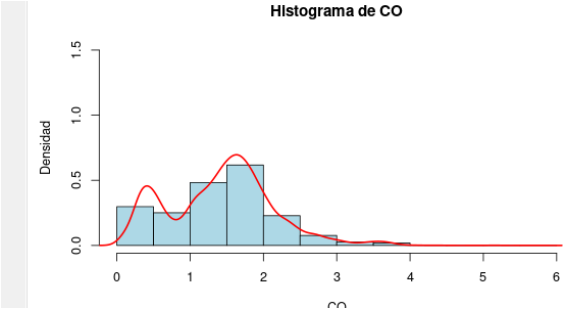


Figura 2.8: Histograma datos de CO estación noreste

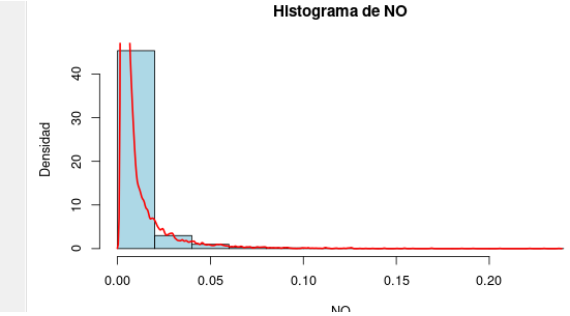


Figura 2.9: Histograma datos de NO estación noreste

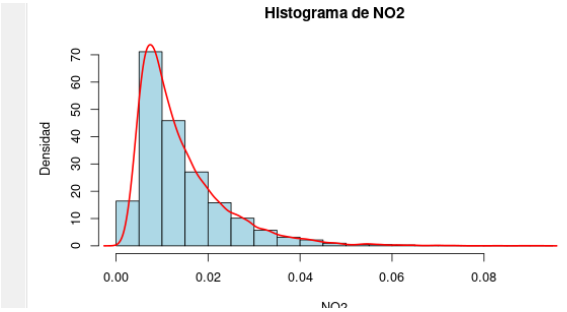


Figura 2.10: Histograma datos de NO2 estación noreste

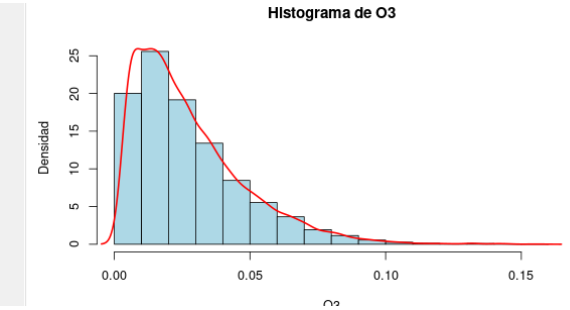
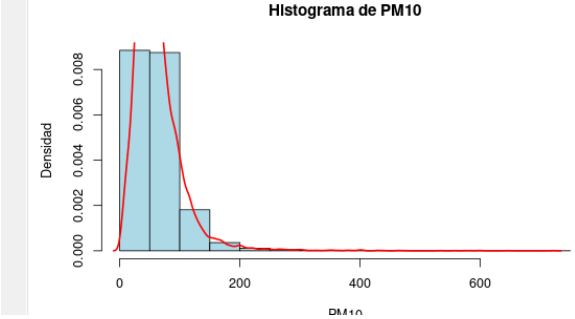


Figura 2.11: Histograma datos de O3 estación noreste



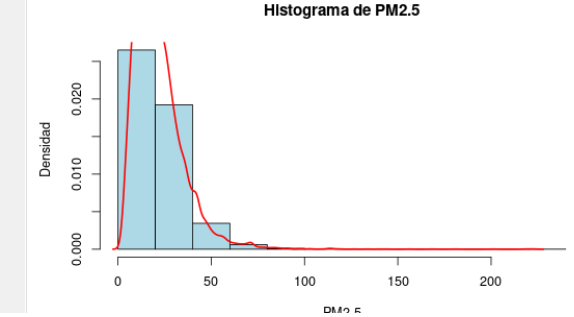


Figura 2.12: Histograma datos de PM10 estación noreste

Figura 2.13: Histograma datos de PM2.5 estación noreste

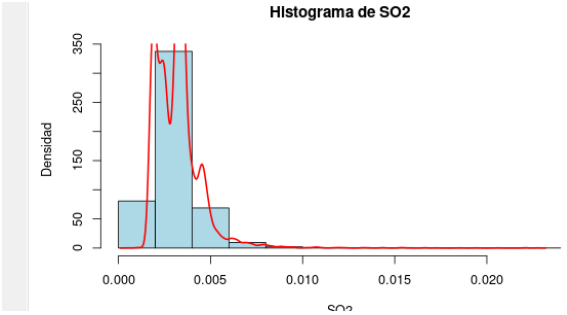


Figura 2.14: Histograma datos de SO2 estación noreste

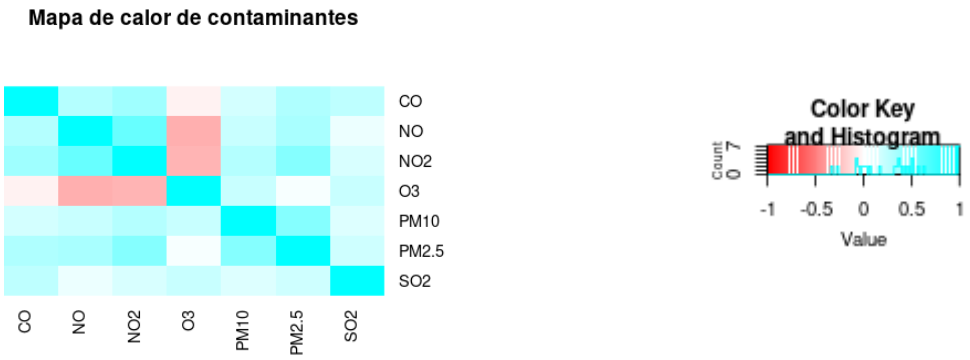


Figura 2.15: Matriz de correlación de los contaminantes estación noroeste

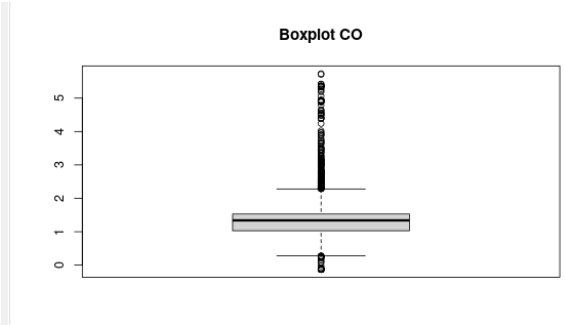


Figura 3.1: Gráfico de caja y bigotes para CO estación sureste

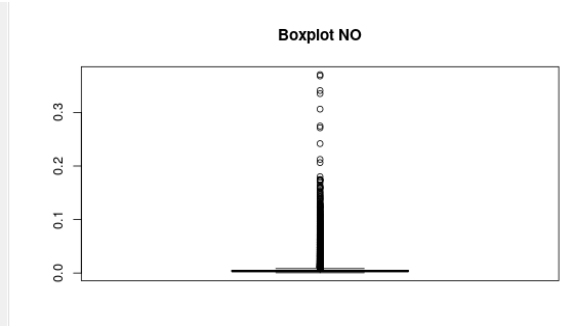


Figura 3.2: Gráfico de caja y bigotes para NO estación sureste

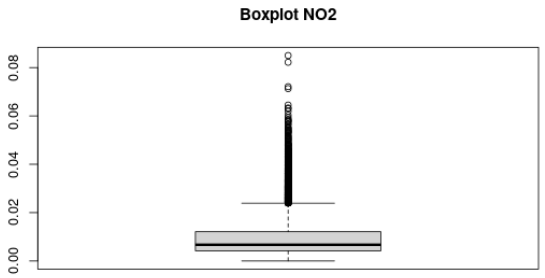


Figura 3.3: Gráfico de caja y bigotes para NO2 estación sureste

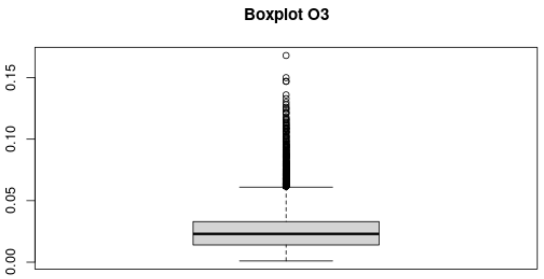
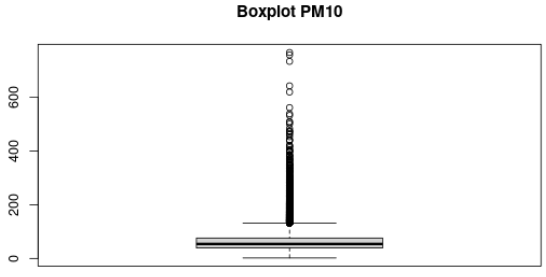


Figura 3.4: Gráfico de caja y bigotes para O3 estación sureste



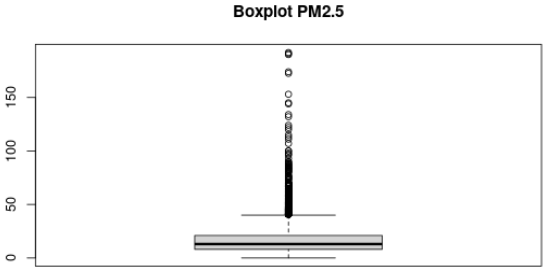


Figura 3.5: Gráfico de caja y bigotes para PM10 estación sureste

Figura 3.6: Gráfico de caja y bigotes para PM2.5 estación sureste

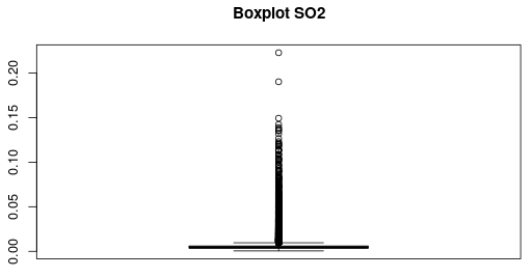
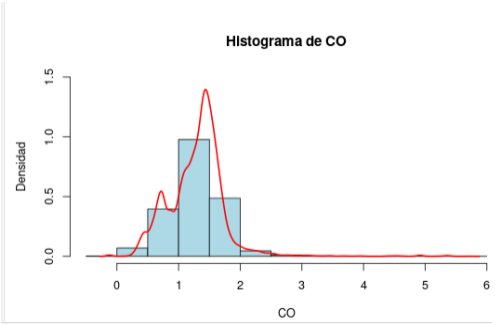


Figura 3.7: Gráfico de caja y bigotes para SO2 estación sureste



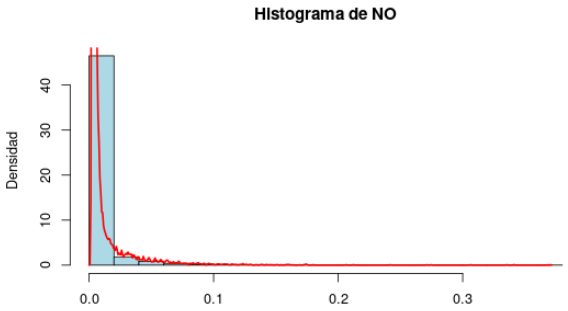


Figura 2.8: Histograma datos de CO estación sureste

Figura 2.9: Histograma datos de NO estación sureste

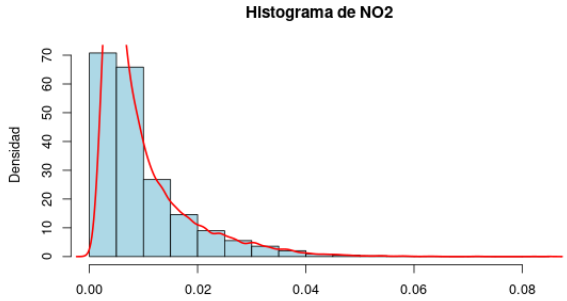


Figura 2.10: Histograma datos de NO2 estación sureste

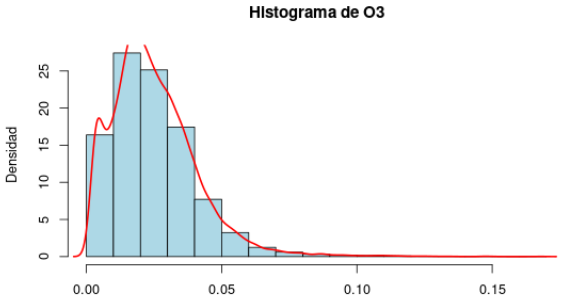


Figura 2.11: Histograma datos de O3 estación sureste

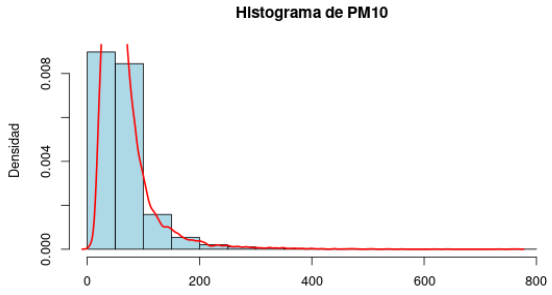


Figura 2.12: Histograma datos de PM10 estación sureste

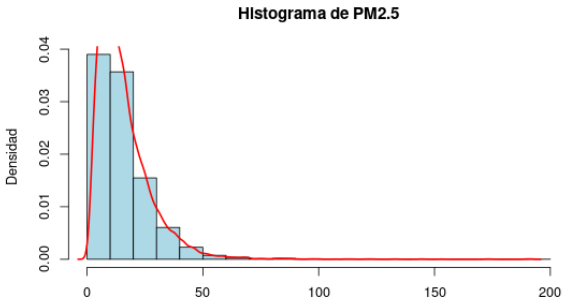


Figura 2.13: Histograma datos de PM2.5 estación sureste

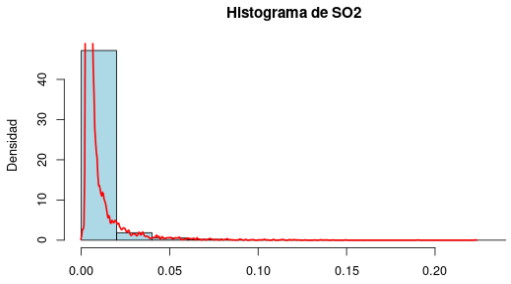


Figura 2.14: Histograma datos de SO2 estación sureste

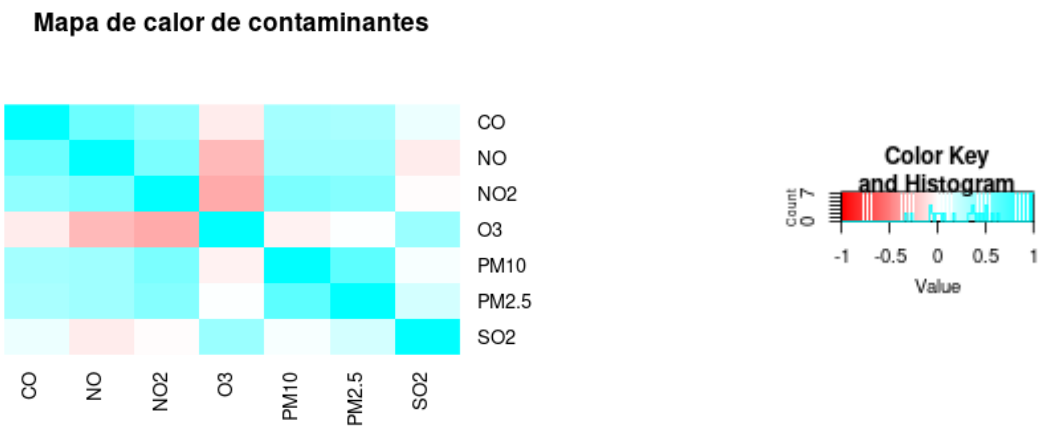


Figura 3.15: Matriz de correlación de los contaminantes estación sureste

```
Coefficients :
               Estimate Std. Error z-value  Pr(>|z|)
(Intercept):2  6.936313   0.189698 36.5650 < 2.2e-16 ***
(Intercept):3  3.088842   0.236579 13.0563 < 2.2e-16 ***
PM10:2        15.355692   0.427775 35.8966 < 2.2e-16 ***
PM10:3        34.696135   0.807110 42.9881 < 2.2e-16 ***
PM2.5:2        2.146627   0.142292 15.0861 < 2.2e-16 ***
PM2.5:3        2.876488   0.186724 15.4051 < 2.2e-16 ***
NO2:2          0.410430   0.089786  4.5712 4.849e-06 ***
NO2:3          0.461671   0.123396  3.7414 0.0001830 ***
CO:2          -0.257549   0.071694 -3.5923 0.0003278 ***
CO:3          -0.374931   0.120911 -3.1009 0.0019294 **
SO2:2          1.311173   0.093313 14.0514 < 2.2e-16 ***
SO2:3          1.607390   0.115618 13.9026 < 2.2e-16 ***
O3:2           0.855452   0.065550 13.0504 < 2.2e-16 ***
O3:3           0.859372   0.093220  9.2187 < 2.2e-16 ***
---
```

Figura 4.1: Coeficientes y su significancia para la estación noreste

```
Coefficients :
              Estimate Std. Error z-value  Pr(>|z|)
(Intercept):2  4.696717   0.125015 37.5692 < 2.2e-16 ***
(Intercept):3 -0.948780   0.249353 -3.8050 0.0001418 ***
PM10:2         9.316285   0.250852 37.1386 < 2.2e-16 ***
PM10:3        27.516071   0.731519 37.6150 < 2.2e-16 ***
PM2.5:2        3.103060   0.109585 28.3165 < 2.2e-16 ***
PM2.5:3        4.572821   0.150125 30.4602 < 2.2e-16 ***
NO2:2         -0.022723   0.074884 -0.3034 0.7615499
NO2:3         -0.353455   0.113501 -3.1141 0.0018450 **
CO:2          -0.171514   0.049087 -3.4941 0.0004757 ***
CO:3          -0.291989   0.091921 -3.1765 0.0014905 **
SO2:2          0.023919   0.061871  0.3866 0.6990605
SO2:3          0.120478   0.091490  1.3168 0.1878927
O3:2           1.062207   0.071201 14.9183 < 2.2e-16 ***
O3:3           0.987789   0.103020  9.5883 < 2.2e-16 ***
```

Figura 4.2: Coeficientes y su significancia para la estación noroeste 2

```
Coefficients :
              Estimate Std. Error z-value  Pr(>|z|)
(Intercept):2  8.707063   0.253275 34.3780 < 2.2e-16 ***
(Intercept):3  2.073143   0.410314  5.0526 4.359e-07 ***
PM10:2        16.843787   0.491865 34.2447 < 2.2e-16 ***
PM10:3        49.175175   1.632206 30.1280 < 2.2e-16 ***
PM2.5:2        0.608256   0.110138  5.5226 3.339e-08 ***
PM2.5:3        0.795816   0.166131  4.7903 1.665e-06 ***
NO2:2          0.290056   0.087376  3.3196 0.0009013 ***
NO2:3          0.553972   0.147807  3.7479 0.0001783 ***
CO:2          -0.189392   0.059646 -3.1753 0.0014969 **
CO:3          -0.262113   0.108786 -2.4094 0.0159772 *
SO2:2         10.085343   0.376673 26.7748 < 2.2e-16 ***
SO2:3         10.596091   0.385782 27.4665 < 2.2e-16 ***
O3:2           0.368120   0.073152  5.0323 4.847e-07 ***
O3:3           0.691670   0.125879  5.4947 3.914e-08 ***
```

Figura 4.3: Coeficientes y su significancia para la estación sureste 3

```
       CO       NO2        O3      PM10     PM2.5       SO2
1.765987  2.561685  2.278391 18.087169  3.653306  2.389287
```

Figura 4.4: Valores de la prueba de VIF para la estación noreste

```
       CO       NO2        O3      PM10     PM2.5       SO2
1.653589 2.750671 2.900658 7.556254 3.999814 2.140225
```

Figura 4.5: Valores de la prueba de VIF para la estación noroeste 2

```
       CO       NO2        O3      PM10     PM2.5       SO2
1.515562  2.066725  2.008428 16.335324  2.450741 32.893511
```

Figura 4.6: Valores de la prueba de VIF para la estación sureste 3

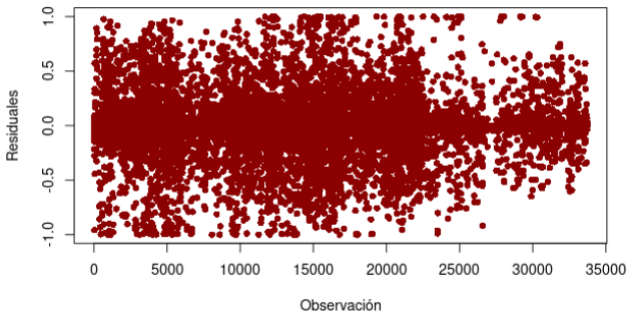


Figura 4.7: Gráfico de residuos de la estación noreste

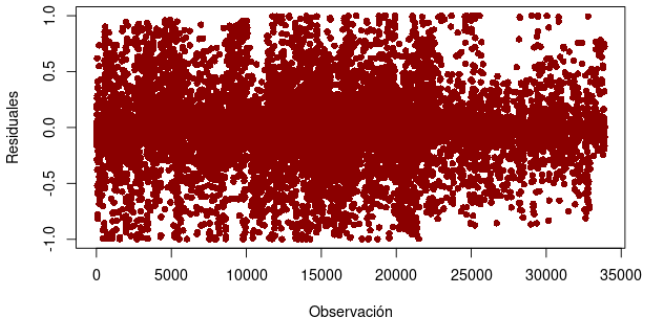


Figura 4.8: Gráfico de residuos de la estación noroeste 2

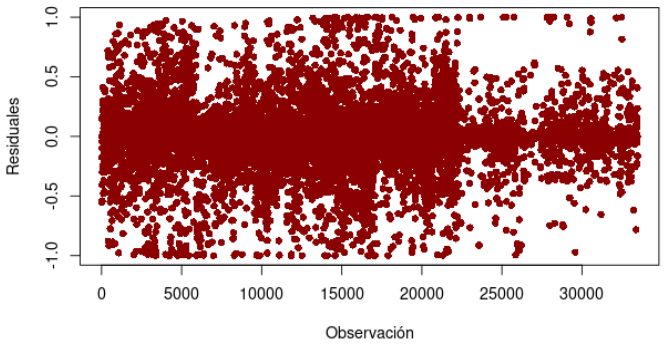


Figura 4.9: Gráfico de residuos de la estación sureste 3

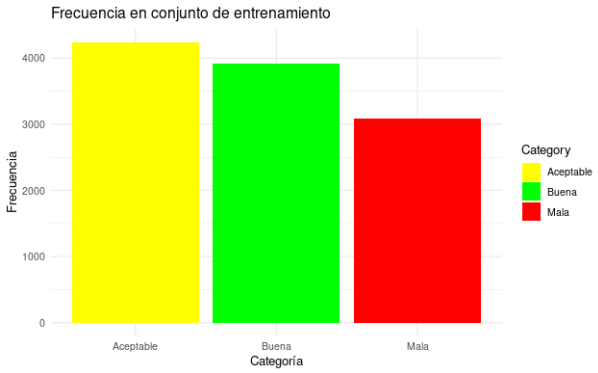


Figura 4.10: Gráfico de cantidad de cada clase en el conjunto de entrenamiento de la estación noreste.

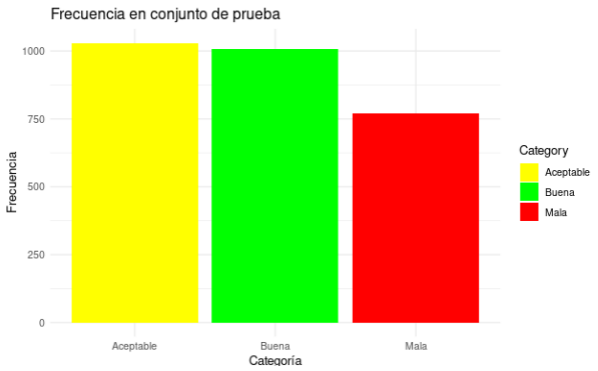


Figura 4.11: Gráfico de cantidad de cada clase en el conjunto de prueba de la estación noreste.

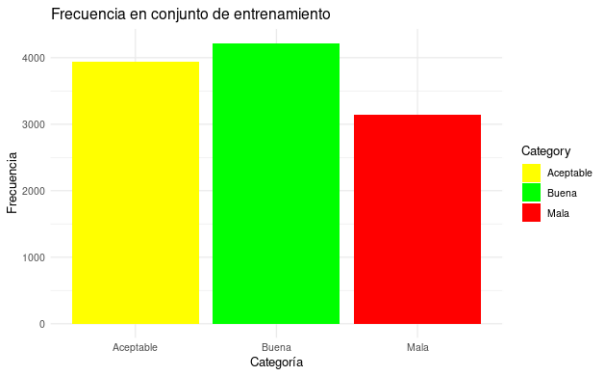


Figura 4.12: Gráfico de cantidad de cada clase en el conjunto de entrenamiento de la estación noroeste 2.

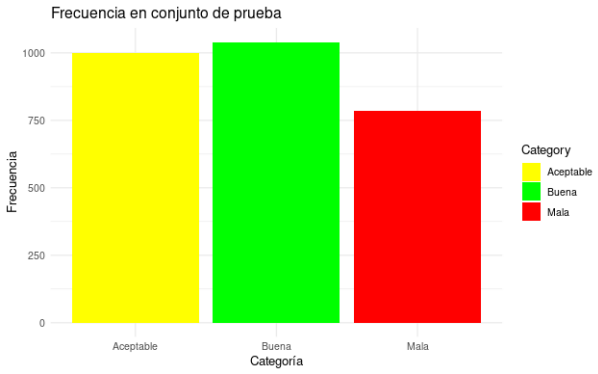


Figura 4.13: Gráfico de cantidad de cada clase en el conjunto de prueba de la estación noroeste 2.

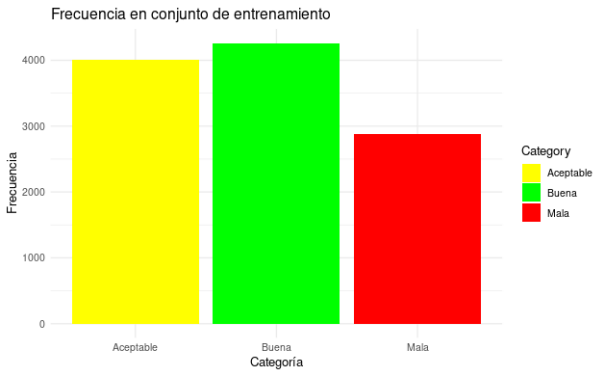


Figura 4.14: Gráfico de cantidad de cada clase en el conjunto de entrenamiento de la estación sureste 3.

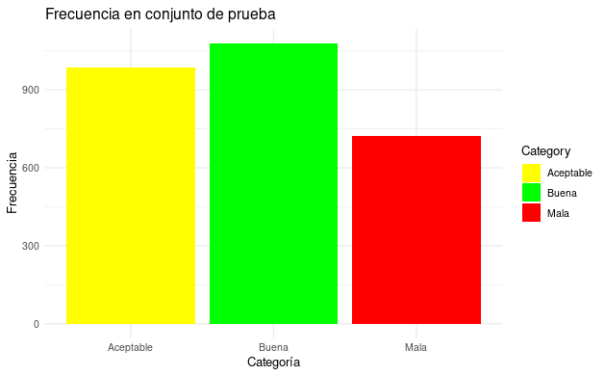


Figura 4.15: Gráfico de cantidad de cada clase en el conjunto de prueba de la estación sureste 3.

```
Coefficients:
  (Intercept)         CO       NO2        O3     PM10    PM2.5      SO2
2   7.6484090 -0.2741527 0.4484481 0.9325788 16.95646 2.350668 1.438665
3   0.9240243 -0.5001854 0.4757288 0.8639782 51.48147 3.587357 1.977008
```

Figura 4.16: Coeficientes de cada contaminante en estación noreste.

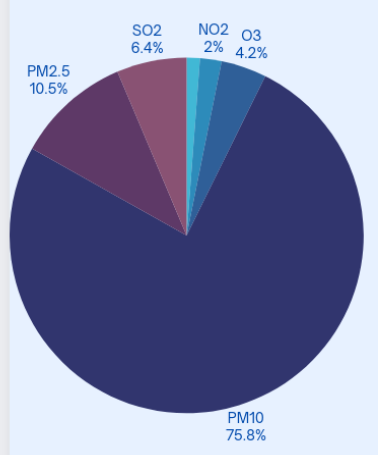


Figura 4.17: Gráfico de pastel de los coeficientes de cada contaminante en la estación noreste.

```
Coefficients:
  (Intercept)         CO         NO2         O3      PM10    PM2.5        SO2
2   4.6942277 -0.1714211 -0.02268051 1.0617656  9.311512 3.101541 0.02407457
3  -0.9482067 -0.2918934 -0.35307053 0.9871972 27.501278 4.570420 0.12056923
```

Figura 4.18: Coeficientes de cada contaminante en estación noroeste 2.

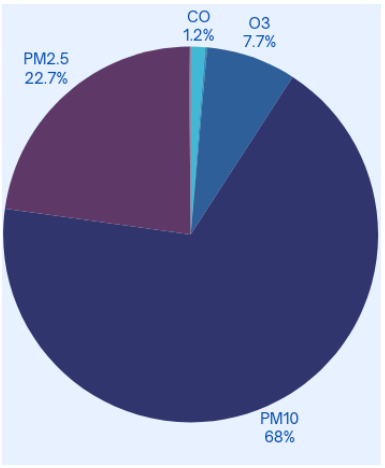


Figura 4.19: Gráfico de pastel de los coeficientes de cada contaminante en la estación noroeste 2.

```
Coefficients:
  (Intercept)         CO       NO2         O3     PM10     PM2.5      SO2
2    8.714637 -0.1895452 0.2902798 0.3684661 16.85803 0.6087202 10.09500
3    1.061331 -0.2598839 0.5887874 0.7446400 54.23093 0.8348314 10.67199
```

Figura 4.20: Coeficientes de cada contaminante en estación sureste 3.

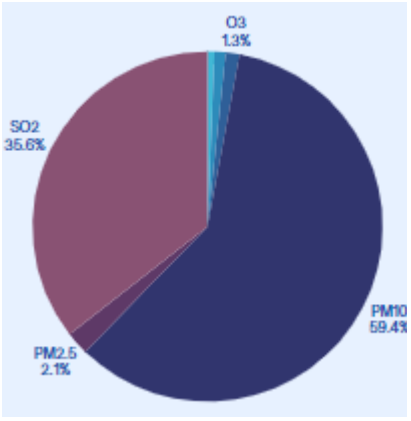


Figura 4.21: Gráfico de pastel de los coeficientes de cada contaminante en la estación sureste 3.

```
predictions   1   2   3
          1 943  56   0
          2  65 962  19
          3   0  12 752
Accuracy: 95.16476 %
Clasificados correctamente 1° clase: 93.55159 %
Clasificados correctamente 2° clase: 93.39806 %
Clasificados correctamente 3° clase: 97.53567 %
```

Figura 4.22: Tabla de predicciones contra resultados reales para la estación noreste.

```
predictions   1   2   3
          1 961  85   0
          2  79 891  32
          3   0  23 753
Accuracy: 93.30229 %
Clasificados correctamente 1° clase: 92.40385 %
Clasificados correctamente 2° clase: 89.18919 %
Clasificados correctamente 3° clase: 95.92357 %
```

Figura 4.23: Tabla de predicciones contra resultados reales para la estación noroeste 2.

```
predictions    1    2    3
          1 1015   65    0
          2   65  912   10
          3    0   10  711
Accuracy: 94.48424 %
Clasificados correctamente 1° clase: 93.98148 %
Clasificados correctamente 2° clase: 92.40122 %
Clasificados correctamente 3° clase: 98.61304 %
```

Figura 4.24: Tabla de predicciones contra resultados reales para la estación sureste 3.

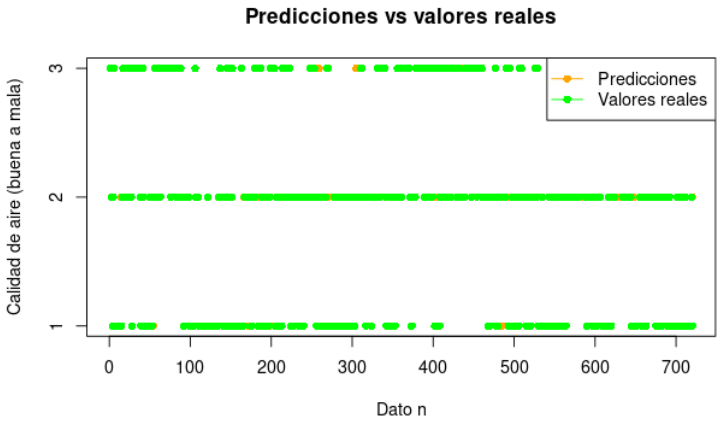


Figura 4.25: Gráfico de comparación de predicciones contra valores reales en estación noreste.

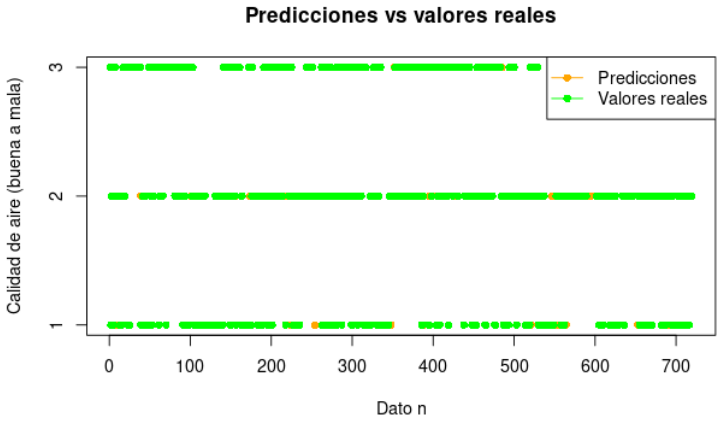


Figura 4.26: Gráfico de comparación de predicciones contra valores reales en estación noroeste 2.

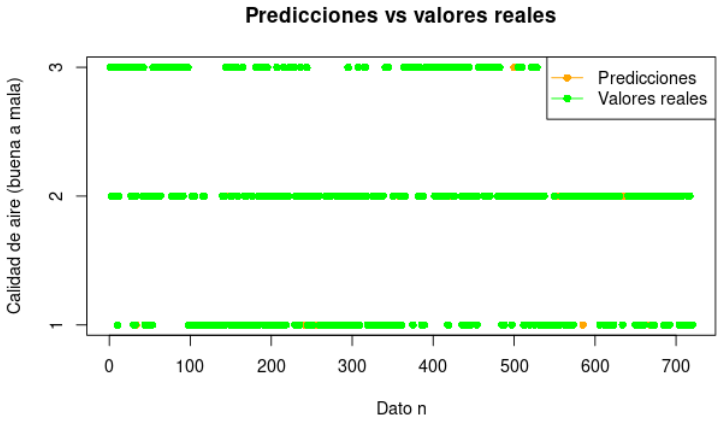


Figura 4.27: Gráfico de comparación de predicciones contra valores reales en estación sureste 3.

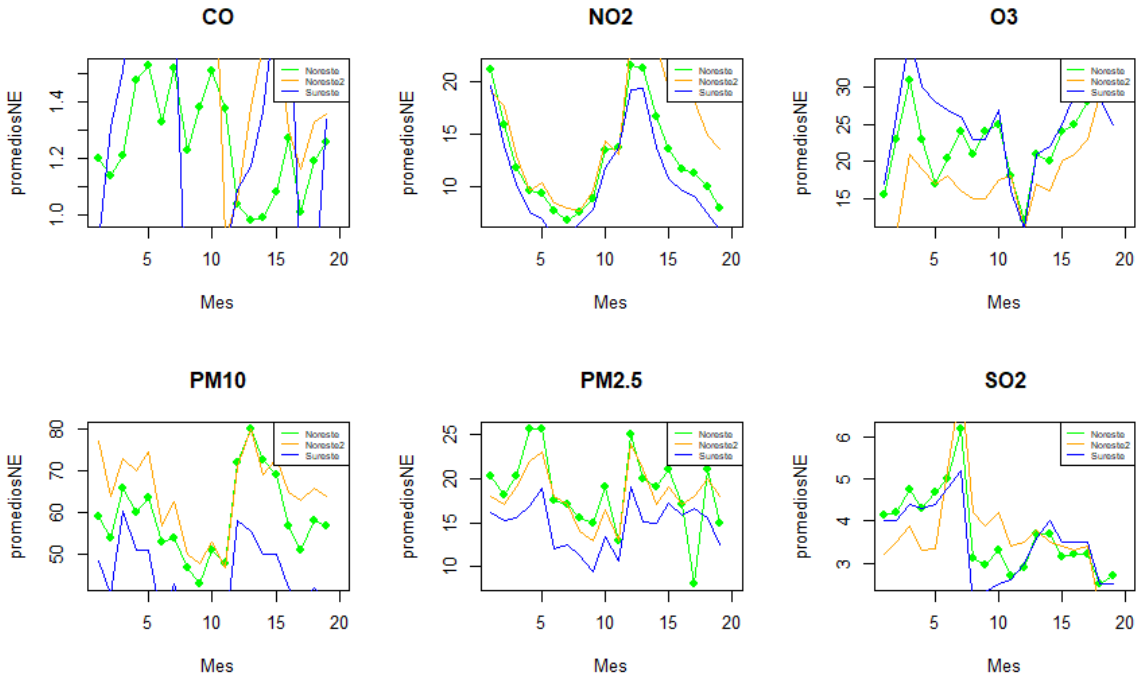


Figura 5.1: Gráfico da valor de la mediana cada 30 días para cada contaminante en cada estación.

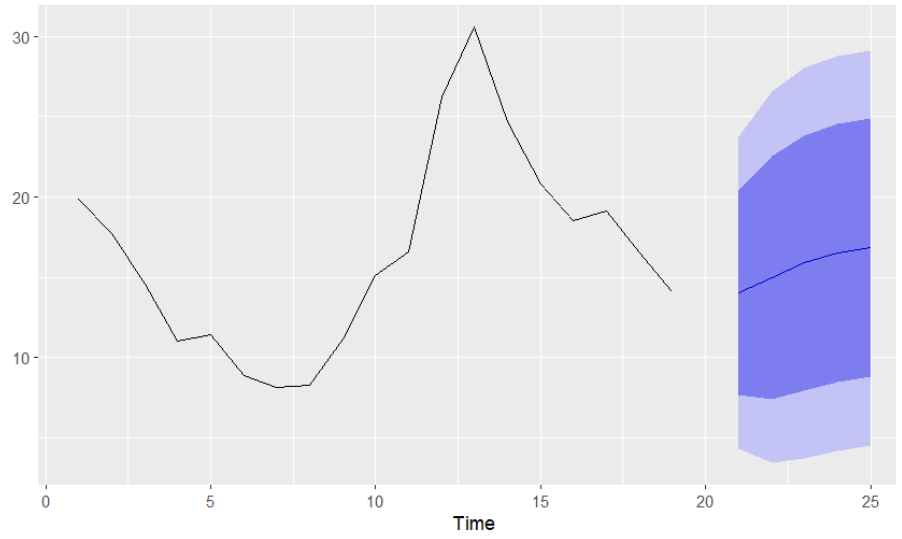


Figura 5.2: Gráfico del modelo de ARIMA de la mediana de cada 30 días.

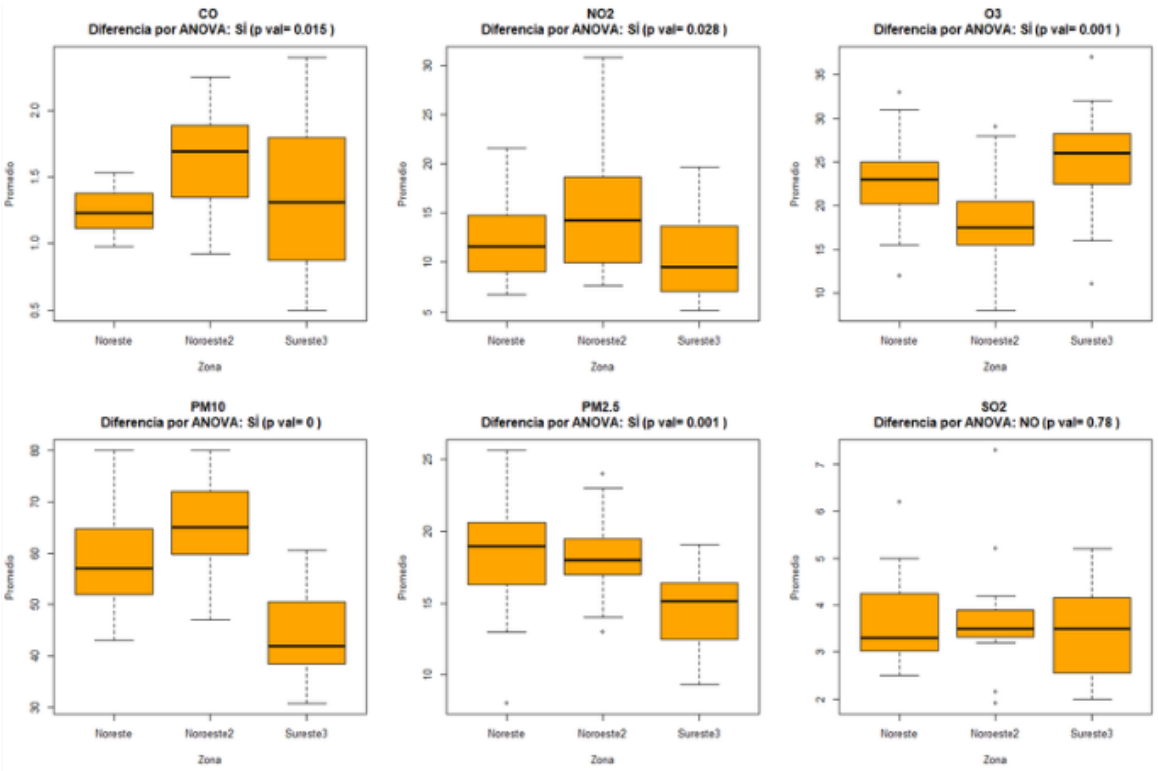


Figura 6.1: Gráfico de caja y bigotes para las concentraciones de cada contaminante comparando cada estación.

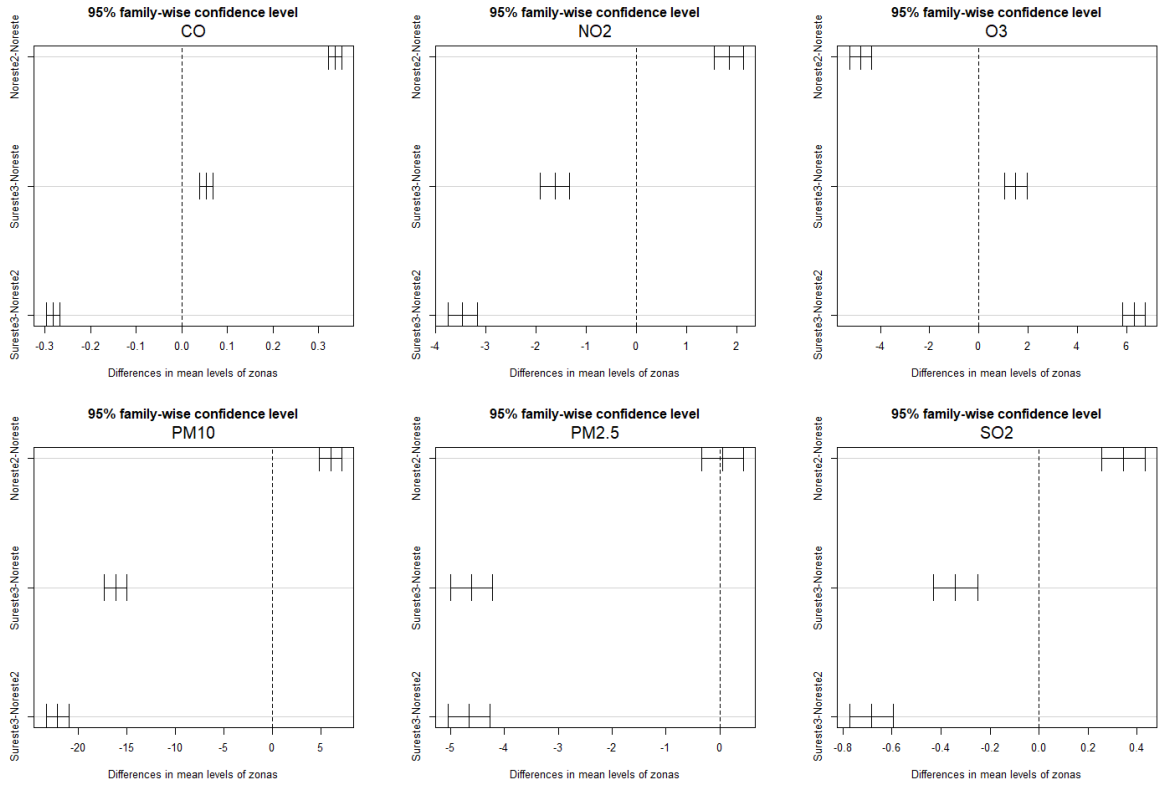


Figura 6.2: Intervalos de confianza de diferencia de medias de cada estación por cada contaminante.

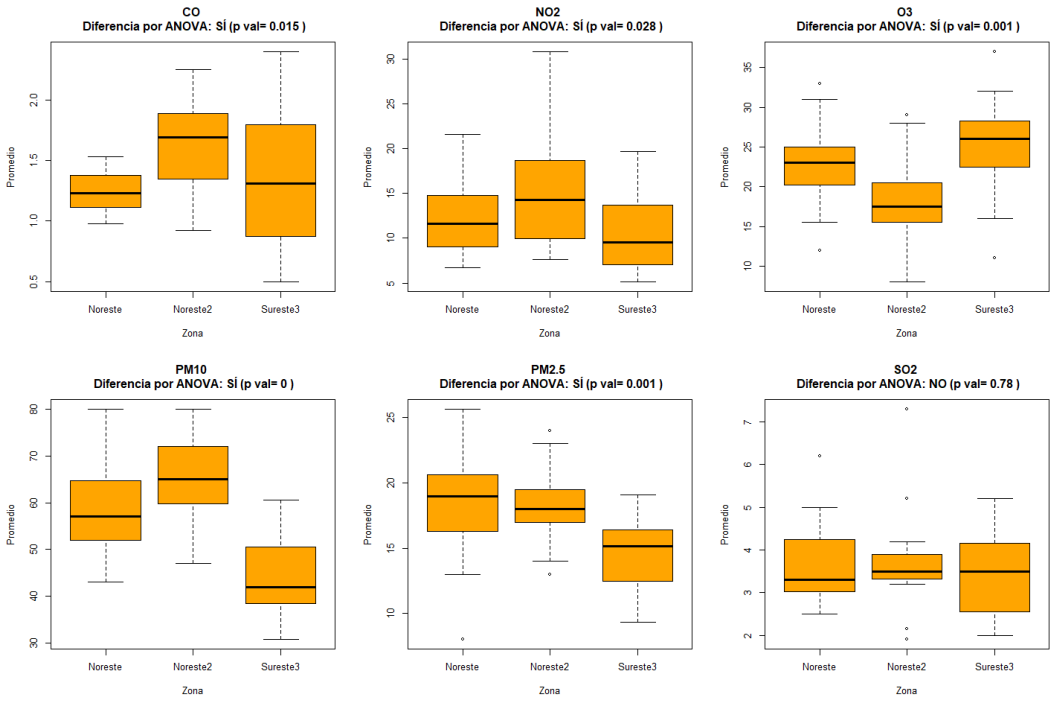


Figura 6.3: Gráfico de caja y bigotes para las concentraciones de cada contaminante luego de la

aplicación de suavizamiento de medianas móviles mensuales comparando cada estación.

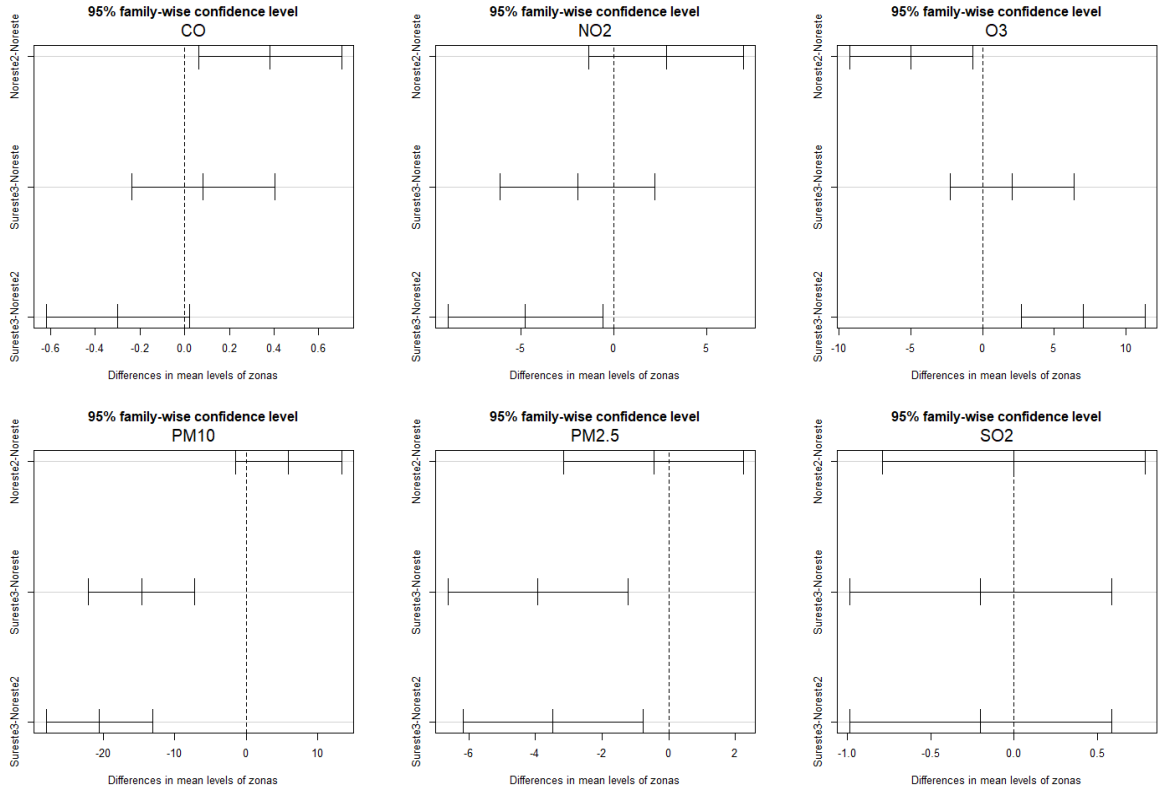


Figura 6.4: Intervalos de confianza de diferencia de medias de cada estación por cada contaminante.

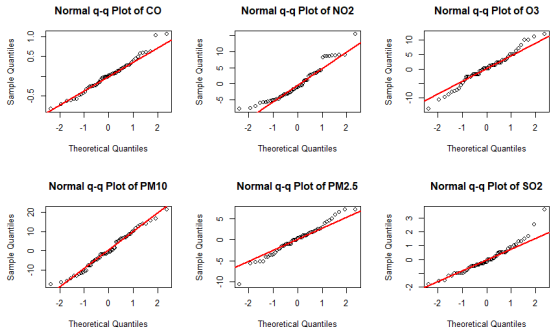


Figura 7.1: Gráfico QQ plot para verificar normalidad de residuos.

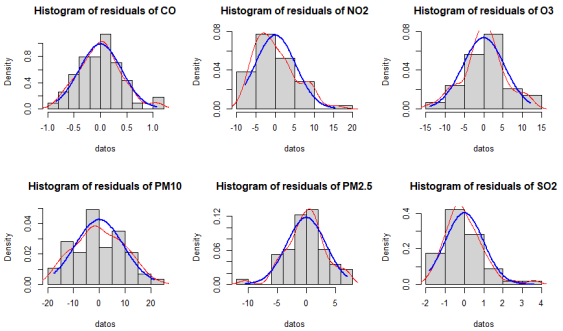


Figura 7.2: Histograma de los residuos del modelo ANOVA por cada contaminante.

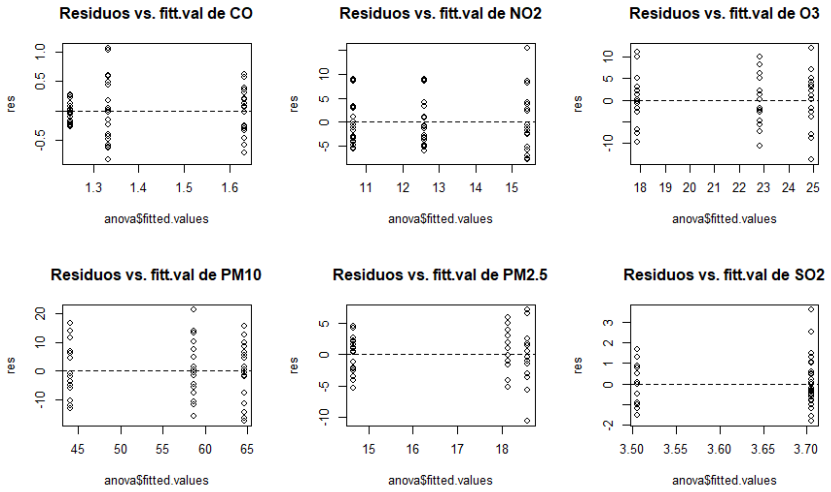


Figura 7.3: Gráfico de residuos de cada contaminante para verificar homocedasticidad.

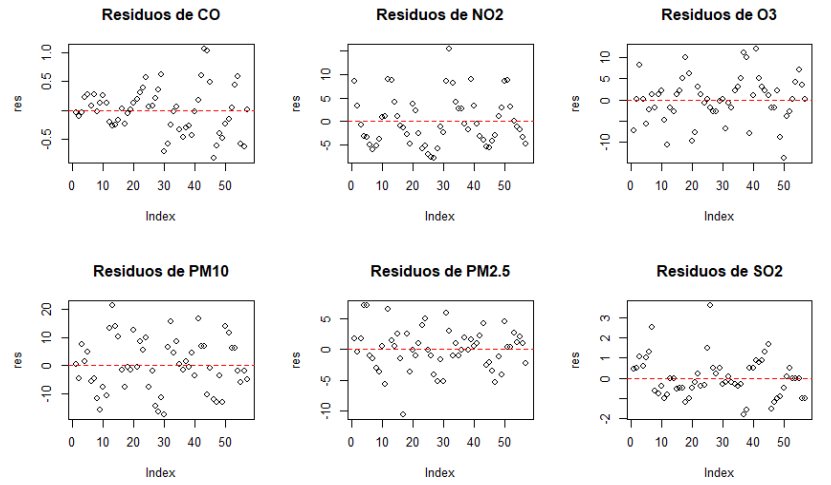


Figura 7.4: Gráfico de independencia de los residuos para cada contaminante para verificar independencia.