



INSTITUTO TECNOLÓGICO Y DE ESTUDIOS SUPERIORES DE
MONTERREY

**Conocimiento de la naturaleza de contaminantes
que influyen en la calidad del aire y sus
interrelaciones en San Nicolás**

Etapa 2. Comprensión y Preparación de los datos

APLICACIÓN DE MÉTODOS MULTIVARIADOS EN CIENCIA DE DATOS

Cantú Rodríguez Pamela A01285128

Ferreira Guadarrama Emiliano A01654418

Núñez López Daniel Isaac A01654137

Ugalde Jiménez Ana Sofía A01702639

Vázquez Álvarez César Guillermo A01197857

en conjunto con SIMA

Supervisado por

José Armando Albert Huerta

Rubí Isela Gutiérrez López

Monterrey, Nuevo León, 9 de noviembre de 2022

Índice

| | |
|--|-----------|
| 1. Resumen | 3 |
| 2. Introducción | 4 |
| 3. Resumen de la revisión de bibliografía | 7 |
| 4. Descripción del problema específico (preguntas de investigación) | 10 |
| 5. Objetivos | 11 |
| 6. Justificación | 11 |
| 7. Mercado potencial (descripción general) | 14 |
| 8. Identificación de clientes/consumidor y usuarios | 15 |
| 9. Descripción de las fuentes de información (datos) | 16 |
| 10. Selección del Modelo | 16 |
| 11. Descripción de la solución | 17 |
| 12. Propuesta de valor | 18 |
| 13. Nombre Detallado del proyecto | 18 |
| 14. Nombre corto o comercial del proyecto | 18 |
| 15. Impacto Social Principal | 19 |
| 16. Impacto hacia los Objetivos de Desarrollo Sostenible | 19 |
| 17. Línea del tiempo | 20 |
| 18. Exploración y preparación de los datos | 20 |
| 18.1. Descripción de las variables | 20 |
| 18.2. Limpieza y transformación de los Datos | 23 |
| 18.3. Imputación de Datos | 24 |
| 19. Análisis Gráfico | 25 |
| 19.1. Comportamiento general de Parámetros Meteorológicos | 25 |
| 19.2. Comportamiento general de los contaminantes criterio respecto a la normativa . . . | 29 |
| 19.3. Análisis de correlación entre parámetros meteorológicos y contaminantes criterio . . | 36 |
| 19.4. Análisis del comportamiento de los parámetros meteorológicos por día de la semana | 37 |
| 19.5. Análisis del comportamiento de los contaminantes criterio por día de la semana . . | 38 |
| 19.6. Índice de Aire y Salud de los Contaminantes Criterio | 39 |

| | |
|---|-----------|
| 20. Adecuación y validación del Modelo | 41 |
| 20.1. Multivariedad | 41 |
| 20.2. Análisis por Componentes Principales | 41 |
| 20.3. Análisis factorial | 43 |
| 20.4. Regresión lineal múltiple | 46 |
| 21. Resultados | 49 |
| 22. Conclusiones | 50 |
| A. Análisis del comportamiento de los parámetros meteorológicos por día de la semana | 53 |
| A.1. PRS - Presión Atmosférica | 53 |
| A.2. RH - Humedad Relativa | 54 |
| A.3. SR - Radiación Solar | 55 |
| A.4. TOUT - Temperatura | 57 |
| A.5. WSR - Velocidad del Viento | 58 |
| A.6. WDR - Dirección del viento | 59 |
| B. Análisis del comportamiento de los contaminantes criterio por día de la semana | 61 |
| B.1. CO - Monóxido de Carbono | 61 |
| B.2. NO_2 - Dióxido de Nitrógeno | 62 |
| B.3. O_3 - Ozono | 63 |
| B.4. $PM_{2.5}$ - Materia Particulada 2.5 | 65 |
| B.5. PM_{10} - Materia Particulada 10 | 66 |
| B.6. SO_2 - Dióxido de Azufre | 67 |
| C. Índice de Aire y Salud por día de la semana | 69 |

1. Resumen

La OMS le atribuye a la contaminación atmosférica alrededor de 7 millones de muertes cada año OMS, 2021a. Asimismo, esta es la causa de muchos padecimientos como el cáncer pulmonar y derrames cerebrales. Por esta razón se realizan estudios con el fin de analizar la situación actual, realizar predicciones y generar propuestas para salvaguardar la salud de la población. Con este fin, se han establecido normas para describir la peligrosidad del nivel de los seis contaminantes criterio: $PM_{2.5}$, Pm_{10} , NO_2 , SO_2 , CO y O_3 .

En este proyecto se presenta un análisis generado a partir de una base de datos proporcionada por SIMA con mediciones de parámetros meteorológicos y de los contaminantes criterio. Cabe destacar que se trabajó exclusivamente con la estación meteorológica de San Nicolás. El objetivo del proyecto es detectar las tendencias en el comportamiento de los contaminantes criterios, así como su relación con actividades antropogénicas y parámetros meteorológicos. Asimismo, se buscó realizar modelos que expliquen la variabilidad del ozono y que sean capaces de predecir su comportamiento.

En primera instancia se realizó un análisis exploratorio con los datos en bruto. En esta etapa también se realizaron gráficas del comportamiento por día para cada contaminante y parámetro meteorológico. Se encontraban diversas tendencias, entre ellas el aumento de $PM_{2.5}$, Pm_{10} y O_3 a causa de los incendios en el mes de marzo. Asimismo, se detectó un incremento de $PM_{2.5}$ y Pm_{10} a finales de año debido a la pirotecnia. También se generaron gráficos para ilustrar el índice de aire y salud de cada contaminante a través del año donde se vio reflejado un aumento de algunos contaminantes como $PM_{2.5}$ y Pm_{10} debido a la sequía que comenzó en noviembre de 2021.

Finalmente se realizaron tres modelos para facilitar el análisis de la relación entre el ozono y los parámetros meteorológicos. Se realizó un análisis por componentes principales capaz de explicar el 72 % utilizando 3 componentes. De la misma manera, se realizó una análisis factorial, resultando en tres factores latente que explican el 54 % de la variabilidad de los datos. Por último, se realizó un modelo de regresión lineal múltiple modelando a O_3 como variable dependiente de los parámetros meteorológicos con un coeficiente de determinación de 0.54.

2. Introducción

En el año 2019, 11.65 % de las muertes a nivel mundial fueron atribuidas a la contaminación atmosférica (Ritchie y Roser, 2021). Actualmente, la Organización Mundial de la Salud (OMS) le atribuye alrededor de 7 millones de muertes cada año, haciendo así la contaminación del aire una de las mayores causas de mortalidad (OMS, 2022b).

La calidad del aire juega un papel determinante en la salud y el desarrollo de cualquier ser vivo. Algunas de las principales enfermedades vinculadas a la respiración de aire con una alta contaminación incluyen cardiopatías, enfermedad pulmonar crónica, cáncer de pulmón, neumonía y derrames cerebrales; sin embargo, una mala calidad del aire puede afectar a cualquier órgano, pues las partículas de algunos contaminantes son capaces de entrar al flujo sanguíneo a través de los pulmones (OMS, 2022a). Además, también se ha encontrado que tiene un gran efecto negativo durante el embarazo, causando así mayor mortalidad, menor peso al nacer, deterioro del desarrollo pulmonar, mayor morbilidad respiratoria posterior y alteraciones tempranas en el desarrollo inmunitario (Proietti, 2013).

Por esa razón, el Sistema Integral de Monitoreo Ambiental (SIMA) se encarga tanto de medir las concentraciones de los contaminantes así como condiciones meteorológicas con los objetivos de descubrir relaciones entre actividades atropogénicas y la contaminación atmosférica, relacionar el estado climatológico con la calidad del aire, advertir a la población de contingencias ambientales y analizar la interacción entre diferentes localidades. Este sistema inició su operación el 20 de noviembre de 1992 con 5 estaciones. Actualmente, cuenta con 14 estaciones en el área Metropolitana de Monterrey (SIMA, 2015). SIMA se basa en el Índice de Aire y Salud, que es un indicador que homologa la difusión de los niveles de contaminación en México dividiendo la población en tres grupos: Población General, Población Vulnerable y Población Escolar; asimismo, se reporta la calidad del aire cada hora a lo largo de todo el año en cada una de sus estaciones y con base en ello realizan recomendaciones de precauciones a tomar para reducir la contaminación ambiental y sus efectos sobre la salud.



Figura 1: Red de Monitoreo SIMA (SIMA, 2015)

De acuerdo a la Organización Mundial de la Salud, los principales aspectos a tomar en cuenta para medir la calidad del aire son las concentraciones de materia particulada ($PM_{2.5}$ y PM_{10}), ozono (O_3), dióxido de nitrógeno (NO_2), dióxido de azufre (SO_2) y monóxido de carbono (CO) (OMS, 2021b). El 21 de septiembre de 2021, la OMS actualizó las directrices internacionales sobre los niveles recomendados de los 6 contaminantes mencionados anteriormente con el objetivo de salvaguardar la salud de las ciudadanos (OMS, 2021a):

- **Materia particulada fina ($PM_{2.5}$)**
 - Media anual de $5 \mu g/m^3$
 - Media diaria de $15 \mu g/m^3$
- **Materia particulada gruesa (PM_{10})**
 - Media anual de $15 \mu g/m^3$
 - Media diaria de $45 \mu g/m^3$
- **Ozono (O_3)**
 - Media anual de $60 \mu g/m^3$
 - Media diaria de $100 \mu g/m^3$
- **Dióxido de nitrógeno (NO_2)**
 - Media anual de $10 \mu g/m^3$
 - Media diaria de $25 \mu g/m^3$
- **Dióxido de azufre (SO_2)**
 - No se establece una media anual
 - Media diaria de $40 \mu g/m^3$
- **Monóxido de carbono (CO)**
 - No se establece una media anual
 - Media diaria de $4 mg/m^3$

A pesar de que todos estos contaminantes son considerados un riesgo para la salud, la materia particulada ($PM_{2.5}$ y PM_{10}) es considerada como un mayor peligro al ser capaces de penetrar profundamente los pulmones. La $PM_{2.5}$ puede llegar al torrente sanguíneo, causando un daño no solo a los pulmones, sino a todos los órganos (OMS, 2021b).

Por todo lo anteriormente mencionado, es de gran relevancia la medición la calidad del aire. Para esto se utiliza el Índice de Calidad del Aire (ICA). El ICA relaciona la cantidad de los 6 contaminantes criterio con un factor de riesgo y se establece un valor de la calidad del aire entre 0 y 500. Mientras mayor sea el valor, menor es la calidad del aire y mayor es el riesgo para la salud. Los valores de ICA se clasifican de la siguiente manera:



Figura 2: Clasificación del ICA (Javier, 2022)

Además de tomar en cuenta las directrices internacionales para cada contaminante, también es imprescindible considerar las Normas Oficiales Mexicanas de la Calidad del Aire (SIMA, 2021):

■ **NOM-025-SSAI-2021**

1. El indicador límite del promedio anual del contaminante $PM_{2.5}$ es de $10 \mu\text{g}/m^3$, mientras que para un periodo de 24 horas es de $41 \mu\text{g}/m^3$
2. El indicador límite del promedio anual del contaminante PM_{10} es de $36 \mu\text{g}/m^3$, mientras que para un periodo de 24 horas es de $70 \mu\text{g}/m^3$

- **NOM-023-SSAI-2021:** la concentración de dióxido de nitrógeno (NO_2) no debe superar las 0.021 ppm anualmente. Mientras que para un tiempo de exposición de 24 horas se tiene un límite de 0.106 ppm.

- **NOM-020-SSAI-2021:** la concentración del ozono (O_3) con una exposición de 1 hora no debe superar las 0.090 ppm; mientras que para un tiempo de exposición de 8 horas el límite es de 0.065 ppm.
- **NOM-022-SSAI-2019:** el indicador límite del dióxido de azufre (SO_2) es de 0.040 ppm para un tiempo de exposición de 24 horas; mientras que para uno de 1 hora es de 0.075 ppm.
- **NOM-021-SSAI-2021:** el monóxido de carbono (CO) no debe de superar las 9.000 ppm en 8 horas y las 26.000 ppm en 1 hora.

Como antes mencionado, SIMA se encarga de monitorear constantemente las concentraciones de dichos contaminantes en el aire de la Zona Metropolitana de Monterrey. Finalmente, es importante destacar que existen factores externos y antropológicos que afectan la recopilación de dichos datos, así como el manejo que se les debe dar al analizarlos. Entre estos factores están la falta de datos debido a la calibración del equipo, la estacionalidad de los datos, condiciones meteorológicas, entre otros.

3. Resumen de la revisión de bibliografía

La OMS define la contaminación del aire como la presencia de agentes, ya sea físicos, químicos o biológicos en el aire, por ejemplo humo, polvo, gas, niebla, entre otras cosas, que alteran las características naturales de la atmósfera. Esta se puede dividir en dos contextos: de interiores (doméstico) y de exteriores. Se reconocen el monóxido de carbono (CO), el ozono (O_3), el dióxido de nitrógeno (NO_2), el dióxido de azufre (SO_2) y las partículas suspendidas (PM_{10} y $PM_{2.5}$) como los contaminantes más preocupantes para la salud. Estos contaminantes son factores de enfermedades cardíacas, accidentes cerebrovasculares, infecciones de las vías respiratorias, cáncer de pulmón, diabetes y enfermedad pulmonar obstructiva crónica (EPOC). La vía principal de exposición a la contaminación del aire es a través del tracto respiratorio. Respirar estos contaminantes provoca inflamación, estrés oxidativo, inmunosupresión y mutagenicidad en las células de todo nuestro cuerpo, lo que afecta los pulmones, el corazón, el cerebro, entre otros órganos. Claramente, la polución puede afectar a la población en diversas medidas, pues la genética, las comorbilidades, la nutrición

y los factores sociodemográficos son factores que afectan a la susceptibilidad de dicho riesgo OMS, 2022a. En la gráfica posterior, realizada por el Institute for Health Metrics and Evaluation (IHME) en su estudio Global Burden of Disease en 2019, se estima el número de muertes de manera global producidas por cada uno de los factores de riesgo atribuidas a la enfermedad en la que se puede observar que la contaminación del aire de exteriores es la causa de aproximadamente 4.51 millones de muertes al año y la contaminación del aire de interiores es la causa de aproximadamente 2.31 millones de muertes al año.

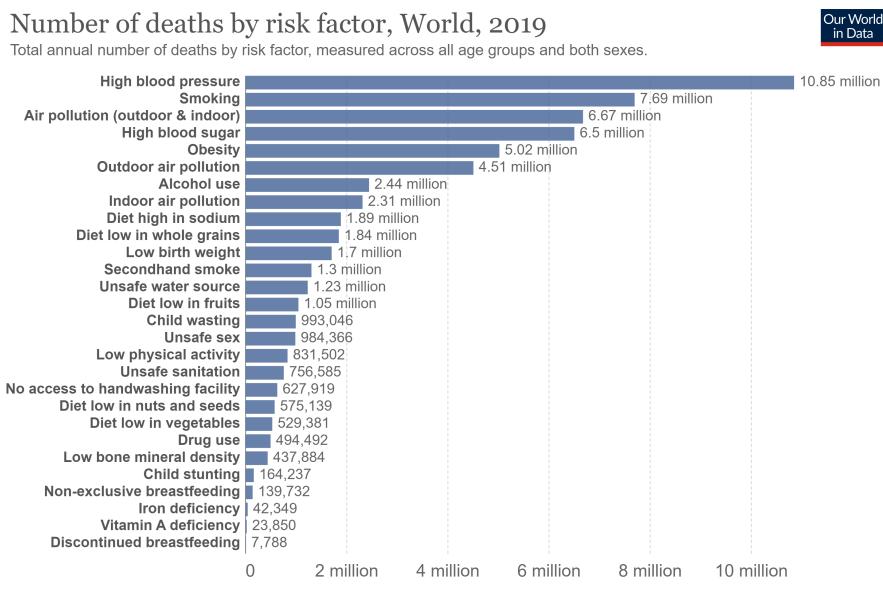


Figura 3: Muertes por factor de riesgo

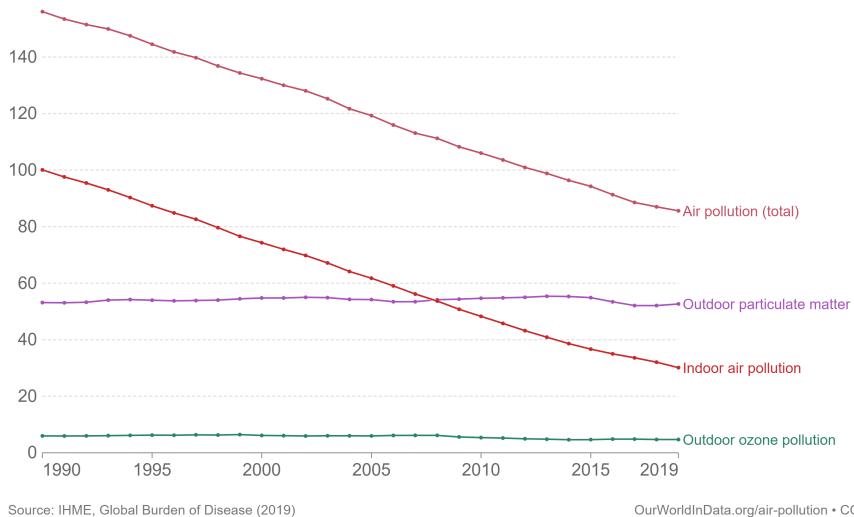
En México, la tasa de morbilidad por contaminación del aire en 2019 era de casi 44 personas por cada 100,000. La tasa de morbilidad por esta razón de cada país va variando conforme los ingresos del país varían, de manera que entre más desarrollado o más ingresos genere el país menor será la tasa de morbilidad del mismo.

Comparando los resultados obtenidos de 1990 a 2019, se ha visto un descenso de la tasa de morbilidad global causada contaminación del aire y esto esta estrechamente relacionado a la mejora que ha tenido la polución de interiores como se muestra en el siguiente grafo Ritchie y Roser, 2021.

Death rate from air pollution, World, 1990 to 2019

Our World
in Data

Death rates are given as the number of attributed deaths from pollution per 100,000 population. These rates are age-standardized, meaning they assume a constant age structure of the population: this allows for comparison between countries and over time.



Source: IHME, Global Burden of Disease (2019)

OurWorldInData.org/air-pollution • CC BY

Figura 4: Tasa de muertes por polución

Además de ser la polución un problema para nuestra salud por las enfermedades que los contaminantes provocan en el cuerpo humano, esta está también relacionada a la problemática del cambio climático y la vida de los ecosistemas, pues la mayor parte de lo que conlleva la producción de los contaminantes emiten gases de efecto invernadero. Por esta razón, si se reduce la contaminación del aire, se estaría combatiendo también al cambio climático tanto a corto como a largo plazo y se mitiga la carga de morbilidad relacionada a ello OMS, 2022b.

A partir de la problemática de contaminación del aire, se ha desarrollado el Sistema Integral de Monitoreo Ambiental (SIMA) que monitorea la concentración de los contaminantes en el aire dictándose por el Índice de Calidad del Aire (ICA), una regla que va de 0 a 500, en la cual entre más alto es el valor del Índice, mayor es el nivel de contaminación atmosférica. Este cálculo utiliza la metodología de la Agencia de Protección Ambiental (EPA) y se muestra a continuación Javier, 2022.

En base a los datos recuperados por organizaciones como SIMA, la ciencia de datos puede ser de

$$I_p = \frac{I_{hi} - I_{lo}}{BP_{hi} - BP_{lo}} (C_p - BP_{lo}) + I_{lo}$$

I_p = Valor del índice para el contaminante p

C_p = Concentración del contaminante

BP_{hi} = Límite superior de concentración del índice de calidad del aire según contaminante y periodo de permanencia

BP_{lo} = Límite inferior de concentración del índice de calidad del aire según contaminante y periodo de permanencia

I_{hi} = Valor de AQI (Air Quality Index) superior según el rango de concentración

I_{lo} = Valor de AQI (Air Quality Index) inferior según el rango de concentración

Figura 5: Cálculo del índice de calidad del aire

gran utilidad comprender la dinámica de la contaminación del aire y construir modelos estadísticos confiables para pronosticar los niveles de contaminación del aire. Liu et al., 2021

4. Descripción del problema específico (preguntas de investigación)

Con la aceleración de la urbanización, la contaminación atmosférica se ha extendido por todo el mundo y se ha convertido en una de las mayores amenazas para la salud humana. La ciencia de los datos puede ayudar a comprender la dinámica de la contaminación atmosférica y a construir modelos estadísticos fiables para prever los niveles de contaminación atmosférica. Para lograr estos objetivos, es necesario aprender los modelos estadísticos que pueden capturar la dinámica de los datos históricos y predecir la contaminación atmosférica en el futuro. (Liu et al., 2021)

Los contaminantes atmosféricos más comunes incluyen principalmente partículas (PM2,5, PM10) y gases como el dióxido de azufre (SO₂), el dióxido de nitrógeno (NO₂), el monóxido de carbono (CO) y el ozono (O₃) (Belavadi et al., 2020).

La exposición a largo plazo a estos contaminantes tiene efectos adversos en la salud física y mental del ser humano, y perjudica su salud de forma global. Por ejemplo, la exposición de contamina-

ción atmosférica puede provocar diversas enfermedades respiratorias y cardiovasculares, y también se ha demostrado que es un factor de riesgo de cáncer. Además, la contaminación atmosférica puede provocar una serie de cambios en el cerebro humano, como la reducción del volumen de la materia gris y de la integridad de la materia blanca.

Se tienen niveles establecidos de concentración y tiempos de exposición de los contaminantes donde el ciudadano promedio puede desenvolverse sin que se afecte de forma significativa su salud. Estos niveles son establecidos por las Normas Oficiales Mexicanas para la Calidad del Aire. Los niveles de la norma ambiental para cada contaminante son diferentes, ya que la vulnerabilidad del ser humano es diferente ante cada tipo de contaminante. («Aire NL», 2022)

5. Objetivos

El objetivo general de este informe es detectar las diferentes relaciones que existen entre los agentes contaminantes medibles y los parámetros meteorológicos; así como descubrir tendencias en el comportamiento de los contaminantes criterio explicadas por diferentes actividades antropogénicas y factores climáticos. Asimismo se tiene el objetivo específico de generar modelos que relacionen y predigan la concentración atmosférica del ozono utilizando variables meteorológicas.

6. Justificación

Un total de 6 de los 13 municipios del Área Metropolitana de Monterrey lideran, en el promedio anual, el ranking de las 25 ciudades más contaminadas de México por partículas suspendidas en el aire $PM_{2.5}$ durante 2021 según Airvisual y la ciudad de San Nicolás es la numero 12 siendo esta la de nuestro interés, la misma fuente de Airvisual nos dice que Monterrey es la ciudad número 9 en contaminación por partículas suspendidas. («Latin America Air Pollution City», s.f.)

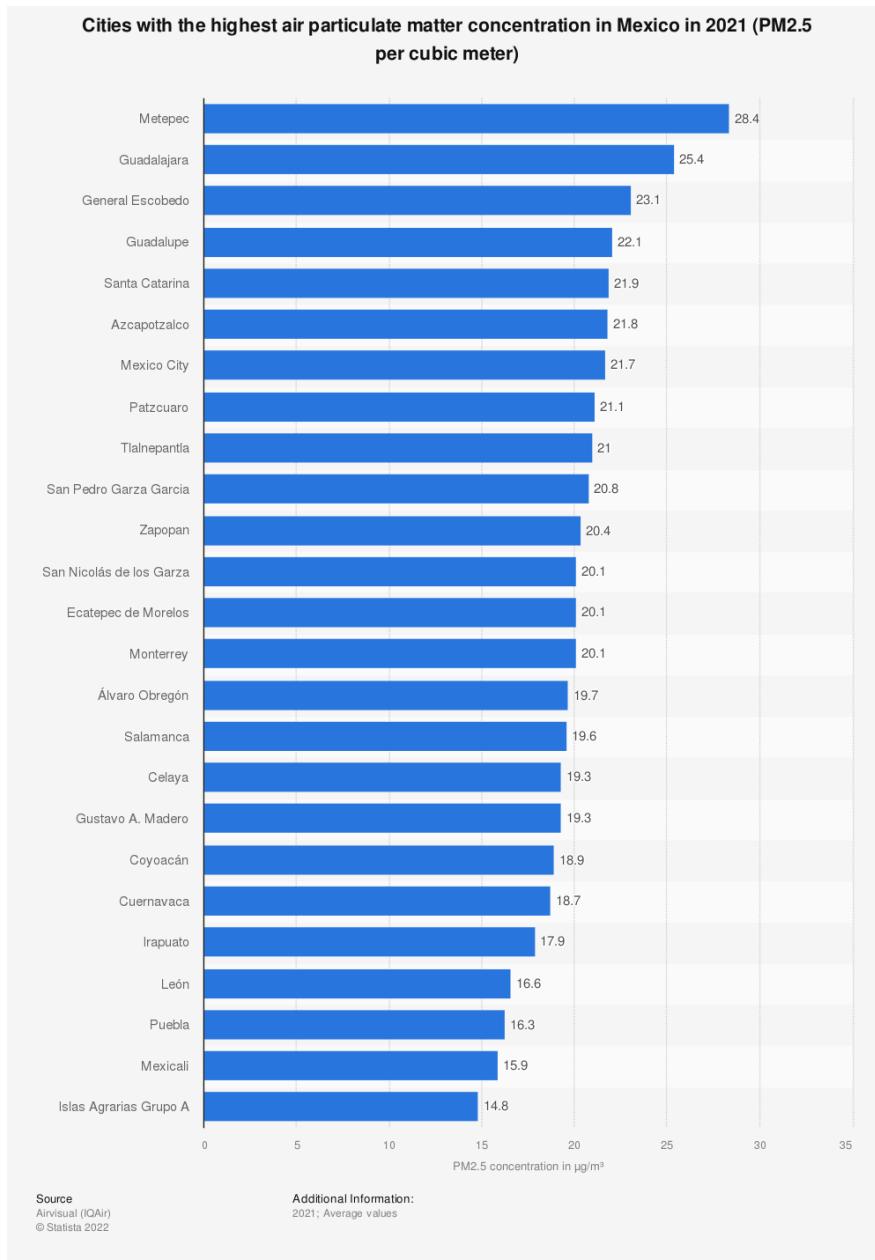


Figura 6: Ranking de México de partículas suspendidas en el aire(«Latin America Air Pollution City», s.f.)

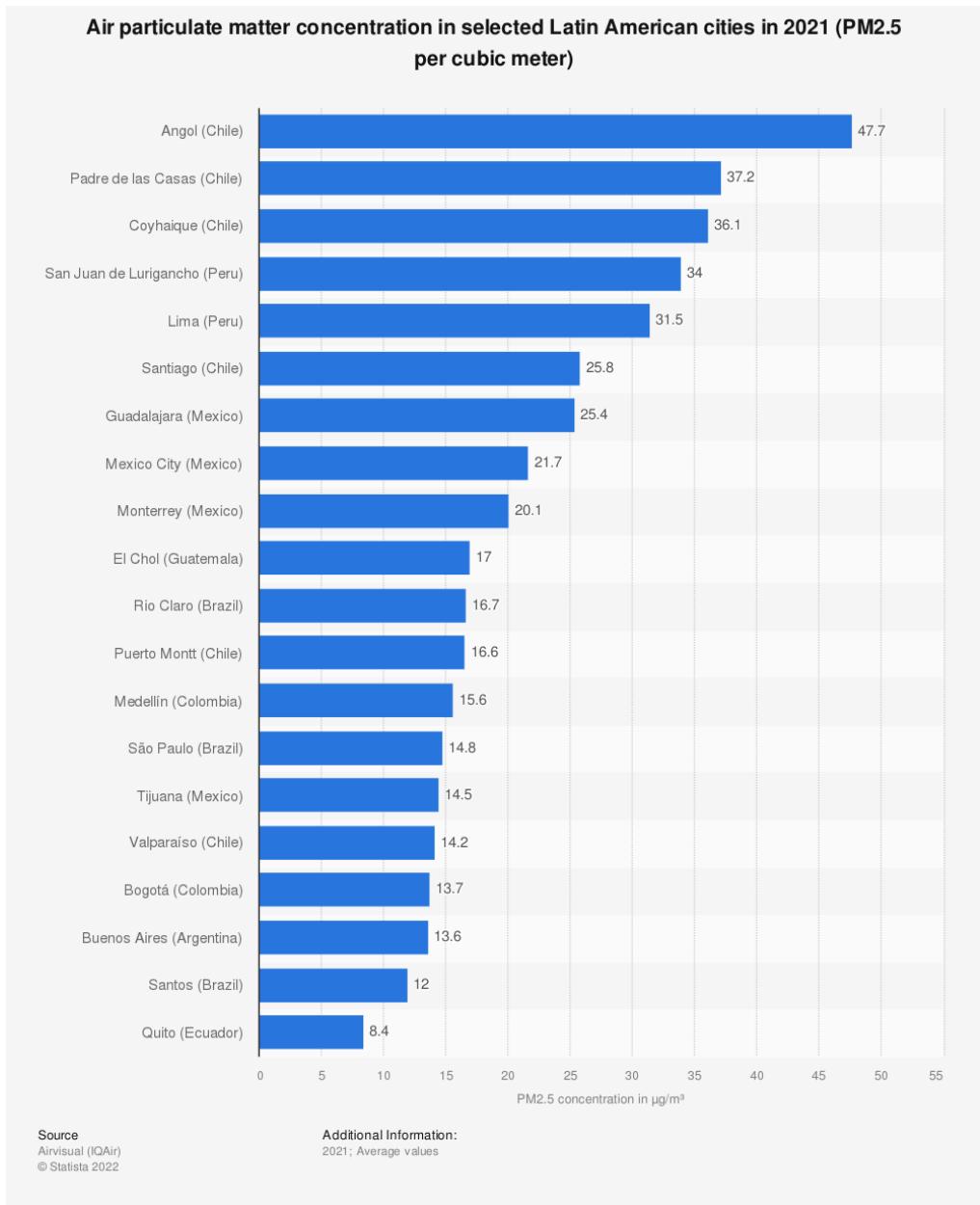


Figura 7: Ranking de Latino América de partículas suspendidas en el aire («Latin America Air Pollution City», s.f.)

De acuerdo con un informe de la OCDE, a menos que se emprendan a partir del gobierno y la

industria las debidas acciones, en 2060, la contaminación atmosférica en exteriores podría causar de 6 a 9 millones de muertes prematuras al año y costará el 1% del PIB global alrededor de 2.6 billones de dólares al año como resultado de faltas por enfermedad, costos médicos y menor producción agrícola («En 2060, la contaminación atmosférica causará de 6 a 9 millones de muertes prematuras al año y tendrá un costo de 1% del PIB – OCDE - OECD», s.f.).

El INSP y el INECC (2017) analizan las muertes que se podrían evitar si se alcanzaran los niveles recomendados por la OMS, valiéndose de una metodología propuesta por la Agencia de protección ambiental (EPA por sus siglas en inglés). Según sus cálculos, aproximadamente 12 mil muertes podrían evitarse con regulaciones que se acerquen a los estándares internacionales, lo cual tendría un beneficio social de 20 mil millones de pesos («En 2060, la contaminación atmosférica causará de 6 a 9 millones de muertes prematuras al año y tendrá un costo de 1% del PIB – OCDE - OECD», s.f.).

7. Mercado potencial (descripción general)

En este reporte el principal cliente es SIMA, sin embargo, la metodología puede ser aplicada a cualquier otro organismo que se encargue de monitorear el aire y que esté interesado en analizar cómo es que afectan las condiciones meteorológicas a las concentraciones de contaminantes.

Un buen modelo de monitoreo en la calidad del aire puede ser comercializado a entidades tanto públicas como privadas, según un reporte de *Fortune Business Insights*, se estima que el mercado crezca de 5.02 MMDD en 2021 a 8.33 MMDD en 2028 (FortuneBusinessInsights, 2020), por lo que hay gran potencial en el mercado.

Un ejemplo de compañías privadas que están desarrollando sistemas de monitoreo de aire son: 3M, Siemens, Honeywell y Merck. En el reporte se indica que en el 2020, el mercado se lideró de la siguiente manera:

1. Comercial y residencial (32.7%)

2. Infraestructura pública
3. Plantas de generación de energía
4. Industria farmacéutica
5. Otros

En el 2020, el mercado fue liderado por la región asiática, sin embargo, no es 100 % infalible, ya que existen factores que lo restringen, en especial por los altos costos de mantenimiento de las estaciones de monitoreo y las políticas de control existentes en cada nación.

8. Identificación de clientes/consumidor y usuarios

El cliente al que está enfocada esta investigación es SIMA, sin embargo en la industria del monitoreo de calidad de aire, existe una mayor cantidad de consumidores y usuarios que podrían beneficiarse de esto.

De acuerdo con *Fortune Business Insights*, el tipo de monitoreo de aire que lidera la industria es en interiores (a diferencia del de exteriores) debido al aumento de casas inteligentes y edificios verdes (*FortuneBusinessInsights*, 2020), en donde la salud y el medio ambiente se vuelve prioridad.

El monitoreo en interiores es el líder ya que con el paso del tiempo se ha ido buscando tener una mejor calidad de vida a nivel general y la mala calidad del aire es un riesgo a la salud. Según un reporte del *Institute of Medicine (IOM)* los riesgos son mayores en niños, personas de la tercera edad y personas con condiciones de cuidado en el sistema respiratorio y cardiaco. (Fisk, 2015)

Si bien la humanidad se ha enfocado en controlar la temperatura del aire, debido al aumento de contaminantes, el enfoque ha cambiado y se han creado sensores que monitorean contaminantes así como condiciones climatológicas para cambiar la calidad de aire en el interior de una construcción realizando ajustes en la ventilación, enfriamiento o calentamiento del aire, así como la apertura o cerrado de ventanas, persianas o aditamentos que permitan mayor protección ante la readiación solar. (Schieweck et al., 2018)

9. Descripción de las fuentes de información (datos)

Con el fin de realizar este proyecto, el Sistema Integral de Monitoreo Ambiental proporcionó dos bases de datos: una sobre las condiciones meteorológicas reportadas y otra sobre la cantidad de los contaminantes presente en cada medición. Asimismo, se facilitó una clave con las definiciones de cada parámetro y variable, así como las unidades utilizadas para las mediciones y descripción de las banderas.

10. Selección del Modelo

Para la selección del modelo se considerarán tanto modelos estadísticos como de clasificación. Entre los modelos estadísticos destaca la regresión lineal por su sencillez, eficiencia e interpretabilidad, aunque esta misma premisa de una relación lineal puede ser limitante. También es un modelo sensible a puntos atípicos y como se asume que las variables son independientes cuando esto es incorrecto también lo son las interpretaciones. Algunas técnicas de clasificación que ya han sido utilizadas para predecir el clima y otros factores atmosféricos anteriormente incluyen:

1. Support Vector Machines
2. Naive Bayes
3. K Nearest Neighbors
4. Perceptrones multicapa

Generalmente se intenta predecir con base en los datos de los últimos 4-7 días, aunque en algunos casos también se utiliza información como promedios históricos, mensuales, etc. (Naveen L., 2019)

Adicionalmente, se indagará en utilizar:

- Análisis de Componentes Principales (PCA): Se evaluará si es posible reducir el espacio muestral por medio del modelo PCA con la finalidad de explicar un la variabilidad con menos

dimensiones. Así mismo, podremos identificar tendencias, saltos, clústers y valores atípicos, así como descubrir las relaciones entre las observaciones y las variables, entre las variables. El análisis PCA es relevante ya que permite realizar un análisis flexible, donde la muestra puede contener multicolinealidad, valores faltantes y datos categóricos. En resumen, este análisis ignora las etiquetas de cada variable y encuentra los principales componentes que explican la varianza de un conjunto de datos.

- Análisis de Discriminante Lineal (LDA): Se evaluará si es posible clasificar las variables en diferentes grupos, siendo este un análisis categórico. El LDA, a diferencia del PCA (aprendizaje no-supervisado), utiliza un algoritmo de aprendizaje supervisado, el cual encuentra los discriminantes lineales y representa aquellos ejes que maximicen la separación entre las distintas variables, así, reduciendo igualmente la dimensionalidad del espacio muestral.
- Regresión Lineal Múltiple: Como mencionado anteriormente, las regresiones lineales son bastante efectivas cuando se cumplen las condiciones que permiten realizar un análisis lineal. En este caso, al con muchas variables, aplicaremos métodos estadísticos para identificar adecuadamente la relación entre las variables independientes, mediante el uso de pruebas de multivariadas, pruebas de normalidad y significancia, análisis de errores (Mean Squared Error, Mean Absolute Error, etc.) y análisis residual.

11. Descripción de la solución

Para realizar un análisis completo de los datos y llegar a una buena predicción se utilizarán distintos modelos estadísticos como lo son la regresión lineal múltiple para utilizar más información en la construcción del modelo haciendo las estimaciones más precisas, el análisis de discriminante para clasificar las variables objetivo en varios grupos y el análisis de componentes principales para simplificar el espacio muestral.

12. Propuesta de valor

La propuesta de valor que se busca dar con este proyecto está compuesta de distintos puntos:

- Identificar las principales causas de contaminantes atmosféricos (de origen antropogénico) y su relación con las mediciones que se cuentan.
- Encontrar qué relación hay entre los contaminantes atmosféricos y las condiciones meteorológicas en el municipio de San Nicolás en Nuevo León, México. Sin embargo el modelo puede emplearse en otras zonas.
- Dar recomendaciones de qué se debería modificar si se busca encontrar una mejora en la calidad del aire de la zona: nuevas normativas, políticas a empresas, etc.
- Explorar las ventajas y desventajas de actuar y no actuar activamente ante los resultados, respectivamente.
- Identificar posibles aplicaciones del modelo en otras regiones y otros usos que no sean para fines de infraestructura pública.
- Mencionar impactos a la salud debido a la exposición prolongada de contaminantes.

13. Nombre Detallado del proyecto

Aplicación de métodos estadísticos multivariados para el conocimiento de la naturaleza de contaminantes que influyen en la calidad del aire y sus interrelaciones en el área de San Nicolás.

14. Nombre corto o comercial del proyecto

Polución en San Nicolás: Contaminantes e interrelaciones.

15. Impacto Social Principal

Como anteriormente mencionado, la calidad del aire es juega un papel determinante en la salud de la población. Además una contaminación atmosférica elevada reduce la productividad laboral, aumenta los gastos médicos y enfatiza las consecuencias del calentamiento global. El impacto social de este proyecto yace en la identificación de posibles causas y relaciones entre diferentes factores (antropogénicos y meteorológicos) y la calidad del aire, contribuyendo así a generar propuestas, normativas y proyectos que permitan mejorar la calidad de aire y, por ende, permitan el sano y completo desarrollo de la sociedad.

16. Impacto hacia los Objetivos de Desarrollo Sostenible

Este proyecto apoya en distintos Objetivos de Desarrollo Sostenible (ODS) establecidos por la ONU, sin embargo, los principales ODSs impactados incluyen:

- **ODS 3 Salud y bienestar:** la calidad del aire tiene un gran impacto en la salud de la población, por lo que al generar propuestas a partir del análisis desarrollado en este proyecto, el daño a la salud de las personas por la contaminación atmosférica disminuirá.
- **ODS 11 Ciudades y Comunidades Sostenibles:** para llegar a tener comunidades sostenibles, es imprescindible medir el progreso o retroceso de la sociedad, así como generar nuevas ideas a partir del análisis detallado de los diversos factores involucrados en el cambio climático y la contaminación ambiental.
- **ODS 13 Acción por el clima:** la contaminación atmosférica juega un rol principal en el cambio climático, por lo que cualquier proyecto enfocado, ya sea directa o indirectamente, en la reducción de los contaminantes en el ambiente, impacta el ODS 13.
- **ODS 15 Vida de ecosistemas terrestres:** la contaminación atmosférica no impacta solamente a los humanos, sino a cualquier ser vivo. Por lo tanto, el proyecto ayudaría a mejorar la vida de ecosistemas terrestres.

17. Línea del tiempo

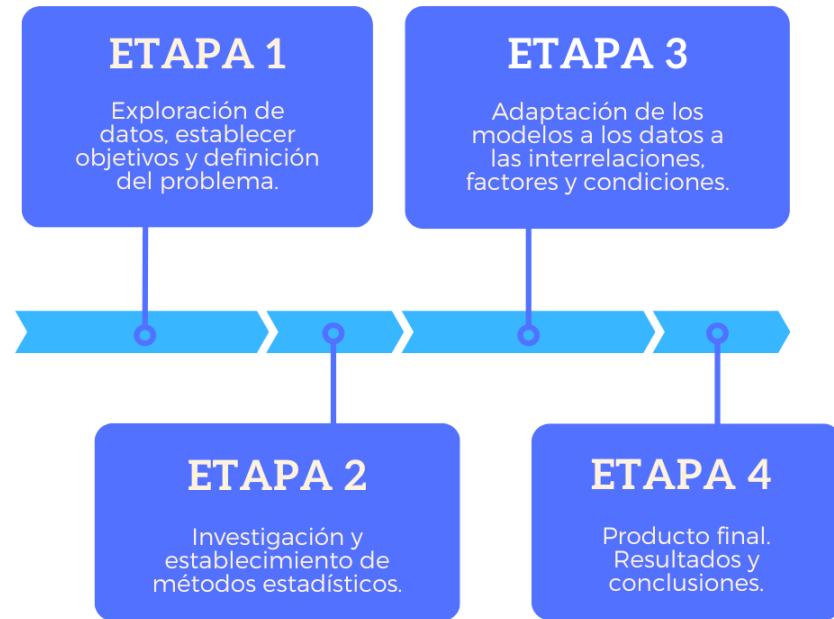


Figura 8: Línea del tiempo de las etapas del proyecto

18. Exploración y preparación de los datos

18.1. Descripción de las variables

Los datos proporcionados por SIMA se dividen en dos bases de datos, una para los contaminantes y otra para las condiciones meteorológicas. La de contaminantes contiene cerca de 66,000 filas y la de condiciones meteorológicas 76,000. Cada fila representa una medición y cada hora se toman 6 mediciones de contaminantes (los mismos 6 contaminantes mencionados anteriormente) y 7 de condiciones (Temperatura, Humedad Relativa, Radiación Solar, Precipitación, Presión Atmosférica, Velocidad del Viento y Dirección del Viento). Esto significa que hay alrededor de 11,000 observaciones desde el 1o de Julio de 2021 hasta el 30 de Septiembre de 2022. Ambas bases tienen

31 columnas, un identificador, la hora de la medición, el parámetro al que corresponde, 14 columnas para estaciones de monitoreo y 14 para sus respectivas banderas. Algunas mediciones horarias parecen no estar presentes, pero esto será abordado en la transformación y limpieza de las bases de datos en la siguiente entrega.

A continuación, se encuentra la descripción de cada estación meteorológica en la tabla 1.

| Abreviatura | Descripción | Estación | Municipio |
|-------------|-------------|----------------|---------------------------|
| SE | Sureste | La Pastora | Guadalupe |
| NE | Noreste | San Nicolás | San Nicolás de los Garzas |
| CE | Centro | Obispado | Monterrey |
| NO | Noroeste | San Bernabé | Monterrey |
| SO | Suroeste | Santa Catarina | Santa Catarina |
| NO2 | Noroeste 2 | García | García |
| NTE | Norte | Escobedo | Escobedo |
| NE2 | Noreste 2 | Apodaca | Apodaca |
| SE2 | Sureste 2 | Juárez | Juárez |
| SO2 | Suroeste 2 | San Pedro | San Pedro Garza García |

Cuadro 1: Descripción de las estaciones meteorológicas

Asimismo, en la tabla 2 se encuentra la definición de cada contaminante criterio (variable), así como las unidades en las que se mide.

| Abreviatura | Contaminante | Unidad |
|------------------------|--|-------------------|
| PM10 | Material Particulado menor a 10 micrométros | µg/m ³ |
| PM2.5 | Material Particulado menor a 2.5 micrométros | µg/m ³ |
| O₃* | Ozono | ppb |
| SO₂* | Dióxido de Azufre | ppb |
| NO₂* | Dioxido de Nitrógeno | ppb |
| CO | Monóxido de Carbono | ppm |
| NO | Monóxido de Nitrógeno | ppb |
| NO_x | Óxidos de Nitrógeno | ppb |

Cuadro 2: Contaminantes criterio y sus unidades

La definición de los parámetros meteorológicos es la siguiente:

| Abreviatura | Contaminante | Unidad |
|--------------|----------------------|--------|
| TOUT | Temperatura | °C |
| RH | Humedad Relativa | % |
| SR | Radiación Solar | kW/m2 |
| RAINF | Precipitación | mm/Hr |
| PRS | Presión Atmosférica | mm Hg |
| WSR | Velocidad del Viento | Km/hr |
| WDR | Dirección del Viento | ° |

Cuadro 3: Parámetros Meteorológicos

En la tabla 4, se encuentra la descripción de las banderas de contaminantes y parámetros meteorológicos.

| Bandera | Descripción | Hora |
|----------|---|----------|
| P | Falla eléctrica | Válida |
| p | Falla eléctrica | Inválida |
| C | Calibración | Válida |
| c | Calibración | Inválida |
| D | Apagado | Válida |
| d | Apagado | Inválida |
| B | Malas condiciones | Válida |
| b | Malas condiciones | Inválida |
| m | Positivo sobre el rango | Inválida |
| l | Negativo sobre el rango | Inválida |
| z | Ceros y negativos | Inválida |
| o | PM10 mayor a 900 ug/m3 | Inválida |
| s | Valores repetidos | Inválida |
| r | comparativo PM10 vs PM2.5 | Inválida |
| e | Eliminar datos NO y Nox | Inválida |
| a | Eliminar PM menor a 5 ug/m3 y 0.05 ppm en CO | Inválida |
| s | Valores iguales consecutivos | Inválida |
| f | Valores 3 veces mayor que el valor anterior para PM10 | Inválida |
| h | Valores de temperatura con más de 10 grados o 10 mmHg de diferencia de una hora | Inválida |

Cuadro 4: Banderas de parámetros meteorológicos y contaminantes

18.2. Limpieza y transformación de los Datos

Debido a la disposición de los datos se requiere hacer una transformación que facilite las operaciones y el análisis futuro. Primero, se seleccionaron las columnas de fecha y hora, parámetro medido, la columna correspondiente a la estación de San Nicolás (Noreste 1) y la columna que contiene sus banderas. Después se filtraron únicamente las filas correspondientes a un contaminante y el vector de valores resultantes conforma una columna de la matriz transformada, mientras que el vector de banderas conforma otra columna. Se repite esto para cada contaminante, obteniendo así la matriz transformada donde cada fila es una hora y tiene 12 columnas: 6 para cada contaminante, 6 para sus respectivas banderas. De manera similar la matriz transformada de parámetros meteorológicos tiene 14 columnas y el mismo número de filas. Al realizar esta transformación es necesario tener cuidado con los valores faltantes ya que si no se toman en cuenta se desfasan las observaciones.

Finalmente, en la tabla 5 se pueden observar las medidas estadísticas obtenidas para cada parámetro meteorológico, mientras que en la tabla 6 se encuentra la frecuencia de las banderas.

| | PRS | RAINF | RH | SR | TOUT | WDR | WSR |
|---------|-------|-------|-------|--------|-------|-------|-------|
| Min | 684.7 | 0 | 2.00 | -0.006 | 0.38 | 1.0 | 0.7 |
| 1st Qu | 716.7 | 0 | 47.0 | 0.000 | 20.62 | 92.0 | 6.0 |
| Mediana | 718.9 | 0 | 65.0 | 0.004 | 24.97 | 127.0 | 9.3 |
| Media | 719.1 | 0 | 61.91 | 0.1726 | 24.34 | 135.8 | 9.798 |
| 3rd Qu | 721.0 | 0 | 79.00 | 0.3020 | 29.24 | 151.0 | 13.1 |
| Max | 734.6 | 0 | 96.00 | 0.8340 | 41.49 | 360.0 | 27.8 |
| NA's | 151 | 138 | 158 | 59 | 154 | 155 | 140 |

Cuadro 5: Medidas estadísticas de los parámetros meteorológicos

| | PRS | RAINF | RH | SR | TOUT | WDR | WSR |
|----------|-------|-------|-------|------|-------|-------|-------|
| x | 126 | 126 | 130 | 47 | 139 | 143 | 128 |
| Sin Flag | 10815 | 10828 | 10808 | 5831 | 10812 | 10811 | 10826 |
| k | 13 | 0 | 0 | 0 | 0 | 0 | 0 |
| n | 4 | 4 | 4 | 4 | 4 | 4 | 4 |
| s | 0 | 0 | 15 | 0 | 1 | 0 | 0 |
| l | 0 | 0 | 0 | 5076 | 0 | 0 | 0 |
| NA's | 9 | 9 | 10 | 9 | 11 | 9 | 9 |

Cuadro 6: Frecuencias de banderas en mediciones de parámetros meteorológicos

De la misma manera en la tabla 7 se pueden observar las medidas estadísticas de cada contaminante; en la tabla 8, la frecuencia de las banderas en los contaminantes criterio.

| | CO | NO2 | SO2 | O3 | PM10 | PM2.5 |
|---------|-------|-------|-------|-------|-------|--------|
| Min | 0.130 | 0.30 | 0.600 | 1.00 | 2.0 | 2.19 |
| 1st Qu | 1.190 | 7.30 | 3.800 | 13.00 | 40.0 | 12.45 |
| Mediana | 1.460 | 10.60 | 4.800 | 22.00 | 53.0 | 17.87 |
| Media | 1.637 | 14.46 | 5.462 | 25.06 | 60.8 | 21.30 |
| 3rd Qu | 2.020 | 17.60 | 6.000 | 34.00 | 71 | 26.16 |
| Max | 5.580 | 96.20 | 55.40 | 128.0 | 551.0 | 406.52 |
| NA's | 137 | 147 | 579 | 307 | 218 | 450 |

Cuadro 7: Medidas estadísticas de los contaminantes criterio

| | CO | NO2 | SO2 | O3 | PM10 | PM2.5 |
|----------|-------|-------|-------|-------|-------|-------|
| Sin Flag | 10829 | 10811 | 10387 | 10659 | 10748 | 10516 |
| a | 124 | 0 | 0 | 0 | 10812 | 23 |
| n | 4 | 0 | 4 | 4 | 4 | 0 |
| x | 1 | 71 | 126 | 125 | 137 | 386 |
| e | 0 | 57 | 1 | 0 | 0 | 0 |
| u | 0 | 8 | 0 | 0 | 0 | 0 |
| s | 0 | 6 | 440 | 170 | 4 | 17 |
| r | 0 | 0 | 0 | 0 | 0 | 12 |
| o | 0 | 0 | 0 | 0 | 65 | 0 |
| NA's | 9 | 9 | 10 | 9 | 9 | 9 |

Cuadro 8: Frecuencia de banderas en mediciones de contaminantes criterio

18.3. Imputación de Datos

Para poder realizar algunos análisis posteriores, fue necesario realizar imputación de datos, ya que no se debía contar con datos nulos.

Se revisó en la página de SIMA las estaciones más cercanas a la asignada, primero se eligió la estación Norte 2, ubicada en el mismo municipio. Sin embargo, aún había valores nulos, por lo que se utilizaron los datos de la estación Noreste 2. Para finalizar se optó por utilizar backfilling para rellenar los datos faltantes.

| | NAs en Contaminantes | NAs en Meteorología |
|---------------------|----------------------|---------------------|
| NAs | 1784 | 889 |
| NAs 1er imputación | 298 | 116 |
| NAs 2nda imputación | 124 | 96 |
| NAs backfilling | 6 | 2 |
| NAs totales | 0 | 0 |

Cuadro 9: Imputación de Valores Faltantes

19. Análisis Gráfico

Con el objetivo de tener una mejor comprensión del comportamiento de los datos, se realizaron algunas gráficas de las variables y parámetros de la estación de San Nicolás.

19.1. Comportamiento general de Parámetros Meteorológicos

En la figura 9 se puede observar un diagrama de violín y un diagrama de caja de las variables meteorológicas con escala logarítmica. En los parámetros TOUT (temperatura), RH (humedad relativa) y WDR (dirección del viento) se perciben una gran cantidad de valores atípicos. Asimismo, en todos los parámetros se puede notar una concentración mayor alrededor del tercer cuartil. También es importante destacar que, tanto los diagramas de violín como los de caja de los parámetros SR (radiación solar) , TOUT (temperatura)y WDR (dirección del viento) son relativamente largos, lo que implica un mayor rango entre las observaciones.

Violin y box plot de variables meteorológicas con escala logarítmica

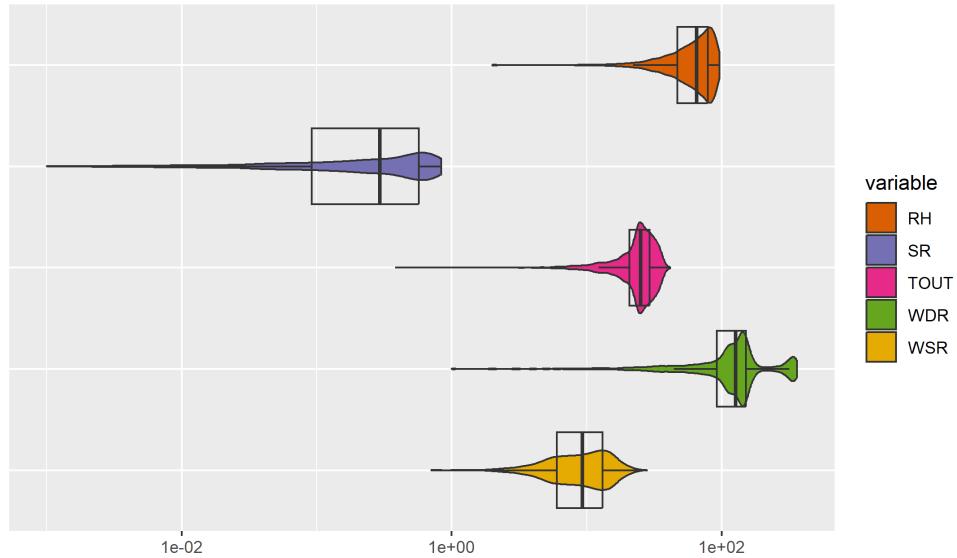


Figura 9: Diagramas de violín y de caja de parámetros meteorológicos

Asimismo, en la figura 10 se puede observar el histograma realizado para ilustrar la distribución de las variables meteorológicas. En esta imagen se puede notar que el rango de SR (radiación solar) es pequeño, a diferencia de lo que observó en la figura 10. Esto se debe a la escala de la gráfica, pues, como se mencionó anteriormente, la figura 9 tiene una escala logarítmica. Asimismo se puede observar que el parámetro RH (humedad relativa) tiene un ligero sesgo positivo. Finalmente, importante notar la gran cantidad de valores atípicos presentes en el parámetro WDR (dirección de viento), lo que coincide con la gráfica anterior.

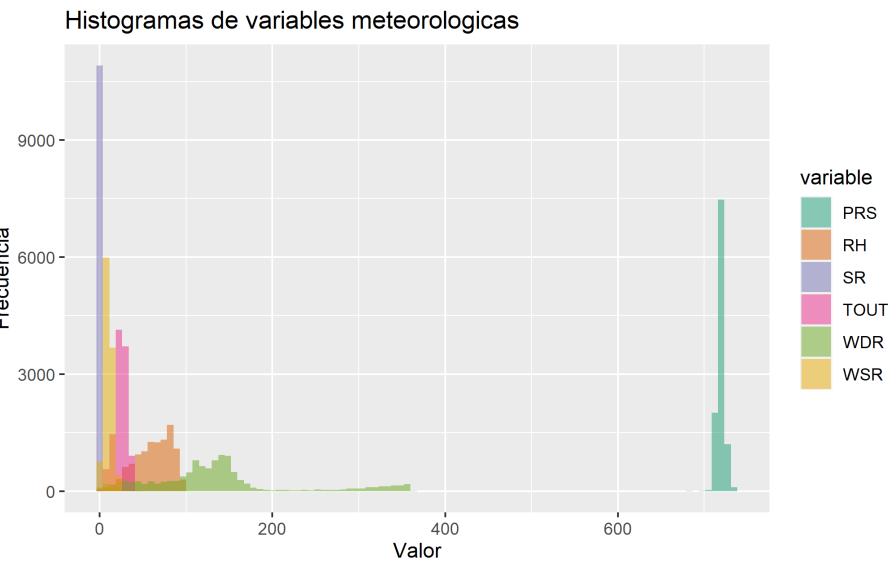


Figura 10: Histograma de parámetros meteorológicos

También se realizó un análisis de frecuencia de las banderas en cada parámetro. Se puede notar una gran presencia de la bandera “x” en todos los parámetros. Sin embargo, en el parámetro SR (radiación solar), la bandera “l” se puede encontrar 5,076 veces. Por otro lado, la bandera más presente en las mediciones de los contaminantes además de “x”, es “s” que indica la presencia de valores repetidos.

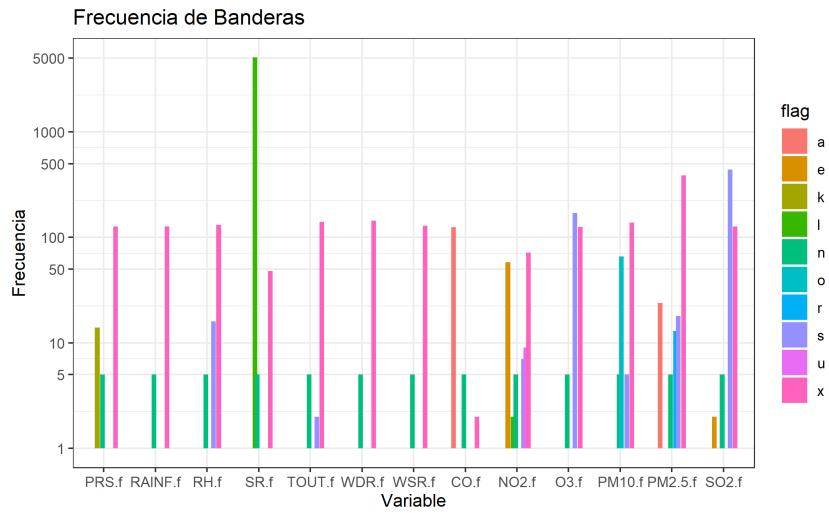


Figura 11: Frecuencia de banderas

Por otro lado, en la figura 12 se encuentra el mapa de correlaciones entre los parámetros meteorológicos. Entre las correlaciones más significativas se puede destacar la correlación negativa de -0.63 entre TOUT (temperatura) y PRS (presión atmosférica). De la misma manera, el parámetro RH (humedad relativa) tiene correlaciones significativas con SR (radiación solar), TOUT (temperatura) y WSR (velocidad del viento). También es importante notar la correlación positiva de TOUT (temperatura) con WSR (velocidad del viento) y SR (radiación solar).

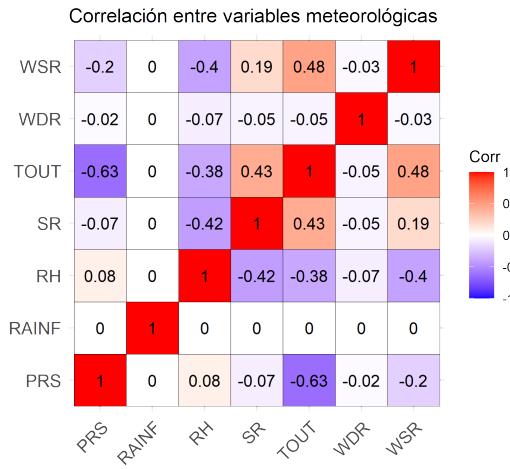


Figura 12: Correlación entre parámetros meteorológicos

19.2. Comportamiento general de los contaminantes criterio respecto a la normativa

En la figura 13 se pueden observar los niveles de PM_{10} a lo largo del último año y se muestra también el límite de los niveles recomendado, el cual está representado por una linea punteada roja. En esta gráfica se puede observar que en la gran mayoría de los meses los niveles sobrepasan la normativa. Se puede apreciar que justo en enero del 2022 los niveles incrementaron más de lo que antes y luego los niveles se hicieron mucho mayores entre marzo y mayo del mismo año.

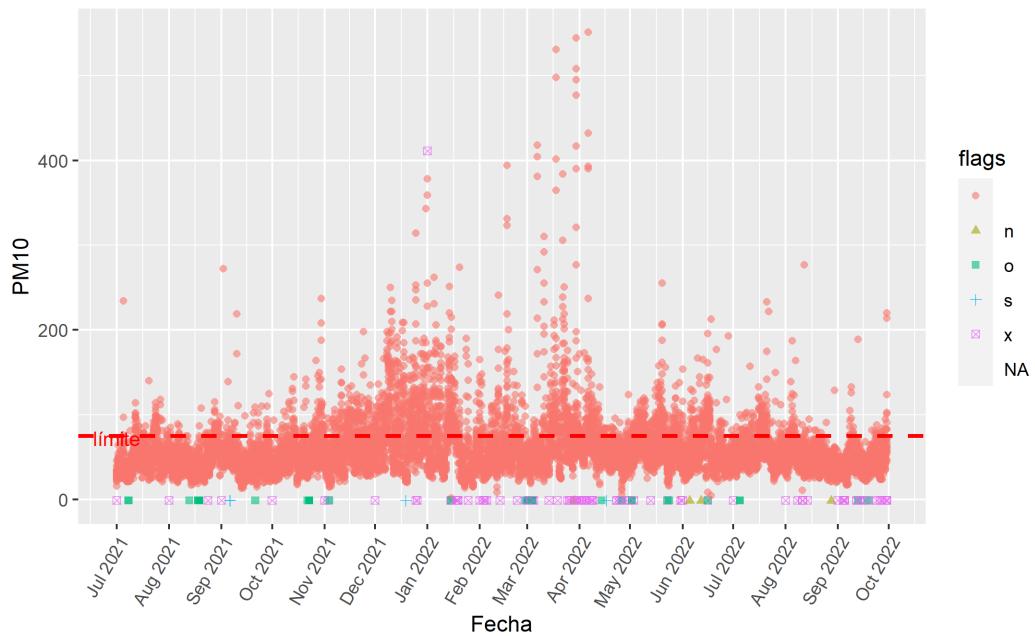


Figura 13: Niveles de PM_{10} sobrepasando la normativa

En cuanto a los niveles de $PM_{2.5}$, mostrados en la figura 14, se observa que por lo regular los niveles de este contaminante son muy cercanos al límite, pero que desde aproximadamente mediados de octubre del 2021 hasta junio del 2022 los niveles crecieron exponencialmente, especialmente al rededor de diciembre del 2021 y febrero del 2022.

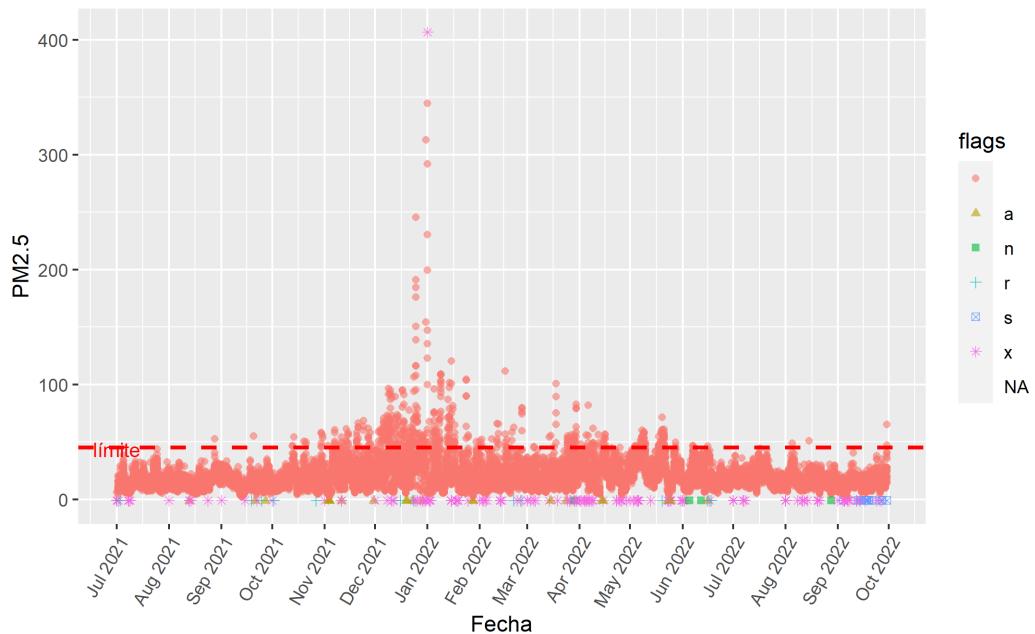


Figura 14: Niveles de $PM_{2.5}$ sobrepasando la normativa

Para el contaminante O_3 , se puede decir que la mayor parte del tiempo los niveles están por debajo del límite, aunque hay algunas excepciones, bastante reducidas, en que los niveles lo sobre-pasan.

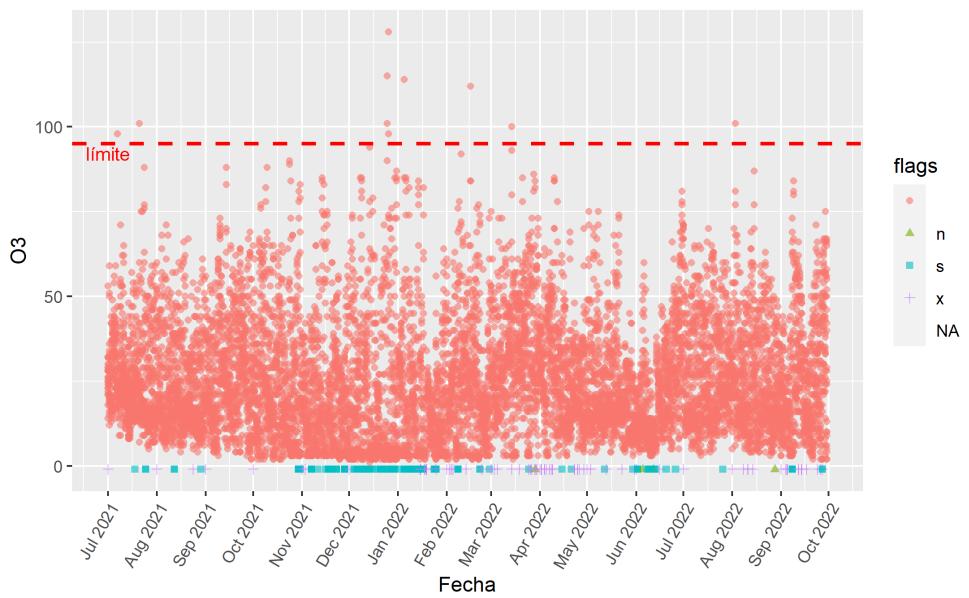


Figura 15: Niveles de O_3 sobrepasando la normativa

En el caso de los niveles del NO_2 , los niveles del componente están muy por debajo del límite, lo cual es bastante bueno. Aunque, aún por debajo del límite se distingue un aumento en los niveles del NO_2 al rededor de diciembre del 2021.

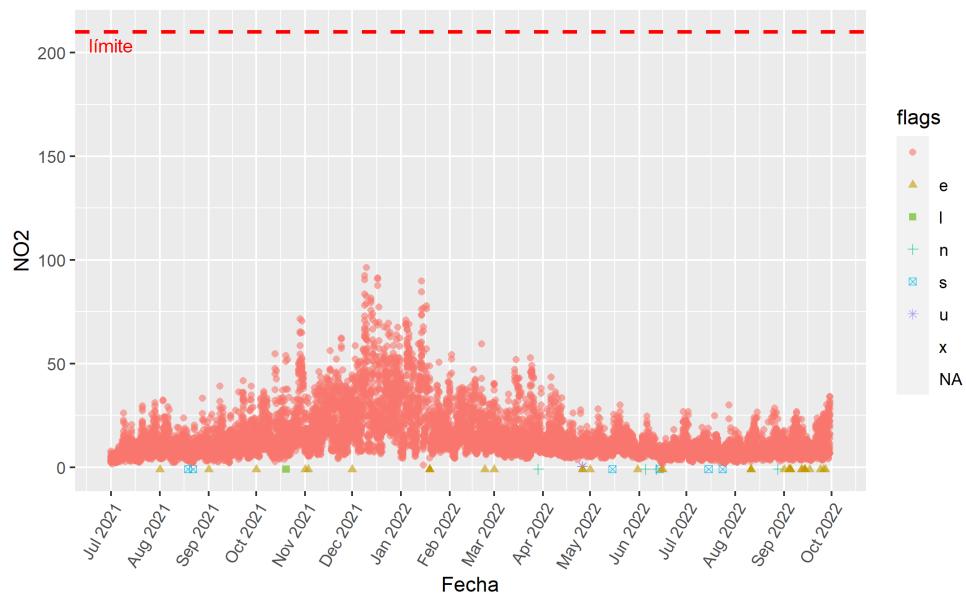


Figura 16: Niveles de NO_2 sobrepasando la normativa

Comparando los niveles de SO_2 contra los niveles de los otros contaminantes, el SO_2 es el contaminante con menores niveles y los más alejados al límite de las normativas.

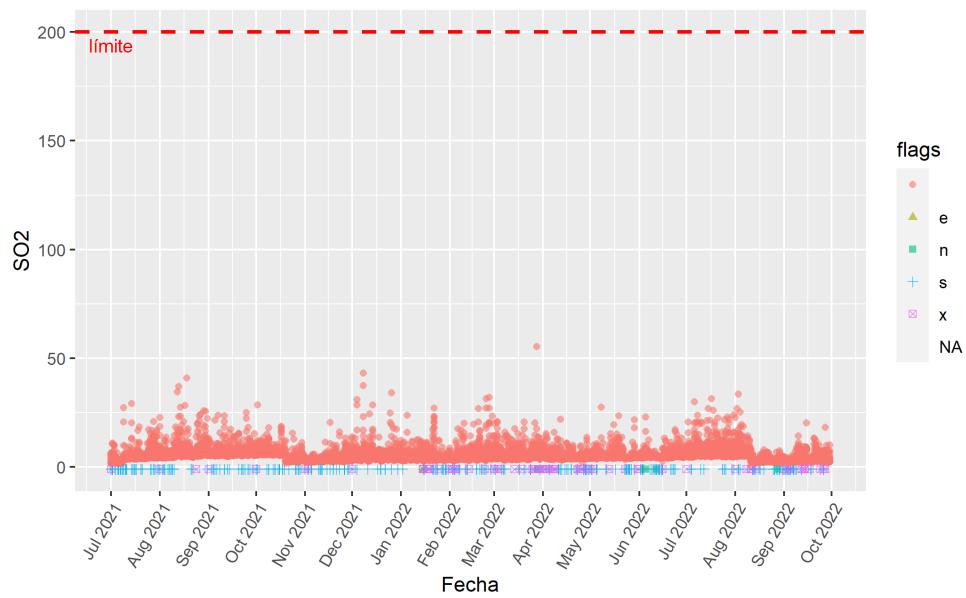


Figura 17: Niveles de SO_2 sobre pasando la normativa

Por último, el CO no sobrepasa en ningún punto el límite establecido y se puede distinguir también que en los últimos meses los niveles del contaminante se han reducido.

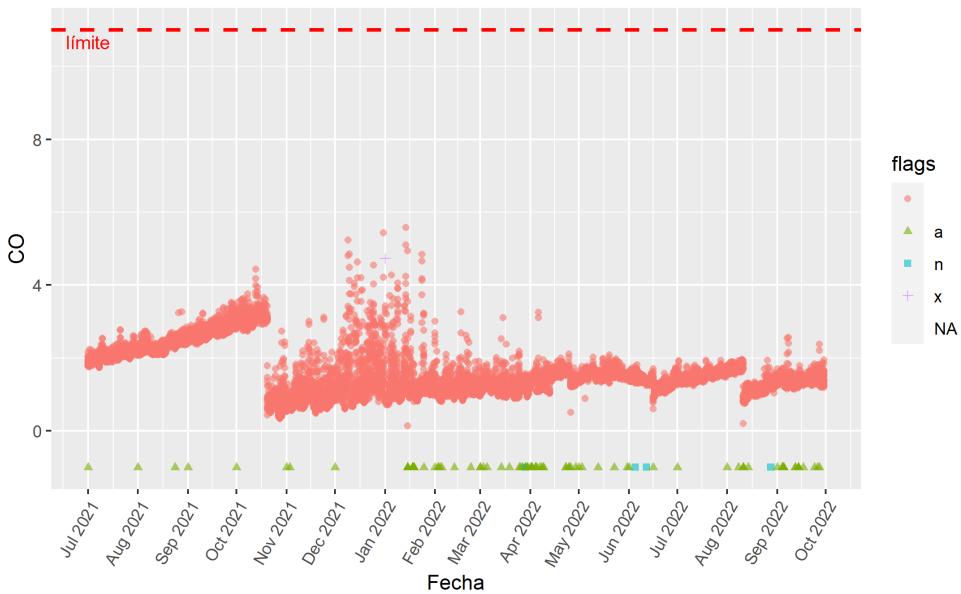


Figura 18: Niveles de CO sobrepasando la normativa

De la misma manera se generó un mapa de correlaciones entre los contaminantes criterio. En este caso, es importante destacar la correlación entre PM_{10} y $PM_{2.5}$. Asimismo, se podría considerar significativa la relación entre NO_2 y ambos tipos de materia particulada.

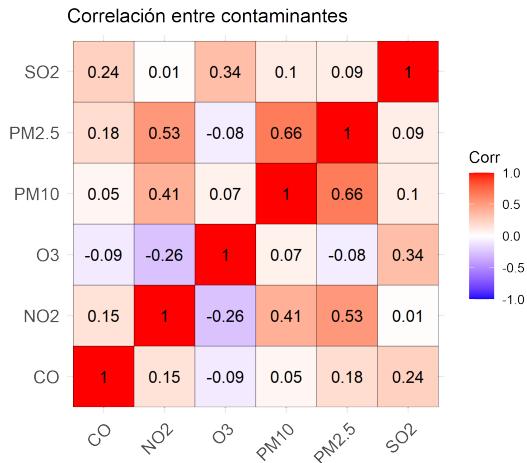


Figura 19: Correlación entre contaminantes criterio

19.3. Análisis de correlación entre parámetros meteorológicos y contaminantes criterio

Gran parte del objetivo de este proyecto yace en encontrar correlaciones entre variables meteorológicas y los contaminantes criterio. En la figura 20, se observa la matriz de correlación entre dichos parámetros y los contaminantes. En este diagrama destaca el contaminante O_3 , pues tiene correlaciones significativas con la Humedad Relativa (-0.55), Radiación Solar (0.64) y la Temperatura (0.42). También se observa una correlación relativamente alta de -0.52 entre el contaminante NO_2 y la velocidad del viento. Finalmente también es importante destacar la correlación de 0.3 entre el contaminante SO_2 y la radiación solar.

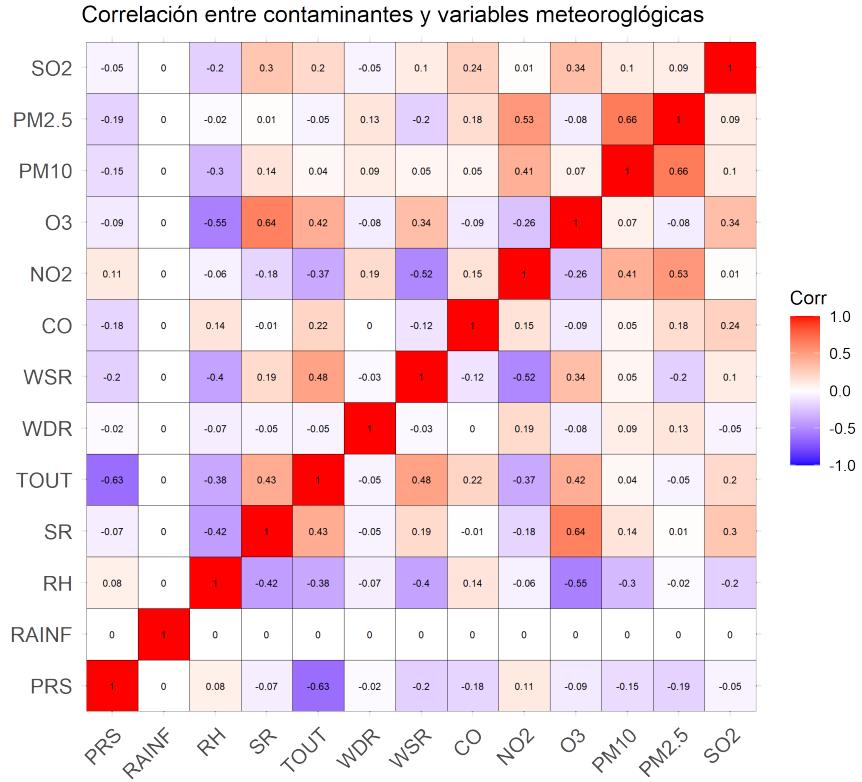


Figura 20: Correlación entre parámetros meteorológicos y contaminantes criterio

Cabe recalcar que para realizar las correlaciones se utilizó el método de Pearson, esto por que el método no es tan sensible a los outliers, lo que en este caso es de suma utilidad ya que dentro de los datos se cuenta con algunos datos atípicos. Por su parte, las correlaciones de Spearman toman en cuenta la ordinalidad de los datos y son monótonos, por esta razón este tipo de correlaciones fue descartado.

19.4. Análisis del comportamiento de los parámetros meteorológicos por día de la semana

Se realizó una análisis por día de la semana para cada parámetro meteorológico (véase Anexo A). En la figura 21 se puede observar cómo la dirección del viento cambia conforme las estaciones del año, por lo que las gráficas con colores parecidos se comportan de manera similar.



Figura 21: Dirección del viento en jueves

En esta gráfica se puede apreciar como la temperatura cambia a lo largo del año: los meses correspondientes a la estación de invierno tienden a tener temperaturas más bajas (alrededor de 15° en promedio), mientras que en los meses de verano, estas temperaturas aumentan (hasta aproximadamente 35° en promedio).

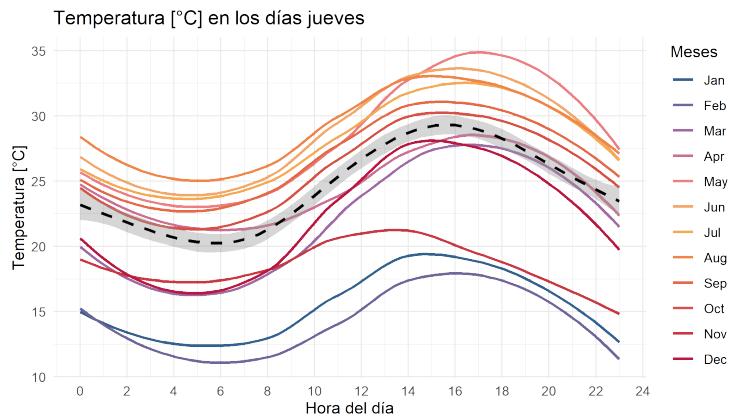


Figura 22: Temperatura en jueves

19.5. Análisis del comportamiento de los contaminantes criterio por día de la semana

De la misma manera, se realizó un análisis sobre el comportamiento de los contaminantes criterio cada día (véase Anexo B). En la figura 23 se puede observar cómo la cantidad de ozono en el ambiente comienza incrementar alrededor de la 8 hrs hasta las 14 hrs aproximadamente. Posterior a esto comienza a decrecer. Este patrón se repite diariamente y podría explicarse por el tráfico vehicular.

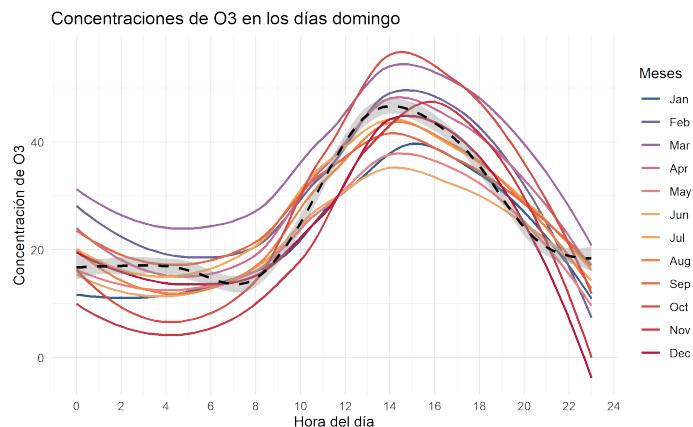


Figura 23: Concentración del ozono en domingo

Otro comportamiento relevante en este proyecto es el del contaminante PM_{10} . Como se observa en la figura 24, existe un pico en su concentración entre las 10 y las 14 hrs durante el mes de marzo. Valdría la pena investigar a fondo la causa de esta variación; sin embargo, los festejos durante la Semana Santa.

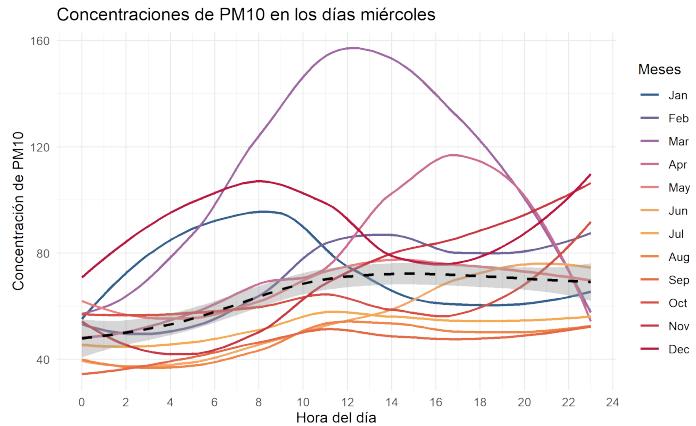


Figura 24: Concentración de PM_{10} en domingo

19.6. Índice de Aire y Salud de los Contaminantes Criterio

Finalmente, se realizaron gráficas para observar el índice salud y aire a lo largo de años (véase C). Entre ellas se destacan las correspondientes al O_3 , $PM_{2.5}$ y PM_{10} .

En la figura 25 se observa que el índice de aire y salud correspondiente al O_3 se mantiene constante durante todo el año, variando entre el color verde y el color amarillo principalmente. Esto indica que el nivel de ozono es aceptable a lo largo de casi todo el año.

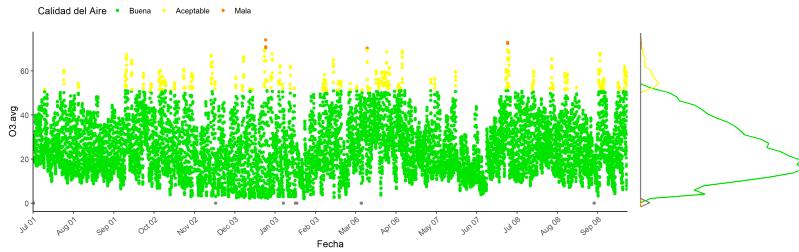


Figura 25: O_3 : Índice aire y salud

Por otro lado, en la gráfica correspondiente al índice de aire y salud del contaminante $PM_{2.5}$ se puede notar un pico en las mediciones entre diciembre y enero donde alcanzan la categoría morada (extremadamente mala). Tomando en cuenta las fechas, esto podría deberse a la pirotecnia utilizada durante Navidad y Año Nuevo.

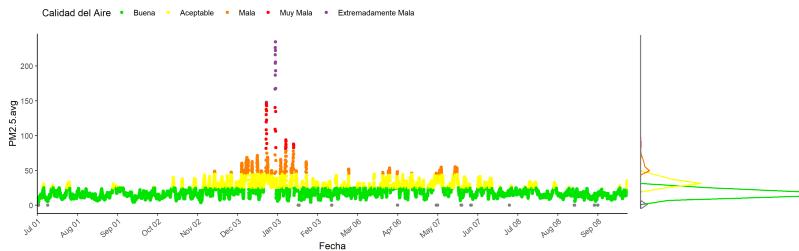


Figura 26: $PM_{2.5}$: Índice aire y salud

De la misma manera, en la figura 27 se observa un pico en el índice de PM_{10} entre diciembre y enero, así como entre marzo y abril; en ambas ocasiones llegando a una calidad del aire extremadamente mala. Al igual que con la materia particulada 2.5, estos picos podrían ser explicados por la pirotecnia utilizada por Navidad, Años Nuevo y Semana Santa. Asimismo, en noviembre del 2021, comenzó una época de sequía, en la que se puede ver un aumento de las partículas, tanto en $PM_{2.5}$ como PM_{10} . Además de que, en los meses de marzo y abril del 2022, se observa un pico en la gráfica de PM_{10} , que puede ser explicado por los incendios que hubo durante ese periodo.

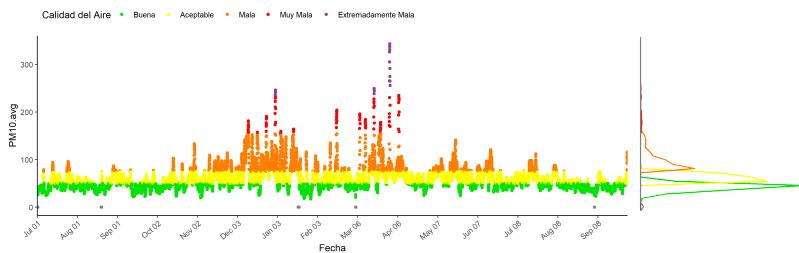


Figura 27: PM_{10} : Índice aire y salud

20. Adecuación y validación del Modelo

20.1. Multivariedad

Se realizó la prueba de Mardia de multivariedad y no se identificó alguna, igualmente, no se presentó normalidad en ninguna de nuestras variables a analizar.

20.2. Análisis por Componentes Principales

En esta ocasión se optó por el Análisis de Componentes Principales puesto a que este ayuda a simplificar la complejidad de los espacios muestrales con muchas dimensiones mientras que se conserva la información Amat Rodrigo, 2017. A su vez esto facilita la comprensión de la variabilidad de los datos, ya que en este caso se cuenta con demasiadas variables tanto meteorológicas como de contaminantes. Por lo tanto el objetivo de utilizar este tipo de análisis es reducir la cantidad de variables involucradas en la investigación para así poder obtener un modelo más sencillo y eficaz a partir de los componentes establecidos durante esta etapa.

Primero se realizó una gráfica de variables, en la cual se puede interpretar a los vectores que se acerquen más a la circunferencia como los más influyentes, por esta razón se puede decir que TOUT, PRS, O3 y RH son de los más relevantes.

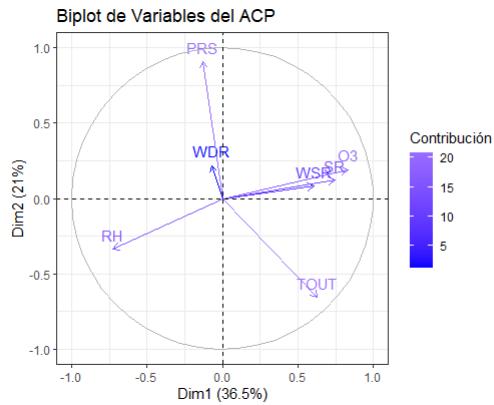


Figura 28: Biplot de Variables del ACP

La mayor parte de los individuos se encuentran en un rango positivo o cercano a cero (aunque hay algunas excepciones), mientras que el dominio negativo tiene una mayor concentración de individuos en un menor espacio y el dominio positivo tiene una cantidad similar de individuos pero están más distribuidos en el área.

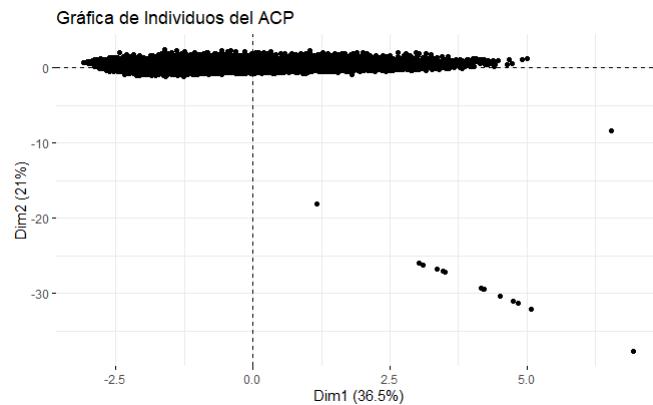


Figura 29: Gráfica de Individuos del ACP

Observando los porcentajes de varianza de los principales componentes se puede deducir que la variabilidad es mayormente explicada por el primer componente principal y de manera significativa por el 2°, 3° y el 4°, pero la aportación del 6° y 7° son cuestionables. Por su parte, empleando porcentajes acumulados de las varianzas se puede observar que utilizando los siete componentes se

llega a explicar el modelo al cien por ciento, aunque desde el tercer componente se obtiene un 72 % , que es un buen porcentaje de variabilidad explicada.

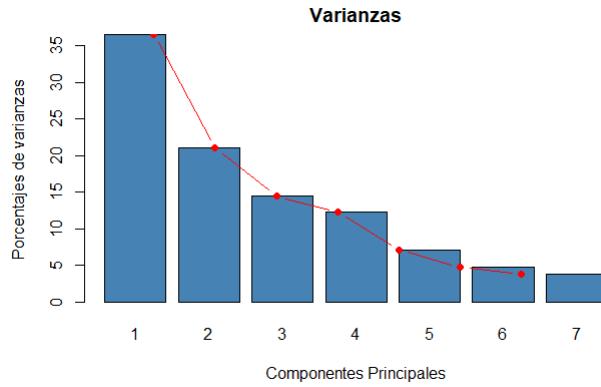


Figura 30: Varianzas

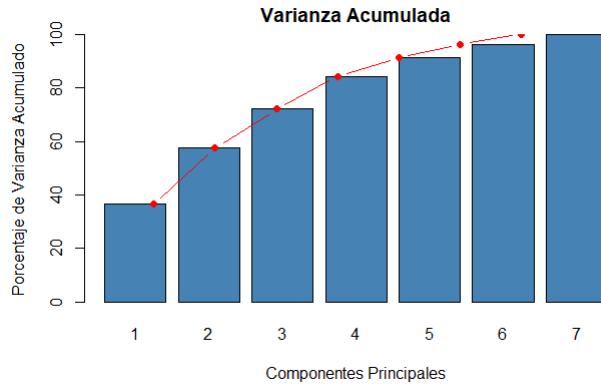


Figura 31: Varianza Acumulada

20.3. Análisis factorial

Al igual que en el análisis por componentes principales, el análisis factorial es una herramienta de gran utilidad para reducir la cantidad de variables utilizadas en el modelo. A través de este método se pueden obtener variables no observadas (factores latentes) que expliquen la variabilidad y las correlaciones de las variables observadas. A partir de este procedimiento se obtuvieron un

total de cinco factores, sin embargo, con un solo factor se puede llegar a explicar el noventa por ciento de los datos. Para visualizar esto, se puede observar el siguiente diagrama:

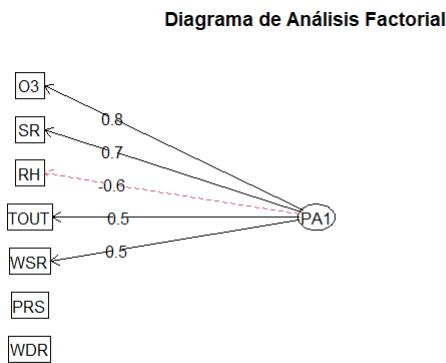


Figura 32: Diagrama Análisis Factorial

Para evaluar la efectividad de este método, se realizó la prueba KMO y se obtuvo un resultado de 0.61 en promedio, el cual indica un nivel de aceptabilidad "Medioocre", sin embargo, al realizar la prueba de Esfericidad, se obtuvo una correlación lo suficientemente aceptable para realizar el análisis factorial, con un valor p significante.

Adicionalmente se realizó la prueba de Bartlett, el cual sugería una correlación significante de los datos en el análisis factorial. Se obtuvo también que la hipótesis nula se rechaza porque el valor p es menor a 0.05.

En el siguiente gráfico se muestra la fuerza de correlación entre la variable y el factor. En esta se utilizaron valores absolutos para no desperdiciar espacio, por lo que las barras con colores fríos (tirándole al azul) denotan una correlación positiva mientras que las barras con colores cálidos (tirándole al rojo) denotan una correlación negativa. Es prudente recalcar también que las correlaciones cercanas a cero tienen colores que no resaltan ya que no son tan relevantes como los demás. En este caso la variable con mayor correlación es el O3 y la de menor correlación es la de RH.

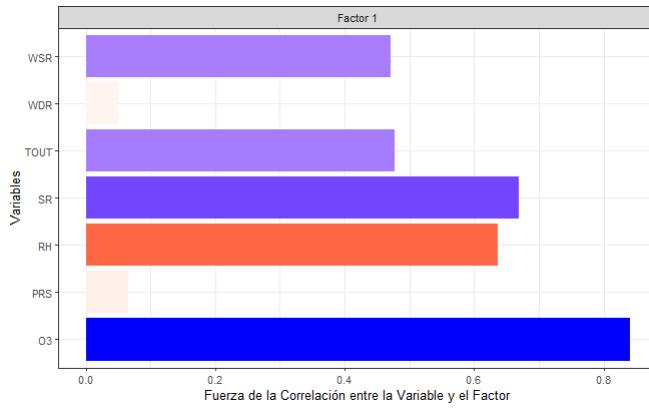


Figura 33: Fuerzas de Correlación

También se muestra en el gráfico a continuación una comparativa de la sedimentación utilizando el Análisis de Componente Principal y el Análisis Factorial. Esta gráfica se puede leer como que los factores que estén por encima del eigen valor de 1.0 son los que se requieren para tener un buen porcentaje de varianza explicada. Por ello se puede concluir a partir de ella que el análisis de componentes principales tres factores son suficientes para explicar la variabilidad mientras que con el análisis factorial solo se requiere uno.

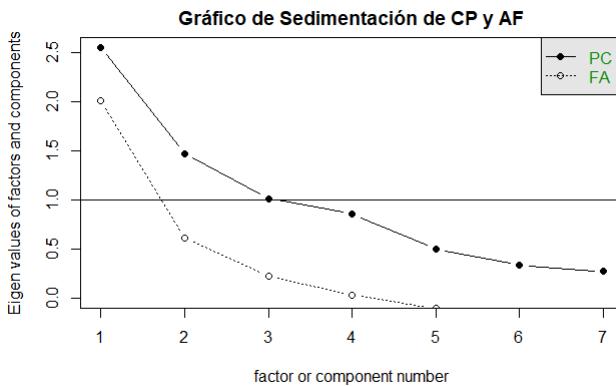


Figura 34: Gráfico de Sedimentación

20.4. Regresión lineal múltiple

En el análisis de correlación realizado anteriormente en el análisis gráfico se obtuvo que el ozono es uno de los contaminantes criterio con correlaciones de mayor relevancia respecto a los parámetros meteorológicos, entre ellos se destacan la humedad relativa, radiación solar y temperatura. Por esta razón, se consideró pertinente generar un modelo capaz de predecir cierta variabilidad en las mediciones del ozono. Cabe destacar que a pesar de que los tres parámetros anteriormente mencionados fueron los más destacados, después de haber utilizado el criterio Akaike (modelo matemático para evaluar la calidad del modelo generado) se optó por utilizar todos los parámetros meteorológicos.

En las gráficas a continuación se pueden observar las relaciones lineales entre los predictores numéricos y las variables dependientes. En la mayoría de estos se pueden observar datos atípicos, aunque en algunas de las gráficas como en la de SR, WSR y WDR se puede apreciar mejor la linealidad del modelo.

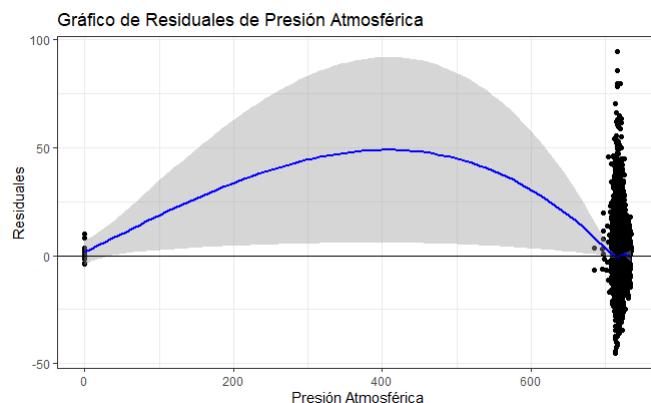


Figura 35: Gráfica de Residuales de Presión Atmosférica

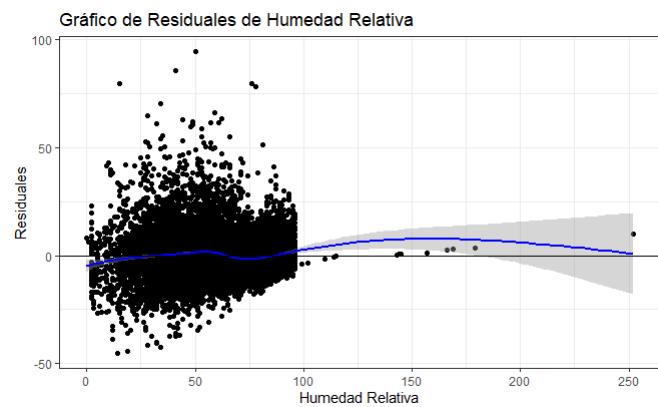


Figura 36: Gráfica de Residuales de Humedad Relativa

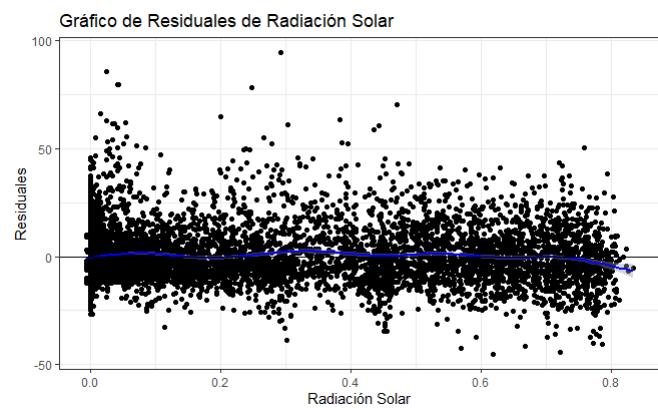


Figura 37: Gráfica de Residuales de Radiación Solar

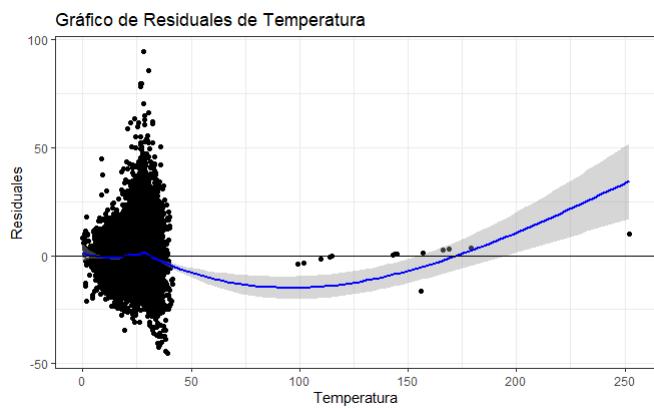


Figura 38: Gráfica de Residuales de Temperatura

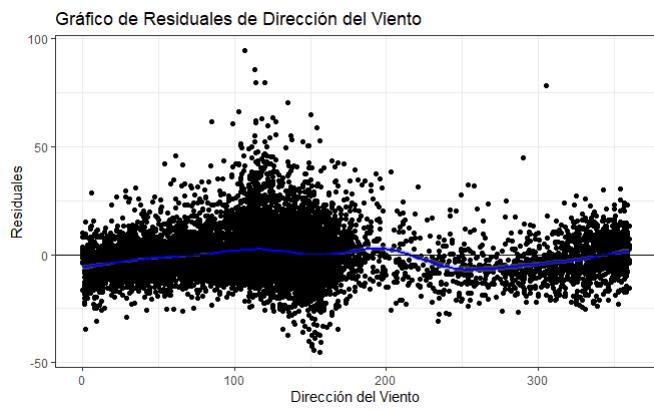


Figura 39: Gráfica de Residuales de Dirección del Viento

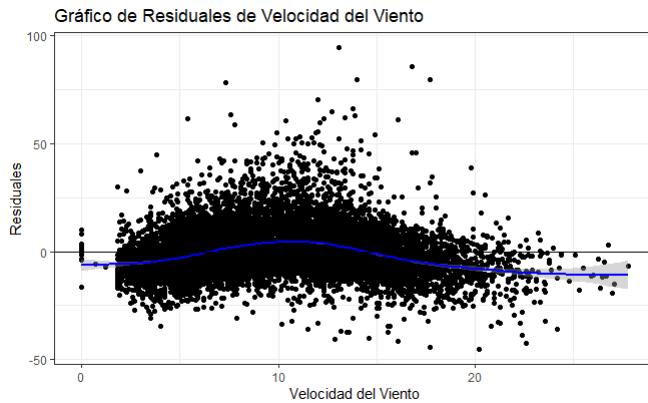


Figura 40: Gráfica de Residuales de Velocidad del Viento

Y la ecuación del modelo de regresión lineal múltiple sería la siguiente:

$$\hat{Y} = 12.63 + 0.02 * \text{Presión Atmosférica} - 0.23 * \text{Humedad Relativa} + 30.37 * \text{Radiación Solar} \\ + 0.15 * \text{Temperatura} - 0.02 * \text{Dirección del Viento} + 0.39 * \text{Velocidad del Viento} + \epsilon$$

Se obtuvo una R^2 ajustada de 0.54, con error estándar de 11.17 y valor p menor al 0.05, por lo que si bien el modelo no es muy preciso, es significante.

21. Resultados

Analizando las gráficas anexadas en la sección anterior, se muestran varios comportamientos, ya sean recurrentes u ocasionales, que se pueden relacionar con las condiciones de nuestro día a día y festividades. Entre ellos podemos ver que se produce un aumento en el ozono a partir de las 8 de la mañana hasta las 6 de la tarde, una posible explicación a esto es que este aumento es causado por el tráfico por ser este un horario de circulación regular. Otra de las observaciones que se hicieron fue que tanto el $PM_{2.5}$ como el PM_{10} tienen picos al rededor de fechas como Navidad y Año Nuevo. Además de ese pico, el PM_{10} también tiene un aumento aún mayor durante Semana Santa, lo cual podría estar relacionado a las celebraciones religiosas tal como el miércoles de ceniza.

Adicionalmente, se considera prudente remarcar que, basado en las distintas estaciones del año, la dirección del viento varia, por ende esto puede causar que los contaminantes sean arrastrados a otras locaciones.

22. Conclusiones

La contaminación es un problema que pone en riesgo la salud de los seres vivos, por ende es indispensable cuantificarla para poder medir los niveles de la misma. La solución, específicamente considerando la problemática atmosférica, tiene varios contaminantes principales, entre ellos el CO , O_3 , NO_2 , SO_2 , PM_{10} y $PM_{2.5}$. Para analizar los comportamientos de estos contaminantes se utilizaron el análisis de correlación, análisis de componentes principales, análisis factorial y regresión lineal múltiple. A partir de los gráficos generados con los diversos tipos de análisis, se facilitó la interpretación de los mismos y la comprensión de los comportamientos de cada uno de los componentes. Gracias a estos gráficos, se hicieron ciertas observaciones en las cuales se detectaron ciertos comportamientos que pueden ser explicados por algunas fechas festivas o actividades que desarrollamos en el día a día. Asimismo, se encontraron tendencias dependientes de la estación o mes del año en el que se tomaron las mediciones. Finalmente, cabe recalcar la importancia de realizar estos estudios pues permiten generar estrategias específicas así como detectar nuevas problemáticas ambientales.

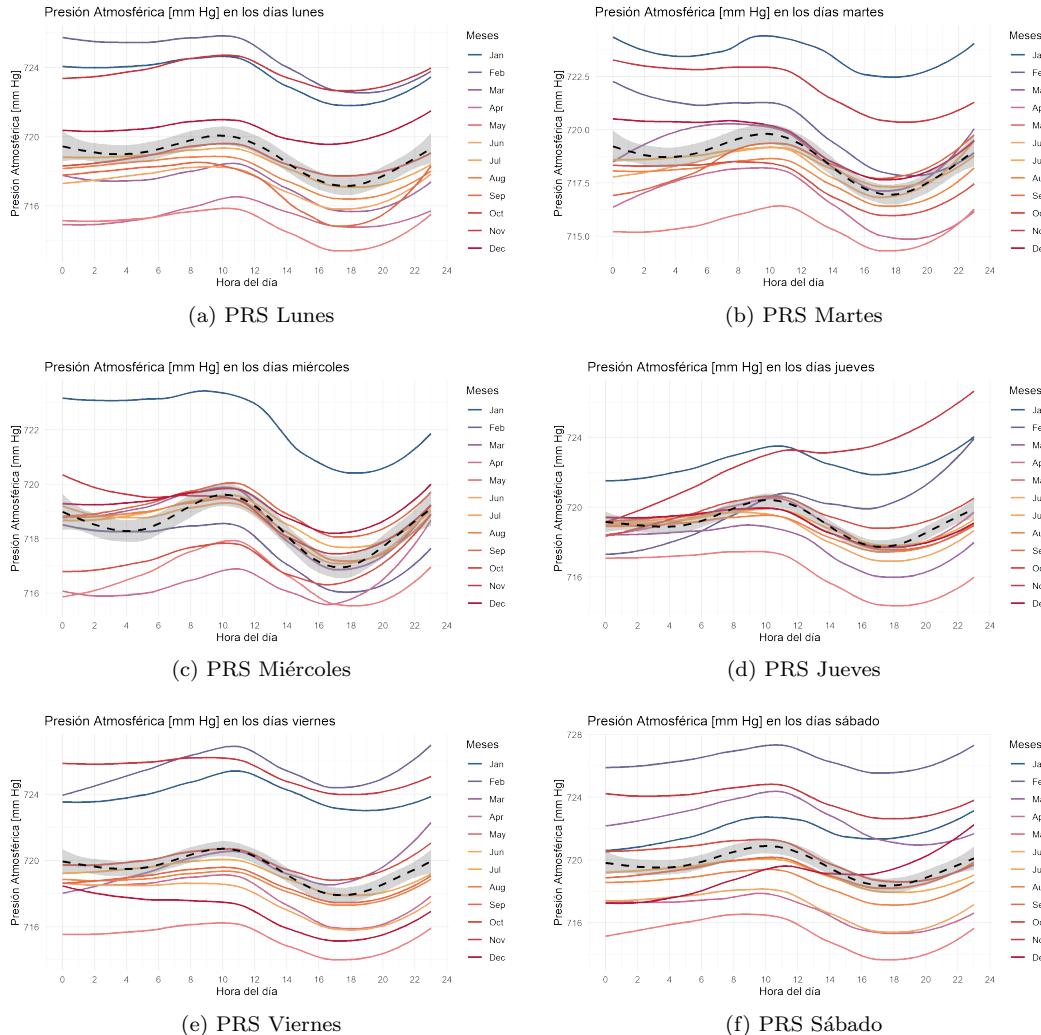
Referencias

- Aire NL. (2022). <http://aire.nl.gob.mx/index.html>
- Amat Rodrigo, J. (2017). Análisis de Componentes Principales (Principal Component Analysis, PCA) y t-SNE. https://www.cienciadedatos.net/documentos/35_principal_component_analysis
- Belavadi, S. V., Rajagopal, S., R, R., & Mohan, R. (2020). Air Quality Forecasting using LSTM RNN and Wireless Sensor Networks [The 11th International Conference on Ambient Systems, Networks and Technologies (ANT) / The 3rd International Conference on Emerging Data and Industry 4.0 (EDI40) / Affiliated Workshops]. *Procedia Computer Science*, 170, 241-248. <https://doi.org/https://doi.org/10.1016/j.procs.2020.03.036>
- En 2060, la contaminación atmosférica causará de 6 a 9 millones de muertes prematuras al año y tendrá un costo de 1 % del PIB – OCDE - OECD. (s.f.). <https://www.oecd.org/centrodemexico/medios/en-2060-la-contaminacion-atmosferica-causara-de-6-a-9-millones-de-muertes-prematuras-al-ao-y-tendra-un-costo-de-1-del-pibocde.htm>
- Fisk, W. J. (2015). Review of some effects of climate change on indoor environmental quality and health and associated no-regrets mitigation measures. *Building and Environment*, 86, 70-80. <https://doi.org/https://doi.org/10.1016/j.buildenv.2014.12.024>
- FortuneBusinessInsights. (2020). Air Quality Monitoring System Market Size, Share & COVID-19 Impact Analysis, By Type (Indoor Monitors and Outdoor Monitors) By End-User (Commercial & Residential, Public Infrastructure, Power Generation Plants, Pharmaceutical Industry, and Others), and Regional Forecast, 2021-2028. <https://www.fortunebusinessinsights.com/air-quality-monitoring-system-market-105614>
- Javier, C. (2022). Instituto Nacional de Sismología, Vulcanología, Meteorología e Hidrología. <https://insivumeh.gob.gt/?p=61234>
- Latin America Air Pollution City. (s.f.). <https://0-www-statista-com.biblioteca ils.tec.mx/statistics/1029132/latin-america-air-pollution-city/>
- Liu, X., Ngai, E., & Zachariah, D. (2021). Scalable Belief Updating for Urban Air Quality Modeling and Prediction. *ACM/IMS Transactions on Data Science*, 2(1), 1-19. <https://doi.org/10.1145/3402903>

- Naveen L., M. H. (2019). Atmospheric Weather Prediction Using various machine learning Techniques: A Survey. <https://ieeexplore.ieee.org/abstract/document/8819643>
- OMS. (2021a). *Directrices mundiales de la OMS sobre la calidad del aire Resumen ejecutivo*. <https://apps.who.int/iris/bitstream/handle/10665/346062/9789240035461-spa.pdf>
- OMS. (2021b). Las nuevas Directrices mundiales de la OMS sobre la calidad del aire tienen como objetivo evitar millones de muertes debidas a la contaminación del aire. <http://aire.nl.gob.mx/>
- OMS. (2022a). Air Quality and Health. <https://www.who.int/teams/environment-climate-change-and-health/air-quality-and-health/health-impacts>
- OMS. (2022b). Contaminación Atmosférica. https://www.who.int/es/health-topics/air-pollution#tab=tab_1
- Proietti, E. (2013). Air pollution during pregnancy and neonatal outcome: a review. <https://pubmed.ncbi.nlm.nih.gov/22856675/#:~:text=Harmful%5C%20effects%5C%20of%5C%20exposure%5C%20to,early%5C%20alterations%5C%20in%5C%20immune%5C%20development.>
- Ritchie, H., & Roser, M. (2021). Air Pollution. <https://ourworldindata.org/air-pollution>
- Schiweck, A., Uhde, E., Salthammer, T., Salthammer, L. C., Morawska, L., Mazaheri, M., & Kumar, P. (2018). Smart homes and the control of indoor air quality. *Renewable and Sustainable Energy Reviews*, 94, 705-718. <https://doi.org/https://doi.org/10.1016/j.rser.2018.05.057>
- SIMA. (2015). SISTEMA INTEGRAL DE MONITOREO AMBIENTAL. <http://aire.nl.gob.mx/>
- SIMA. (2021). NORMATIVIDAD. http://aire.nl.gob.mx/nor_normatividad.html

A. Análisis del comportamiento de los parámetros meteorológicos por día de la semana

A.1. PRS - Presión Atmosférica



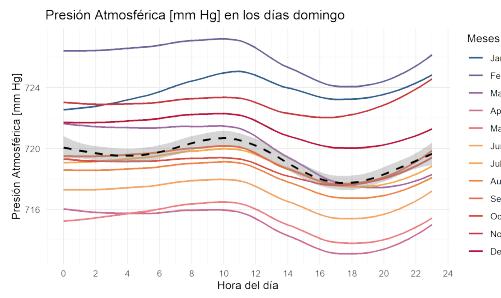
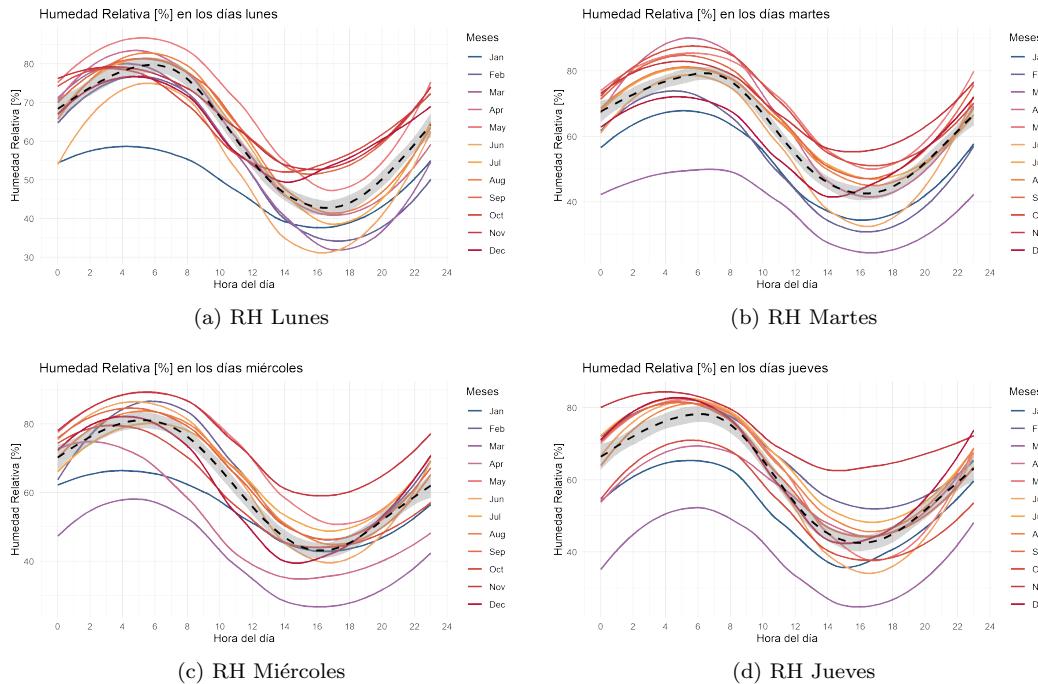
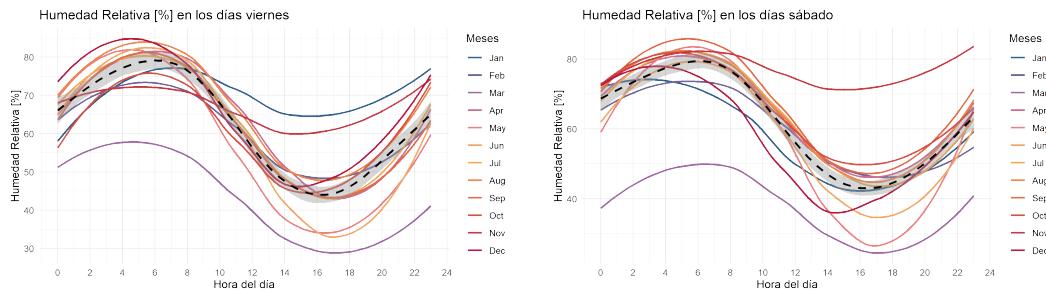


Figura 41: PRS Domingo

A.2. RH - Humedad Relativa





(e) RH Viernes

(f) RH Sábado

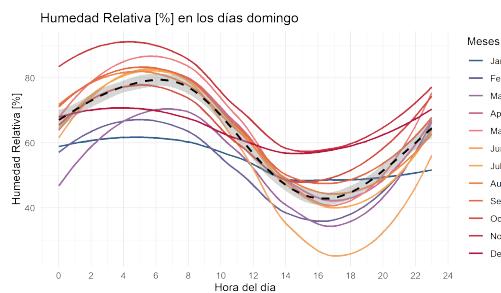
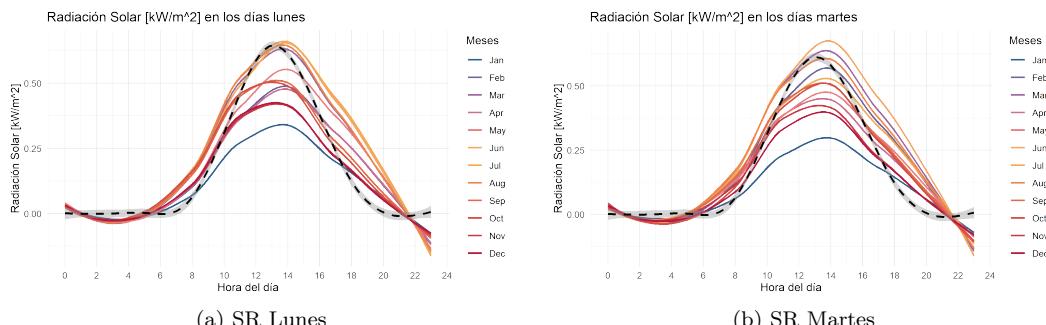


Figura 42: RH Domingo

A.3. SR - Radiación Solar



(a) SR Lunes

(b) SR Martes

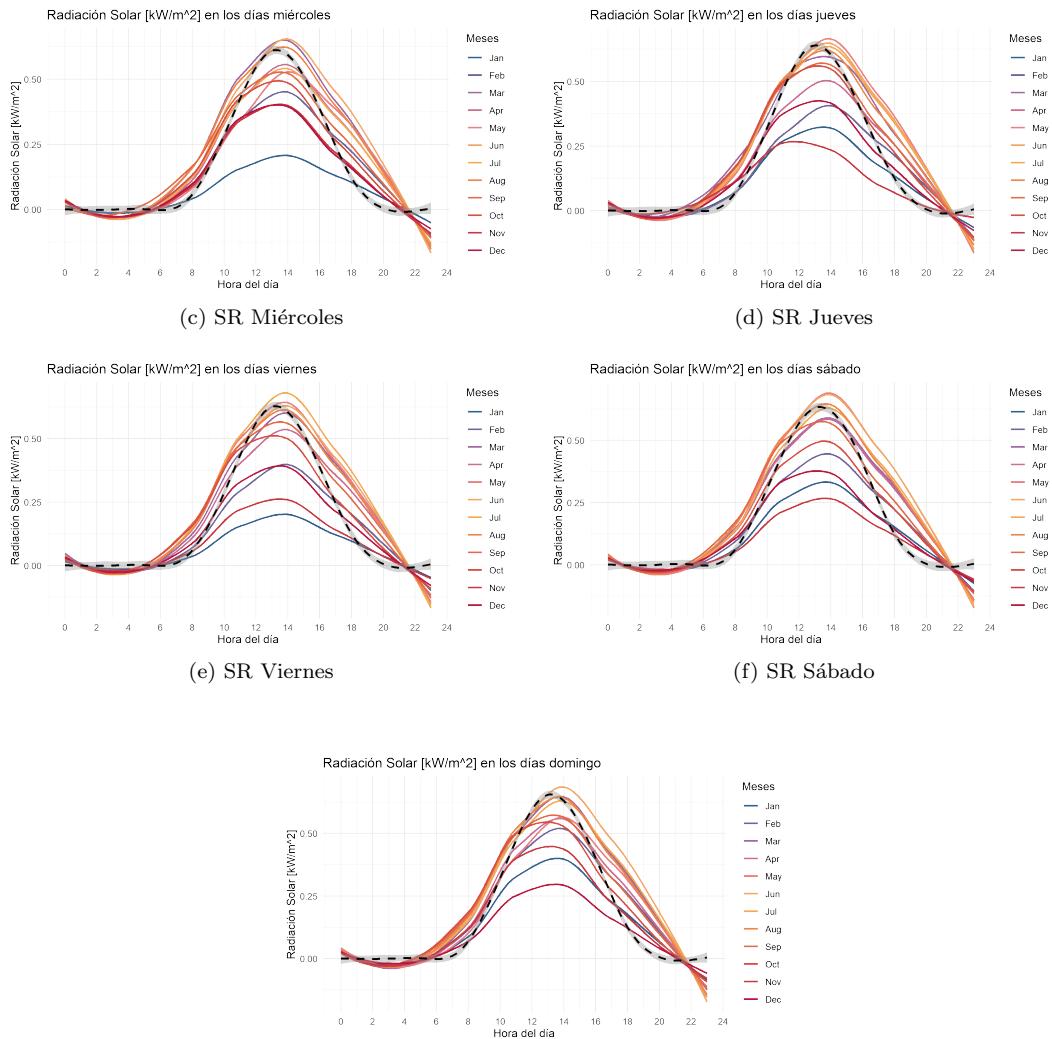
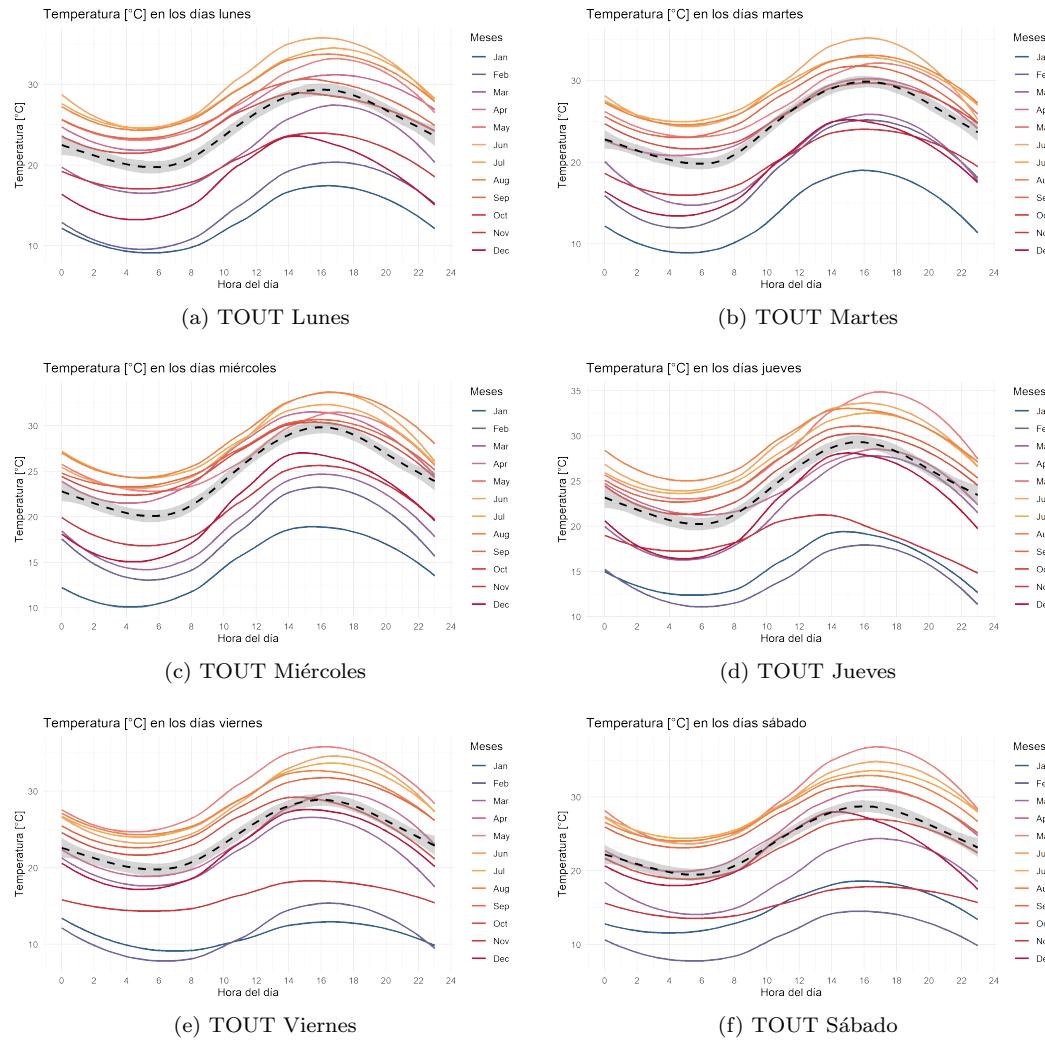


Figura 43: SR Domingo

A.4. TOUT - Temperatura



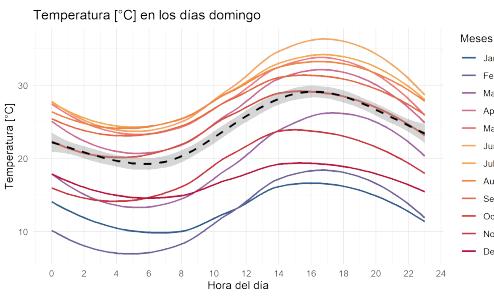
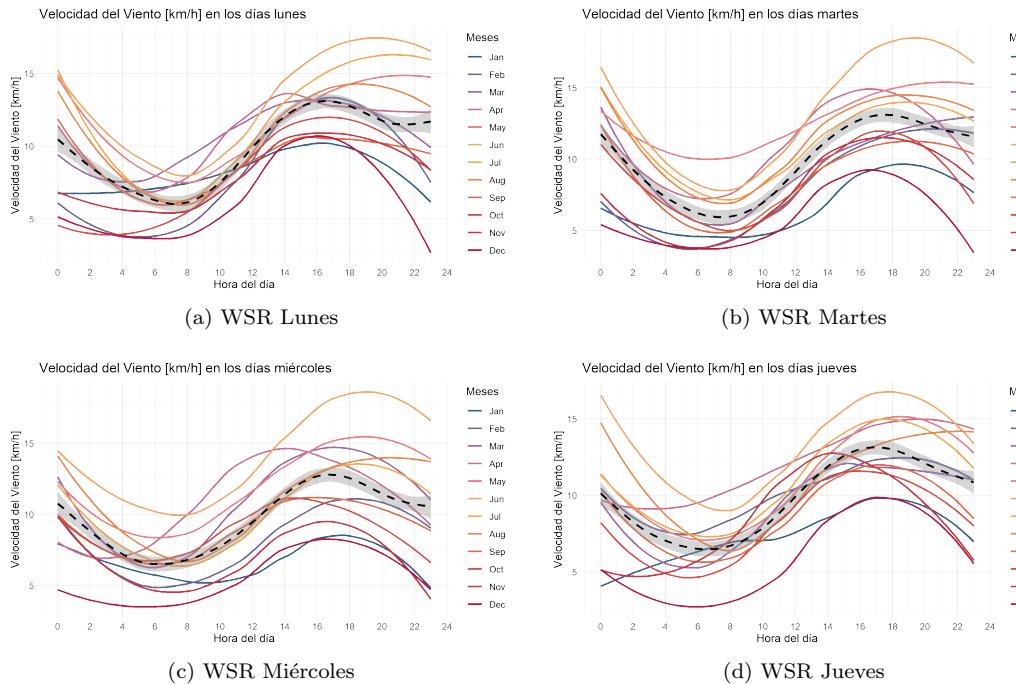


Figura 44: TOUT Domingo

A.5. WSR - Velocidad del Viento



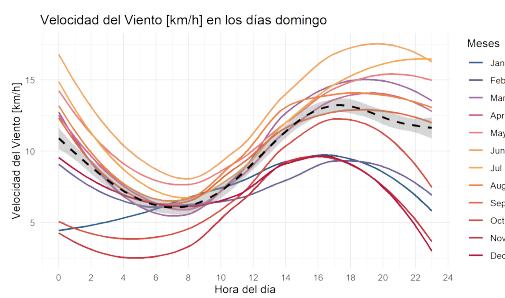
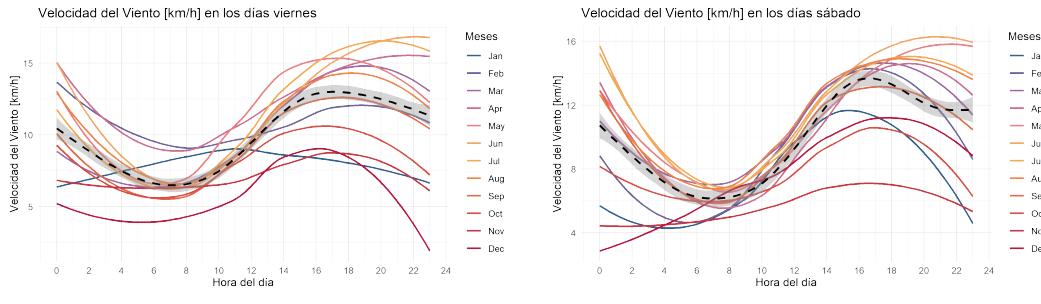
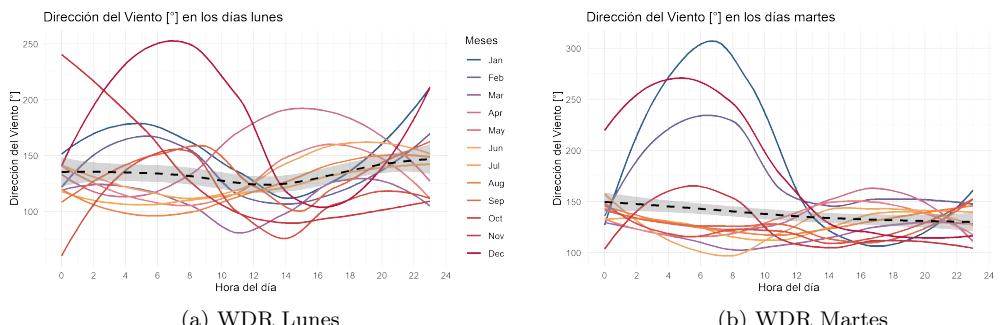


Figura 45: WSR Domingo

A.6. WDR - Dirección del viento



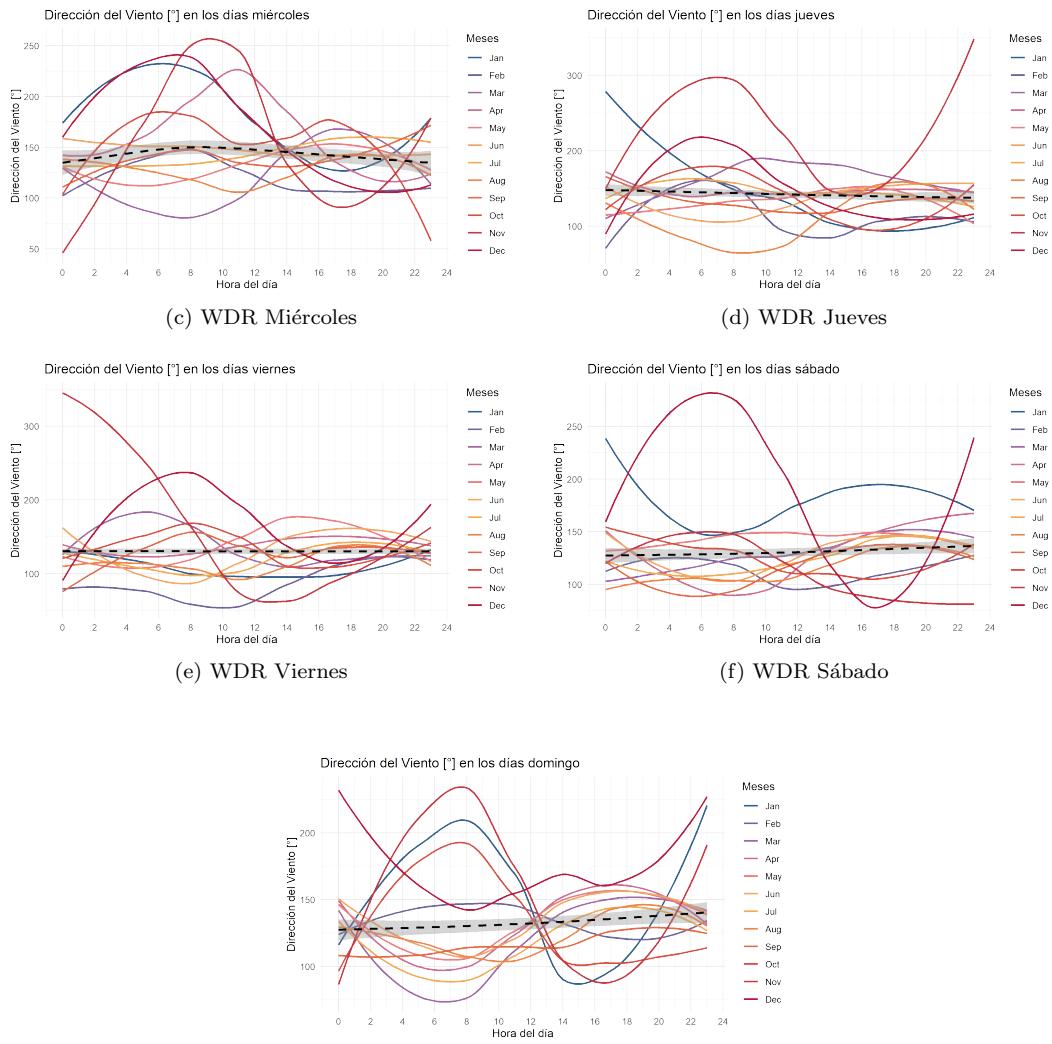
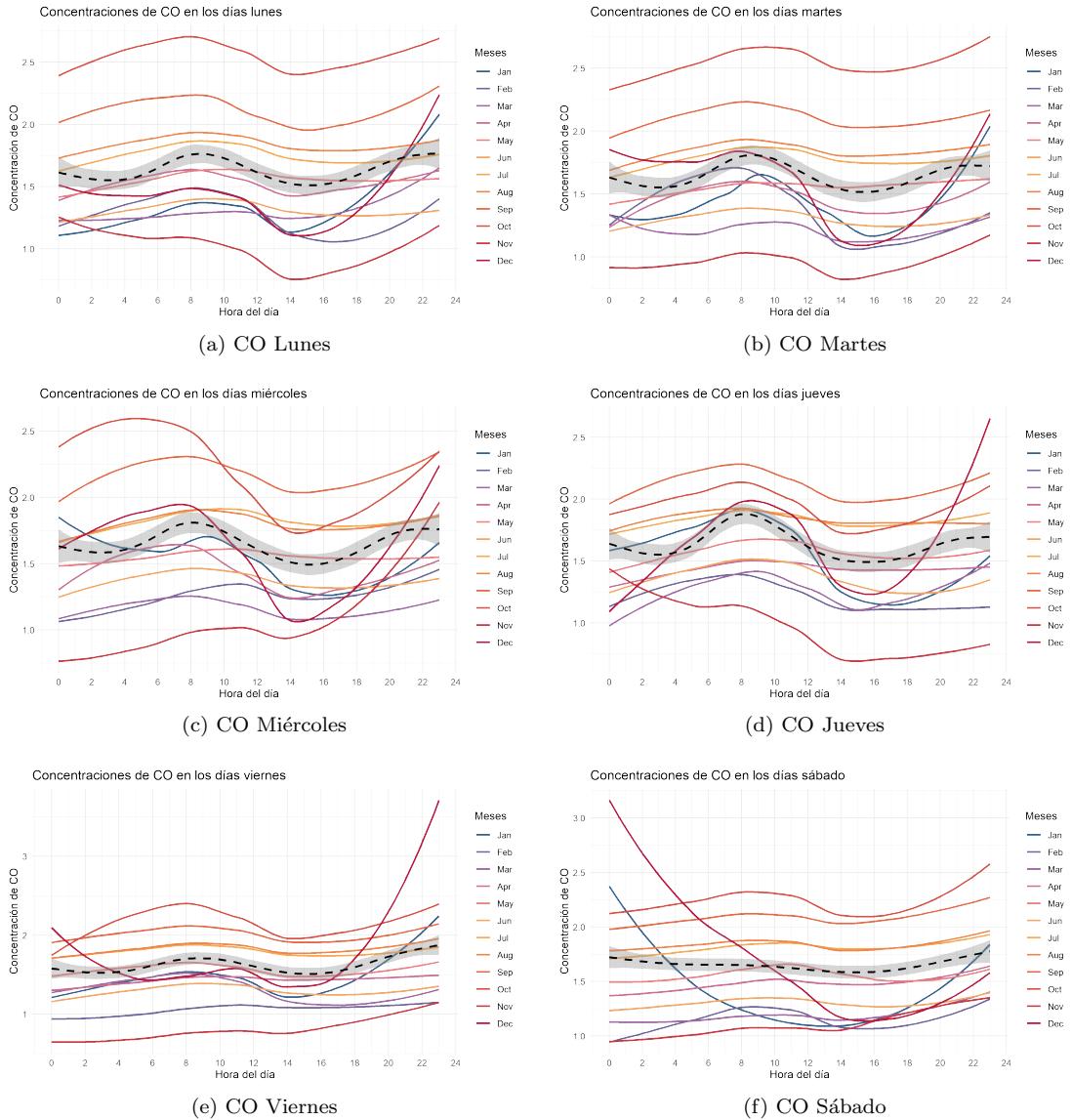


Figura 46: WDR Domingo

B. Análisis del comportamiento de los contaminantes crítico por día de la semana

B.1. CO - Monóxido de Carbono



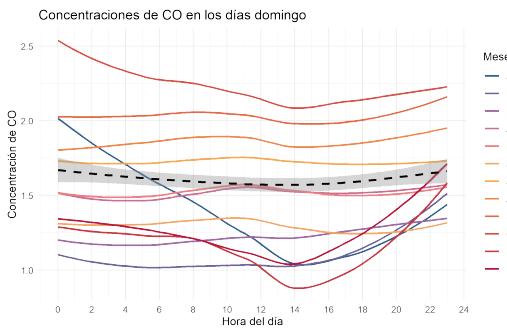
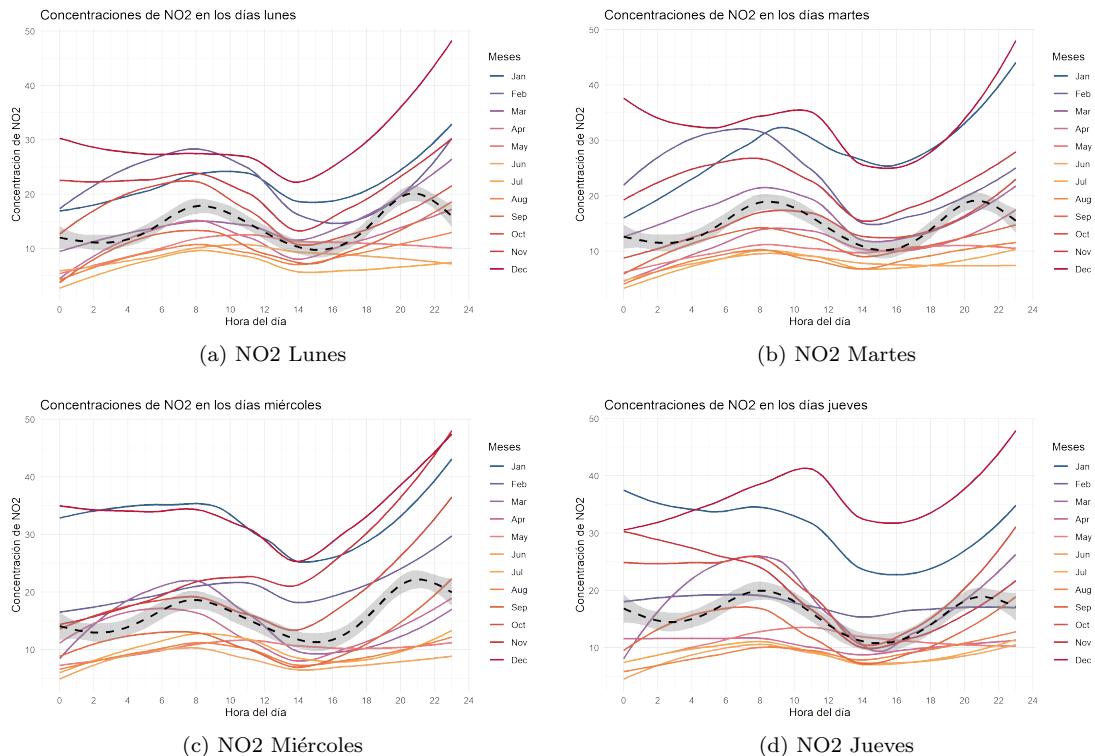


Figura 47: CO Domingo

B.2. NO_2 - Dióxido de Nitrógeno



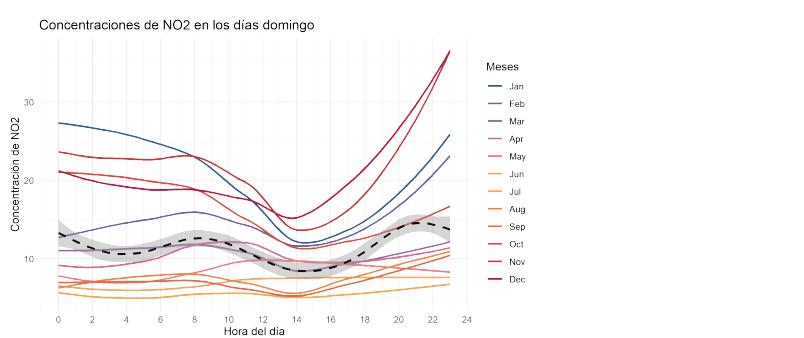
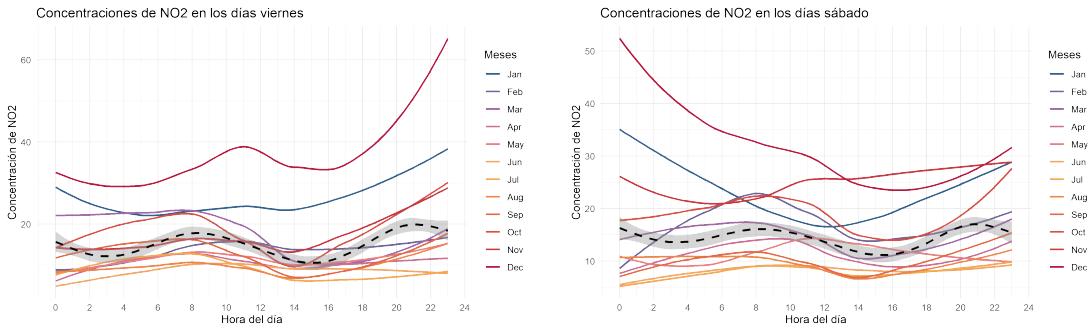
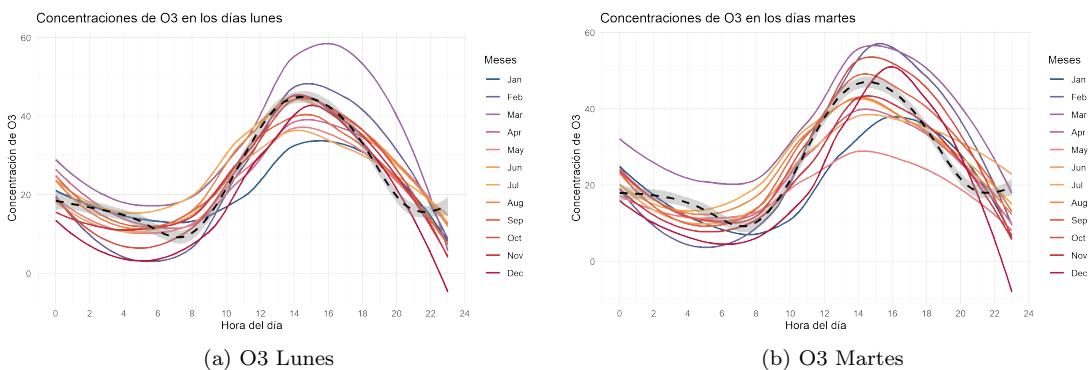


Figura 48: NO₂ Domingo

B.3. O₃ - Ozono



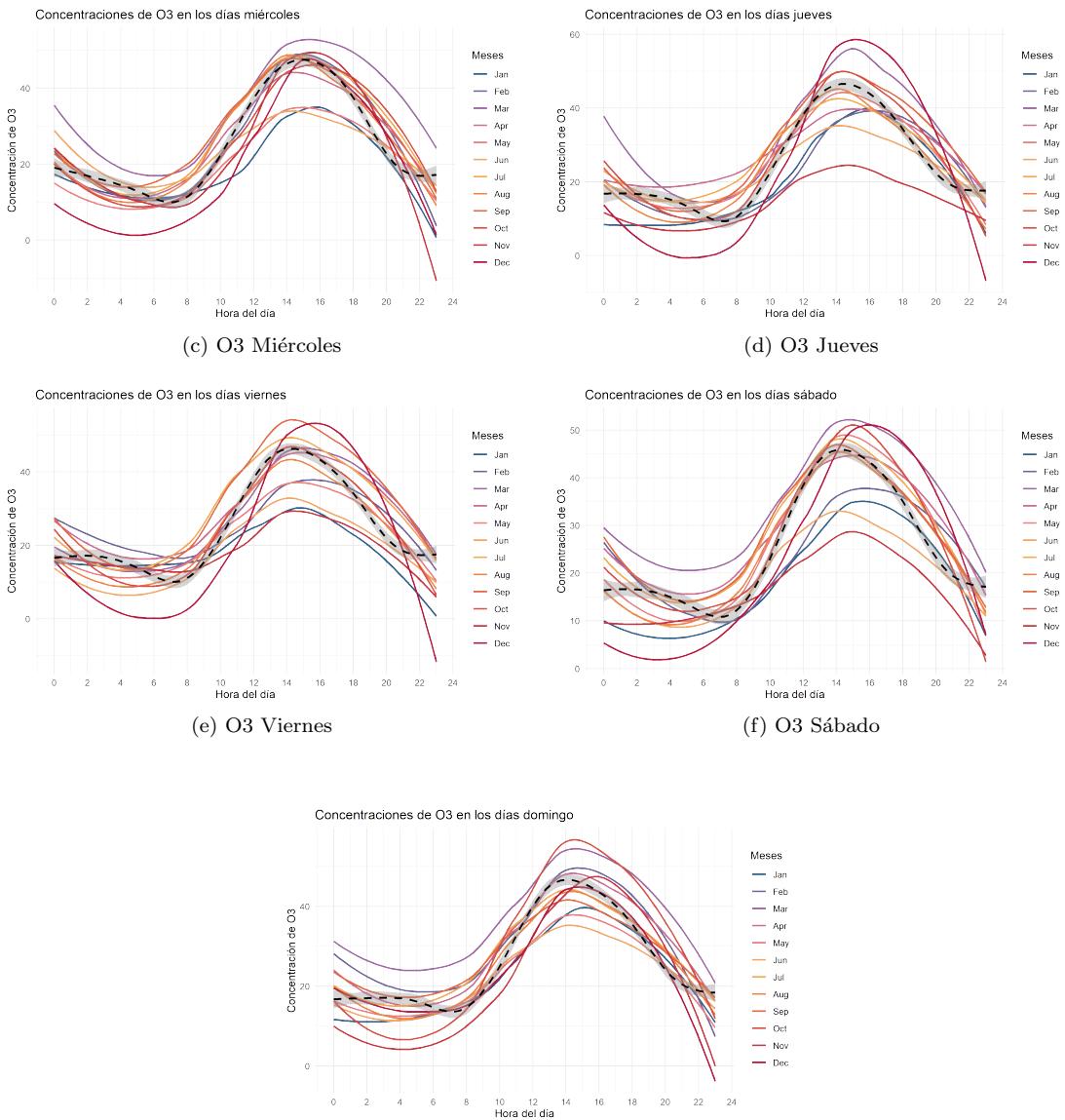
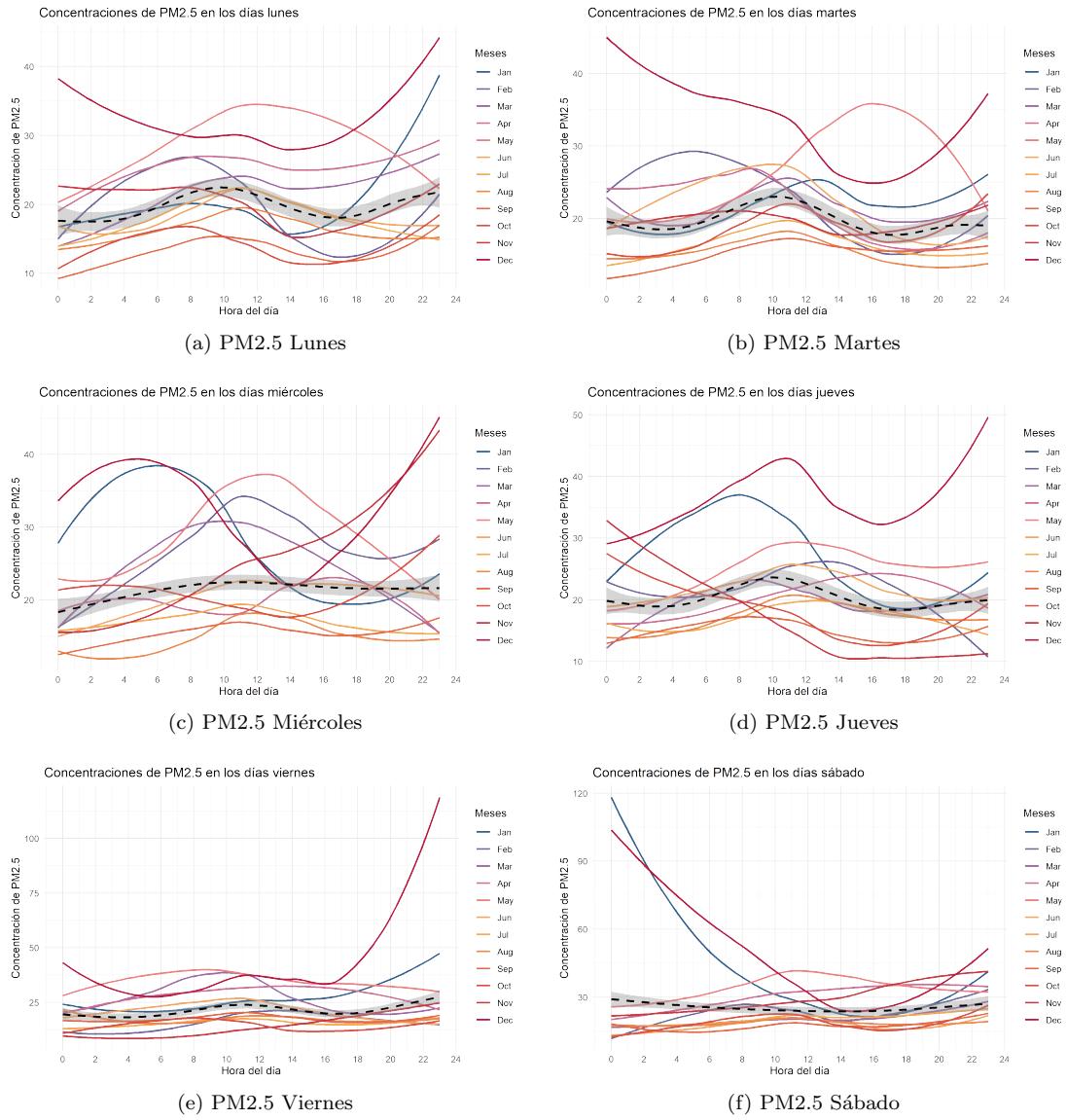


Figura 49: O₃ Domingo

B.4. $PM_{2.5}$ - Materia Particulada 2.5



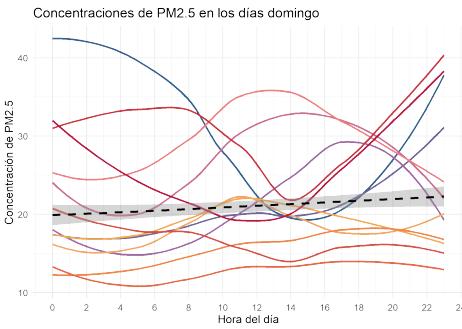
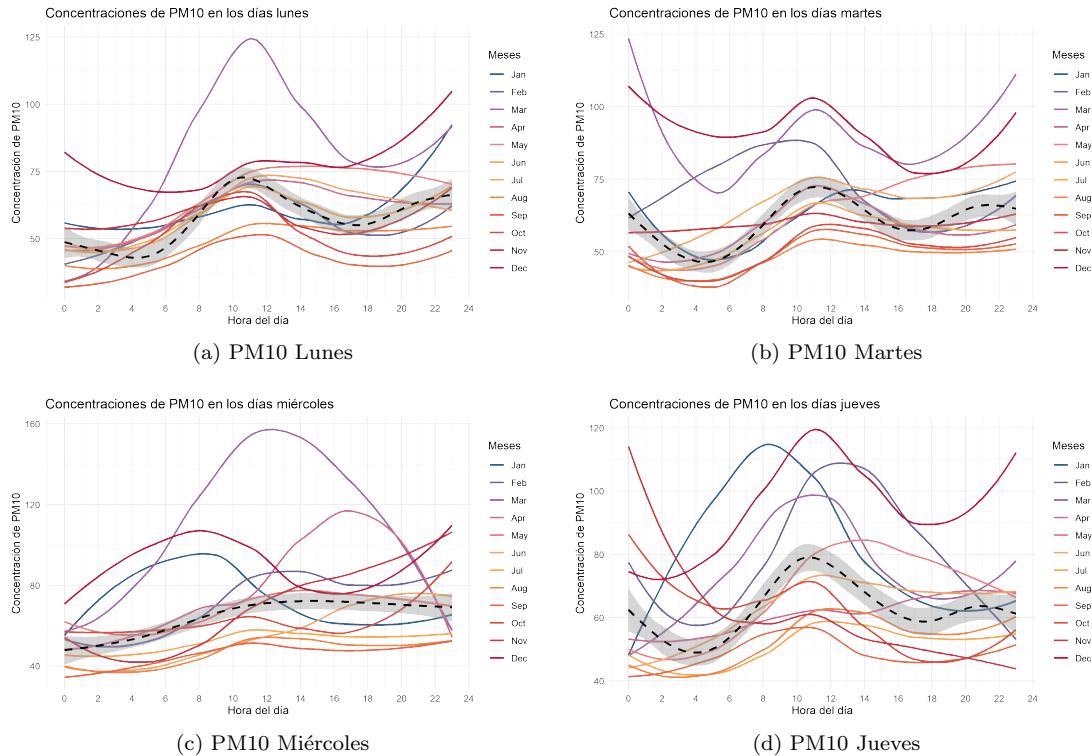


Figura 50: PM2.5 Domingo

B.5. PM_{10} - Materia Particulada 10



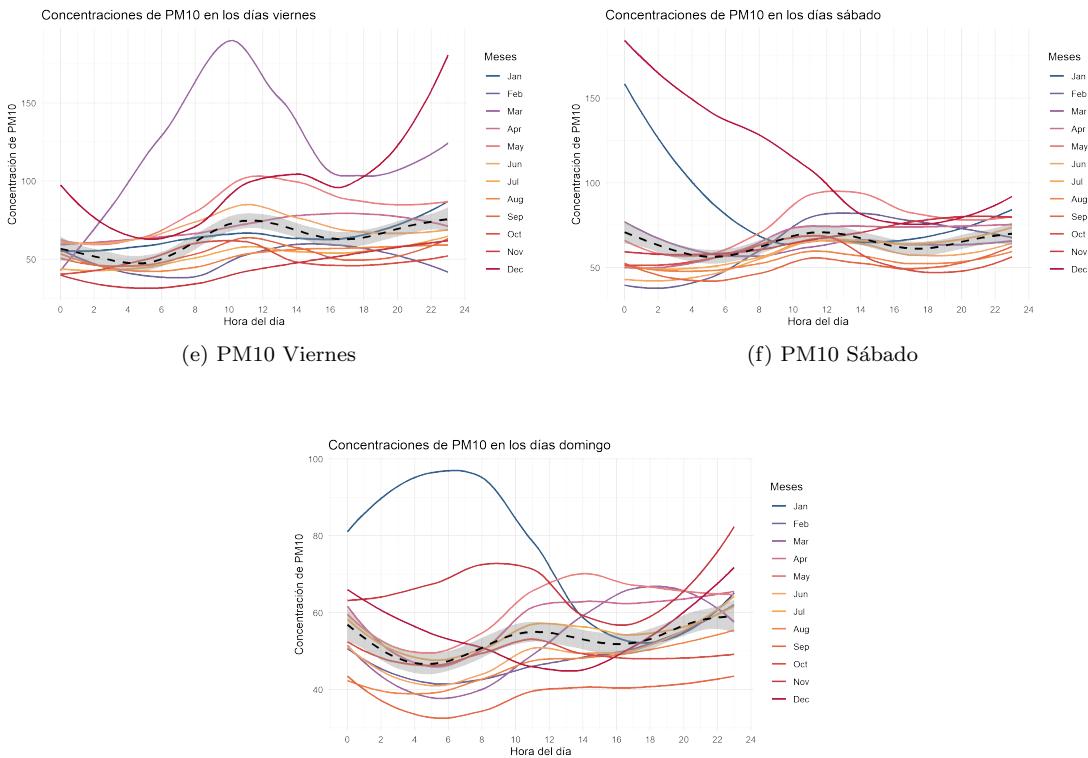
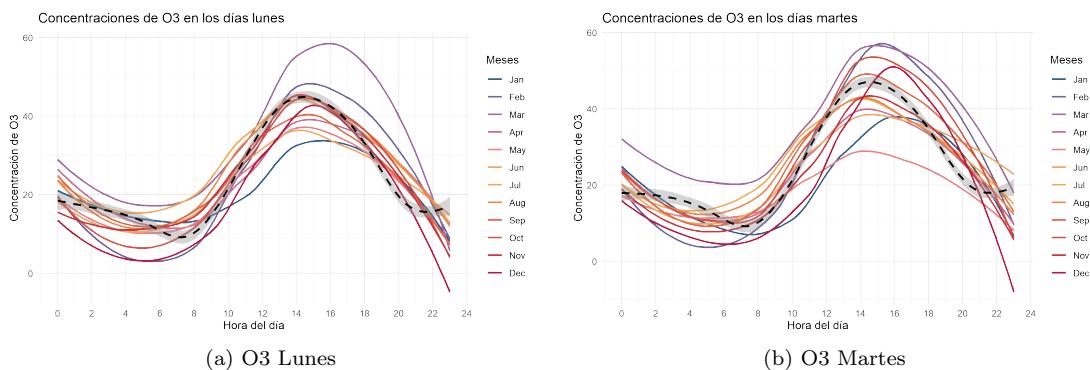


Figura 51: PM10 Domingo

B.6. SO_2 - Dióxido de Azufre



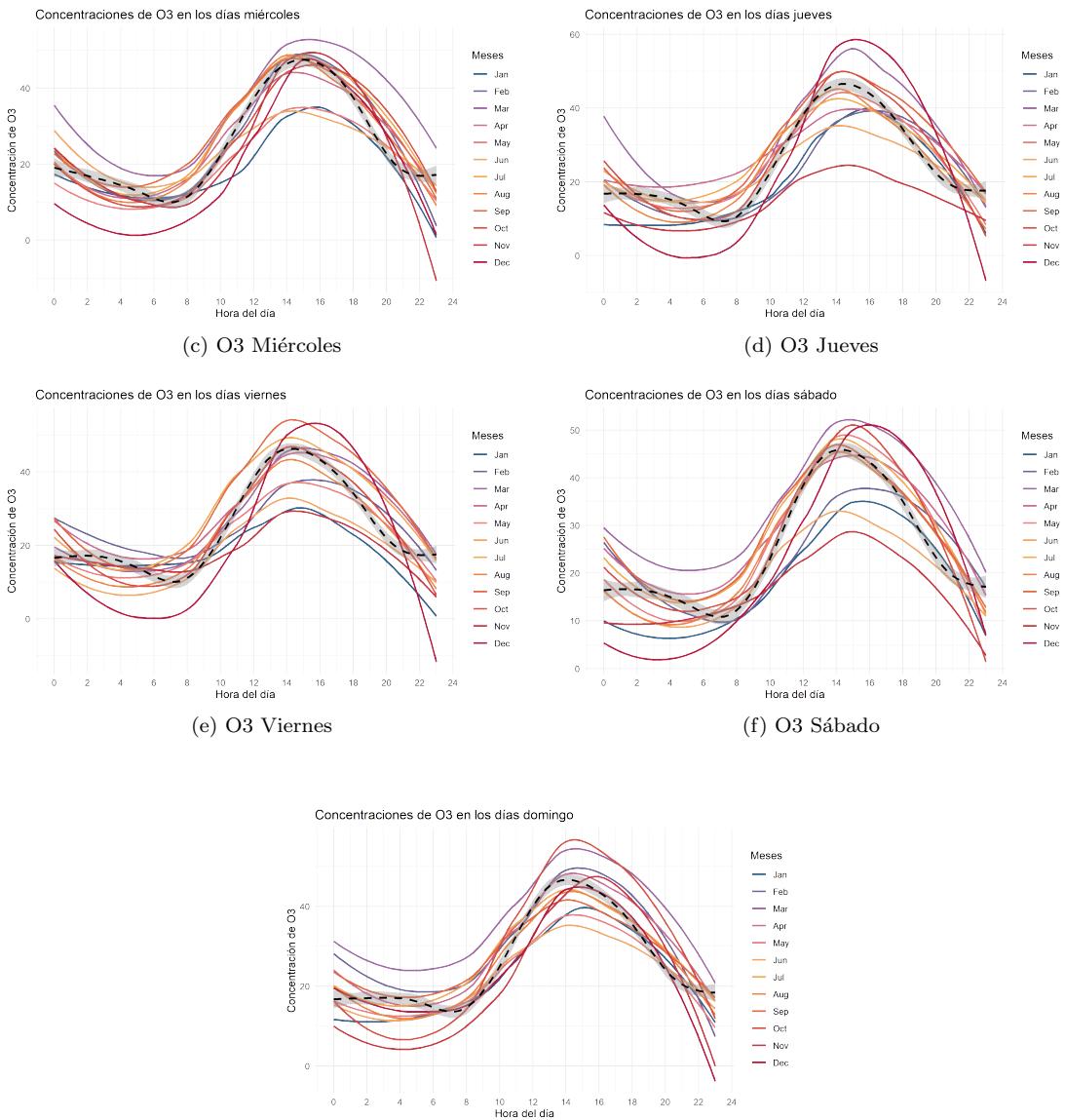


Figura 52: O3 Domingo

C. Índice de Aire y Salud por día de la semana

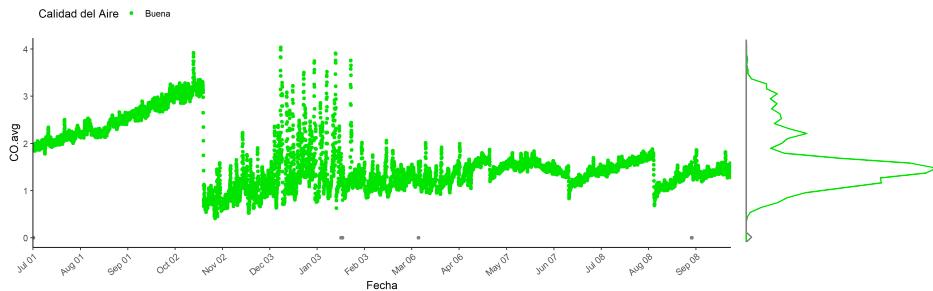


Figura 53: CO -Índice de aire y salud

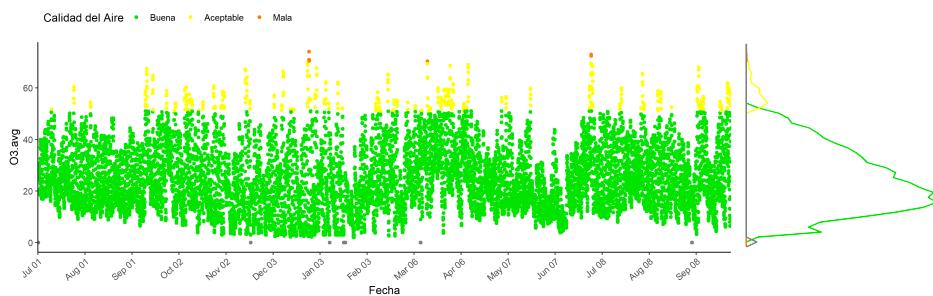


Figura 54: O3 -Índice de aire y salud

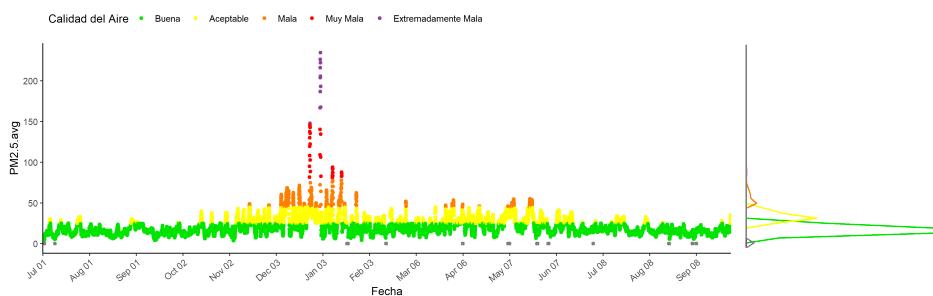


Figura 55: PM2.5 -Índice de aire y salud

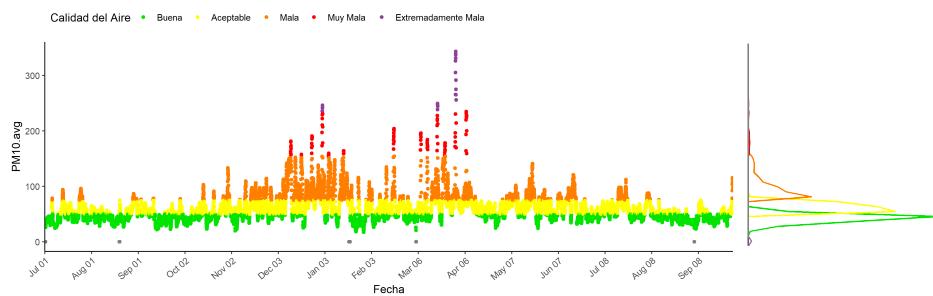


Figura 56: PM10 -Índice de aire y salud

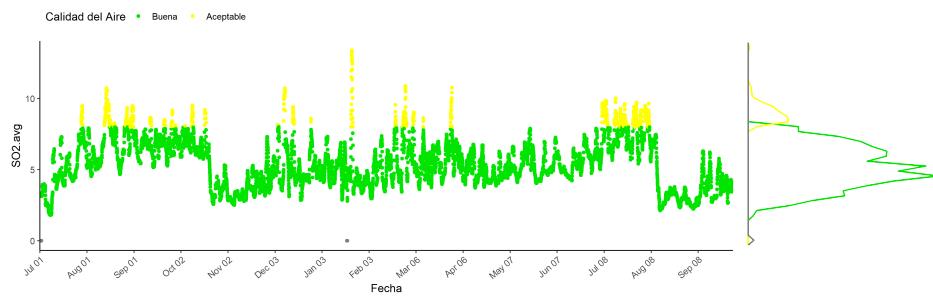


Figura 57: SO2 -Índice de aire y salud