



Modelación del promedio octohorario de Ozono Troposférico usando regresión cuantílica.

Lina Jhoana Ramírez Marulanda

Universidad del Valle
Facultad de Ingeniería
Escuela de Estadística
Colombia, Santiago de Cali
2025

Modelación del promedio octohorario de Ozono Troposférico usando regresión cuantílica.

Lina Jhoana Ramírez Marulanda

Trabajo de grado presentado como requisito parcial para optar al título de:
Estadística

Director:
Javier Olaya Ochoa

Codirector:
Álvaro José Florez Poveda

Universidad del Valle
Facultad de Ingeniería
Escuela de Estadística
Colombia, Santiago de Cali
2025

“Siempre parece imposible hasta que se hace.”

-Nelson Mandela

Esta tesis está dedicada a:

Mi hermana y mis primas, quienes son mi mayor motivación.

Para que sigamos haciendo sentir orgullosa a nuestra familia y juntas sanemos el linaje familiar. Que este logro sea un ejemplo de perseverancia y las anime alcanzar sus sueños con valentía y determinación.

Agradecimientos

A lo largo de la realización de esta tesis, he contado con el apoyo y la guía de algunas personas a quienes me gustaría expresar mi más profundo agradecimiento.

En primer lugar, agradezco a mis directores de tesis, Álvaro José Florez y Javier Olaya, por su invaluable orientación y paciencia. Su conocimiento y experiencia ha sido importante y fundamental en la culminación de este trabajo.

A mi novia, Valeria Londoño por brindarme con mucho amor su constante apoyo, que ha sido esencial para alcanzar este logro.

También quiero agradecer, a mi familia, amigos y demás conocidos que con sus palabras inyectaron motivación en los momentos de mayor dificultad durante este proceso.

A los lectores de esta tesis, les extiendo mi más sincero agradecimiento. Espero que este trabajo les sea de gran utilidad y contribuya significativamente a su conocimiento y comprensión del tema abordado. Su interés hace que este documento tenga valor.

Finalmente, agradezco profundamente a la universidad pública, por brindarme la oportunidad de formarme dentro y fuera de sus aulas. Ser hija de la universidad pública es un orgullo y un privilegio que me ha permitido crecer tanto profesional como personalmente.

Resumen

El presente trabajo tiene como objetivo modelar el promedio octohorario del Ozono Troposférico ($P8O_3$) en la estación de monitoreo Pance en Cali durante el año 2019, utilizando regresión cuantílica. Ante la presencia de datos faltantes, se implementó la imputación múltiple mediante el algoritmo MICE. Posteriormente se implementaron modelos de regresión cuantílica paramétrica y no paramétrica para evaluar la variabilidad de las concentraciones del contaminante a través de diferentes cuantiles, permitiendo capturar patrones en niveles bajos, medios y altos. El análisis incluyó variables ambientales como el material particulado (PM10) y meteorológicas relevantes como la temperatura, la lluvia, la radiación solar y la velocidad del viento, debido a su influencia en la formación y transporte de Ozono Troposférico. Se observó que el enfoque no paramétrico con B-splines cúbicos, proporciona un ajuste más flexible y captura adecuadamente la no linealidad y la heterocedasticidad presente en la distribución de las concentraciones de $P8O_3$ a lo largo del día.

Palabras clave: Regresión cuantílica, Ozono Troposférico, Promedio octohorario, Imputación múltiple, Algoritmo MICE.

Abstract

This study aims to model the eight-hour average of Tropospheric Ozone ($P8O_3$) at the Pance monitoring station in Cali during 2019 using quantile regression. Because some data were missing, we used multiple imputation with the MICE algorithm. Then we applied both parametric and non-parametric quantile regression models to examine how pollutant levels change at different quantiles, capturing patterns at low, medium, and high levels. The analysis included environmental variables such as particulate matter (PM10) and important weather factors like temperature, rain, solar radiation, and wind speed, since they affect ozone formation and transport. We found that the non-parametric approach with cubic B-splines gives a more flexible fit and correctly captures the nonlinearity and heteroscedasticity in $P8O_3$ concentrations throughout the day.

Keywords: Quantile regression, Tropospheric Ozone, Eight-hour average, Multiple imputation, Algorithm MICE.

Contenido

Agradecimientos	vi
Resumen	vii
Lista de figuras	xi
Lista de tablas	1
1. Introducción	2
2. Planteamiento del problema, justificación y objetivos	4
2.1. Planteamiento del problema	4
2.2. Justificación	6
2.3. Objetivos	8
2.3.1. Objetivo general	8
2.3.2. Objetivos específicos	8
3. Antecedentes y marco teórico	9
3.1. Antecedentes	9
3.2. Marco teórico	11
3.2.1. Ozono Troposférico y su medición	11
3.2.2. Aspectos metodológicos de los datos	12
3.2.3. Regresión Cuantílica	20
4. Metodología	25
4.1. Datos	25
4.1.1. Organización de la base de datos	25
4.1.2. Análisis exploratorio	26
4.2. Imputación de datos	27
4.2.1. Métodos y modelos de imputación	27
4.2.2. Prueba de bondad de ajuste	29
4.3. Ajuste de los modelos de regresión cuantílica	30
4.3.1. Ajuste del modelo de regresión cuantílica paramétrica	31
4.3.2. Estimación de los coeficientes e intervalos de confianza	31
4.3.3. Métricas de evaluación del modelo	32

4.3.4. Ajuste del modelo de regresión cuantílica no paramétrica	32
5. Resultados	35
5.1. Análisis exploratorio de datos	35
5.1.1. Análisis datos faltantes	39
5.2. Imputación de datos	41
5.3. Modelo de regresión cuantílica paramétrica	44
5.3.1. Métricas y test de Wald	46
5.4. Modelo de regresión cuantílica no paramétrica de <i>P8O3</i> durante el día . . .	47
6. Conclusiones y recomendaciones	50
6.1. Conclusiones	50
6.2. Recomendaciones	51
A. Anexos	53
A.1. Resultados	53
A.1.1. Análisis exploratorio de datos	53
A.1.2. Imputación de datos	55
A.1.3. Métricas del modelo de regresión cuantílica paramétrico	59
Bibliografía	61

Lista de Figuras

2-1. Localización de las estaciones de monitoreo en Santiago de Cali. Tomado del informe de Análisis O_3 - Santiago de Cali - 2015 - 2022 (DAGMA, 2023). . .	7
4-1. Árbol de clasificación basado en CART empleado para la imputación de la variable Rad Solar según la hora del día y la temperatura	29
5-1. Distribución porcentual de la variable Lluvia en el año 2019.	36
5-2. Distribución de O_3 bajo condiciones de lluvia y sin lluvia.	36
5-3. Rosa de los vientos.	37
5-4. Rosa de contaminación de O_3	37
5-5. Rosa de contaminación de O_3 cada hora.	38
5-6. Diagrama de cajas de O_3 por hora.	39
5-7. Distribución del promedio octohorario de Ozono Troposférico durante el año en contraste con las directrices de la OMS.	40
5-8. Identificación del patrón de datos faltantes.	41
5-9. Comparación de las distribuciones originales e imputadas para las variables ambientales y meteorológicas.	43
5-10. Serie temporal anual de O_3 (8 760 registros horarios) para el primer conjunto imputado: en gris, valores observados que se conservaron y en azul, valores imputados.	43
5-11. Coeficientes de regresión cuantílica de $P8O3$ para los cuantiles 0.1, 0.25, 0.5, 0.75, 0.9 e IC 95 % por variable explicativa.	45
5-12. Curvas diarias de $P8O3$ para cuantiles entre 0.1 y 0.9, comparando la presencia lluvia vs. sin lluvia.	48
A-1. Diagramas de caja de los contaminantes del aire estandarizados.	53
A-2. Diagramas de caja de las variables meteorológicas estandarizadas.	53
A-3. Diagrama de correlación de las variables.	54
A-4. Serie temporal anual de O_3 (8 760 registros horarios) para el segundo conjunto imputado.	55
A-5. Serie temporal anual de O_3 (8 760 registros horarios) para el tercero conjunto imputado.	56
A-6. Serie temporal anual de O_3 (8 760 registros horarios) para el cuarto conjunto imputado.	56

A-7. Serie temporal anual de O_3 (8 760 registros horarios) para el quinto conjunto imputado.	57
A-8. Comparativa de frecuencias observadas e imputadas para la variable categórica Lluvia.	58

Lista de Tablas

4-1.	Definición de variables del conjunto de datos	26
5-1.	Medidas de tendencia central y dispersión de las variables del conjunto de datos.	35
5-2.	Cantidad y proporción de datos faltantes en cada una de las variables de la base de datos.	41
5-3.	Medidas de tendencia central y dispersión de las variables después de la imputación.	42
5-4.	Coeficientes estimados del modelo de regresión cuantílica por cuantil τ	44
5-5.	Coeficientes estimados del modelo de regresión cuantílica por cuantil τ	46
5-6.	Pseudo- R^2 y MAE por cuantil del modelo de regresión cuantílica no paramétrica con B-splines, diferenciados por condición de lluvia.	49
A-1.	Comparación de distribuciones mediante pruebas de bondad de ajuste en los cinco conjuntos de datos imputados.	55
A-2.	Estimaciones por cada cuantil y variable de la regresión cuantílica.	59

1. Introducción

La contaminación del aire, según el Instituto de Hidrología, Meteorología y Estudios Ambientales (IDEAM) es un problema ambiental que afecta la salud pública, ya que la exposición prolongada a contaminantes atmosféricos puede generar efectos adversos en la salud humana. Entre estos contaminantes, se encuentra el Ozono Troposférico (O_3) que es capaz de generar daños respiratorios, empeorar las cardiopatías existentes y afectar los ecosistemas. Su formación se da a partir del dióxido de nitrógeno (NO_2) en reacción con la radiación solar y es influenciado por diversas variables meteorológicas como la temperatura, la humedad y la velocidad del viento, lo que hace que sus niveles fluctúen significativamente a lo largo del día.

La Organización Mundial de la Salud (OMS) establece el promedio octohorario del O_3 como una medida adecuada para evaluar la exposición a este contaminante, debido a que sus efectos adversos en la salud no dependen exclusivamente de los picos horarios, sino de la cantidad total de ozono inhalado durante varias horas (Organización Mundial de la Salud, 2021). Siguiendo estas directrices, la normatividad colombiana adoptó esta medida a través de la Resolución 2254 de 2017, la cual establece un límite de $100\mu g/m^3$ como concentración permisible para la media móvil de 8 horas. Por lo tanto, modelar el comportamiento del promedio octohorario del O_3 permite evaluar el cumplimiento de los estándares de calidad del aire.

En la ciudad de Santiago de Cali, el monitoreo de la calidad del aire es llevado a cabo en nueve estaciones distribuidas en la ciudad, permitiendo evaluar las concentraciones de contaminantes atmosféricos. Una de estas estaciones es la Estación Pance, ubicada en el sur de la ciudad, la cual permite la recolección de datos sobre O_3 y otras variables ambientales en la zona de expansión urbana. Sin embargo, el análisis de estos datos enfrenta dos desafíos principales en esta investigación.

El primer desafío consiste en la presencia de valores faltantes en las mediciones, estos pueden deberse a fallos en los equipos de monitoreo, interrupciones en la transmisión de datos o mantenimientos en las estaciones de medición. La omisión de estos datos podría generar estimaciones sesgadas o perder información valiosa para los análisis (Saeipourdizaj et al., 2021). Para abordar este problema, en este estudio se realizará imputación múltiple mediante la librería `mice` en R, utilizando la técnica de Especificación Totalmente Condicional con siglas en inglés FCS. En particular, la imputación se implementará sobre las mediciones

horarias de O_3 , lo que permitió mantener la dinámica real del contaminante a lo largo del día y minimizar posibles sesgos en los análisis posteriores.

El segundo desafío se presenta en la dificultad de modelar adecuadamente las concentraciones de la media móvil de 8 horas del O_3 debido a su naturaleza no lineal y variable (Olaya et al., 2014). Es por esto, que se optó por utilizar regresión cuantílica en lugar de modelos tradicionales basados en la media. A diferencia de la regresión de mínimos cuadrados ordinarios, la regresión cuantílica permite analizar diferentes partes de la distribución condicional del O_3 , lo que proporciona una visión más detallada de su comportamiento en distintos niveles de concentración. Lo anterior, es de gran importancia para evaluar los efectos de las variables meteorológicas no solo en la mediana de la distribución, sino también en sus cuantiles extremos, que representan los niveles más críticos para la calidad del aire y la salud pública (Le Cook and Manning, 2013).

En este contexto, el presente trabajo de grado tiene como objetivo modelar el promedio octohorario del O_3 en la estación de monitoreo Pance en Santiago de Cali durante el año 2019, utilizando regresión cuantílica. Para ello, este documento desarrolla en el Capítulo 1 la introducción y la motivación del estudio. En el Capítulo 2, expone el planteamiento del problema, la justificación y los objetivos de la investigación. El Capítulo 3 aborda la revisión de antecedentes y el marco teórico. En el Capítulo 4, detalla la metodología utilizada, incluyendo la imputación de datos y la aplicación de la regresión cuantílica. En el Capítulo 5, se presentan los resultados obtenidos y su análisis. Finalmente, en el Capítulo 6, se exponen las conclusiones y recomendaciones para futuras investigaciones y aplicaciones de este estudio.

2. Planteamiento del problema, justificación y objetivos

2.1. Planteamiento del problema

La contaminación del aire es un problema ambiental que afecta nuestro planeta de diversas maneras, presentando sus niveles más altos en centros urbanos, debido a las emisiones provenientes principalmente del tráfico vehicular, las actividades industriales y otras acciones humanas (IDEAM, 2002). Entre los contaminantes atmosféricos de mayor impacto se encuentra el O_3 , un gas que deteriora la calidad del aire y genera consecuencias adversas para la salud de la población si se encuentra expuesta de forma prolongada. Santiago de Cali no está exenta de esta problemática, ya que desde la década de los 1970 experimenta un rápido crecimiento urbano que ha incrementado la emisión de contaminantes atmosféricos.

Pese a los avances tecnológicos y metodológicos en el monitoreo ambiental, comprender el comportamiento de las concentraciones de O_3 continúa siendo un desafío, puesto que varía notablemente durante el día por las condiciones meteorológicas, lo que dificulta el uso de modelos tradicionales de regresión basados en la media, pues no capturan adecuadamente la variabilidad en distintos niveles de concentración debido a que los datos no se comportan de manera lineal, y presentan valores extremos.

En consecuencia, el reto se encuentra en modelar la contaminación por O_3 y resulta necesario explorar métodos más flexibles que vayan más allá de la tendencia central, donde estudios previos han abordado esta problemática utilizando técnicas no paramétricas debido a su capacidad para modelar patrones no lineales en la distribución del O_3 (Olaya et al., 2014). Adicionalmente, otro desafío en el análisis de los datos de calidad del aire es la presencia valores faltantes, los cuales son ocasionados por diferentes factores, entre ellos las fallas técnicas en los equipos de monitoreo, interrupciones en el fluido eléctrico o periodos de mantenimiento de las estaciones de medición.

Lo anterior, genera vacíos en la información, que si no se maneja de manera adecuada puede afectar la validez de los resultados obtenidos, sesgar las estimaciones y limitar la capacidad de los modelos para capturar la variabilidad real del O_3 . Para mitigar estas limitaciones, es necesario aplicar una técnica de imputación sobre las mediciones horarias de O_3 que permita

estimar los valores ausentes conservando la estructura de los datos y la dinámica real del contaminante a lo largo del día.

En este contexto, la presente investigación busca describir y modelar el comportamiento del promedio octohorario del O_3 en Santiago de Cali durante el año 2019. Para ello, se utilizará la regresión cuantílica, una técnica estadística adecuada para abordar este fenómeno debido a su capacidad para captar la variabilidad en distintos niveles de concentración. Además, presenta la ventaja de ser robusta frente a observaciones extremas y no requiere supuestos específicos sobre la distribución de los datos (Koenker, 2005). Dicha técnica otorgará la posibilidad de analizar la distribución completa del O_3 , incluyendo concentraciones críticas que podrían pasar inadvertidas con métodos estadísticos basados en la media.

2.2. Justificación

El Ozono Troposférico (O_3) es uno de los principales contaminantes atmosféricos, ya que, al superar determinados niveles de concentración, puede desencadenar problemas cardiorrespiratorios y reducir la productividad de los ecosistemas. En atención a estos riesgos, tanto la OMS como la normativa colombiana recomiendan monitorear este contaminante utilizando el promedio octohorario. Esta métrica resulta más pertinente desde el enfoque de salud pública, pues permite evaluar la exposición acumulada a lo largo de varias horas, lo que no sería posible mediante registros horarios aislados.

En la ciudad de Santiago de Cali, el Departamento Administrativo de Gestión del Medio Ambiente (DAGMA) es responsable de monitorear, analizar y reportar la calidad del aire mediante el Sistema de Vigilancia de Calidad del Aire (SVCA). Dicho sistema cuenta con nueve estaciones de medición continua distribuidas en la ciudad, tal y como se presenta en la figura **2-1**; sin embargo, solo seis captan datos específicos de O_3 : Base Aérea, Compartir, ERA-Obrero, Flora, Pance y Univalle. Entre ellas, se seleccionó la estación de Pance como fuente principal de información para el año 2019, dado que los datos de 2021 a 2025 son limitados y los de 2020 carecen de representatividad por el confinamiento. Es importante recordar que el O_3 se forma en presencia de compuestos orgánicos volátiles e hidrocarburos, por lo que las restricciones de movilidad y actividad durante la pandemia habrían alterado significativamente su dinámica habitual.

De acuerdo con informes del DAGMA (2023), las estaciones de monitoreo ubicadas en Pance y Univalle registran excedencias en las concentraciones de O_3 , las cuales se atribuyen al acelerado crecimiento urbano y al incremento en la movilidad vehicular hacia el sur de la ciudad. Esta situación se ve agravada por la dirección predominante del viento, que favorece el arrastre de contaminantes hacia dicha zona (Alcaldía de Santiago de Cali, 2014). Asimismo, se ha identificado la presencia recurrente de valores faltantes en los registros, producto de fallas en los equipos o interrupciones del servicio, lo que dificulta una caracterización precisa del comportamiento del contaminante.

Para superar esta limitación, se optará por emplear un enfoque de imputación múltiple mediante la Especificación Totalmente Condicional, que conserva la variabilidad natural del fenómeno y asegura la representatividad del conjunto de dato. Lo anterior se encuentra en afinidad con lo planteado por Saeipourdizaj et al. (2021), quienes resaltan la necesidad de preservar las propiedades estadísticas de los datos ambientales al imputar, y sugieren el uso de técnicas que respeten la estructura y la distribución subyacente de las variables.



Figura 2-1.: Localización de las estaciones de monitoreo en Santiago de Cali. Tomado del informe de Análisis O_3 - Santiago de Cali - 2015 - 2022 (DAGMA, 2023).

Una vez refinada la base de datos, se aplicará un modelo de regresión cuantílica para analizar el promedio octohorario de O_3 . Esta técnica permite explorar la distribución condicional del contaminante en distintos niveles y ofrece una mayor robustez frente a valores atípicos.

A diferencia de la regresión lineal tradicional, no impone supuestos distribucionales sobre los errores, ya que reemplaza la función de pérdida cuadrática por una función de pérdida asimétrica, que asigna distintos pesos a los errores de subestimación y sobrestimación en función del cuantil modelado. Lo anterior, resulta especialmente útil para caracterizar la relación de variables meteorológicas, como la presencia de lluvia o la radiación solar, en diferentes rangos de concentración. Gracias a esta flexibilidad, la regresión cuantílica facilita la identificación de patrones en las concentraciones octohorarias de O_3 que podrían acercarse o superar los límites normativos a lo largo del día.

2.3. Objetivos

2.3.1. Objetivo general

Modelar el promedio octohorario de Ozono Troposférico durante un día usando regresión cuantílica, para la estación de monitoreo Pance del año 2019.

2.3.2. Objetivos específicos

- Implementar una técnica de imputación de datos faltantes para el promedio horario de Ozono Troposférico.
- Analizar el impacto de diferentes variables meteorológicas en el comportamiento de las concentraciones octohorarias de Ozono Troposférico.

3. Antecedentes y marco teórico

3.1. Antecedentes

A continuación, se describen de manera breve las investigaciones que han abordado, por un lado, la problemática de los valores faltantes en los datos ambientales, y por otro, la aplicación de la regresión cuantílica para modelar el comportamiento del Ozono Troposférico. Dichos estudios resultaron fundamentales para el desarrollo metodológico de esta tesis.

En primer lugar, el estudio realizado por Saeipourdizaj et al. (2021) evaluó distintas técnicas de imputación aplicadas a concentraciones horarias de PM10 y O_3 en la ciudad de Tabriz, Irán. Los autores compararon métodos univariados y multivariados como la interpolación lineal, el promedio móvil, los vecinos más cercanos (KNN), el algoritmo EM, los árboles de clasificación y regresión (CART), y el emparejamiento por media predictiva (PMM), estos dos últimos dentro del marco del algoritmo de imputación múltiple por ecuaciones encadenadas, conocido por sus siglas en inglés como MICE. A través de simulaciones con porcentajes de valores faltantes del 10 %, 20 % y 30 %, se observó que los métodos que incorporan la dimensión temporal como el promedio móvil y KNN ofrecieron mejores resultados en términos de R^2 , MAE y RMSE. Aunque, PMM presentó un menor desempeño predictivo en comparación con otros métodos, se destacó su utilidad en escenarios multivariados con relaciones complejas. El estudio concluye que la elección del método de imputación debe basarse en el comportamiento particular de cada variable, la dependencia temporal del fenómeno y el contexto del análisis.

Complementariamente, el estudio desarrollado por Kebalepile et al. (2024) abordó la imputación de datos faltantes de contaminantes atmosféricos como NO_2 , SO_2 , O_3 y PM10 en diversas estaciones de monitoreo de Sudáfrica, bajo un escenario en el que los datos faltantes ocurrían de forma completamente aleatoria (MCAR), como consecuencia de fallas técnicas, mantenimientos imprevistos y limitaciones operativas en las estaciones. Los autores aplicaron el algoritmo MICE, comparando los métodos de PMM y CART. Para ello, construyeron modelos tanto por contaminante de forma individual como combinados entre sí y por estación, incorporando además variables meteorológicas como predictores. En cada modelo se generaron entre 5 y 10 conjuntos de datos imputados, sobre los cuales se evaluó el desempeño de ambos métodos. Los resultados mostraron que CART presentó un mejor ajuste que PMM, con valores de R^2 significativamente más altos entre 0.3 y 0.7 frente a 0.02 y 0.25, respectiva-

mente, manteniendo el rango y la variabilidad natural de los contaminantes imputados. Este trabajo resaltó la utilidad del enfoque multivariado y adaptativo de MICE, especialmente en el contexto donde, a pesar de la limitación de los recursos, es posible seleccionar predictores con alta correlación que fortalecen el proceso de imputación.

En cuanto a las investigaciones sobre la aplicación de la modelación por regresión cuantílica, el trabajo de Porter et al. (2015) la usó para analizar como las condiciones meteorológicas inciden en los niveles extremos de la contaminación por O_3 y material particulado fino ($PM_{2.5}$) en diferentes regiones de Estados Unidos. En dicho artículo se estimaron modelos de regresión cuantílicos de cada variable para los diferentes cuantiles superiores $\tau = 0,75, 0,90, 0,95$ y $0,98$, comparando los resultados obtenidos con las estimaciones de la regresión convencional por mínimos cuadrados ordinarios (MCO), evidenciando que la regresión cuantílica captura efectos heterogéneos no visibles en la media. También, los resultados revelan que el comportamiento del ozono es positivamente influenciado por la temperatura en temporada de verano y en invierno por la radiación entrante y la estabilidad atmosférica.

De manera similar, diversos autores han demostrado la utilidad de la regresión cuantílica para caracterizar las concentraciones de los contaminantes atmosféricos en diferentes niveles. Por ejemplo, Fix et al. (2018) la emplearon en el cuantil 0.95 para evaluar la sensibilidad del O_3 frente a predictores meteorológicos tras centrar y escalar sus variables, contrastando modelos con datos simulados y observados en tres estaciones de Estados Unidos entre 1996 y 2005. En un enfoque multivariante, Zhang et al. (2022) realizaron lo mismo simultáneamente para $PM_{2.5}$ y O_3 en los cuantiles 0.50 y 0.95 en China entre los años 2013 y 2019, e implementaron una regresión escalonada al percentil 0.90 que permitió seleccionar las variables dominantes según estación y región. Por su parte, Rehan and Munir (2022) analizaron la interacción entre O_3 , NO_2 y otros gases contaminantes (NO , CO , SO_2) junto a variables meteorológicas en los cuantiles 0.05, 0.25, 0.50, 0.75 y 0.95 en Yeda (Arabia Saudita), revelando patrones locales de calidad del aire en distintos niveles de concentración.

Finalmente, en la investigación de Reiss and Huang (2012) se implementó la regresión cuantílica no paramétrica mediante B-splines cúbicos penalizados, gracias a la flexibilidad y versatilidad de esta técnica para modelar curvas suaves y adaptativas en distintos cuantiles, al añadir un término de penalización sobre la segunda derivada de los splines, controlaron la complejidad del ajuste y evitaron el sobreajuste en los extremos, logrando estimaciones más estables y precisas que los métodos convencionales. Este enfoque fue aplicado a datos de resonancia magnética funcional para estudiar la conectividad cerebral en pacientes con TDAH, donde capturaron eficazmente la heterogeneidades y no linealidades sin imponer restricciones rígidas en la forma distribucional.

3.2. Marco teórico

Esta sección aporta una comprensión profunda de los conceptos, métodos analíticos y teóricos que respaldan la investigación de este trabajo de grado, con el fin de abordar de manera práctica la modelación del comportamiento del Ozono Troposférico en un día mediante la utilización de la regresión cuantílica.

3.2.1. Ozono Troposférico y su medición

Ozono Troposférico

El Ozono Troposférico es denominado O_3 y se considera un contaminante criterio, medido a través de fotometría ultravioleta en estaciones de monitoreo de calidad del aire. Este contaminante no se emite directamente a la atmósfera, sino que su proceso de formación comienza con la emisión de dióxido de nitrógeno (NO_2) e hidrocarburos, que actúan como los precursores principales. Estos compuestos reaccionan en presencia de calor y luz solar para producir O_3 . Este es un gas, incoloro, compuesto por tres átomos de oxígeno y es conocido como uno de los principales ingredientes del smog urbano (IDEAM, 2002).

Es así como las variables meteorológicas juegan un papel muy importante en la creación y transporte del O_3 ya que según Rehan and Munir (2022) estos factores influyen en el comportamiento de las concentraciones de ozono.

Concentración octohoraria

Los administradores del Sistema de Vigilancia de la Calidad del Aire (SVCA), junto con personal calificado deben llevar a cabo de manera rutinaria la revisión de la generación de los datos de las estaciones de monitoreo para la Calidad del Aire, aplicando los procedimientos de validación establecidos en el Manual de Operación de los Sistemas de Vigilancia de la Calidad del Aire expedido por el Ministerio de Ambiente, para crear bases de datos confiables antes de ser utilizadas para su análisis.

En cuanto a los registros de los niveles de concentración de los contaminantes son valores promediados cada hora, es decir, valores de concentración horaria, medidos en microgramos por metro cubico ($\mu g/m^3$) para el O_3 . Teniendo los datos horarios validados se calcula el promedio móvil de 8 horas para cada hora determinada, siguiendo las directrices del manual mencionado, en cada grupo para el cálculo de la media móvil se debe contar con el 75 % de los datos válidos.

Normativa sobre las concentraciones de Ozono Troposférico

A continuación, se presenta la normatividad vigente desde las directrices de la OMS y los límites establecidos en la norma colombiana, para los niveles críticos de la concentración de O_3 en la atmósfera.

Las directrices de la calidad del aire por la Organización Mundial de la Salud (2021) establece:

- La media móvil de 8 horas no debe exceder el nivel de $100 \mu g/m^3$.
- La temporada alta está definida por el promedio aritmético más alto de las concentraciones máximas diarias de medias móviles de 8 horas de Ozono Troposférico. Esta no debe exceder un nivel de $60 \mu g/m^3$.

La norma colombiana según el Ministerio de Ambiente y Desarrollo Sostenible (2017), establece:

- La media móvil de 8 horas no debe exceder el nivel de $100 \mu g/m^3$.

Dado que la fecha de publicación de la actualización de las directrices de la OMS es posterior a la norma colombiana, en ella no se encuentra la regulación de la temporada alta, sin embargo, en el manual de operación de los SVCA se debe realizar dicho análisis.

3.2.2. Aspectos metodológicos de los datos

Datos circulares

En el contexto estadístico, un dato circular se define como una medida que se encuentra en un dominio angular, representado típicamente en grados o radianes, en el cual los valores están asociados con posiciones específicas en una circunferencia de radio 1 (Pérez Bote, 2020). A diferencia de los datos lineales, donde la variabilidad se extiende a lo largo de una escala unidimensional, los datos circulares exhiben una naturaleza cíclica al situarse en una dimensión angular cerrada (Pewsey et al., 2013).

Por tal razón, las estadísticas descriptivas para datos circulares difieren de las aplicadas a datos lineales debido a la naturaleza cíclica de los primeros. Así se definen las principales medidas de tendencia central y dispersión.

Ángulo medio

Según Pérez Bote (2020), el ángulo medio indica la orientación preferida de una muestra de n datos angulares. Se parte de un sistema de coordenadas rectangulares x e y con centro

en el origen (O), donde, ϕ_i es uno de los n ángulos. Así, se tendrá por definición de seno y coseno que:

$$x_i = \sin(\phi_i), \quad y_i = \cos(\phi_i) \quad (3-1)$$

Donde la media correspondiente de x e y sería:

$$\bar{x} = \frac{\sum_{i=1}^n \sin(\phi_i)}{n}, \quad \bar{y} = \frac{\sum_{i=1}^n \cos(\phi_i)}{n} \quad (3-2)$$

Dado lo anterior, el ángulo medio se calcula de la siguiente manera:

$$\bar{\phi} = \arctan\left(\frac{\bar{y}}{\bar{x}}\right) \quad (\bar{y} > 0), \quad (3-3)$$

$$\bar{\phi} = 180^\circ + \arctan\left(\frac{\bar{y}}{\bar{x}}\right) \quad (\bar{y} < 0) \quad (3-4)$$

Medida de concentración

A partir del promedio de las coordenadas rectangulares se obtiene la longitud del vector medio r :

$$r = \sqrt{\bar{x}^2 + \bar{y}^2} \quad (3-5)$$

que indica el grado de concentración de los datos sobre algún punto de la circunferencia. Esta medida varía directamente con la dispersión de los datos, es decir, un $r = 1$ indica mayor concentración de datos y $r = 0$ indica que hay menos concentración de los datos.

Datos faltantes

Se denominan datos faltantes a la ausencia o falta de información en las variables u observaciones de un conjunto de datos. En general, la ausencia de datos plantea dificultades para explicar y comprender los fenómenos de estudio, ya que esto afecta la capacidad de inferir o deducir la naturaleza de dichos fenómenos.

Es muy común encontrar datos faltantes en diversas fuentes de información, ya que lo ideal es contar con un conjunto de datos completo, en la realidad de la captura de información es inevitable que haya datos ausentes. Aplicar métodos inapropiados de imputación para completar la base de datos puede generar más problemas de los que resuelve (Medina and Galván, 2007). Por eso, abordar la estimación de valores faltantes se ha convertido en un área de estudio previo que para generar análisis estadísticos precisos, requiere identificar el patrón y el mecanismo de los datos faltantes más apropiado para el tratamiento de la base

de datos.

Rubin (1978) definió tres mecanismos principales para clasificar los datos faltantes, los cuales permiten modelar la forma en que ocurre la ausencia de información. Para esto, se considera un conjunto de datos completo $\mathbf{Y} = (\mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}})$ con n observaciones y p variables, donde $\mathbf{Y}_{\text{obs}}$ representa los valores observados y $\mathbf{Y}_{\text{mis}}$ los valores ausentes. Asimismo, se define una matriz binaria $\mathbf{R}$ de tamaño $n \times p$, indicando que cada elemento r_{ij} toma el valor 1 si el dato fue observado y 0 si está ausente. Bajo esta estructura, se definen los siguientes mecanismos:

- **Faltantes completamente al azar (MCAR):** El mecanismo de datos faltantes completamente al azar (Missing Completely At Random, MCAR), se presenta cuando la probabilidad de que un dato esté ausente no depende ni de los valores observados ni de los valores faltantes. Se expresa de manera formal como:

$$P(\mathbf{R} \mid \mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}}, \boldsymbol{\theta}) = P(\mathbf{R} \mid \boldsymbol{\theta}) \quad (3-6)$$

En este caso, la ausencia de datos ocurre de forma totalmente aleatoria, por lo que no introduce sesgo en el análisis.

- **Faltantes al azar (MAR):** El mecanismo de datos faltantes al azar (Missing At Random, MAR), se da cuando la probabilidad de que un dato esté ausente depende de los valores observados, pero no de los valores faltantes. Se expresa como:

$$P(\mathbf{R} \mid \mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}}, \boldsymbol{\theta}) = P(\mathbf{R} \mid \mathbf{Y}_{\text{obs}}, \boldsymbol{\theta}) \quad (3-7)$$

Es decir, una vez se conoce la información observada, la falta de datos no depende de los valores ausentes. Este mecanismo se considera ignorables bajo ciertos supuestos, ya que es posible modelar adecuadamente la ausencia con las variables disponibles.

- **Faltantes no al azar (MNAR):** El mecanismo de datos faltantes no al azar (Missing Not At Random, MNAR), ocurre cuando la razón por la cual un dato está ausente depende del valor del dato que falta. Es decir, la probabilidad de que falte cierta información está relacionada con la propia información que no se observó. Incluso si conocemos todos los demás datos disponibles, no es suficiente para explicar por qué falta ese valor. Formalmente, esto se expresa como:

$$P(\mathbf{R} \mid \mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}}, \boldsymbol{\theta}) \neq P(\mathbf{R} \mid \mathbf{Y}_{\text{obs}}, \boldsymbol{\theta}) \quad (3-8)$$

En este caso, la ausencia de información está relacionada con la propia variable con datos faltantes, lo que puede introducir sesgos en el análisis si no se modela correctamente.

Imputación de datos faltantes

En las últimas décadas se han desarrollado distintas técnicas para abordar el problema de los datos faltantes, donde una de las estrategias más comunes y sencillas es eliminar las observaciones con datos ausentes, ya sea mediante la eliminación por pares o la eliminación por lista. En la eliminación por pares, solo se excluyen las observaciones que presenten datos faltantes en las variables que participan en un análisis específico, lo cual puede llevar a que el número de observaciones varíe entre análisis. En la eliminación por lista, también conocida como análisis completo, se eliminan todas las observaciones que presenten al menos un valor faltante en cualquiera de las variables del estudio (Rubin, 2004).

Aunque eliminar los datos faltantes permite trabajar con una base de datos aparentemente completa, esta estrategia puede introducir sesgos importantes en los resultados, especialmente cuando el mecanismo de los datos faltantes no es completamente al azar. De hecho, si los valores ausentes se deben a un mecanismo MAR o MNAR, eliminar observaciones reduce el poder estadístico y puede generar estimaciones sesgadas (Saeipourdizaj et al., 2021). Por eso, es fundamental tener en cuenta el mecanismo de datos faltantes antes de decidir si se eliminan o imputan los datos.

Como alternativa a la eliminación, existen diversos métodos de imputación que permiten completar los datos faltantes según supuestos estadísticos. Estas técnicas pueden ser univariadas o multivariadas, y pueden generar un único valor imputado o varios valores simulados, con el objetivo de reflejar la incertidumbre asociada a los datos ausentes. En la imputación única, cada valor faltante se reemplaza por un único valor estimado, como la media, la mediana o un valor cercano. Sin embargo, este enfoque tiende a subestimar la variabilidad del conjunto de datos.

Por otro lado, la imputación múltiple propone generar varias versiones completas del conjunto de datos, imputando varios valores para cada dato ausente. Posteriormente, los análisis se realizan por separado en cada una de estas versiones y se combinan los resultados para obtener estimaciones más robustas y confiables (Little and Rubin, 2019). Este enfoque permite incorporar la incertidumbre del proceso de imputación, lo que resulta especialmente útil cuando el mecanismo de datos faltantes es MAR (Engels and Diehr, 2003).

Imputación múltiple

La imputación múltiple es un método predictivo utilizado para abordar el problema de los datos faltantes en un conjunto de datos. Este método implica la creación de múltiples versiones de los datos completos, cada una con un conjunto diferente de valores imputados para los datos faltantes, utilizando métodos de simulación de Monte Carlo (Van Buuren, 2018). Estas versiones de los conjuntos de datos completos se analizan y combinan para obtener estimaciones más robustas y precisas (Rubin, 1996). Para llevar a cabo este proceso, se toma la matriz indicadora R para determinar los datos faltantes del conjunto de datos Y , donde:

$$\mathbf{R} = \begin{cases} r_{ij} = 1, & \text{si } y_{ij} \text{ es observado,} \\ r_{ij} = 0, & \text{en caso contrario.} \end{cases} \quad (3-9)$$

Con esta matriz R , se puede dividir los datos observados y faltantes de $\mathbf{Y}$ en $\mathbf{Y}_{\text{obs}}$ y $\mathbf{Y}_{\text{mis}}$, respectivamente. El primer conjunto contiene los elementos $y_{i,j}$ cuando $r_{i,j} = 1$, mientras que el segundo conjunto representa los datos faltantes. Así, $\mathbf{Y} = (\mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}})$ representa hipotéticamente los datos completos.

La imputación se realiza mediante un modelo predictivo de $\mathbf{Y}_{\text{mis}} \mid \mathbf{Y}_{\text{obs}}, \boldsymbol{\theta} \sim f(\boldsymbol{\theta} \mid \mathbf{Y}_{\text{obs}})$, donde $\boldsymbol{\theta}$ son los parámetros del modelo estimados a partir de $\mathbf{Y}_{\text{obs}}$. Posteriormente, se utilizan simulaciones de este modelo para predecir los valores faltantes.

Es importante señalar que este método de imputación asume que el mecanismo de datos faltantes es MAR. Este enfoque es esencial para la imputación, ya que al generar múltiples valores para los datos faltantes, se establece una distribución adecuada de posibles valores completos para cada observación ausente. Esta distribución considera la relación entre las variables observadas y los valores faltantes, utilizando los datos existentes para estimar los valores desconocidos (Van Buuren, 2018).

Técnica de especificación totalmente condicional

La imputación múltiple por Especificación Totalmente Condicional (FCS, por sus siglas en inglés), desarrollada por Van Buuren (2018), es una técnica robusta para manejar datos faltantes mediante imputación múltiple. Esta técnica requiere especificar un modelo de regresión para cada una de las variables a imputar utilizando la información disponible de las demás variables en el conjunto de datos. Al considerar las interacciones entre las variables durante el proceso de imputación, FCS puede capturar patrones sutiles y dependencias entre variables que se podrían perder con otros métodos.

Cada variable $\mathbf{y}_j$ con datos faltantes se modela condicionalmente como:

$$P(\mathbf{y}_j \mid \mathbf{Y}_{-j}, \boldsymbol{\theta}_j) \quad (3-10)$$

donde:

- $\mathbf{Y}_{-j}$: Todas las variables excepto la j -ésima.
- $\boldsymbol{\theta}_j$: Vector de parámetros asociado al modelo condicional para $\mathbf{y}_j$.

El algoritmo comienza con las imputaciones iniciales para los valores faltantes, generalmente tomadas aleatoriamente de los datos observados, seguido de la imputación iterativa para cada variable $\mathbf{y}_j$, actualizando los parámetros del modelo $\boldsymbol{\theta}_j$ utilizando datos observados e imputados de la forma:

$$\boldsymbol{\theta}_j^{(t+1)} = \arg \max_{\boldsymbol{\theta}} P(\mathbf{y}_j^{\text{obs}}, \mathbf{y}_j^{\text{mis}(t)} \mid \mathbf{Y}_{-j}, \boldsymbol{\theta}) \quad (3-11)$$

Imputa de nuevo valores faltantes de $\mathbf{y}_j$ basándose en el modelo actualizado:

$$\mathbf{y}_j^{\text{mis}(t+1)} \sim P(\mathbf{y}_j \mid \mathbf{Y}_{-j}, \boldsymbol{\theta}_j^{(t+1)}) \quad (3-12)$$

Se continua iterando a través de todas las variables varias veces, usualmente de 5 a 10 iteraciones (Van Buuren, 2018). Por último, el algoritmo converge cuando las distribuciones de los valores imputados se estabilizan a lo largo de las iteraciones. Los conjuntos de datos imputados se combinan utilizando las reglas de Rubin para obtener las estimaciones finales (Liu et al., 2014).

Métodos de imputación

Dentro de la técnica de FCS, es posible especificar diferentes modelos para imputar cada variable de acuerdo con su naturaleza y el patrón de asociación con las demás variables del conjunto de datos. En este trabajo se emplearon tres métodos de imputación integrados dentro de este enfoque: Coincidencia de medias predictivas (*Predictive Mean Matching*, PMM), los árboles de clasificación y regresión (*Classification and Regression Trees*, CART) y la regresión logística (*Logistic Regression*, logreg).

Coincidencia de medias predictivas (PMM): Es un método de imputación basado en la idea de emparejar los valores faltantes con valores observados similares. Es una técnica semi-paramétrica que combina elementos de regresión y emparejamiento lo que la hace especialmente útil para mantener la distribución de los datos imputados dentro del rango observado y la relación entre las variables. Dicho método funciona de la siguiente manera:

Primero, se ajusta un modelo de regresión lineal múltiple utilizando las observaciones completas para generar valores predichos de los datos faltantes. Luego, los valores imputados se seleccionan aleatoriamente entre los valores observados más cercanos a las predicciones obtenidas del modelo.

Árboles de Clasificación y Regresión (CART): Es un método no paramétrico que se utiliza tanto para tareas de clasificación como de regresión. Dicho método se basa en un algoritmo que divide el espacio de las variables de entrada en subconjuntos homogéneos en función de un criterio de selección, construyendo un modelo de árbol de decisión para predecir los valores faltantes. Este método es flexible, ya que no requiere de suposición sobre la distribución de las variables, captura las relaciones no lineales en ellas y trata adecuadamente los valores extremos. Adicionalmente, puede manejar tanto variables continuas como categóricas.

Regresión logística (logreg): Es un método de imputación utilizado para variables categóricas binarias. Se basa en la construcción de un modelo de regresión logística que estima la probabilidad de que una observación pertenezca a una de las dos categorías posibles, en función de las demás variables del conjunto de datos. Dicho método funciona de la siguiente manera: primero, se ajusta el modelo utilizando únicamente las observaciones completas para predecir la variable con datos faltantes. Luego, las imputaciones se generan asignando valores de acuerdo con las probabilidades estimadas por el modelo, permitiendo conservar la estructura binaria de la variable y su relación con las variables auxiliares.

Pruebas de bondad de ajuste

Las pruebas de bondad de ajuste permiten evaluar si una muestra de datos proviene de una distribución específica o si dos muestras pueden considerarse provenientes de la misma distribución. Se basan en comparar las funciones de distribución empíricas con funciones teóricas o con distribuciones obtenidas a partir de otra muestra. Estas pruebas son ampliamente utilizadas en estadística aplicada debido a su capacidad para contrastar la coherencia entre datos observados y modelos o imputaciones de datos faltantes (D'Agostino and Stephens, 1986).

Entre dichas pruebas no paramétricas se encuentran: Cramér–von Mises (CvM), Anderson–Darling (AD) y Chi-cuadrado, seleccionadas por su pertinencia para este estudio en la comparación entre valores observados e imputados.

Prueba de Cramér–von Mises (CvM): La prueba de Cramér–von Mises compara la función de distribución empírica $F_n(x)$ de una muestra con otra función de referencia $F(x)$ teórica o empírica. El estadístico se define como:

$$W^2 = n \int_{-\infty}^{\infty} [F_n(x) - F(x)]^2 dF(x), \quad (3-13)$$

donde n es el tamaño muestral.

En la práctica, el estadístico puede calcularse como:

$$W^2 = \frac{1}{12n} + \sum_{i=1}^n \left(\frac{2i-1}{2n} - F(X_{(i)}) \right)^2, \quad (3-14)$$

siendo $X_{(i)}$ los datos ordenados Stephens (1974).

El *CvM* pondera de manera uniforme todo el rango de la distribución, siendo sensible a discrepancias globales. Por ello es recomendado para validar que la distribución de datos imputados no distorsiona la estructura general de los observados (D'Agostino and Stephens, 1986).

Prueba de Anderson–Darling (AD): La prueba de Anderson–Darling es una modificación de CvM que otorga mayor peso a las colas de la distribución, lo cual resulta útil cuando los valores extremos son relevantes. Su estadístico se expresa como:

$$A^2 = -n - \frac{1}{n} \sum_{i=1}^n \left[(2i-1) (\ln F(X_{(i)}) + \ln (1 - F(X_{(n+1-i)}))) \right], \quad (3-15)$$

donde $F(X)$ es la función de distribución acumulada teórica o de comparación, y $X_{(i)}$ los valores ordenados Anderson and Darling (1954).

La ventaja del test de AD es que detecta diferencias en las colas con mayor potencia que el CvM o el Kolmogorov–Smirnov, siendo recomendable en contextos donde los valores extremos son críticos Stephens (1974).

Prueba Chi-cuadrado (χ^2): La prueba Chi-cuadrado permite comparar distribuciones categóricas mediante el contraste entre frecuencias observadas O_i y frecuencias esperadas E_i .

El estadístico se define como:

$$\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i}, \quad (3-16)$$

donde k es el número de categorías.

Bajo la hipótesis nula, χ^2 se distribuye aproximadamente como una chi-cuadrado con $k - 1$ grados de libertad Plackett (1983).

3.2.3. Regresión Cuantílica

Cuantil

Los cuantiles son una generalización del concepto de mediana, así como la mediana divide a los datos en dos partes con el mismo número de elementos en cada una, de la misma manera se utiliza para dividir en cuatro, diez o cien partes, llegando así al concepto de cuantil.

Koenker (2005) planea una definición formal, donde X es una variable aleatoria caracterizada por su función de distribución acumulada:

$$F(x) = P(X \leq x). \quad (3-17)$$

Mientras que, se define el τ -ésimo cuantil de X a :

$$Q_\tau(x) = \inf\{x : F(x) \geq \tau\}, \quad 0 < \tau < 1 \quad (3-18)$$

Por ejemplo la mediana esta definida como $F^{-1}(\frac{1}{2})$ que juega el papel central.

Regresión cuantílica

Koenker and Bassett Jr (1978) extendieron el modelo de regresión lineal para modelar no la media condicional, sino los cuantiles condicionales de la variable de interés Y . Para cada $\tau \in (0, 1)$, la regresión cuantílica describe la función

$$Q_Y(\tau | \mathbf{x}) = \inf\{y : F_{Y|\mathbf{x}}(y | \mathbf{x}) \geq \tau\},$$

que corresponde al cuantíl de orden τ de Y dado el vector de covariables explicativas

$$\mathbf{x} = (x_1, x_2, \dots, x_p)^\top.$$

Este enfoque captura asimetrías, permite modelar heterocedasticidad y ofrece robustez frente a valores atípicos.

Regresión cuantílica paramétrica

La regresión cuantílica busca estimar el cuantil τ condicional de la variable dependiente en función de un conjunto de covariables, proporcionando así una caracterización más completa de la distribución condicional. Sea y_i la variable dependiente y $\mathbf{x}_i$ el vector de covariables para

$i = 1, 2, 3, \dots, n$. En lugar de modelar la media condicional $\mathbb{E}(y_i|\mathbf{x}_i)$, la regresión cuantílica modela el τ -ésimo cuantil condicional de y_i dado $\mathbf{x}_i$, denotado como $Q_y(\tau|\mathbf{x}_i)$. Formalmente, el modelo de regresión cuantílica paramétrica se expresa como:

$$Q_y(\tau | \mathbf{x}_i) = \mathbf{x}_i^\top \boldsymbol{\beta}(\tau) \quad (3-19)$$

donde $\boldsymbol{\beta}(\tau)$ es el vector de parámetros a estimar para el cuantil τ . Este enfoque permite que los coeficientes varíen para diferentes niveles de τ , proporcionando una representación más flexible y detallada de la relación entre las covariables y la variable dependiente. La estimación de $\boldsymbol{\beta}(\tau)$ se realiza minimizando una función de pérdida asimétrica basada en los residuales. Para un nivel de cuantil $\tau \in (0, 1)$ la función de pérdida se define como:

$$\rho_\tau(u) = u(\tau - \mathbb{I}(u < 0)) \quad (3-20)$$

Donde $\mathbb{I}(\cdot)$ es la función indicadora. Esta función de pérdida penaliza de forma asimétrica los residuales, ponderando de forma diferente los valores positivos y negativos según el valor de τ . El estimador de $\boldsymbol{\beta}(\tau)$ se obtiene resolviendo el siguiente problema de optimización:

$$\hat{\boldsymbol{\beta}}(\tau) = \arg \min_{\boldsymbol{\beta} \in \mathbb{R}^p} \sum_{i=1}^n \rho_\tau(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}) \quad (3-21)$$

Esta formulación convierte el problema en una optimización convexa que se puede resolver eficientemente mediante técnicas de programación lineal, lo que facilita su implementación en la práctica. Por otra parte, según Koenker and Hallock (2001) el estimador de $\boldsymbol{\beta}(\tau)$ también se puede obtener como la solución al siguiente problema de minimización:

$$\hat{\boldsymbol{\beta}}_\tau = \underset{\boldsymbol{\beta}_\tau \in \mathbb{R}}{\text{Min}} \left[\sum_{y_i \geq \mathbf{x}_i^\top \boldsymbol{\beta}_\tau} \tau |y_i - \mathbf{x}_i^\top \boldsymbol{\beta}_\tau| + \sum_{y_i < \mathbf{x}_i^\top \boldsymbol{\beta}_\tau} (1 - \tau) |y_i - \mathbf{x}_i^\top \boldsymbol{\beta}_\tau| \right] \quad (3-22)$$

Métricas y test de Wald

En la regresión cuantílica, las métricas de evaluación juegan un papel fundamental para determinar el ajuste del modelo y su interpretación. A continuación, se describen las principales métricas utilizadas:

Pseudo R^2

El Pseudo R^2 en el contexto de la regresión cuantílica se utiliza para evaluar el ajuste del modelo, de manera análoga al R^2 tradicional en regresión lineal, pero adaptado para cuantiles específicos. Una definición que comparte Koenker (2005) es:

$$\text{Pseudo-}R^2 = 1 - \frac{\sum_{i=1}^n \rho_\tau(y_i - \hat{Q}_{y_i}(\tau | \mathbf{x}_i))}{\sum_{i=1}^n \rho_\tau(y_i - \tilde{Q}_y(\tau))} \quad (3-23)$$

donde $\tilde{Q}_y(\tau)$ es el cuantil empírico τ -ésimo. Valores cercanos a 1 indican mejor ajuste.

Error Absoluto Medio (MAE)

El Error Absoluto Medio (MAE) es una métrica robusta para evaluar la precisión del modelo, ya que mide la magnitud promedio del error de predicción

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (3-24)$$

Koenker (2005) destaca que, a diferencia de la regresión de mínimos cuadrados, que minimiza los errores cuadráticos, la regresión cuantílica se enfoca en minimizar los errores absolutos ponderados, lo que la hace menos sensible a valores atípicos y proporciona un ajuste robusto. Esto está en línea con la filosofía del MAE, que también es robusto ante observaciones extremas debido a que no amplifica el impacto de los errores grandes como lo haría el error cuadrático medio (MSE).

Test de Wald

La prueba de Wald se utiliza para evaluar hipótesis sobre los coeficientes del modelo de regresión cuantílica, específicamente para evaluar combinaciones lineales de los coeficientes considerando diferentes cuantiles. La hipótesis es la siguiente:

$$H_0 : \mathbf{C}\boldsymbol{\beta}(\tau) = \mathbf{r} \quad \text{vs.} \quad H_a : \mathbf{C}\boldsymbol{\beta}(\tau) \neq \mathbf{r} \quad (3-25)$$

donde $\mathbf{C}$ es una matriz de restricciones de dimensión $p \times q$, con q restricciones lineales y p parámetros estimados. Dicha matriz, especifica cuáles coeficientes se están evaluando y $\mathbf{r}$ es el vector de valores específicos bajo la hipótesis nula. Después, la estadística de prueba de Wald se define como:

$$W = (\mathbf{C}\hat{\boldsymbol{\beta}}(\tau) - \mathbf{r})^\top (\mathbf{C}\hat{\boldsymbol{\Sigma}}\mathbf{C}^\top)^{-1} (\mathbf{C}\hat{\boldsymbol{\beta}}(\tau) - \mathbf{r}) \quad (3-26)$$

$\hat{\boldsymbol{\beta}}(\tau)$ corresponde al vector de coeficientes estimados y $\hat{\boldsymbol{\Sigma}}$ a su matriz de covarianza estimada. Koenker (2005) destaca que la estadística de Wald sigue una distribución χ^2 con grados de libertad iguales al rango de la matriz $\mathbf{C}$ y se utiliza para determinar si existe evidencia estadísticamente significativa para rechazar la hipótesis nula.

Regresión cuantílica no paramétrica

Sea Y la variable respuesta y $\mathbf{x} \in \mathbb{R}^p$ el *vector* de covariables. Cuando la relación entre Y y $\mathbf{x}$ presenta no linealidad que un modelo lineal no captura, el τ -ésimo cuantil condicional se describe mediante una función desconocida g_τ :

$$Q_Y(\tau \mid \mathbf{x}) = g_\tau(\mathbf{x}), \quad 0 < \tau < 1.$$

Aquí $g_\tau \in \mathcal{G}$ pertenece a una familia flexible de funciones como los splines, núcleos o bosques aleatorios, evitando imponer una forma funcional específica. El estimador se define como

$$\hat{g}_\tau = \arg \min_{g \in \mathcal{G}} \sum_{i=1}^n \rho_\tau(Y_i - g(\mathbf{x}_i)), \quad \rho_\tau(u) = u(\tau - \mathbb{I}\{u < 0\}),$$

donde ρ_τ es la función de pérdida asimétrica de Koenker y Bassett. La complejidad de $\mathcal{G}$ suele regularizarse incorporando un término de penalización.

Spline de suavizado

En la regresión cuantílica no paramétrica, se utiliza frecuentemente el suavizado en bases spline para estimar la función $g(x)$. Las bases spline son funciones que dividen el rango de la covariable en segmentos mediante puntos llamados nodos, dentro de los cuales la relación se modela utilizando polinomios de bajo grado, generalmente cúbicos que es continua hasta la segunda derivada. Matemáticamente, un spline cúbico $s(x)$ con k nodos puede expresarse como:

$$s(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \sum_{k=1}^K \gamma_k (x - \xi_k)_+^3 \quad (3-27)$$

donde

- $\beta_0, \beta_1, \beta_2, \beta_3$: Coeficientes del polinomio base.
- γ_k : Coeficientes asociados a los nodos.
- $(x - \xi_k)_+^3$: Función cúbica truncada si $x > \xi_k$ y 0 en otro caso .

Los splines permiten una gran flexibilidad en el ajuste de datos complejos y no lineales, ya que la función puede adaptarse localmente a la estructura de los datos.

Regresión cuantílica no paramétrica con Splines

Koenker (2005), combina la regresión cuantílica no paramétrica con splines como herramienta para estimar los cuantiles condicionales de forma flexible a través de diferentes partes de la distribución condicional. Asimismo, destaca la importancia de la penalización de suavidad para evitar el sobreajuste del modelo. El problema de estimación se formula como

$$\hat{g}_\tau = \arg \min_{g \in \mathcal{G}} \sum_{i=1}^n \rho_\tau(y_i - g(\mathbf{x}_i)) + \lambda J(g) \quad (3-28)$$

donde:

$$g(x) = \sum_{j=1}^p \beta_j B_j(x), \quad B_j(x) = \text{son las funciones base spline}, \quad (3-29)$$

$$J(g) = \int (g''(x))^2 dx, \quad (3-30)$$

en el que, $J(g)$ es el término de penalización más común que generalmente mide la rugosidad del spline y que para un spline cúbico se define como la integral de la segunda derivada.

Así, el parámetro de suavizado λ controla el compromiso entre el equilibrio del ajuste a los datos y la suavidad de la función estimada. Valores altos de λ producen un ajuste más suave, mientras que valores bajos permiten que el modelo se ajuste más estrechamente a los datos. A su vez, en este caso la función $g(\mathbf{x}_i)$ es representada utilizando una base spline y describe la relación entre x y la variable de respuesta y de manera flexible.

Esta técnica permite modelar datos de manera flexible, adaptándose a diferentes patrones sin imponer una estructura lineal estricta, lo que es particularmente útil cuando la relación entre las variables es desconocida o se sospecha de no linealidad. También, en estudios donde se observa heterocedasticidad, debido a la capacidad para modelar cambios locales en la distribución de la variable respuesta.

4. Metodología

A continuación, se describe la metodología utilizada para modelar el comportamiento de la concentración promedio octohoraria del O_3 mediante regresión cuantílica, evaluando distintos percentiles a lo largo del día con el propósito de identificar el efecto de las variables meteorológicas que interactúan con dicho comportamiento. Todos los procedimientos estadísticos desarrollados en este estudio fueron implementados en el software libre R (R Core Team, 2023), empleando principalmente las librerías `mice` (Van Buuren and Groothuis-Oudshoorn, 2011) para la imputación de datos faltantes y `quantreg` (Koenker et al., 2018) para el ajuste de los modelos de regresión cuantílica. En esta sección se define la base de datos empleada y se detallan de manera clara y estructurada los procesos estadísticos aplicados a lo largo del análisis.

4.1. Datos

La base de datos utilizada para este estudio proviene del DAGMA, entidad que administra el SVCA de Santiago de Cali y se encuentra pública en la plataforma de Datos Abiertos del Gobierno Nacional (DAGMA, 2022). Los registros fueron captados en la estación de monitoreo de Pance para el año 2019 que contiene las variables de los contaminantes en niveles de concentración promedio horaria de material particulado (PM10) y O_3 . También, cuenta con variables meteorológicas como temperatura, humedad, radiación solar, lluvia, velocidad y dirección del viento.

Dicho lo anterior, se partió de la organización de la base de datos para así poder calcular el promedio octohorario de O_3 y realizar un exhaustivo análisis exploratorio.

4.1.1. Organización de la base de datos

La organización de la base de datos incluyó, en primer lugar, la separación de la fecha y la hora en dos columnas independientes, con el fin de manejar adecuadamente la dimensión temporal de los registros. Posteriormente, la variable Lluvia, originalmente continua, fue transformada en una variable binaria que indica la presencia o ausencia de precipitación, categorizada como “Lluvia” y “Sin lluvia”. A continuación, se presenta una tabla con la descripción general de las variables que conforman la base de datos utilizada en este estudio:

Nombre de la variable	Descripción
No	Identificador secuencial que numera los registros horarios del año 2019.
Fecha	Fecha calendario correspondiente a cada registro, expresada en el formato año-mes-día.
Hora	Hora del día expresada entre las 00:00 y las 23:00 horas.
O ₃	Concentración horaria de Ozono Troposférico en microgramos por metro cúbico ($\frac{\mu g}{m^3}$).
PM10	Concentración horaria de material particulado menor a 10 micras en microgramos por metro cúbico ($\frac{\mu g}{m^3}$).
Vel Viento	Velocidad del viento en metros por segundo ($\frac{m}{s}$).
Dir Viento	Dirección del viento en grados sexagesimales ($^{\circ}$).
Temperatura	Temperatura del aire en grados Celsius ($^{\circ}C$).
Humedad	Humedad relativa del aire expresada en porcentaje (%).
Rad Solar	Radiación solar en vatios por metro cuadrado ($\frac{W}{m^2}$).
Lluvia	Variable binaria que indica la presencia (1) o ausencia (0) de lluvia en el registro correspondiente.

Tabla 4-1.: Definición de variables del conjunto de datos

Finalmente, la variable Dirección del Viento, dada su naturaleza cíclica, se abordó mediante técnicas de análisis circular, lo que permitió explorar su asociación con la concentración de O₃ y la velocidad del viento.

4.1.2. Análisis exploratorio

En esta parte del estudio, se llevaron a cabo análisis descriptivos de la base de datos previamente descrita. En primer lugar, se calcularon medidas de tendencia central y dispersión para las variables de contaminantes ambientales y variables meteorológicas con el fin de caracterizar su distribución. Para la variable *Dirección del Viento*, se aplicaron métodos de análisis de datos circulares, entre ellos la elaboración de la rosa de los vientos. Adicionalmente, se construyó la rosa de contaminación de O₃, que permitió explorar la relación entre la concentración del contaminante y la dirección del viento.

También, se realizaron análisis de correlación con el propósito de identificar asociaciones entre variables que podrían ser incluidas en los modelos de imputación y de regresión cuantílica. Además, se elaboraron gráficos que representan las concentraciones de O₃ a lo largo del tiempo para identificar patrones en el comportamiento diario. Por último, se exploró la estructura de los valores faltantes mediante representaciones gráficas, lo que permitió obtener una visión preliminar del patrón de ausencia de datos, aunque no fue posible identificar con certeza el mecanismo subyacente.

4.2. Imputación de datos

Para desarrollar el primer objetivo de este trabajo se implementó el método de imputación múltiple para abordar el problema de los datos faltantes en la variable O_3 , permitiendo capturar la incertidumbre inherente a los valores ausentes y mejorar la robustez de los análisis posteriores. Es importante destacar que la aplicación de dicha técnica fue extendida a las demás variables exceptuando la dirección del viento y la humedad debido a razones metodológicas específicas.

En primer lugar, se excluyó la variable circular *Dirección del Viento*, que no se puede tomar de manera lineal debido a su naturaleza cíclica. Por tal razón, su tratamiento requiere de una transformación a coordenadas cartesianas de seno y coseno que podría llevar a desafíos en relación a la multicolinealidad y la interpretación de los modelos de imputación (Fisher, 1995).

Por su parte, la variable Humedad se excluyó debido a la alta correlación con la variable Temperatura. Dado que ambas variables comparten una proporción considerable de información, prescindir de una de ellas contribuye a evitar problemas de colinealidad sin comprometer la calidad de los análisis (Rubin, 1978).

Aclarando lo anterior, la función `mice()` permitió definir los métodos más adecuados para la imputación con base al tipo y la distribución de cada variable, al igual que la cantidad de imputaciones m a realizar para crear los diferentes conjuntos de datos completos. Lo anterior, se evaluó realizando análisis descriptivos, histogramas, gráficos de densidad y pruebas formales, comparando las distribuciones de los valores observados y los imputados.

4.2.1. Métodos y modelos de imputación

Se seleccionaron tres métodos existentes dentro la gran variedad métodos de la función `mice()`, por su capacidad para manejar los diferentes tipos de variables y sus características específicas.

El algoritmo `mice()` es iterativo y aplica imputaciones por cadena. Esto quiere decir que, cada variable con datos faltantes se imputa por turnos, una tras otra, utilizando los valores completos y los imputados en las iteraciones anteriores. Es por esta razón, que se comenzó a imputar por la variable O_3 que tiene el mayor número de datos faltantes hasta la menor que corresponde a la radiación solar. Dicho principio busca minimizar el sesgo y maximizar el uso de información disponible, de este modo reducir la propagación de errores y mejorar la estabilidad numérica del proceso de imputación.

Dicho lo anterior, teniendo en cuenta los métodos *PMM*, *CART* y *Logreg* seleccionados para las diferentes variables, se aplicaron los modelos de imputación correspondientes, de acuerdo con la implementación de la función `mice()`. Lo anterior permitió seguir conservando el mayor rango de información y hacer una imputación robusta. Adicionalmente, como la imputación se realiza en ciclos, y en cada ciclo, se imputa la variable, este proceso fue iterado 5 veces de acuerdo a la investigación de Van Buuren (2018), siguiendo el orden especificado de los modelos ajustados a continuación:

Modelos de Regresión *PMM*

En el marco del método de imputación *PMM*, se ajustan modelos de regresión para las variables O_3 , $PM10$, $VelViento$ y $Temperatura$, en función de las demás variables observadas, con el fin de preservar la distribución original de los datos y respetar posibles relaciones no lineales o asimetrías presentes. Los modelos utilizados se presentan a continuación:

$$O_3 = \beta_0 + \beta_1 \cdot Hora + \beta_2 \cdot PM10 + \beta_3 \cdot Vel Viento + \beta_4 \cdot Temperatura \\ + \beta_5 \cdot Lluvia + \beta_6 \cdot Rad Solar + \epsilon$$

$$PM10 = \beta_0 + \beta_1 \cdot Hora + \beta_2 \cdot O_3 + \beta_3 \cdot Vel Viento + \beta_4 \cdot Temperatura \\ + \beta_5 \cdot Lluvia + \beta_6 \cdot Rad Solar + \epsilon$$

$$Vel Viento = \beta_0 + \beta_1 \cdot Hora + \beta_2 \cdot O_3 + \beta_3 \cdot PM10 + \beta_4 \cdot Temperatura \\ + \beta_5 \cdot Lluvia + \beta_6 \cdot Rad Solar + \epsilon$$

$$Temperatura = \beta_0 + \beta_1 \cdot Hora + \beta_2 \cdot O_3 + \beta_3 \cdot PM10 + \beta_4 \cdot Vel Viento \\ + \beta_5 \cdot Lluvia + \beta_6 \cdot Rad Solar + \epsilon$$

Modelos de Regresión *Logreg*

El modelo logístico especificado estima la probabilidad de ocurrencia de lluvia en función de las covariables $Hora$, O_3 , $PM10$, velocidad del viento, temperatura y radiación solar.

$$\text{logit } P(\text{Lluvia}=1) = \beta_0 + \beta_1 \cdot Hora + \beta_2 \cdot O_3 + \beta_3 \cdot PM10 + \beta_4 \cdot Vel Viento \\ + \beta_5 \cdot Temperatura + \beta_6 \cdot Rad Solar + \epsilon$$

Modelos de Regresión *CART*

La construcción del árbol de decisión se realiza mediante divisiones recursivas basadas en los predictores disponibles, eligiendo en cada división el predictor y el punto de corte que minimicen la suma del Error Cuadrático Medio (MSE) dentro de los nodos resultantes. Para el entrenamiento del modelo, se ajusta el modelo de regresión mediante el algoritmo CART. Una vez construido el árbol, para imputar cada valor faltante, se utiliza el árbol para predecir y asignar valores imputados basándose en el promedio de los valores en el nodo final. Este proceso se repite 5 veces para crear varios conjuntos de datos imputados. Finalmente, se validan y refinan los modelos imputados utilizando validación cruzada para asegurar su precisión y evitar el sobreajuste (Breiman, 2017).

Los modelos de regresión utilizados en la imputación CART para la radiación solar no se expresan en términos de fórmulas algebraicas tradicionales, ya que este método consiste en algoritmos que generan reglas de decisión basadas en los datos (Fernández-Casal et al., 2024), de la siguiente manera:

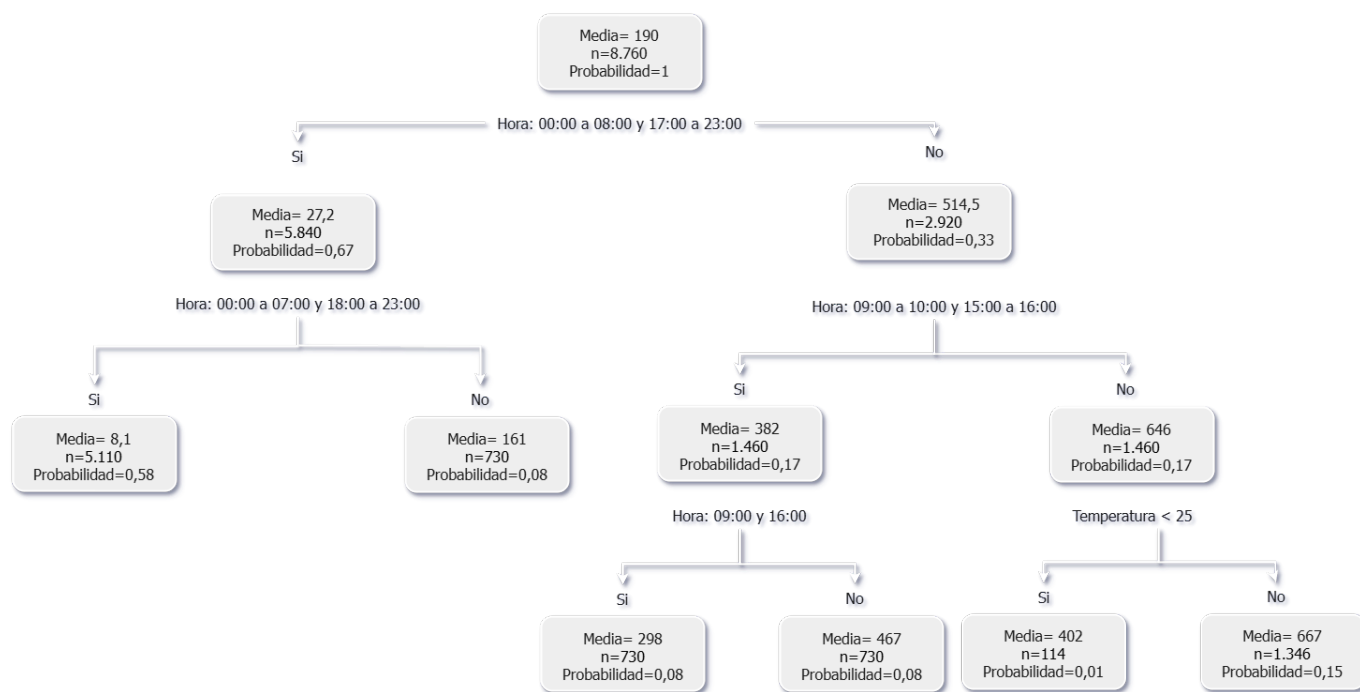


Figura 4-1.: Árbol de clasificación basado en CART empleado para la imputación de la variable Rad Solar según la hora del día y la temperatura

4.2.2. Prueba de bondad de ajuste

Con el propósito de evaluar la coherencia entre los datos originales y los conjuntos completos tras la imputación múltiple, se aplicaron pruebas de bondad de ajuste basadas en la

comparación de distribuciones. Para las variables continuas PM10, Temperatura, Velocidad del Viento y Radiación Solar se utilizó la prueba de Cramér–von Mises (CvM), la cual contrasta la discrepancia global entre la función de distribución empírica de los datos observados y la de los conjuntos imputados, otorgando un peso uniforme a todo el rango de valores. En el caso de la variable O_3 , se empleó la prueba de Anderson–Darling (AD), debido a que proporciona mayor sensibilidad en las colas de la distribución, aspecto relevante en el análisis de este contaminante atmosférico. Finalmente, para la variable binaria Lluvia se aplicó la prueba de Chi-cuadrado de Pearson, adecuada para comparar proporciones categóricas. En todos los contrastes se consideró como criterio de aceptación de la imputación la no existencia de diferencias estadísticamente significativas ($p > 0,05$) entre los datos observados y los completados.

4.3. Ajuste de los modelos de regresión cuantílica

Con los 5 conjuntos de datos imputados se procedió a dar cumplimiento al segundo objetivo de la investigación, sin antes asegurar la coherencia temporal entre la variable de respuesta y los predictores incluidos en los modelos de regresión cuantílica. Dado que el interés de este estudio se centra en el promedio octohorario de O_3 (P8O3), el cual representa una medida acumulada, no resulta metodológicamente adecuado utilizar como explicativas variables meteorológicas en su forma horaria, ya que esto podría afectar la validez de las estimaciones e inducir interpretaciones erróneas sobre la relación entre las variables, debido a la discrepancia en las escalas temporales. Para resolver esta inconsistencia, se aplicó el mismo cálculo de la media móvil de ocho horas a las covariables continuas, las cuales se denominaron: *PM10_P8*, *VelViento_P8*, *Temperatura_P8*, y *RadSolar_P8*. Así, cada observación del conjunto de datos representa condiciones atmosféricas promedio correspondientes al mismo intervalo de tiempo que la variable de respuesta, lo que permite una interpretación más adecuada y consistente de los coeficientes obtenidos.

Siguiendo el mismo principio de consistencia temporal aplicado a las covariables continuas, la precipitación se recodificó con las mismas categorías Lluvia o Sin Lluvia, pero como un indicador dicotómico asignado a cada bloque de ocho horas cuando al menos uno de los registros presentó precipitación. Esta estrategia no solo mantiene la coherencia en la escala de tiempo, sino que también incrementó la proporción de registros en la categoría “Lluvia”, al capturar episodios breves que en la escala horaria podían quedar dispersos o subrepresentados. Asimismo, el enfoque adoptado no busca capturar eventos puntuales de lluvia, sino estimar el efecto de haber tenido lluvia en algún momento dentro de las últimas 8 horas, lo cual resulta metodológicamente coherente y razonable desde el punto de vista estadístico.

Una vez transformadas y alineadas temporalmente todas las variables, se prescindió de los primeros siete registros de cada variable debido a que el cálculo de las medias móviles de

ocho horas, al estar alineado a la derecha, requiere de al menos ocho observaciones previas completas para generar el primer valor válido. Adicionalmente, se centraron las covariables continuas en cada uno de los cinco conjuntos de datos imputados, con el propósito de mejorar la estabilidad numérica de los modelos, facilitar la interpretación del intercepto y mitigar posibles problemas de colinealidad entre predictores.

De esta manera, con las variables debidamente preparadas, se procedió al ajuste de los modelos de regresión cuantílica. Se consideraron dos enfoques: primero una especificación paramétrica lineal, orientada a estimar el efecto de las variables meteorológicas sobre los distintos cuantiles de $P8O3$; y segundo, una alternativa no paramétrica basada en funciones spline, utilizada para caracterizar su comportamiento diario. Ambos enfoques se describen a continuación.

4.3.1. Ajuste del modelo de regresión cuantílica paramétrica

Para cada uno de los cinco conjuntos de datos imputados, se ajustó un modelo de regresión cuantílica lineal múltiple utilizando la función `rq()` del paquete `quantreg` para cada cuantil de interés: $\tau = 0.10, 0.25, 0.50, 0.75$ y 0.9 con el objetivo de analizar el impacto de distintas variables meteorológicas suavizadas sobre diferentes cuantiles de la distribución de la concentración del $P8O3$.

La estructura general del modelo ajustado para los diferentes cuantiles τ tiene la siguiente forma:

$$Q_{P8O3}(\tau \mid \mathbf{X}) = \beta_0(\tau) + \beta_1(\tau) \cdot PM10_P8 + \beta_2(\tau) \cdot VelViento_P8 + \beta_3(\tau) \cdot Temperatura_P8 + \beta_4(\tau) \cdot RadSolar_P8 + \beta_5(\tau) \cdot \mathbb{I}(\text{Lluvia}) \quad (4-1)$$

Donde $\mathbb{I}(\text{Lluvia})$ es la función indicadora que vale 1 si en al menos una de las últimas 8 horas hubo lluvia, y 0 en el caso contrario.

4.3.2. Estimación de los coeficientes e intervalos de confianza

A partir de la funcionalidad del paquete `quantreg` en R, se estimaron los coeficientes de los modelos de regresión cuantílica para cada cuantil. La estimación se basa en la minimización de la suma ponderada de los residuos absolutos, lo que permite obtener los coeficientes $\beta(\tau)$ que describen la relación condicional entre las concentraciones de $P8O3$ y las covariables del modelo para cada nivel de la distribución.

Posteriormente, se optó por utilizar el método bootstrap para la estimación de los errores estándar (SE) de los coeficientes, debido a la presencia de heterocedasticidad y valores atípicos en los datos ya que pueden afectar la precisión si se emplean métodos tradicionales de inferencia. Para ello, se generaron 500 muestras aleatorias con reemplazo a partir de cada conjunto de datos imputado, manteniendo el tamaño del conjunto original ($n=43.765$). En cada muestra, se ajustó el modelo de regresión cuantílica para cada cuantil τ , obteniendo coeficientes estimados $\hat{\beta}_b(\tau)$ para cada iteración b del bootstrap. Así, los errores estándar se calcularon como la desviación estándar de estos coeficientes a lo largo de las 500 replicaciones.

Los intervalos de confianza al 95 % se construyeron utilizando los errores estándar obtenidos y asumiendo una distribución aproximadamente normal de los coeficientes estimados. Finalmente, se aplicó el enfoque propuesto por Rubin (1978) para combinar los resultados obtenidos en los cinco conjuntos imputados que consistió en promediar para cada cuantil, los coeficientes estimados, sus errores estándar y las métricas de ajuste, incorporando la variabilidad tanto intra como entre imputaciones.

4.3.3. Métricas de evaluación del modelo

Para evaluar el desempeño de los modelos ajustados en cada cuantil, se calcularon tres métricas: el pseudo- R^2 , el error absoluto medio (MAE) y el estadístico de Wald.

El pseudo- R^2 se estimó para cada conjunto imputado y cuantil utilizando la definición propuesta por Koenker, y luego se combinaron los valores resultantes siguiendo la regla de Rubin para imputación múltiple. Esta métrica permitió evaluar la calidad relativa del ajuste del modelo en cada nivel de la distribución de $P8O_3$. Mientras que el MAE se calculó como el promedio del valor absoluto de los residuos del modelo ajustado, proporcionando una medida directa de la precisión predictiva en unidades originales. Por su parte, el estadístico de Wald se utilizó para evaluar la significancia individual de los coeficientes estimados. En este caso, se calculó un estadístico de prueba a partir del cociente entre cada coeficiente y su error estándar. Bajo la hipótesis nula de que el coeficiente es igual a cero, este estadístico sigue asintóticamente una distribución normal estándar. A partir de dicha distribución se derivaron los valores- p , los cuales permitieron determinar qué efectos resultaron estadísticamente significativos en cada cuantil.

4.3.4. Ajuste del modelo de regresión cuantílica no paramétrica

Con el objetivo de modelar el comportamiento del $P8O_3$ a lo largo del día, inicialmente se implementó una regresión cuantílica paramétrica utilizando la variable *Hora* como única covariable. Esta decisión respondió al reconocimiento de la influencia significativa que ejercen

los cambios horarios sobre la variabilidad del $P8O_3$, derivada principalmente de los ciclos diarios de radiación solar y las condiciones atmosféricas asociadas. No obstante, durante el análisis se identificó un cruce de las curvas de cuantiles, lo que reveló problemas de monotonía y puso en evidencia las limitaciones del modelo paramétrico para captar adecuadamente relaciones no lineales complejas.

Como respuesta a esta problemática, se implementó un enfoque de regresión cuantílica no paramétrica, permitiendo mayor flexibilidad al no imponer una forma funcional específica entre $P8O_3$ y la *Hora*. Esta técnica facilitó la captura de patrones no lineales más complejos en los datos. No obstante, la regresión no paramétrica sola no fue suficiente para evitar el cruce de cuantiles, por lo que se decidió incorporar un suavizado mediante B-splines cúbicos. Esta técnica, comúnmente utilizada según Koenker (2005), permitió modelar de manera flexible la relación no lineal entre $P8O_3$ y la *Hora*, garantizando una transición suave entre los segmentos definidos por los puntos de corte o nodos, lo que evitó el cruce de cuantiles y proporcionó una estimación más robusta (Hastie et al., 2009).

Así, se utilizó una base de B-splines cúbicos, cuya implementación fue realizada mediante la función `bs()` del paquete `splines` en R. De acuerdo a las recomendaciones de Koenker (2005) el modelo se construyó con 15 grados de libertad como punto medio entre 10 y 20 que suelen hacerse en la práctica, lo que equivale a utilizar 15 funciones base spline para capturar adecuadamente la variación diaria de $P8O_3$ sin sobreajustar los datos. Esta decisión fue fundamentada tanto en la literatura como en ensayos exploratorios, buscando un equilibrio entre suavidad y capacidad de representación del comportamiento real de la variable.

Cabe señalar que no se incorporó explícitamente un parámetro de penalización λ sobre la curvatura del spline, como ocurre en los modelos de P-splines. En su lugar, la complejidad y suavidad del modelo fueron reguladas indirectamente mediante el número de grados de libertad, lo cual es conceptualmente equivalente a asumir un $\lambda = 0$, tal como se contempla en enfoques no penalizados.

El análisis se llevó a cabo de manera separada para los registros clasificados en dos condiciones: “Lluvia” y “Sin lluvia”. Para ello, se utilizaron como base los cinco conjuntos de datos imputados, sobre los cuales se ajustaron modelos independientes de regresión cuantílica no paramétrica para los cuantiles $\tau = 0,10, 0,25, 0,50, 0,75, 0,90$. La especificación general del modelo adoptado se presenta a continuación:

$$Q_{P8O_3}(\tau \mid \text{Hora}) = \beta_0(\tau) + \sum_{k=1}^{15} \beta_k(\tau) B_k(\text{Hora}), \quad (4-2)$$

donde $B_k(\text{Hora})$ representa la k -ésima función base spline cúbica.

Las predicciones fueron obtenidas para cada hora del día (desde las 0 hasta las 23 horas) y para cada cuantil dentro de cada conjunto de datos imputado, generando así cinco matrices independientes de predicción por cada régimen de precipitación. Posteriormente, estas matrices fueron promediadas con el fin de obtener una estimación final suavizada, diferenciada por cuantil y por condición de precipitación. Este procedimiento permitió describir claramente el comportamiento promedio del PM_{10} durante el día, en función de la presencia o ausencia de lluvia.

Finalmente, se elaboraron gráficos comparativos en los cuales las curvas estimadas para cada cuantil fueron superpuestas sobre las observaciones reales, diferenciando claramente entre ambas condiciones meteorológicas. Esta visualización facilitó el análisis de las variaciones en los cuantiles del PM_{10} a lo largo del día, permitiendo contrastar con claridad los patrones observados según la condición de lluvia o ausencia de la misma.

5. Resultados

En esta sección se desarrolla la metodología planteada iniciando con los resultados del análisis exploratorio de datos, la imputación de datos faltantes y se presenta el ajuste del modelo de regresión por cuantiles. Finalmente, se analiza el impacto de variables meteorológicas sobre las concentraciones promedio octohorarias de Ozono Troposférico.

5.1. Análisis exploratorio de datos

La tabla 5-1 muestra que la mayoría de las observaciones de los contaminantes atmosféricos presentan valores por debajo del promedio, lo que sugiere una distribución sesgada a la izquierda y una dispersión considerable, de acuerdo con los coeficientes de variación calculados. Entre estos, la concentración horaria de O_3 sobresale por su alta variabilidad, con un coeficiente de variación del 106 %. En la figura A-2 y la figura A-2 de los anexos, se observa la distribución y la presencia de valores atípicos en el extremo superior de los diagramas de caja correspondientes a las concentraciones de O_3 y PM_{10} .

En cuanto a las variables meteorológicas, se evidencia una alta dispersión en las observaciones de velocidad del viento y radiación solar, con coeficientes de variación de 78 % y 140 %, respectivamente. Además, esta última presenta una mayor proporción de registros en cero, atribuible a su dependencia de la hora del día. Por otro lado, la temperatura y la humedad muestran una distribución más homogénea, con coeficientes de variación del 16 % y 24 %, correspondientemente.

Variables	Mínimo	Cuartil 1 (25 %)	Mediana (50 %)	Media	Cuartil 3 (75 %)	Máximo	Desviación estándar	Coefficiente de variación
O_3	0	6.5	18.5	35.9	58.5	207	37.98	1.06
PM_{10}	1	20	31	33.1	43	194	18.53	0.56
Vel Viento	0	1.3	2.2	2.77	3.5	18	2.17	0.78
Temperatura	15.2	20.6	22.6	23.75	27	35.4	3.84	0.16
Humedad	16.9	58.3	79	73.82	89.6	97.6	17.66	0.24
Rad Solar	0	0	21.5	189.6	348.6	1000.3	265.9	1.4

Tabla 5-1.: Medidas de tendencia central y dispersión de las variables del conjunto de datos.

Al respecto de la variable *Lluvia*, el (IDEAM, 2019) reportó que la ocurrencia del fenómeno de “El Niño” provocó un déficit significativo en la precipitación durante los años 2018 y 2019.

Como se observa en la figura 5-1, el 87.4 % de los registros corresponden a condiciones sin lluvia, mientras que solo el 8.7 % presentan lluvia y un 3.9 % corresponden a valores faltantes. En relación con las concentraciones de O_3 , la figura 5-2 muestra que, en ausencia de lluvia, estas son más elevadas y dispersas, con una alta presencia de valores atípicos. En contraste, cuando hay lluvia, las concentraciones tienden a ser más bajas y con menor variabilidad.

Distribución de la Variable Lluvia

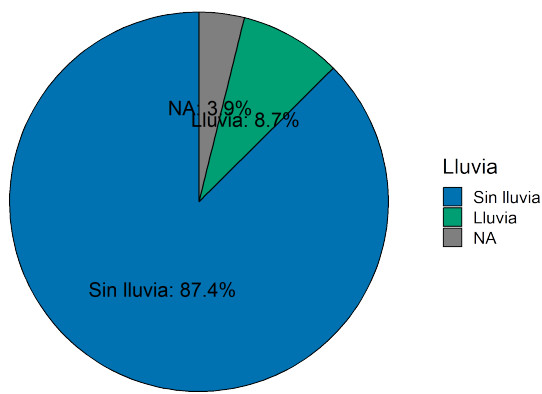


Figura 5-1.: Distribución porcentual de la variable Lluvia en el año 2019.

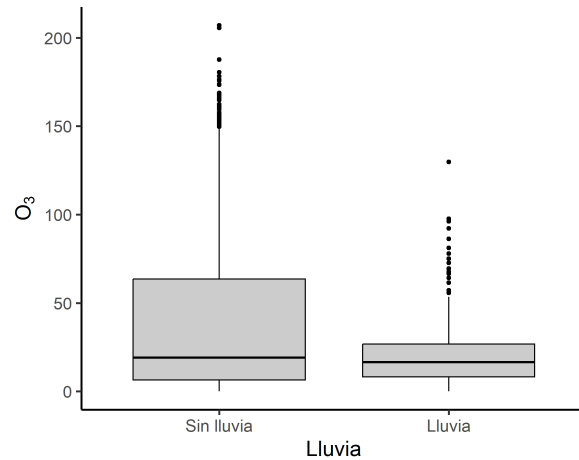
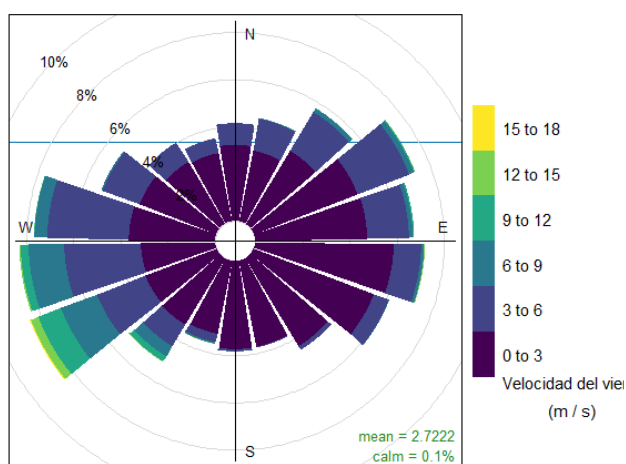
Comportamiento de O_3 según Lluvia

Figura 5-2.: Distribución de O_3 bajo condiciones de lluvia y sin lluvia.

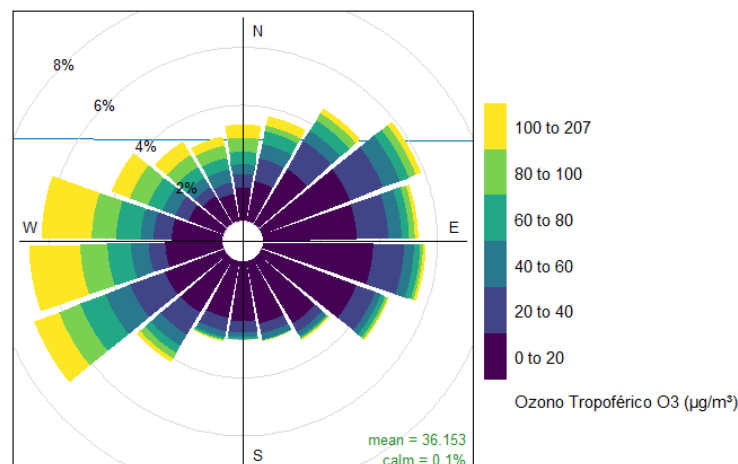
Por otra parte, las correlaciones entre los pares de variables presentadas en la figura A-3 de los anexos, se observa que las concentraciones de O_3 tienen una fuerte relación con la temperatura y la humedad, mientras que con la velocidad del viento y la radiación solar es moderada. Por su parte, entre la temperatura y la humedad se muestra una fuerte correlación negativa de -0.96, lo que indica que estadísticamente, una puede ser explicada en términos de la otra. También se destacan las correlaciones positivas de la radiación solar con la temperatura y la humedad con valores de 0.72 y -0.71 respectivamente.

En el análisis de dirección del viento, se obtuvo una dirección promedio de $254,7^\circ$, apuntando hacia el suroeste respecto a la estación de monitoreo. Mientras, la concentración angular presentó un estadístico de $r = 0.0117$ lo cual señala una notable dispersión en las direcciones del viento, en lugar de una concentración definida hacia un único punto cardinal. Lo anterior se puede observar en la rosa de los vientos de la figura 5-3, donde paralelamente se expone que las velocidades en el rango de 15 a $18 \frac{m}{s}$ son menos frecuentes, mientras que los vientos con velocidades de 3 a $6 \frac{m}{s}$ son más comunes. Además, se observa que la velocidad media es $2.7 \frac{m}{s}$ y que la condición de calma, correspondiente a velocidades del viento cercana a $0 \frac{m}{s}$ ocurren el 10 % del tiempo.

Así mismo, la figura 5-4 muestra la distribución desigual de las concentraciones de O_3 , resaltando la dirección predominante de las concentraciones más altas entre 100 a 207 $\frac{\mu g}{m^3}$ hacia el suroeste, lo que podría indicar posibles fuentes de emisiones o transporte del contaminante desde esa dirección, sin embargo, determinar la procedencia requeriría de un análisis mas detallado. Por otro lado, las concentraciones más bajas se distribuyen de manera uniforme en diferentes direcciones del viento.



Frequency of counts by wind direction (%)



Frequency of counts by wind direction (%)

Figura 5-3.: Rosa de los vientos.

Figura 5-4.: Rosa de contaminación de O_3 .

En la figura 5-5, se analiza detalladamente el comportamiento horario de las concentraciones de O_3 en relación con la dirección y velocidad del viento. Claramente se identifica un patrón diario consistente, donde las concentraciones más bajas de O_3 predominan durante las primeras horas del día, específicamente entre 00:00 y 08:00, con valores inferiores a 20 $\frac{\mu g}{m^3}$ y velocidades del viento relativamente bajas provenientes principalmente del este-noreste.

A partir de las 09:00, se observa un cambio en la dirección del viento y un incremento gradual en las concentraciones de ozono, que alcanzan sus valores máximos entre las 14:00 y las 17:00 horas, particularmente marcando las direcciones que ya se habían identificado como la oeste y suroeste. Durante estas horas, las concentraciones superan con frecuencia los 100 $\frac{\mu g}{m^3}$, coincidiendo con mayor velocidad del viento, lo que sigue indicando la influencia de posibles fuentes precursoras ubicadas hacia estas áreas, intensificados por la radiación solar.

Déspués del máximo alcanzado, al finalizar la tarde las concentraciones comienzan a descender progresivamente, volviendo nuevamente a valores inferiores alrededor de las 22:00 y manteniéndose así durante la noche y la madrugada siguiente. En esta misma hora, la dirección predominante del viento cambia nuevamente hacia el este y noreste, acompañada por una reducción en la velocidad del viento.

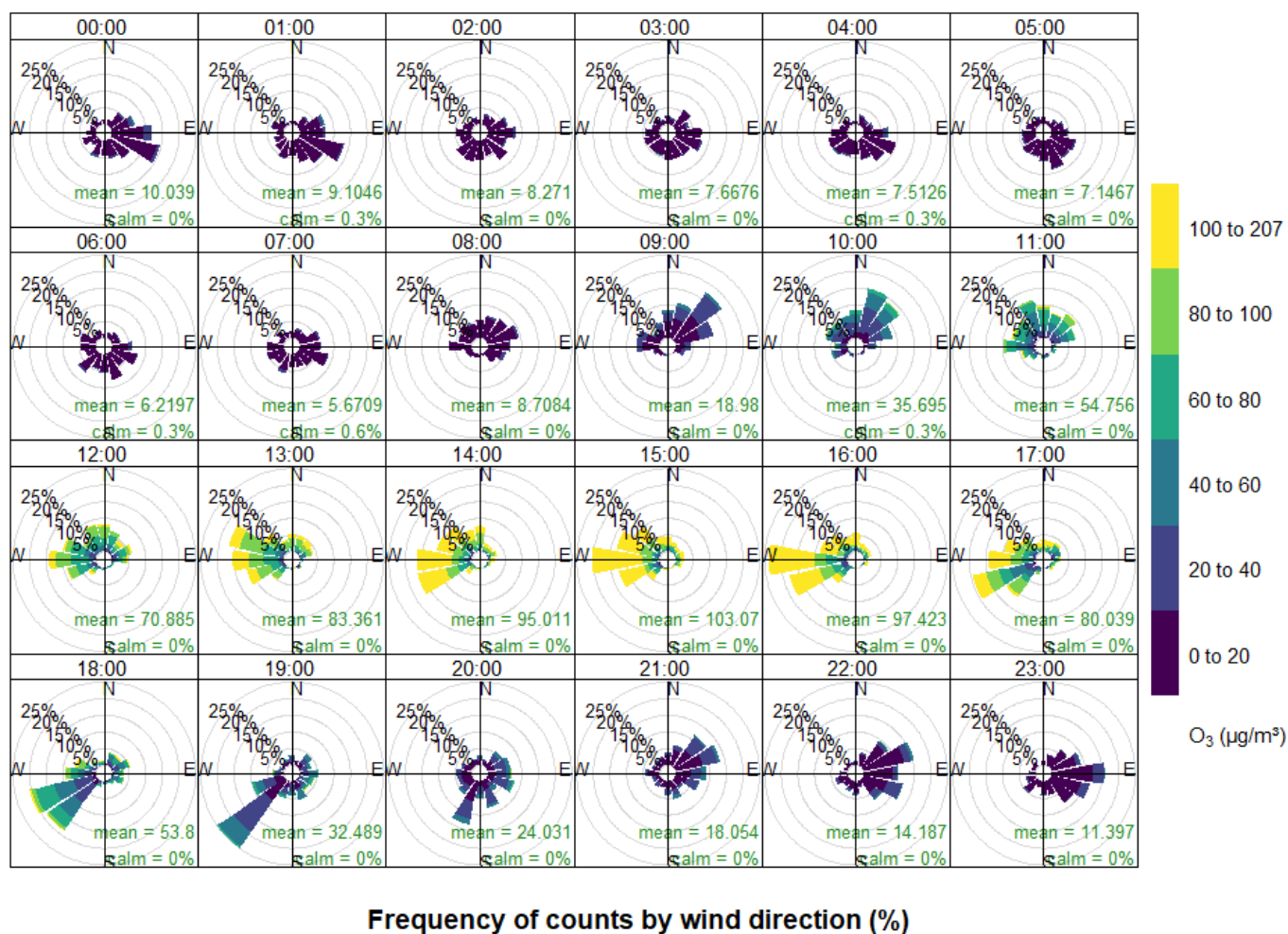


Figura 5-5.: Rosa de contaminación de O_3 cada hora.

Este patrón horario en las concentraciones de O_3 está estrechamente vinculado con la incidencia de la radiación solar sobre su formación, como se ha detallado previamente al analizar la rosa de contaminantes por hora.

En la figura 5-6, se presenta una perspectiva adicional mediante diagramas de cajas, destacando no solo las variaciones medias y máximas por hora, sino también la dispersión y la presencia de valores extremos. Este enfoque complementario permite visualizar claramente la magnitud y frecuencia de episodios pico durante las horas críticas de la tarde, específicamente entre las 12:00 y 17:00 horas, así como la consistencia y estabilidad relativa de los niveles mínimos durante la madrugada.

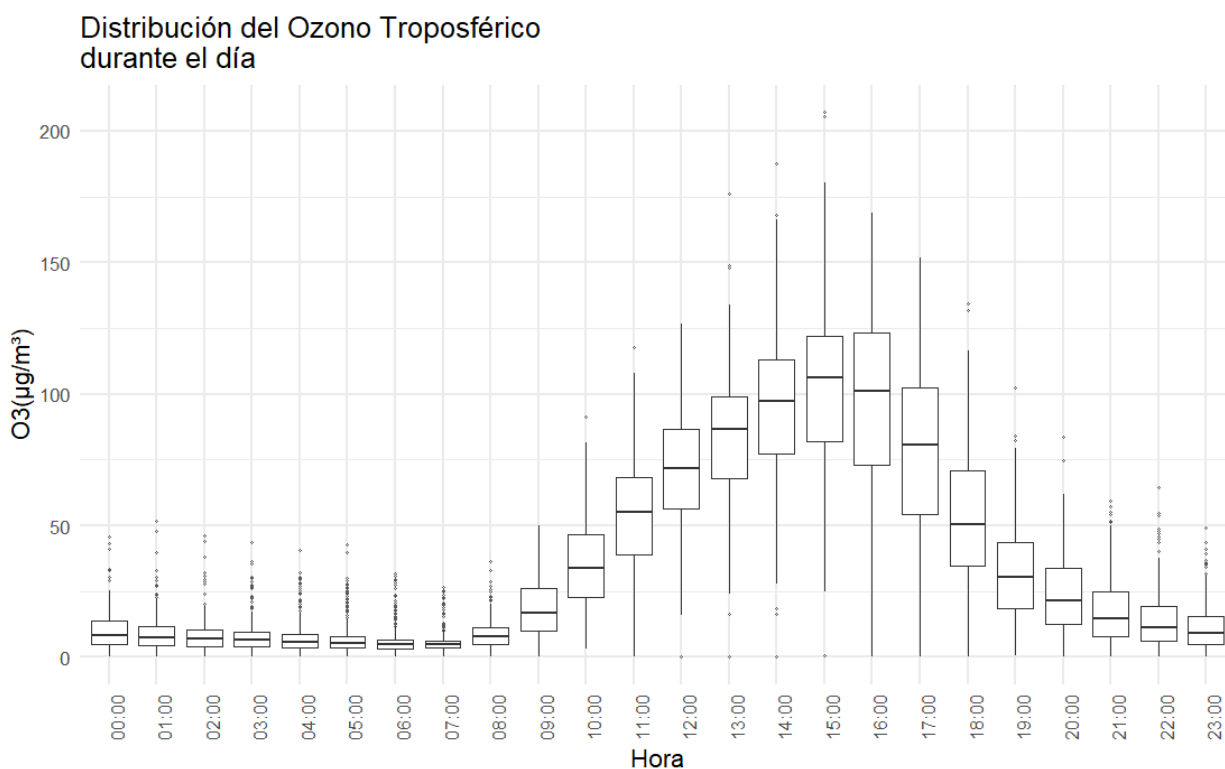


Figura 5-6.: Diagrama de cajas de O_3 por hora.

5.1.1. Análisis datos faltantes

En la figura 5-7, se realiza una comparación entre el límite establecido por la Resolución 2254 del Ministerio de Ambiente y Desarrollo Sostenible (2017) y las concentraciones horarias de $P8O3$ registradas durante el año 2019. Se evidencia que dicho umbral de $100 \frac{\mu g}{m^3}$, representado por la línea discontinua roja, es superado 162 veces a lo largo del año.

Estos picos se presentan con mayor frecuencia durante los meses de enero, abril, junio a octubre y diciembre, lo que podría estar asociado a condiciones atmosféricas favorables para la formación de ozono, como mayor radiación solar y menor precipitación. En contraste, durante los meses de marzo, mayo y noviembre se observa una disminución en la magnitud de las concentraciones, posiblemente influenciada por una mayor cobertura nubosa o eventos de lluvia.

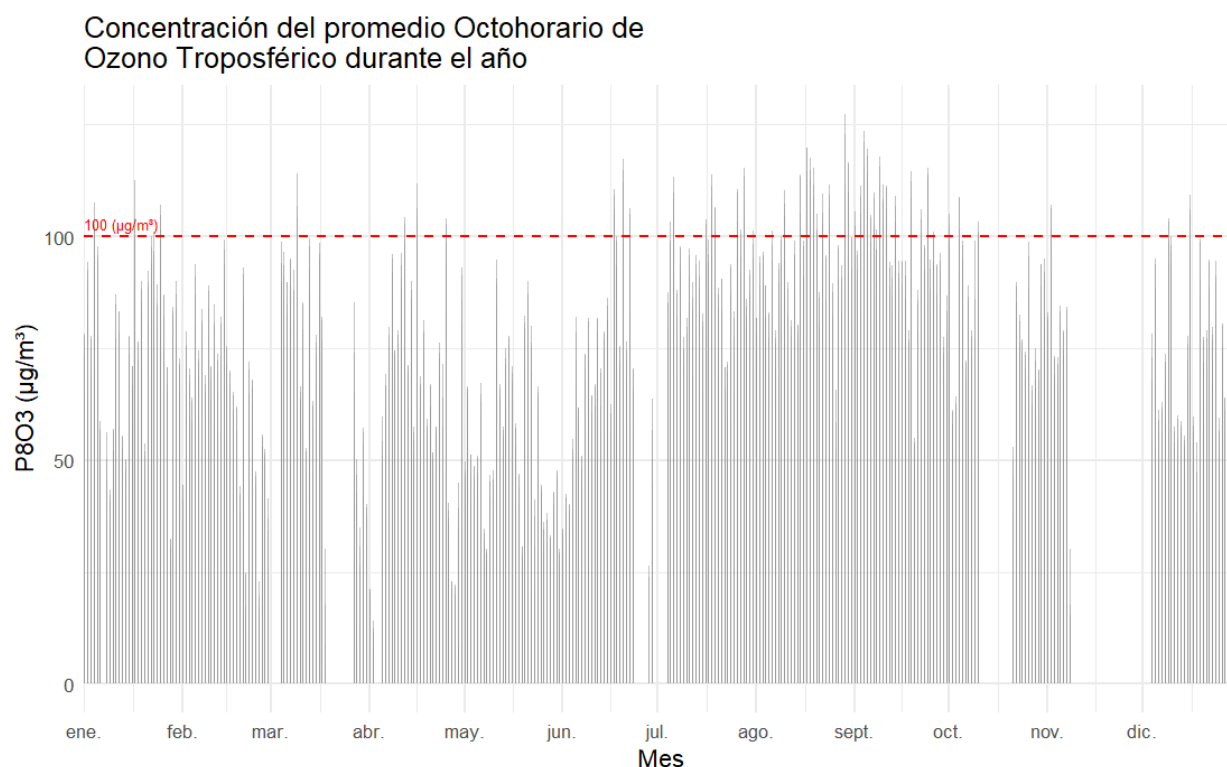


Figura 5-7.: Distribucción del promedio octohorario de Ozono Troposférico durante el año en contraste con las directrices de la OMS.

Adicionalmente, se identifican vacíos de información en algunos tramos del año, particularmente entre marzo y abril, y hacia noviembre y diciembre, lo cual afecta tanto a la variable O_3 como a su promedio móvil octohorario $P8O3$. Esta última, aunque se calcula para representar de forma más adecuada la exposición continua al contaminante, no reduce la cantidad de datos faltantes, por el contrario, tiende a aumentarla. Esto se debe a que el cálculo del promedio móvil de ocho horas solo es válido cuando al menos el 75 % de las observaciones horarias dentro de cada intervalo están disponibles, de modo que la presencia de varias horas consecutivas sin información impide su cálculo. Como resultado, los vacíos se extienden y pueden abarcar incluso días completos, amplificando la pérdida de información respecto a la serie original. En la figura 5-7 puede observarse este efecto, donde, a pesar del incremento en los valores perdidos, el comportamiento general de $P8O3$ no se aleja significativamente al observado en O_3 .

Por último, en la figura 5-8 se presenta el patrón de los datos faltantes para cada una de las variables del conjunto estudiado. La proporción global de valores perdidos es del 7.2 %, aunque estos se distribuyen de forma desigual entre las variables. Particularmente, la variable O_3 concentra el porcentaje más alto de 19.7 %, como se muestra en la tabla 5-2. Este comportamiento podría relacionarse con situaciones habituales en sistemas de monitoreo ambiental,

como cortes en el suministro eléctrico, fallas en los sensores o actividades de mantenimiento en las estaciones de medición. Si bien tales indicios apuntan a una pérdida de datos con cierto grado de estructura, la información disponible no permite establecer con certeza el mecanismo específico involucrado, por lo que debe abordarse cuidadosamente al momento de aplicar técnicas de imputación.

Variables	NA	%NA
O_3	1727	19.7
Vel Viento	343	3.91
Dir Viento	686	7.83
Temperatura	338	3.86
Humedad	338	3.86
Lluvia	338	3.86
Rad Solar	62	0.71

Tabla 5-2.: Cantidad y proporción de datos faltantes en cada una de las variables de la base de datos.

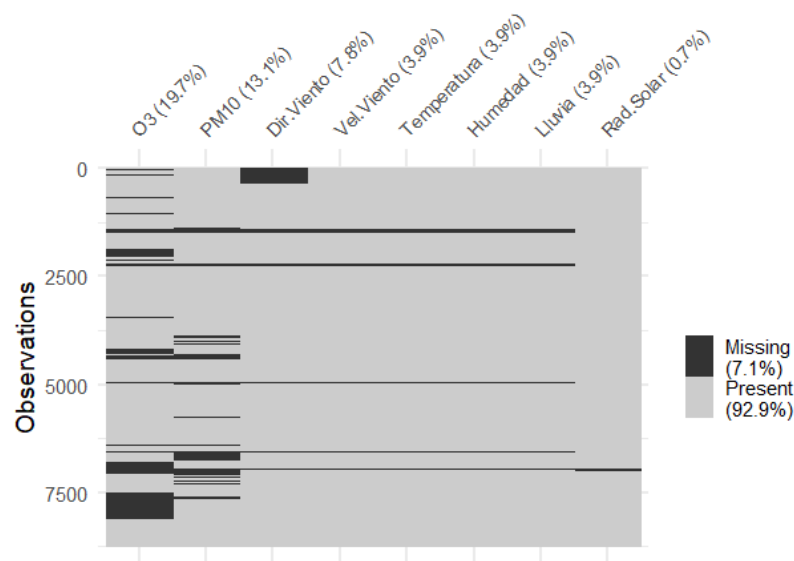


Figura 5-8.: Identificación del patrón de datos faltantes.

5.2. Imputación de datos

En cuanto a los resultados de la imputación de datos faltantes, se generaron cinco conjuntos completos de datos siguiendo el enfoque de Especificación Condicional Completa (FCS), lo que permitió capturar la incertidumbre asociada al proceso. Para cada variable se seleccionó el método más adecuado: Coincidencia de Medias Predictiva (PMM) para las variables O_3 , PM10, *VelViento*, y *Temperatura*. Árboles de Clasificación y Regresión (CART) en variable *RadSolar*, y Regresión Logística (Logreg) en variable *Lluvia*. De esta forma, no se buscó la convergencia hacia un único método, sino una estrategia diferenciada según el tipo y comportamiento de cada variable.

La tabla 5-3 presenta las estadísticas descriptivas de los conjuntos imputados, las cuales pueden compararse con las estadísticas originales de la tabla 5-1. Se observa que las medidas de tendencia central y dispersión permanecen en rangos similares, lo que indica que la

imputación no modificó de manera significativa la estructura de los datos. Para respaldar esta afirmación, se aplicaron pruebas estadísticas formales de comparación entre distribuciones imputadas y observadas, que confirmaron la ausencia de diferencias significativas en la mayoría de los casos. Esta evidencia se presenta en la tabla **A-1** donde refuerza la robustez del procedimiento de imputación implementado.

Variabes	Mínimo	Cuartil 1 (25 %)	Mediana (50 %)	Media	Cuartil 3 (75 %)	Máximo	Desviación estándar	Coficiente de variación
O_3	0	6.44	17.9	35.1	56.2	207	37.5	1.07
PM_{10}	1	20	30,6	32.9	43	194	18.6	0.56
Vel Viento	0	1.3	2.2	2.77	3.5	18	2.17	0.78
Temperatura	15.2	20.6	22.6	23.8	27	35.4	3.83	0.16
Rad Solar	0	0	21.7	190	349	1000	266	1.4

Tabla 5-3.: Medidas de tendencia central y dispersión de las variables después de la imputación.

Este comportamiento puede visualizarse en los gráficos que complementan el análisis estadístico, proporcionando evidencia clave para evaluar la calidad y efectividad del proceso de imputación. La figura **5-9** muestra las distribuciones originales e imputadas para cada una de las variables meteorológicas y contaminantes. En la figura **5-9**, se observa una alta concordancia entre los histogramas de los datos observados y las curvas de densidad correspondientes a los cinco conjuntos imputados, lo cual indica que hay poca variabilidad entre las imputaciones, lo que preserva adecuadamente tanto la forma como la dispersión de las distribuciones originales.

Asimismo, la figura **5-10** ilustra la distribución temporal de los valores imputados para la concentración de O_3 a lo largo del año, considerando un índice horario completo de 8760 observaciones. Se evidencia en el primer conjunto de datos que los valores imputados se integran de manera coherente en la serie temporal original, sin alterar los patrones estacionales ni la dinámica horaria del contaminante. Esta integración natural confirma que la imputación respetó el comportamiento temporal de la variable, manteniendo la continuidad de la serie y evita la introducción de anomalías artificiales. Los cuatro conjuntos imputados restantes se incluyen en los anexos **A-4** y muestran un comportamiento muy similar.

Finalmente, en la figura **A-8** se compara la frecuencia de las observaciones de lluvia entre los datos observados y los conjuntos imputados. Se aprecia que la distribución proporcional de los registros con lluvia y sin lluvia es consistente entre los conjuntos imputados y los datos originales, evidenciando que la imputación respeta la distribución categórica original y evita distorsionar la frecuencia relativa de eventos de lluvia, fundamentales para la modelación posterior.

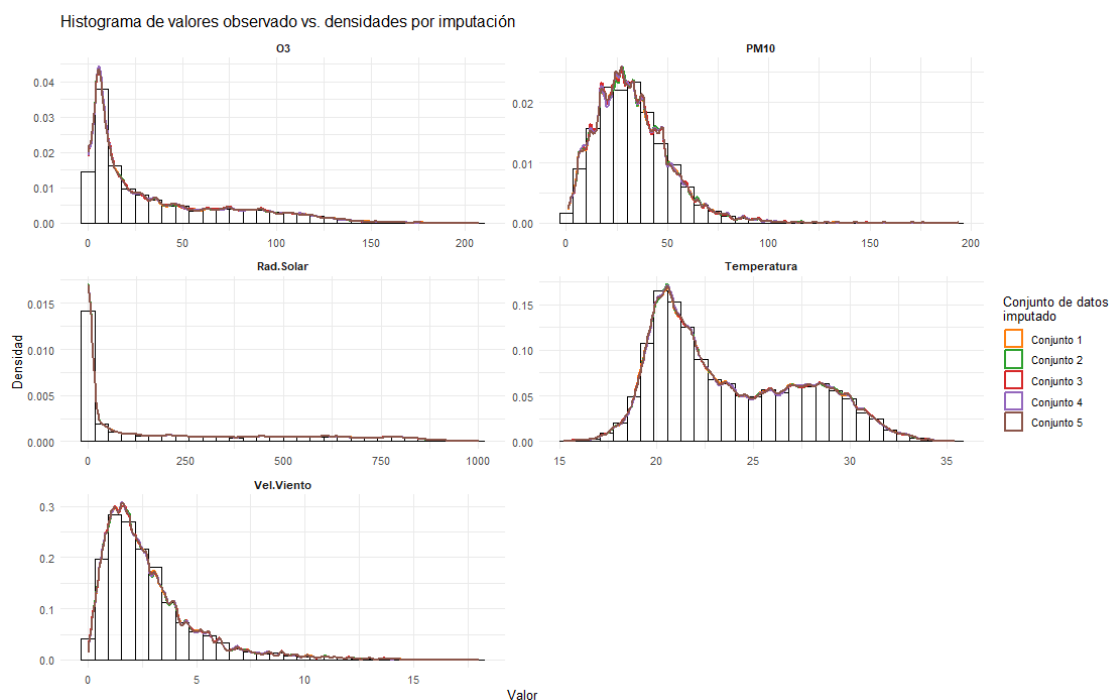


Figura 5-9.: Comparación de las distribuciones originales e imputadas para las variables ambientales y meteorológicas.

Concentración de Ozono Troposférico
con el conjunto 1 de valores imputados

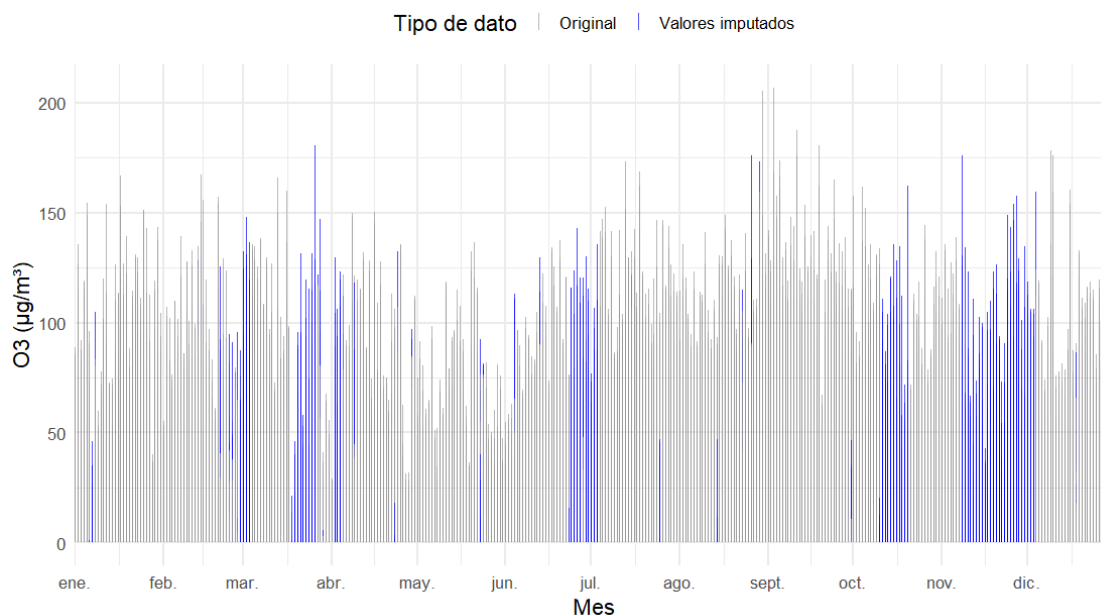


Figura 5-10.: Serie temporal anual de O_3 (8 760 registros horarios) para el primer conjunto imputado: en gris, valores observados que se conservaron y en azul, valores imputados.

5.3. Modelo de regresión cuantílica paramétrica

Al estimar los coeficientes globales mediante los cinco modelos de regresión cuantílica agrupados por cuantil, en la figura 5-11 se evidencia un comportamiento heterogéneo de los efectos que ejercen las covariables sobre las concentraciones del $P8O3$. Cabe especificar que estos coeficientes representan la contribución, en microgramos por metro cúbico ($\frac{\mu g}{m^3}$), que cada covariable aporta al promedio octohorario del Ozono Troposférico. Un resumen de esto mismo es la siguiente tabla:

Variables	τ				
	0.10	0.25	0.50	0.75	0.90
Intercepto	21.09	27.27	33.90	41.50	49.17
PM10_P8	-0.06	-0.02	0.03	0.09	0.10
VelViento_P8	1.98	2.01	2.28	3.52	4.93
Temperatura_P8	3.98	4.96	5.67	5.93	6.22
RadSolar_P8	0.04	0.04	0.04	0.04	0.03
Lluvia	-0.27	0.76	2.18	3.18	3.48

Tabla 5-4.: Coeficientes estimados del modelo de regresión cuantílica por cuantil τ .

En primer lugar, el coeficiente asociado a $PM10$ presenta una evolución notable a lo largo de la distribución del ozono, pasando de valores ligeramente negativos en los cuantiles bajos (entre $\tau = 0,10$ y $0,25$) a valores claramente positivos en los cuantiles medios y altos, alcanzando hasta $0,10 \mu g/m^3$ por unidad de $PM10$ en $\tau \geq 0,90$. Cabe destacar que los intervalos de confianza para esta variable incluyen el cero en los cuantiles inferiores y medios ($\tau = 0,25$ y $0,50$), indicando que su efecto no es estadísticamente significativo en estas zonas de la distribución, mientras que sí lo es en cuantiles superiores, donde condiciones atmosféricas extremas parecen potenciar su influencia.

Respecto a las variables meteorológicas, se inició el análisis por la temperatura, la cual exhibe una relación positiva fuerte y estable en todos los cuantiles analizados, agregando aproximadamente $4,2 \mu g/m^3$ en $\tau = 0,10$ hasta superar $6 \mu g/m^3$ en $\tau = 0,90$ a las concentraciones de $P8O3$. Los intervalos de confianza, siempre alejados del cero, confirman que su efecto es consistentemente significativo. Este hallazgo sugiere claramente que incrementos en la temperatura favorecen sistemáticamente aumentos en las concentraciones octohorarias de ozono, especialmente en sus niveles más elevados.

Asimismo, el coeficiente estimado para la radiación solar también resulta positivo y relativamente estable, oscilando entre $0,033$ y $0,037 \mu g/m^3$ por unidad desde los cuantiles bajos hasta los centrales ($\tau = 0,10$ a $0,75$). No obstante, presenta una ligera disminución en el cuantil más alto ($\tau = 0,90$), coincidiendo con el período en el cual el ozono acumulado alcanza niveles elevados, pero la radiación solar empieza a disminuir rápidamente. En todos los

casos, los intervalos de confianza permanecen por encima del cero, subrayando una influencia estadísticamente significativa a lo largo de toda la distribución.

En cuanto a la Velocidad del Viento, el análisis muestra un coeficiente moderadamente estable alrededor de $2 \mu g/m^3$ hasta $\tau = 0,50$. Sin embargo, en los cuantiles superiores, el valor del coeficiente incrementa notablemente hasta acercarse a $5 \mu g/m^3$ en $\tau = 0,90$. Los intervalos de confianza excluyen consistentemente el cero, indicando que su efecto es significativo a lo largo de toda la distribución analizada. Este patrón sugiere que, en condiciones de elevadas concentraciones de ozono, incrementos en la velocidad del viento podrían intensificar procesos de dispersión y transporte de precursores, reforzando así el efecto neto sobre las concentraciones observadas.

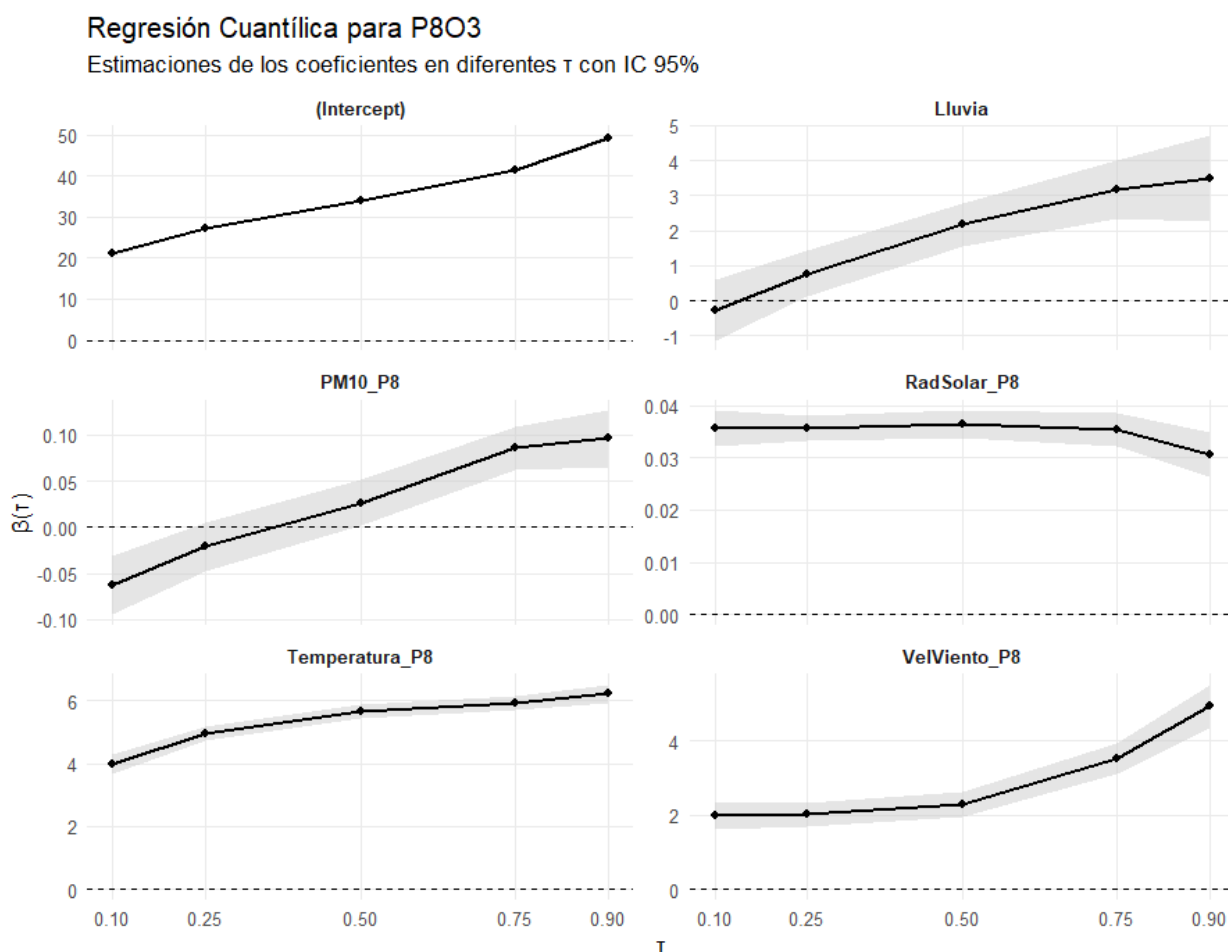


Figura 5-11.: Coeficientes de regresión cuantílica de P8O3 para los cuantiles 0.1, 0.25, 0.5, 0.75, 0.9 e IC 95 % por variable explicativa.

Por otro lado, aunque la precipitación suele asociarse con mecanismos efectivos de remoción

de contaminantes, los resultados del modelo muestran, de forma llamativa, un coeficiente positivo y creciente a lo largo de la distribución del $P8O3$. Este inicia con valores próximos a cero en $\tau = 0,10$, supera $\approx 1, \frac{\mu g}{m^3}$ en $\tau = 0,25$ y alcanza más de $3,5, \frac{\mu g}{m^3}$ en $\tau = 0,90$. Una interpretación plausible es que esta variable captura, además del evento de lluvia en sí, los efectos indirectos de la reducción simultánea de la radiación solar y la temperatura durante episodios húmedos. Estas disminuciones cambian la forma en que la luz afecta la descomposición de los precursores del O_3 y cómo esas sustancias reaccionan entre sí.

Sin embargo, al cesar la precipitación, la atmósfera podría quedar enriquecida temporalmente con precursores acumulados y humedad residual, favoreciendo procesos posteriores de formación de O_3 . En consecuencia, el coeficiente positivo probablemente refleja un efecto compensado, donde la lluvia funciona como indicador de condiciones meteorológicas complejas, más que como agente directo de remoción. Este resultado enfatiza la necesidad de interpretarlo cuidadosamente en conjunto con las demás covariables, pues su contribución podría estar confundida con las variaciones que estas variables experimentan en presencia de lluvia.

5.3.1. Métricas y test de Wald

Revisando la calidad del ajuste de los modelos propuestos para los diferentes cuantiles se observa en la tabla 5-5 que los valores de Pseudo R^2 son elevados en los cuantiles 0.5 a 0.90 con una capacidad explicativa de 0.60, 0.64 y 0.62, mientras que en los cuantiles inferiores como 0.1 y 0.25, los niveles son menores correspondientes a 0.38 y 0.5, lo que indica que el modelo describe de manera adecuada la variación de la media a los extremos superiores de la distribución.

Métrica	τ				
	0.10	0.25	0.50	0.75	0.90
Pseudo R^2	0.3739	0.5027	0.6039	0.6361	0.6175
MAE	15.2091	10.9030	9.1201	11.1832	16.3582
Test Wald	3917.403	13717.064	25969.213	29167.116	20207.158

Tabla 5-5.: Coeficientes estimados del modelo de regresión cuantílica por cuantil τ .

Por otra parte, el Error Absoluto Medio (MAE) varía notablemente en los diferentes cuantiles, iniciando en el cuantil 0.1 con un MAE de 15.21, reduciendo hasta el cuantil 0.5 con un MAE de 9.12 y aumentando nuevamente en el cuantil 0.75 a un MAE de 11.18, llegando en el cuantil 0.90 con un MAE de 16.36, donde indica que el modelo es más preciso en el centro de la distribución que en los extremos, lo cual es esperable en el comportamiento de las concentraciones de $P8O3$.

Finalmente, de acuerdo con los resultados del test de Wald de la Tabla **A-2**, las variables velocidad del viento, temperatura y radiación solar son estadísticamente significativas en todos los cuantiles analizados con $p < 0,001$, e intervalos de confianza que no incluyen el cero, lo que evidencia un efecto robusto sobre $P8O3$ a lo largo de toda la distribución. En cuanto a $PM10_P8$, su significancia no es uniforme: resulta significativo en los extremos (0.10 con $p = 0,01$ y 0.75–0.90 con $p < 0,001$), pero pierde significancia en la parte central (0.25 con $p = 0,2246$ y 0.50 con $p = 0,0898$), sugiriendo una asociación más marcada en cuantiles bajos y altos. Para *Lluvia*, el patrón es inverso: no es significativa en 0.10 ($p = 0,5714$) y es marginal en 0.25 ($p = 0,0638$), pero a partir del cuantil 0.50 se vuelve claramente significativa y mantiene dicha significancia en 0.75 y 0.90 ($p < 0,001$), indicando que su efecto se manifiesta principalmente desde la mediana hacia los niveles superiores de $P8O3$.

5.4. Modelo de regresión cuantílica no paramétrica de $P8O3$ durante el día

En la figura 5-12 se presentan las curvas estimadas mediante regresión cuantílica no paramétrica para las concentraciones octohorarias del $P8O3$, en los cuantiles 0.10, 0.25, 0.50, 0.75 y 0.90. Estas curvas están diferenciadas según la presencia de la precipitación, el panel izquierdo corresponde a registros con lluvia, mientras que el derecho representa los registros sin precipitación. La representación permite identificar claramente la dinámica diaria del $P8O3$, mostrando no solo su comportamiento promedio, sino también la dispersión y la variabilidad asociadas a valores extremos.

Durante la madrugada y las primeras horas de la mañana, entre las 0 y 8 horas, todas las curvas registran valores mínimos, con concentraciones entre 15 y 20 $\frac{\mu g}{m^3}$ en ambas condiciones meteorológicas. Posteriormente, a medida que transcurre la mañana, se observa un aumento sostenido en los niveles de ozono en todos los cuantiles, coincidiendo con la intensificación de la radiación solar y el incremento de la temperatura, condiciones que favorecen las reacciones fotoquímicas que generan el O_3 . El máximo diario se alcanza entre las 16 y 19 horas, siendo considerablemente más alto en ausencia de lluvia, aproximadamente 120 $\frac{\mu g}{m^3}$ en el cuantil 0.90 en comparación con los días lluviosos que no supera los 90 $\frac{\mu g}{m^3}$ para el mismo cuantil. A partir de este máximo, las concentraciones disminuyen gradualmente hasta alcanzar niveles menores hacia la medianoche, aunque mantienen diferencias marcadas entre cuantiles que reflejan una variabilidad residual producto de condiciones atmosféricas locales.

Se muestra de manera clara que la presencia de lluvia conduce a una reducción general en las concentraciones de $P8O3$, tanto en los valores más bajos como en los máximos diarios. Además, la amplitud entre los cuantiles extremos 0.10 y 0.90 aumenta moderadamente desde aproximadamente 30 $\frac{\mu g}{m^3}$ durante la madrugada hasta cerca de 70 $\frac{\mu g}{m^3}$ al atardecer. Sin

embargo, en registros secos esta dispersión, aunque similar, ocurre sobre niveles absolutos considerablemente más altos, resaltando el impacto significativo de la ausencia de precipitación en las concentraciones máximas de $P8O3$ durante las horas de mayor actividad humana.

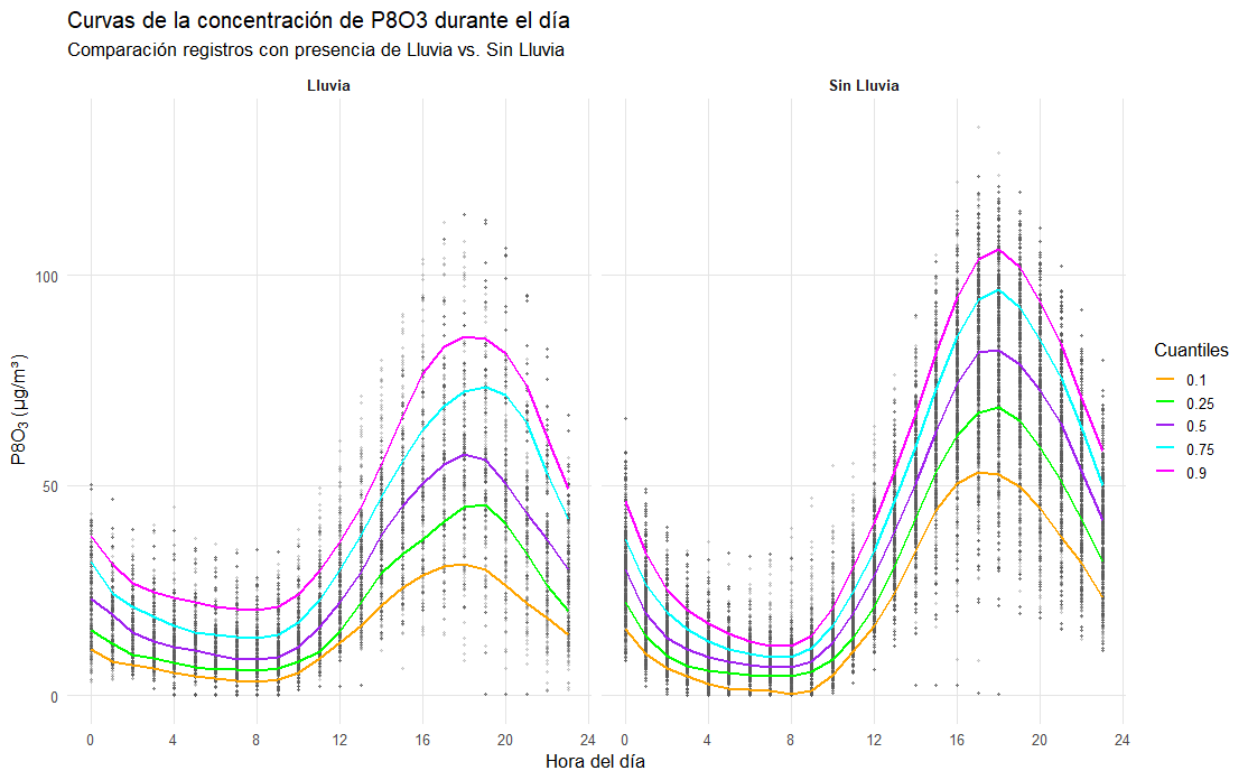


Figura 5-12.: Curvas diarias de $P8O3$ para cuantiles entre 0.1 y 0.9, comparando la presencia lluvia vs. sin lluvia.

Análisis de las métricas de ajuste

En la tabla 5-6 se presenta las métricas con las cuales se realizó la evaluación del modelo de regresión cuantílica no paramétrica mediante el Pseudo- R^2 y el Error Absoluto Medio (MAE), calculadas para los cuantiles $\tau = 0,10, 0,25, 0,50, 0,75, 0,90$, diferenciando entre las condiciones de “Lluvia” y “Sin lluvia”.

En cuanto al Pseudo- R^2 , se observa que los valores son sistemáticamente mayores en el grupo “Sin lluvia” para todos los cuantiles, lo que indica una mejor capacidad explicativa del modelo cuando no hay precipitación. Específicamente, el modelo alcanza un máximo de $R^2 = 0,659$ en el cuantil $\tau = 0,75$, lo que sugiere que las covariables utilizadas capturan con mayor precisión las variaciones del ozono troposférico en los niveles superiores de la distribución bajo condiciones secas. En contraste, en el grupo “Lluvia”, el Pseudo- R^2 varía

entre 0.237 y 0.514, mostrando menor desempeño, especialmente en los cuantiles inferiores ($\tau = 0,10$ y $\tau = 0,25$), donde el ajuste es más limitado.

Respecto al **MAE**, se encuentra que los errores absolutos son generalmente *menores para los cuantiles centrales* ($\tau = 0,50$), tanto en los días con lluvia (MAE = 8.92) como sin lluvia (MAE = 9.31), lo que concuerda con la mayor estabilidad del ozono en la mediana. En cambio, los *extremos de la distribución* ($\tau = 0,10$ y $\tau = 0,90$) presentan errores mayores, alcanzando un máximo de 16.66 en el cuantil 0.90 para días lluviosos y 16.46 en el cuantil 0.10 para días secos. Esto es consistente con la mayor variabilidad del ozono en las colas de la distribución, donde el modelo tiene mayores dificultades para estimar con precisión.

En conjunto, estas métricas reflejan que el modelo tiene un desempeño aceptable, especialmente en los cuantiles centrales y en condiciones sin lluvia, donde se logra mayor poder explicativo y menores errores. No obstante, se evidencia una pérdida de precisión en los cuantiles extremos, lo cual es esperable en contextos ambientales debido a la mayor dispersión y presencia de valores atípicos en las colas de la distribución.

Grupo	τ	Pseudo R^2	MAE
Lluvia	0.10	0.237	14.16
	0.25	0.331	10.66
	0.50	0.425	8.92
	0.75	0.485	11.39
	0.90	0.514	16.66
Sin lluvia	0.10	0.398	16.46
	0.25	0.528	11.31
	0.50	0.632	9.31
	0.75	0.659	11.25
	0.90	0.641	15.73

Tabla 5-6.: Pseudo- R^2 y MAE por cuantil del modelo de regresión cuantílica no paramétrica con B-splines, diferenciados por condición de lluvia.

6. Conclusiones y recomendaciones

6.1. Conclusiones

La imputación múltiple mediante la técnica de Especificación Totalmente Condicional implementada con el paquete `mice`, permitió reconstruir adecuadamente los valores faltantes, preservando la estructura, tendencias y relaciones entre la variable respuesta O_3 y las covariables meteorológicas. Esta estrategia aseguró la validez de las inferencias posteriores, en particular en los modelos de regresión cuantílica, al conservar tanto la variabilidad natural de los datos como los patrones de asociación subyacentes.

A partir de estos datos, los modelos de regresión cuantílica paramétricos revelaron que las variables temperatura, radiación solar y velocidad del viento ejercen efectos positivos y estadísticamente significativos en toda la distribución de $P8O3$. En contraste, el material particulado ($PM10$) y la lluvia mostraron efectos más variables: $PM10$ tuvo un comportamiento no significativo en los cuantiles bajos y medios, pero se volvió positivo y significativo en los cuantiles superiores; mientras que la lluvia, contra lo esperado, presentó efectos positivos crecientes, con mayor intensidad en los cuantiles más altos. Este comportamiento se interpretó como un posible reflejo de condiciones meteorológicas complejas, donde la lluvia podría estar correlacionada con descensos de radiación y temperatura, afectando indirectamente la formación de O_3 .

Los hallazgos anteriores concuerdan con los estudios previos tratados en los antecedentes, donde en ciudades estadounidenses destacan a la temperatura y a la velocidad del viento como impulsores clave en los cuantiles altos de ozono (Fix et al., 2018), mientras que investigaciones en China confirman la influencia conjunta de temperatura y radiación, con efectos de $PM10$ y lluvia que varían según el percentil (Zhang et al., 2022). Además, un antecedente local para Cali registró los mayores niveles de $P8O3$ al finalizar la tarde e iniciar la noche con mayor sensibilidad en la zona sur, corroborando nuestro patrón horario y la zona de mayor impacto (Olaya et al., 2014).

Complementariamente, la implementación de modelos no paramétricos con B-splines penalizados ofreció una representación más flexible de las relaciones entre las covariables y $P8O3$, capturando adecuadamente la no linealidad de los efectos y evitando el cruce de cuantiles, lo cual es fundamental en contextos con heterocedasticidad y distribución asimétrica de los

contaminantes. Esta formulación permitió identificar interacciones complejas entre variables y mejorar la calidad del ajuste a lo largo de la distribución del $P8O3$.

Finalmente, entre los patrones más relevantes identificados, se destaca el pico diario en las concentraciones de $P8O3$ entre las 16:00 y las 19:00 horas, momento en el que la diferencia entre cuantiles extremos alcanzó hasta $70 \mu g/m^3$. Este hallazgo señala la existencia de episodios críticos de contaminación en franjas horarias específicas, lo que resulta clave para el diseño de políticas públicas de mitigación para el perímetro urbano de la ciudad de Santiago de Cali, sobre todo en la zona sur donde se encuentra la estación de monitoreo que se usó en este análisis.

6.2. Recomendaciones

Para futuras investigaciones y aplicaciones en datos ambientales, se recomienda explorar aún más los modelos aditivos generalizados (GAM, por sus siglas en inglés), los cuales permiten incorporar términos suavizados e interacciones no lineales entre covariables complementarias como Dióxido de nitrógeno (NO_2), Monóxido de carbono (CO), Cobertura nubosa, o variables que incluyan la actividad vehicular de la ciudad, entre muchas más que puedan enriquecer el análisis.

La literatura reciente muestra que los modelos aditivos generalizados (GAM) son herramientas robustas y flexibles para detallar la dinámica del Ozono Troposférico, Shin and Lee (2024) demuestran que los GAM permiten separar y cuantificar el efecto meteorológico en las tendencias urbanas de O_3 a largo plazo, mientras que Kawichai et al. (2025) evidencian cómo sus suavizadores e interacciones capturan con precisión las variaciones estacionales impulsadas por la temperatura, radiación y NO_2 .

Asimismo, resulta conveniente ampliar la cobertura espacial y temporal de los datos, integrando registros provenientes de estaciones adicionales y extendiendo el período de análisis a varios años. Esta estrategia no solo mejoraría la representatividad y estabilidad de los modelos, sino que también permitiría aplicar técnicas más robustas de imputación. Por ejemplo, la información de una estación puede ayudar a imputar datos faltantes en otra cercana, reforzando la validez de las estimaciones en presencia de datos faltantes.

De igual manera, es recomendable que en futuros trabajos se incorporen diagnósticos de multicolinealidad en los modelos de regresión cuantílica, complementados con la exploración de posibles soluciones que ya han sido probadas en la literatura reciente. En particular, la regresión ridge, al introducir un parámetro de penalización, ha mostrado ser eficaz para estabilizar los coeficientes y reducir su varianza en presencia de predictores altamente corre-

lacionados, como se observó en aplicaciones médicas de regresión cuantílica Bager (2018).

De forma complementaria, la regresión cuantílica por componentes principales (QPCR) ha demostrado que la síntesis de variables correlacionadas en factores ortogonales elimina la colinealidad sin sacrificar información relevante, lo que mejora la estabilidad y robustez de los modelos Davino et al. (2022). En un contexto ambiental, donde las variables meteorológicas y contaminantes suelen estar fuertemente interrelacionadas, la adopción de estas estrategias podría contribuir significativamente a la fiabilidad y la interpretabilidad de los resultados.

Finalmente, en lo referente a los modelos no paramétricos, resulta pertinente recomendar la realización de un análisis de sensibilidad que evalúe cómo aspectos técnicos de la estimación mediante B-splines penalizados como el tipo de base, el grado, el número de nodos y el valor del parámetro de suavizado (λ) pueden influir en los resultados. Aunque estos métodos son robustos y el parámetro de suavización se selecciona de forma óptima, su exploración sistemática podría aportar mayor evidencia sobre la estabilidad y consistencia del modelo.

A. Anexos

A.1. Resultados

A.1.1. Análisis exploratorio de datos

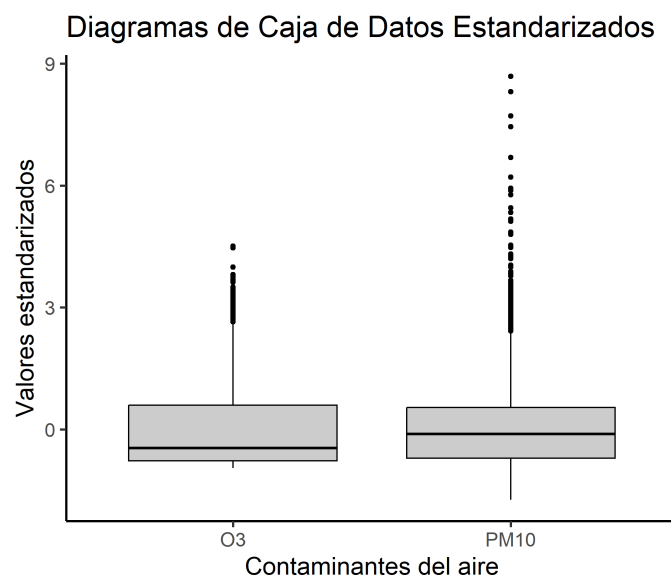


Figura A-1.: Diagramas de caja de los contaminantes del aire estandarizados.

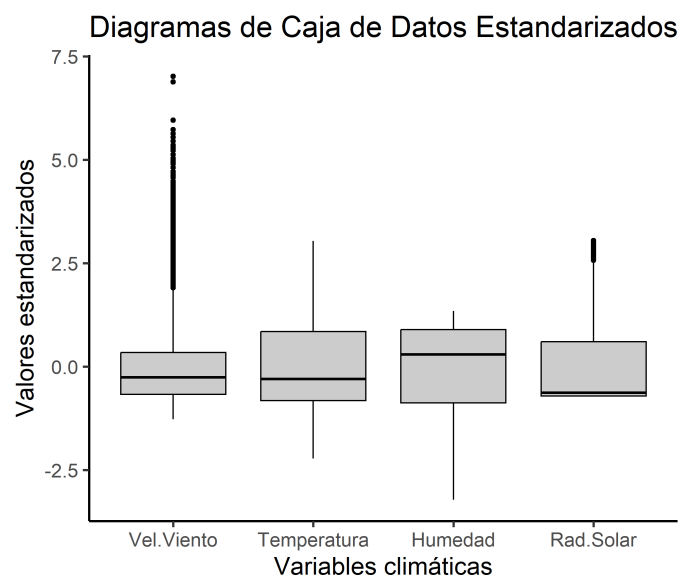


Figura A-2.: Diagramas de caja de las variables meteorológicas estandarizadas.

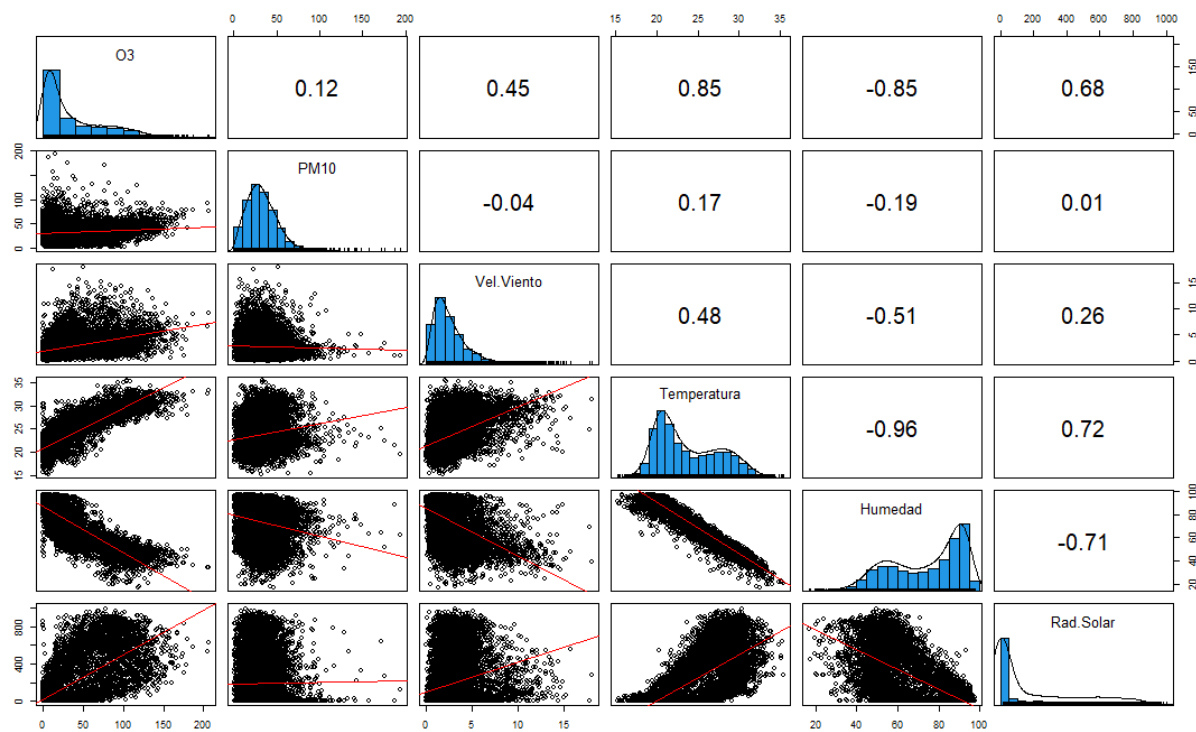


Figura A-3.: Diagrama de correlación de las variables.

A.1.2. Imputación de datos

Prueba	Variable	Conjunto 1	Conjunto 2	Conjunto 3	Conjunto 4	Conjunto 5
1*AD	O ₃	28065 (0.423)	40802 (0.222)	22699 (0.547)	31799 (0.361)	26075 (0.465)
CvM	PM10	0.058 (1)	0.074 (0.997)	0.26 (0.782)	0.066 (1)	0.132 (0.967)
	Vel.Viento	0.013 (1)	0.002 (1)	0.0148 (1)	0.013 (1)	0.012 (1)
	Temperatura	0.009 (1)	0.008 (1)	0.004 (1)	0.022 (1)	0.007 (1)
	Rad.Solar	0.003 (1)	0.001 (1)	0.008 (1)	0.010 (1)	0.011 (1)
	Lluvia	0.004 (1)	0.001 (1)	0.004 (1)	0.022 (1)	0.004 (1)
1* χ^2	Lluvia	0.004 (0.950)	0.029 (0.863)	0.004 (0.950)	0.166 (0.683)	0.001 (0.971)

Tabla A-1.: Comparación de distribuciones mediante pruebas de bondad de ajuste en los cinco conjuntos de datos imputados.

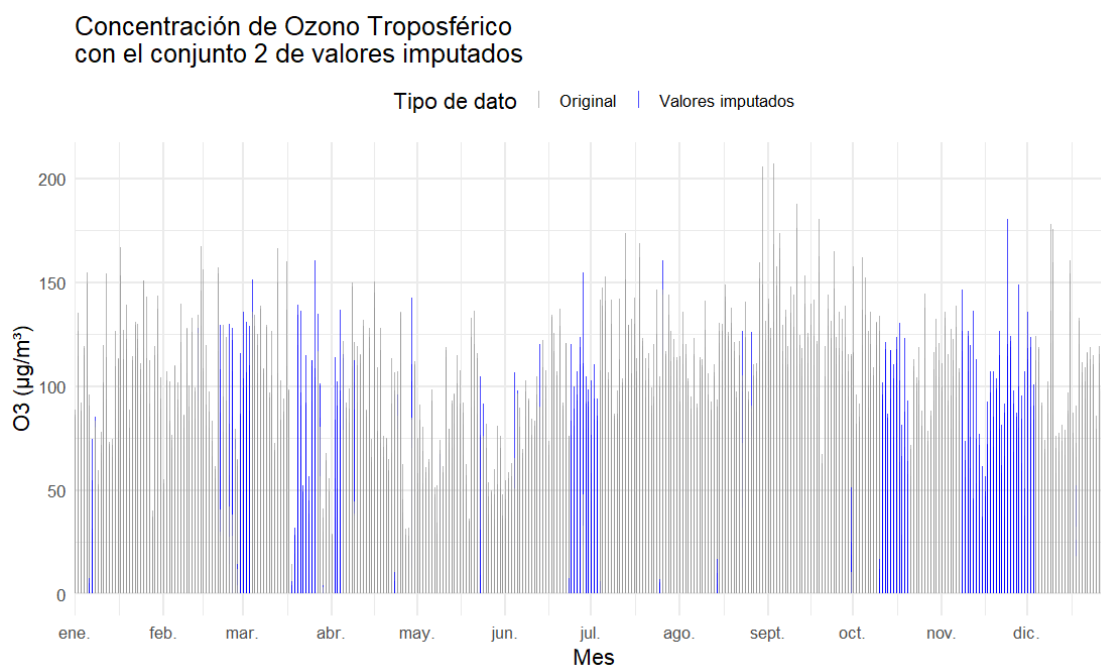


Figura A-4.: Serie temporal anual de O₃ (8 760 registros horarios) para el segundo conjunto imputado.

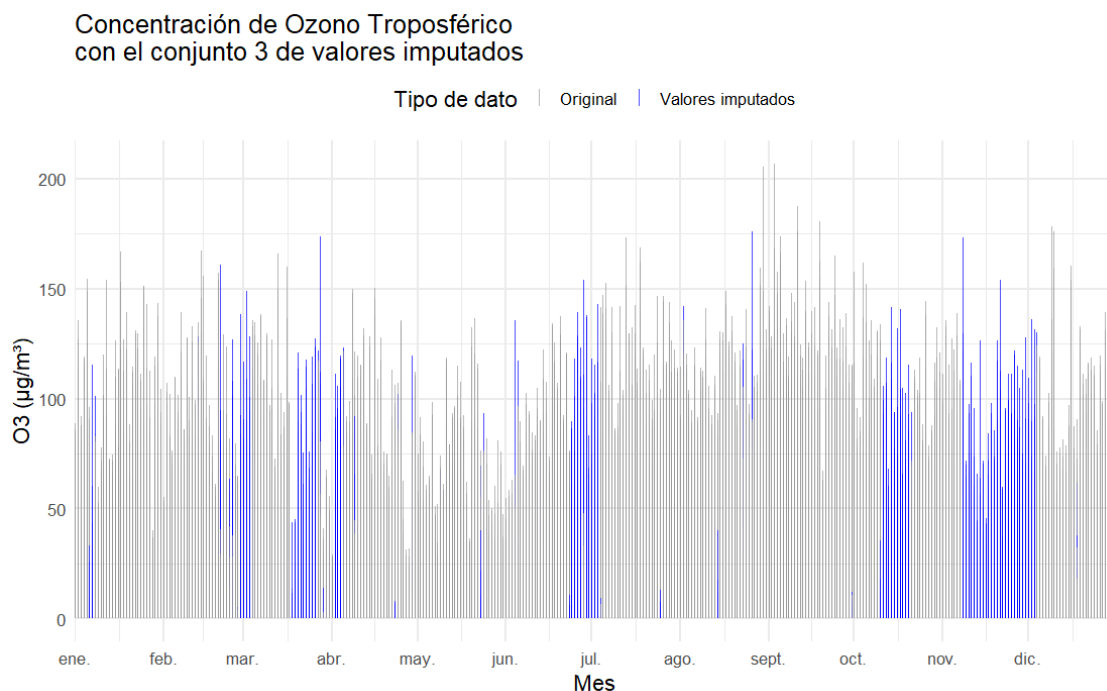


Figura A-5.: Serie temporal anual de O_3 (8 760 registros horarios) para el tercer conjunto imputado.

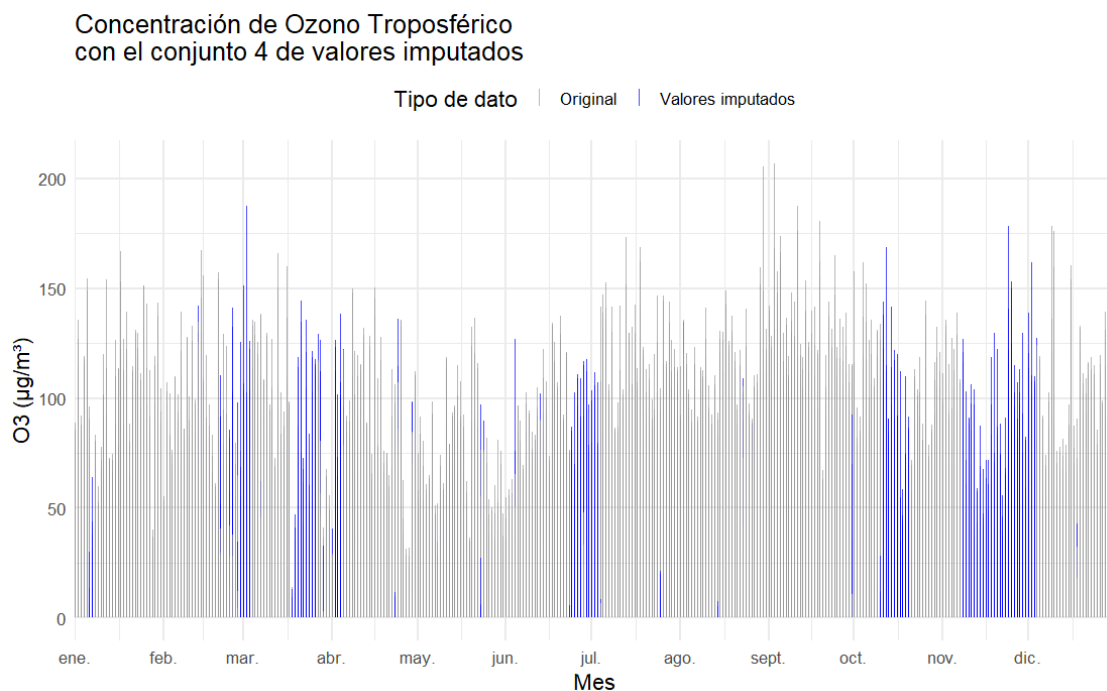


Figura A-6.: Serie temporal anual de O_3 (8 760 registros horarios) para el cuarto conjunto imputado.

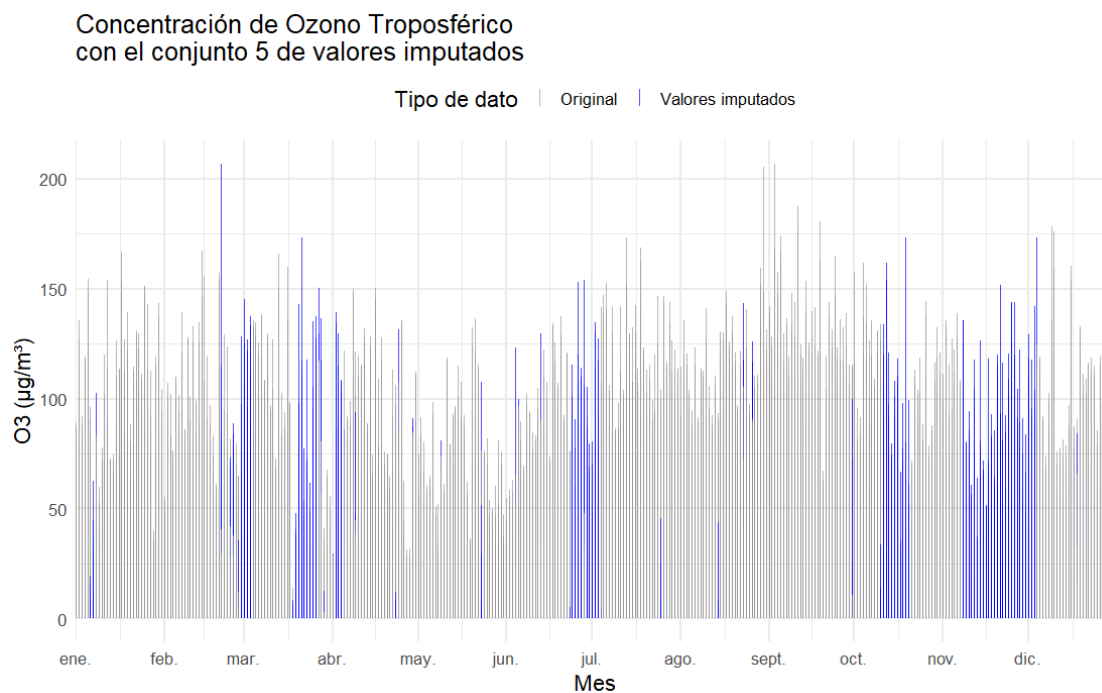


Figura A-7.: Serie temporal anual de O_3 (8 760 registros horarios) para el quinto conjunto imputado.

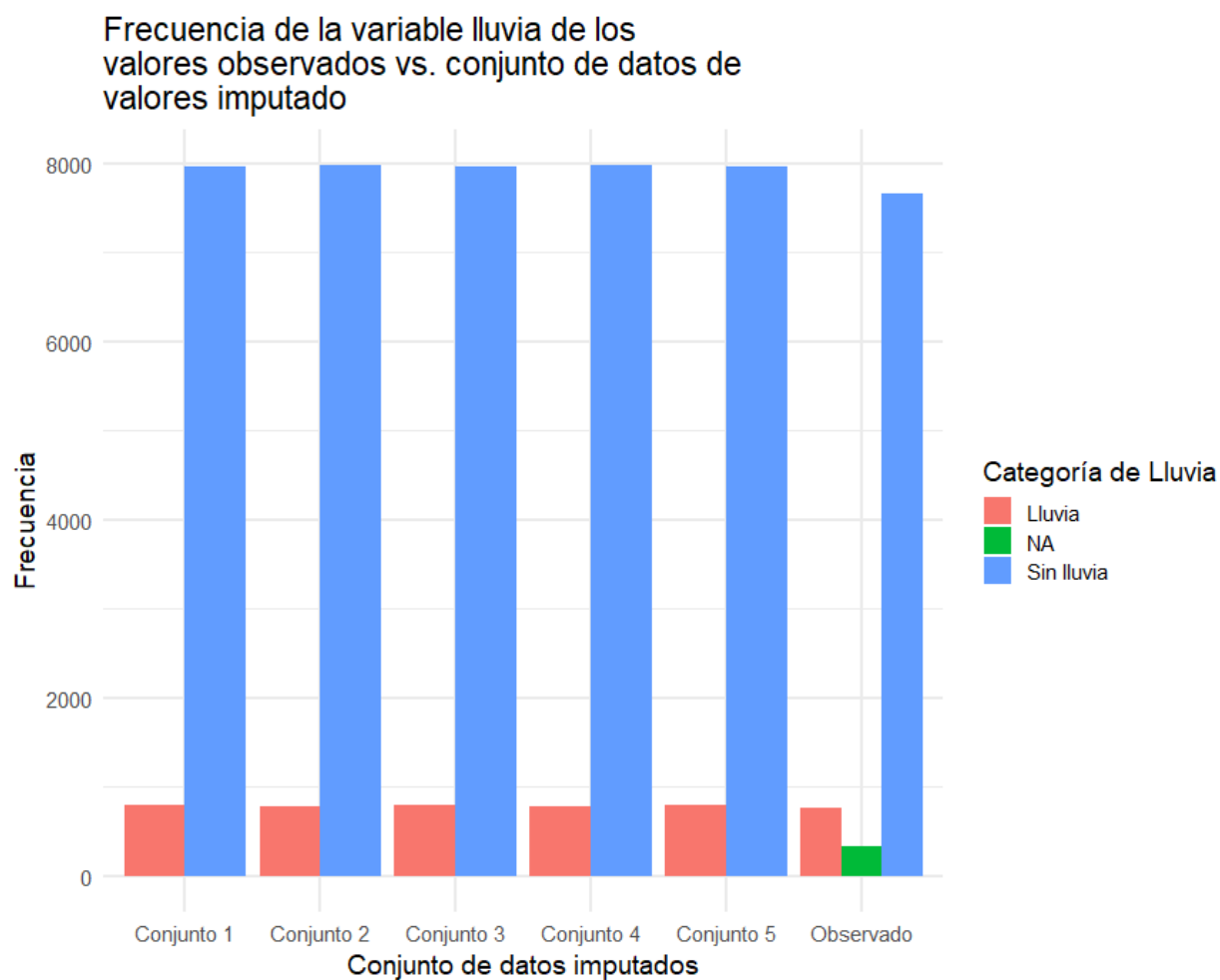


Figura A-8.: Comparativa de frecuencias observadas e imputadas para la variable categórica Lluvia.

A.1.3. Métricas del modelo de regresión cuantílica paramétrico

Resultados test de Wald para cada covariable

τ	Variable	Coefficiente	IC-LI	IC-LS	Error estd.	Wald	Valor-p
0.10	Intercepto	21.09	20.54	21.63	0.28	78.06	0
	<i>PM10_P8</i>	-0.06	-0.09	-0.03	0.02	-3.92	0.01
	<i>VelViento_P8</i>	1.98	1.60	2.37	0.20	10.14	0
	<i>Temperatura_P8</i>	3.98	3.65	4.30	0.17	24.12	0
	<i>RadSolar_P8</i>	0.04	0.03	0.04	0.00	20.42	0
	Lluvia	-0.27	-1.10	-0.56	0.42	-0.60	0.5714
0.25	Intercepto	27.27	26.93	27.61	0.18	158.67	0
	<i>PM10_P8</i>	-0.02	-0.05	0.01	0.01	-1.63	0.2246
	<i>VelViento_P8</i>	2.01	1.69	2.33	0.16	12.30	0
	<i>Temperatura_P8</i>	4.96	4.74	5.18	0.11	44.74	0
	<i>RadSolar_P8</i>	0.4	0.03	0.04	0.00	28.42	0
	Lluvia	0.76	0.12	1.41	0.33	2.28	0.0638
0.50	Intercepto	33.90	33.56	34.23	0.17	201.05	0
	<i>PM10_P8</i>	0.03	0.00	0.05	0.01	2.12	0.0880
	<i>Vel Viento_{P8}</i>	2.28	1.94	2.63	0.18	12.99	0
	<i>Temperatura_P8</i>	5.67	5.45	5.89	0.11	48.83	0
	<i>RadSolar_P8</i>	0.04	0.03	0.04	0.00	27.64	0
	Lluvia	2.18	1.54	2.81	0.32	6.89	0
0.75	Intercepto	41.50	41.08	41.92	0.22	204.14	0
	<i>PM10_P8</i>	0.09	0.06	0.11	0.01	7.36	0
	<i>VelViento_P8</i>	3.52	3.09	3.94	0.22	16.71	0
	<i>Temperatura_P8</i>	5.93	5.70	6.17	0.12	50.37	0
	<i>RadSolar_P8</i>	0.04	0.03	0.04	0.00	21.	0
	Lluvia	3.18	2.33	4.02	0.43	7.59	0
0.90	Intercepto	49.17	48.56	49.78	0.31	160.30	0
	<i>PM10_P8</i>	0.10	0.07	0.13	0.02	6.23	0
	<i>VelViento_P8</i>	4.93	4.33	5.53	0.31	16.51	0
	<i>Temperatura_P8</i>	6.22	5.92	6.51	0.15	41.30	0
	<i>RadSolar_P8</i>	0.03	0.03	0.03	0.00	14.01	0
	Lluvia	3.48	2.25	4.72	0.63	5.57	0

Tabla A-2.: Estimaciones por cada cuantil y variable de la regresión cuantílica.

Bibliografía

- Alcaldía de Santiago de Cali (2014). Acuerdo 0373 del 2014, por el cual se adopta el plan de ordenamiento territorial - pot de santiago de cali. <https://saul.cali.gov.co/pimu/pot/otros/Acuerdo%200373%20de%202014.pdf>. [PDF].
- Anderson, T. W. and Darling, D. A. (1954). A test of goodness of fit. *Journal of the American statistical association*, 49(268):765–769.
- Bager, A. (2018). Ridge parameter in quantile regression models: An application in biostatistics. *International Journal of Statistics and Applications*, 8(2):72–78.
- Breiman, L. (2017). *Classification and Regression Trees*. Routledge.
- DAGMA (2022). Datos de calidad del aire en la estación pance de santiago de cali. <https://www.datos.gov.co/dataset/Datos-de-Calidad-del-Aire-en-la-Estaci-n-Pance-de-/ek65-8kwc>. Última consulta: 2022-12-12.
- DAGMA (2023). Análisis de ozono troposférico - santiago de cali - 2015 - 2022. Technical report, Alcaldía de Cali.
- Davino, C., Romano, R., and Vistocco, D. (2022). Handling multicollinearity in quantile regression through the use of principal component regression. *Metron*, 80(2):153–174.
- D’Agostino, R. B. and Stephens, M. (1986). Tests for normal distribution in goodness-of-fit techniques. *Marcel Decker*.
- Engels, J. M. and Diehr, P. (2003). Imputation of missing longitudinal data: A comparison of methods. *Journal of Clinical Epidemiology*, 56(10):968–976.
- Fernández-Casal, R., Costa, J., and Oviedo, M. (2024). *Métodos predictivos de aprendizaje estadístico*. Servizo de Publicacións, Universidade da Coruña.
- Fisher, N. I. (1995). *Statistical Analysis of Circular Data*. Cambridge University Press.
- Fix, M. J., Cooley, D., Hodzic, A., Gilleland, E., Russell, B. T., Porter, W. C., and Pfister, G. G. (2018). Observed and predicted sensitivities of extreme surface ozone to meteorological drivers in three us cities. *Atmospheric Environment*, 176:292–300.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.

IDEAM (2002). Ozono troposférico. <https://www.ideam.gov.co/web/tiempo-y-clima/ozono-troposferico>. Última consulta: 2023-04-03.

IDEAM (2019). Seguimiento fenómeno el niño. boletín no. 137. http://www.ideam.gov.co/documents/21021/93551524/12_ENSO_IFN_DIC_18_2019.pdf/fa0d9ae5-ba26-4c39-b789-63e94fc28b72?version=1.0. Última consulta: 2023-04-18 [PDF].

Kawichai, S., Kliengchuay, W., Aung, H. W., Niampradit, S., Mingkhwan, R., Niemmanee, T., Srimanus, W., Phonphan, W., Suwanmanee, S., and Tantrakarnapa, K. (2025). The influence of meteorological conditions and seasons on surface ozone in chonburi, thailand. *Toxics*, 13(3):226.

Kebalepile, M. M., Dzikiti, L. N., and Voyi, K. (2024). Using diverse data sources to impute missing air quality data collected in a resource-limited setting. *Atmosphere*, 15(3):303.

Koenker, R. (2005). *Quantile Regression*, volume 38. Cambridge University Press.

Koenker, R. and Bassett Jr, G. (1978). Regression quantiles. *Econometrica*, 46(1):33–50.

Koenker, R. and Hallock, K. F. (2001). Quantile regression. *Journal of Economic Perspectives*, 15(4):143–156.

Koenker, R., Portnoy, S., Ng, P. T., Zeileis, A., Grosjean, P., and Ripley, B. D. (2018). Package ‘quantreg’. *Reference manual available at R-CRAN*.

Le Cook, B. and Manning, W. G. (2013). Thinking beyond the mean: a practical guide for using quantile regression methods for health services research. *Shanghai Archives of Psychiatry*, 25(1):55.

Little, R. J. A. and Rubin, D. B. (2019). *Statistical Analysis with Missing Data*, volume 793. John Wiley & Sons.

Liu, J., Gelman, A., Hill, J., Su, Y.-S., and Kropko, J. (2014). On the stationary distribution of iterative imputations. *Biometrika*, 101(1):155–173.

Medina, F. and Galván, M. (2007). *Imputación de datos: Teoría y práctica*. CEPAL.

Ministerio de Ambiente y Desarrollo Sostenible (2017). Resolución 2254 del 1 noviembre del 2017. <https://www.minambiente.gov.co/wp-content/uploads/2021/10/Resolucion-2254-de-2017.pdf>. Última consulta: 2023-04-15 [PDF].

- Olaya, J., Reina, J., Torres, L. L., Paz, M. C., and Pereira, L. A. (2014). Modelación no paramétrica de la contaminación promedio octohoraria del aire debida al monóxido de carbono y al ozono troposférico. *Ingeniería y Competitividad*, 16(1):147–157.
- Organización Mundial de la Salud (2021). *WHO Global Air Quality Guidelines: Particulate Matter (PM_{2.5} and PM₁₀), Ozone, Nitrogen Dioxide, Sulfur Dioxide and Carbon Monoxide*. World Health Organization. xxi, 273 p.
- Pewsey, A., Neuhäuser, M., and Ruxton, G. D. (2013). *Circular Statistics in R*. Oxford University Press.
- Plackett, R. L. (1983). Karl pearson and the chi-squared test. *International statistical review/revue internationale de statistique*, pages 59–72.
- Porter, W. C., Heald, C. L., Cooley, D., and Russell, B. T. (2015). Investigating the observed sensitivities of air-quality extremes to meteorological drivers via quantile regression. *Atmospheric Chemistry and Physics*, 15(17):10349–10366.
- Pérez Bote, J. L. (2020). *Introducción a la Estadística Circular*. Universidad de Extremadura, Servicio de Publicaciones.
- R Core Team (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Rehan, M. and Munir, S. (2022). Analysis and modeling of air pollution in extreme meteorological conditions: A case study of jeddah, the kingdom of saudi arabia. *Toxics*, 10(7):376.
- Reiss, P. T. and Huang, L. (2012). Smoothness selection for penalized quantile regression splines. *The International Journal of Biostatistics*, 8(1).
- Rubin, D. B. (1978). Multiple imputations in sample surveys: A phenomenological bayesian approach to nonresponse. In *Proceedings of the Survey Research Methods Section of the American Statistical Association*, pages 20–34. American Statistical Association.
- Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91(434):473–489.
- Rubin, D. B. (2004). *Multiple Imputation for Nonresponse in Surveys*, volume 81. John Wiley & Sons.
- Saeipourdizaj, P., Sarbakhsh, P., and Gholampour, A. (2021). Application of imputation methods for missing values of PM₁₀ and O₃ data: Interpolation, moving average and k-nearest neighbor methods. *Environmental Health Engineering and Management Journal*, 8(3):215–226.

- Shin, M. Y. and Lee, H. J. (2024). A comprehensive multi-tiered approach to assessing weather penalties on O_3 levels and exceedance days in seoul using generalized additive models (2001–2019). *Chemosphere*, 368:143687.
- Stephens, M. A. (1974). Edf statistics for goodness of fit and some comparisons. *Journal of the American statistical Association*, 69(347):730–737.
- Van Buuren, S. (2018). *Flexible Imputation of Missing Data*. CRC Press.
- Van Buuren, S. and Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45:1–67.
- Zhang, X., Xiao, X., Wang, F., Brasseur, G., Chen, S., Wang, J., and Gao, M. (2022). Observed sensitivities of PM_{2.5} and O_3 extremes to meteorological conditions in china and implications for the future. *Environment International*, 168:107428.