Tendencia espacio temporal de la concentración de $MP_{2,5}$ y su carga de mortalidad en Bogotá entre 2008 y 2021

David Alejandro González Gutiérrez

Universidad Nacional de Colombia
Facultad de Ciencias, Departamento de Estadística
Bogotá D.C, Colombia
2023

Tendencia espacio temporal de la concentración de $MP_{2,5}$ y su carga de mortalidad en Bogotá entre 2008 y 2021

David Alejandro González Gutiérrez

Trabajo de grado presentado como requisito parcial para optar al título de:
Magíster en Ciencias - Estadística

Directora:
PhD. B. Piedad Urdinola
Profesora Asociada. Facultad de Economía

Codirector:
PhD. Néstor Yesid Rojas
Profesor Asociado. Facultad de Ingeniería

Universidad Nacional de Colombia
Facultad de Ciencias, Departamento de Estadística
Bogotá D.C, Colombia
2023

A mi familia

"¡Presentar lo que es verdad, escribirlo para que quede claro, defenderlo hasta el último aliento!"
Ludwig Boltzmann
(1844-1906)

Agradecimientos

Agradezco a la Universidad Nacional de Colombia y al departamento de estadística por todo lo que me han brindado en el transcurso de estos años, dentro y fuera del aula. A la profesora Piedad Urdinola y al profesor Néstor Rojas por acompañarme y guiarme en el desarrollo de este trabajo.

A mis profesores, compañeros y amigos. En especial a Juan Felipe Latorre y su familia por su apoyo incondicional y sus asesorías en modelos de *aprendizaje de máquina*. A Sergio Rengifo por su apoyo, motivación y trabajo en grupo. A Githza Ruano por acompañarme en el proceso. A mi madre y mis hermanos, que siempre me han apoyado. A mi padre y mi abuela Amelia, muchas gracias, pues siguen motivándome a pesar de su ausencia.

A todas las personas que de una u otra manera me ayudaron en este proceso, muchas gracias. Porque después de todo, esto es más de ellos que mío. Y por supuesto, al lector, quien hace que este documento tenga un motivo de ser.

Índice general

Resumen

Título en español: Tendencia espacio temporal de la concentración de $MP_{2,5}$ y su carga de mortalidad en Bogotá entre 2008 y 2021

La exposición prolongada a material particulado fino, de tamaño inferior a 2,5 micras ($MP_{2,5}$), representa uno de los siete factores de mayor riesgo de muertes prematuras en todo el mundo. Con esta motivación, en el presente trabajo se estimó el número de muertes prematuras asociadas a la exposición prolongada de $MP_{2,5}$ en la ciudad de Bogotá, por localidad y para el período comprendido entre los años 2008 y 2021. Para lograrlo, se realizaron modelos de los niveles de concentración de $MP_{2,5}$ anualmente y se promediaron, utilizando dos enfoques: un modelo *bosque aleatorio (RF)* y un modelo *refuerzo de gradiente extremo (XGBoost)*. Además, se calculó el cociente de riesgo para las muertes cardio metabólicas mediante un modelo proporcional de Cox, tomando como población de referencia la que estuvo expuesta a niveles iguales o menores a $15{,}15 \mu g/m^3$. Los resultados revelaron que un incremento en los niveles de $MP_{2,5}$ está asociado con un aumento en la cantidad de muertes cardio metabólicas, y se identificó que las localidades más afectadas son Kennedy, Bosa y Ciudad Bolívar. Estos hallazgos son coherentes con otros resultados presentados en la literatura. En conclusión, este documento contribuye al análisis del impacto de la contaminación en la salud pública de la ciudad de Bogotá.

Palabras clave: Carga de mortalidad, cociente de riesgo, polución, $MP_{2,5}$, mortalidad, exposición de largo plazo, curva concentración - respuesta .

Abstract

Título en inglés: Spatiotemporal trend of $PM_{2.5}$ and its mortality burden in Bogota between 2008 and 2021

Long-term exposure to fine particulate matter, which is less than 2.5 microns in size ($PM_{2.5}$), is considered one of the seven major risk factors for premature deaths worldwide. As a result, it becomes crucial to investigate its local effects and implement public health policies aimed at reducing premature mortality. This study focuses on estimating the number of premature deaths linked to $PM_{2.5}$ material exposure in Bogota, analyzing data by locality for the years 2008 to 2021. To achieve this, the concentration levels of $PM_{2.5}$ were modeled and averaged annually, using two approaches: a random forest model and an XGBoost model. In addition, the hazard ratio for cardio-metabolic deaths was calculated using a Cox proportional model, taking as the reference population those exposed to levels equal to or less than $15.15 g/m^3$. The findings demonstrate a direct correlation between elevated $PM_{2.5}$ levels and an increase in cardio-metabolic deaths, with Kennedy, Bosa, and Ciudad Bolivar emerging as the most affected localities. These outcomes align with previous research in the field. Consequently, this document contributes to the broader analysis of pollution's impact on public health in Bogota.

Keywords: Mortality burden, Hazard ratio, pollution, $PM_{2.5}$, mortality, long-term exposure, concentration-response curve .

Lista de Figuras

Lista de Tablas

Lista de Símbolos y Siglas

MP_{10} Partículas suspendidas en el aire con un diámetro menor a $10\mu m$

$MP_{2,5}$ Partículas suspendidas en el aire con un diámetro menor a $2{,}5\mu m$

$h(t)$ Función de riesgo en función del tiempo

CR *Curva concentración - respuesta*

GEMM Por sus siglas en inglés. Corresponde a *Modelo de mortalidad por exposición global* o *Global Exposure Mortality Model*

HR Por sus siglas en inglés. Corresponde al *Cociente de riesgo* o *Hazard ratio*.

RF Por sus siglas en inglés. Corresponde a *Bosque aleatorio* o *Random Forest*.

RR *Riesgos relativos*

XGBoost Por sus siglas en inglés. Corresponde a *Refuerzo de gradiente extremo* o *Extreme gradient boosting*.

1. Introducción

El precio de la luz es menor que el costo de la oscuridad

Arthur C. Nielsen (1897-1980)

Los efectos perjudiciales del aumento de la polución del aire sobre la salud humana han sido destacados por diversas investigaciones y organizaciones alrededor del mundo (Gakidou et al. 2017; Pope III y Dockery, 2006). Este tema es de suma importancia para las políticas públicas debido a que representa uno de los principales factores de riesgo para el desarrollo de enfermedades cardiorrespiratorias, tal como lo han demostrado numerosos estudios epidemiológicos (Blanco-Becerra et al. 2014). Un ejemplo significativo es el informe *Carga global de la enfermedad 2019*, que identifica a la contaminación por material particulado menor a $2{,}5\mu m$ de diámetro, o $MP_{2,5}$, como el séptimo factor de riesgo de muerte a nivel mundial, superando al colesterol alto y el alcoholismo (Murray et al. 2020). Por lo tanto, resulta crucial estimar la cantidad de muertes prematuras relacionadas con las concentraciones de $MP_{2,5}$ en áreas densamente pobladas para mejorar la calidad de vida. Sin embargo, debido a la imposibilidad de realizar mediciones directas de este impacto, es necesario recurrir a estimaciones, como las realizadas en el presente trabajo, que se dividen en dos etapas. La primera etapa aborda los niveles de concentración de $MP_{2,5}$, mientras que la segunda se enfoca en la mortalidad por causas específicas para las zonas y periodos de interés.

En cuanto a la estimación de los niveles de concentración de $MP_{2,5}$, se han utilizado diversos métodos. Los más tradicionales incluyen la interpolación espacial, como el kriging ordinario y universal (Sampson et al. 2013), las superficies de tendencia y la interpolación ponderada por el inverso de las distancias. Estos métodos son útiles para conjuntos de datos sin grandes problemas de datos faltantes, aunque en ocasiones la visualización de los resultados puede resultar complicada (G. Zhang et al. 2018). Además, pueden presentar sobreajuste de los niveles de $MP_{2,5}$ cerca de las zonas de monitoreo, mientras que al alejarse de ellas, las estimaciones pierden precisión (H. Zhang et al. 2016). Para evitar tales problemas, algunos autores complementan la información con datos de satélites o utilizan métodos alternos como técnicas de *aprendizaje de máquina*, especialmente los *bosques aleatorios*, lo que conduce a estimaciones más precisas y capaces de capturar efectos locales (Liang et al. 2020). Sin embargo, la efectividad de estos enfoques también depende de la escala de la región de estudio,

ya que los datos satélites pueden no contribuir de manera significativa. En el caso de Bogotá, estos datos no se consideran relevantes debido a que es un área relativamente pequeña en comparación con departamentos o países.

Desde los estudios sobre mortalidad se tiene un amplio espectro en la literatura. Algunos se centran en analizar la relación entre la mortalidad y las condiciones socioeconómicas de grupos poblacionales en general, como se observa en el trabajo de Blanco-Becerra et al. (2014). Por otro lado, existen investigaciones que se adentran en el estudio de la mortalidad por cohortes, lo que implica realizar un seguimiento de diferentes grupos, como menciona (Yang et al. 2020). La elección del enfoque depende de los objetivos específicos de cada investigación y de la disponibilidad de información.

Para relacionar la mortalidad con la concentración de $MP_{2,5}$, se utiliza el cociente de riesgo o cociente de riesgo, que representa la razón de las funciones de riesgo entre una población y otra de referencia o control. A partir de esta relación, se estima la curva de concentración-respuesta (CR). Sin embargo, es necesario hacer supuestos sobre la forma de la curva, ya que los niveles de exposición se obtienen de manera experimental y deben ser incluidos en el modelo (Apte et al. 2015; Murray et al. 2020).

El objetivo de este trabajo consiste en estimar el número de muertes prematuras relacionadas con la exposición a $MP_{2,5}$ en la ciudad de Bogotá. Esta investigación encuentra su motivación en el trabajo de Liang et al. (2020), quienes propusieron modelos alternativos a los de la estadística espacial para lograr niveles más precisos de estimación y para incorporarlos en la curva de concentración-respuesta (CR). Específicamente, el estudio se enfoca en modelar el cociente de riesgo en función de los niveles de concentración de $MP_{2,5}$ para muertes no accidentales, ya que este tipo de muertes puede estar asociado con largos períodos de exposición al material particulado. Para lograr estos objetivos, el trabajo se estructura en varios capítulos. En el Capítulo 2 se presentan el objetivo general y los objetivos específicos de la investigación. En el Capítulo 3 se realiza una revisión de la literatura relacionada con el tema. El Capítulo 4 abarca el marco teórico y los conceptos clave necesarios para desarrollar los modelos utilizados en el estudio. En el Capítulo 5 se detalla la metodología seguida, mientras que en el Capítulo 6 se presentan los resultados obtenidos y se discuten. Finalmente, el trabajo concluye con un resumen de las conclusiones alcanzadas, junto con algunas recomendaciones y sugerencias para futuras investigaciones.

2. Objetivos

2.1. Objetivo General

Estimar el número de muertes prematuras relacionadas con la exposición de material $MP_{2,5}$ para cada una de las zonas cercanas a las estaciones de monitoreo de la Secretaria Distrital de Ambiente de Bogotá del 2008 al 2021.

2.2. Objetivos Específicos

- Conformar la base de datos con la información disponible para Bogotá y sus localidades sobre calidad del aire, condiciones socioeconómicas, distribución de la población y mortalidad.
- Estimar el nivel promedio mensual de $MP_{2,5}$ en la ciudad de Bogotá, para cada una de las estaciones de monitoreo de la Secretaría Distrital de Ambiente, a lo largo de los años 2008 a 2021 usando dos modelos de *aprendizaje de máquina*: *bosque aleatorio* (RF) y *Refuerzo de gradiente extremo* (*XGBoost*).
- Modelar el cociente de riesgo usando modelos de Cox a partir de las variables socio-económicas obtenidas.
- Encontrar la curva de concentración - respuesta (CR) para las muertes no accidentales en cada una de las estaciones de monitoreo de calidad del aire de Bogotá.

3. Revisión de la literatura

3.1. Contexto mundial

Hasta los años 80 y 90, los estudios epidemiológicos que analizaban los efectos de la contaminación del aire en la salud humana se basaban en comparar la mortalidad y el estado de salud de la población entre distintas regiones con niveles significativamente diferentes de $MP_{2,5}$ o MP_{10}. Estos análisis eran descriptivos y cualitativos (Sram et al. 1996). Otras aproximaciones implicaban examinar el número de ingresos hospitalarios en áreas con altos niveles de polución, así como las razones de las consultas médicas para identificar posibles enfermedades o factores de riesgo. Además, el uso de series de tiempo para observar los cambios en las tasas de mortalidad era una práctica común (Dockery et al. 1993). Sin embargo, la mayoría de estos estudios no consideraban directamente las concentraciones de material particulado u otros compuestos químicos presentes en el aire.

A mediados de los años 90, se popularizó el uso de modelos de riesgos proporcionales de Cox, que involucran modelar la función de riesgo h en función del tiempo t, junto con un conjunto de covariables $\vec{X}$ que contienen información socioeconómica y médica de diferentes cohortes de la población de interés (Dockery et al. 1993; Apte et al. 2015). Estos modelos tienen la siguiente forma funcional

$$h(t;\vec{X}) = h_0(t)exp\left(\vec{X}^T\vec{\beta}\right), \tag{3-1}$$

donde $h_0(t)$ es el riesgo base (cuando todas las covariables son 0) y $\vec{\beta}$ el vector de parámetros. Además, estos modelos usan el supuesto que los riesgos son proporcionales, es decir, el efecto de cada covariable es constante en el tiempo (Cox, 1997; Dockery et al. 1993). Sin embargo, el desarrollo y validación de estos supuestos pueden ser un proceso complejo, ya que pueden surgir casos en los que una misma covariable aplique para una población pero no para otra, o su efecto puede variar con el tiempo.

Además, durante los años 90 surge el *Carga global de la enfermedad*, un estudio destinado a cuantificar el impacto en la salud pública de más de 100 enfermedades y factores de riesgo, encargado inicialmente por el Banco Mundial y posteriormente asumido por la Organización Mundial de la Salud. Con el paso del tiempo, estos estudios ganaron impulso y, a partir de

2010, se publican cada dos o tres años. El informe más reciente disponible corresponde al año 2019, a partir del cual Murray et al. (2020) destacan que la contaminación del aire ha incrementado su posición como factor de riesgo. En concreto, ha pasado de ser el décimo tercer factor más relevante en 1990 a convertirse en el séptimo en 2019, como se muestra en la tabla **3-1**.

Número	Mayores Factores de riesgo 1990	Mayores Factores de riesgo 2019
1	Desnutrición infantil	Presión arterial sistólica elevada
2	Bajo peso al nacer	Tabaquismo
3	Gestación prematura	Glucosa plasmática en ayunas elevada
4	Contaminación del aire en el hogar	Bajo peso al nacer
5	Tabaquismo	Sobrepeso
6	Agua contaminada	Gestación prematura
7	Presión arterial sistólica elevada	Contaminación del aire (por material $MP_{2,5}$)
8	Niños con bajo peso	Colesterol elevado
9	Saneamiento inseguro	Alcoholismo
10	Lavado de manos	Contaminación del aire en el hogar

Tabla 3-1.: Mayores factores de riesgo para todas la edades de acuerdo al informe *Carga global de la enfermedad*. Fuente: Tomado de Murray et al. (2020).

Debido al aumento de la contaminación del aire como factor de riesgo y a la necesidad de cuantificar el riesgo asociado a la larga exposición a material $MP_{2,5}$, se han implementado diversas metodologías. Las cuales se basan en calcular el riesgo relativo (RR) de una población expuesta a un nivel de concentración de $MP_{2,5}$, respecto a una población expuesta a concentración de referencia. A partir de ello, se ajusta el riesgo relativo a los niveles de concentración de $MP_{2,5}$, lo que se conoce como curva concentración - respuesta. Es importante notar que en este proceso también se hacen uso de técnicas de *aprendizaje de máquina* para la estimación de los niveles de $MP_{2,5}$ (R. T. Burnett et al. 2014). Otras metodologías consisten en consolidar diferentes estimaciones de riesgos relativos y ajustar una respuesta integrada a la exposición. Sin embargo, el riesgo relativo se mide a partir de las probabilidades absolutas. Así, que otra manera de abordar el problema es verlo desde el cociente de riesgo el cual se calcula a partir de probabilidades instantáneas y presenta menos limitaciones para consolidar diferentes fuentes (Stare y Maucort-Boulch, 2016). Es por ello que R. Burnett et al. (2018) proponen el Modelo de mortalidad por exposición global, (GEMM) por sus siglas en inglés, el cual tiene bastante popularidad pues estandarizada las metodologías desarrolladas anteriormente y usa el cociente de riesgo.

Esta metodología se ha implementado en diversas regiones, como China, Corea, Canadá, Brasil y otros lugares (Liang et al. 2020; Yang et al. 2020; Han et al. 2018; Southerland et al. 2022). Por otro lado, la Organización Mundial de la Salud ha desarrollado herramientas

como AirQ+ para cuantificar el riesgo por la contaminación del aire, aunque su popularidad no es tan extensa en estudios relacionados con el tema (Mudu et al. 2018). En la actualidad, los estudios se centran en analizar la mortalidad asociada a causas de muerte específicas y también a escenarios vinculados con el cambio climático o los incendios forestales (Cheng et al. 2023; Johnston et al. 2021; Chowdhury et al. 2018). En este contexto, también se trabaja en mejorar la calidad de la información disponible con el objetivo de obtener resultados que permitan comparar diferentes escenarios.

3.2. Contexto nacional

En el caso de Colombia, los estudios y análisis han adoptado diferentes enfoques. En la década de los 2000, algunos estudios se centraron en casos particulares, como el modelado de la curva de concentración-respuesta para tres contaminantes químicos en Bogotá durante 1998 (Lozano, 2004). Posteriormente, se abordaron otros aspectos, como los impactos económicos de las muertes prevenibles debido a la exposición prolongada a material MP_{10} en Bogotá durante los años 2010 y 2020 (Ortiz-Durán y Rojas-Roa, 2013). También se aplicaron métodos basados en series de tiempo para investigar la relación entre la mortalidad y los niveles de MP_{10} en Bogotá durante los años 1998 y 2006 (Blanco-Becerra et al. 2014).

Posteriormente, Zafra-Mejía et al. (2020) analizan la relación entre la condición de la atmósfera (AC) y la mortalidad asociada a MP_{10} en Bogotá. Para ello, investigan la estabilidad de la atmósfera en función de los meses y su relación con la mortalidad. Por otro lado, Grisales-Romero et al. (2021) realizan un análisis descriptivo de la carga de mortalidad atribuible a la exposición de $MP_{2,5}$ en Medellín durante el período 2010 al 2016. Además, existen estudios más amplios como el presentado por Rodriguez-Villamizar et al. (2022), donde se estiman las muertes prematuras en Colombia durante el período 2014 a 2019 a nivel municipal utilizando una metodología de riesgos relativos. Asimismo, Arregocés et al. (2023) utilizan el software AirQ+ de la OMS aplicado a la región Caribe para estimar la proporción de mortalidad atribuible a los niveles de $MP_{2,5}$.

De manera paralela, el Instituto Nacional de Salud (INS) ha elaborado una serie de informes denominados *Informe Carga de Enfermedad Ambiental*, donde se analizan los principales factores de riesgo ambiental que pueden impactar la salud de los colombianos en todo el país. El último de estos informes corresponde al año 2018, en el cual se estima que el 8 % de las muertes ocurridas en el país durante 2016 se deben a la exposición al aire y agua de mala calidad (Instituto Nacional de Salud, 2018). Además, Greenpeace realizó un análisis en 2022 sobre la carga de mortalidad en Bogotá durante 2021, y estimó que 3400 muertes prematuras fueron relacionadas con la exposición de material $MP_{2,5}$, lo que representa el 8 % del total de muertes (Farrow et al. 2022).

Por otro lado, existen estudios relacionados con la estimación de material particulado en el aire, en los cuales se emplean diversas herramientas, desde métodos geoespaciales tradicionales utilizados por la Secretaría Distrital de Ambiente, hasta enfoques menos convencionales basados en *aprendizaje de máquina*. Un ejemplo de esto es el trabajo de Casallas et al. (2020), que se enfoca en la predicción de los niveles de $MP_{2,5}$ y MP_{10} durante febrero de 2019 en Bogotá, utilizando tres modelos distintos. Además, Casallas et al. (2021) presentan la implementación de redes neuronales para estimar los niveles de $MP_{2,5}$ en Bogotá con información de los últimos 7 años.

4. Marco Teórico

En este capítulo se presentan los conceptos y desarrollos teóricos relacionados con este trabajo, dividiéndose en cuatro partes. La primera parte aborda la contaminación del aire y los modelos utilizados en la literatura para estimar los niveles de concentración de compuestos químicos. En la segunda parte se examina la mortalidad y se introducen conceptos de análisis de supervivencia. La tercera parte se enfoca en las formas de modelar la curva concentración-respuesta (CR). Finalmente, en la cuarta parte se describen las metodologías para realizar la estimación de muertes prematuras relacionadas con la exposición prolongada a $MP_{2,5}$.

4.1. Contaminación del aire

La contaminación del agua o del aire debido a residuos industriales o biológicos es conocida como polución, y representa un riesgo para la salud humana y otros seres vivos. Aunque se ha convertido en un problema a gran escala con la revolución industrial y la tecnificación de diversos sectores económicos, también existía en momentos anteriores de la historia, como en la antigüedad y la edad media, debido a la quema de fogatas y la falta de sistemas sanitarios adecuados (Urbinato, 1994). Aunque el término polución es amplio, hay muchos compuestos químicos que forman parte de la contaminación del aire, como los óxidos de nitrógeno (NO_x), el dióxido de carbono (CO_2), el dióxido de azufre y otros. Sin embargo, la clasificación más común se basa en el diámetro de las partículas contaminantes suspendidas en la atmósfera. Por lo tanto, se denominan material particulado MP_{10} y $MP_{2,5}$ a las partículas con diámetros menores a $10\mu m$ y $2{,}5\mu m$, respectivamente. Debido a esta diferencia de tamaño, las partículas $MP_{2,5}$ tienen una mayor probabilidad de ingresar al sistema respiratorio y provocar problemas de salud.

4.1.1. Estimación de niveles de $MP_{2,5}$

Teniendo en cuenta lo expuesto anteriormente, en este trabajo se realizaron estimaciones de los niveles de $MP_{2,5}$ con los siguientes dos métodos de aprendizaje de máquina. Los cuales ayudan a abordar los problemas de no linealidad presentes en los datos, pero no contemplan la estructura de dependencia espacial; por lo tanto se utilizan para estimación y no para predicción.

- ***Bosque aleatorio* (RF)**

  Es un algoritmo desarrollado por Leo Breiman en 2001 que funciona construyendo un gran número de árboles de decisión durante el proceso de entrenamiento. Un árbol de decisión es un modelo no paramétrico, donde la relación entre la variable objetivo y sus variables explicativas se describe mediante un conjunto de bifurcaciones organizadas en una estructura de árbol. Este enfoque se puede utilizar tanto para regresiones como para clasificaciones. Este método es una alternativa muy importante cuando se enfrenta a datos con información faltante, ya que proporciona una mayor precisión que los métodos de interpolación espacial (Breiman, 2001; G. Zhang et al. 2018).

  En el caso de las regresiones, cada condición se expresa como $x_j > x_{j,th}$, donde x_j es el valor de una variable regresora en el índice j y $x_{j,th}$ es el valor umbral. Tanto la variable como su umbral son determinados durante la etapa de entrenamiento. El procedimiento inicia con el conjunto de datos completo y un valor predefinido de árboles. En cada árbol, el modelo toma una muestra de los datos y los divide según una de las variables regresoras. Luego, realiza una predicción sobre la variable objetivo y selecciona el valor que minimiza el error cuadrático medio o cualquier otra función de pérdida definida. Posteriormente, se selecciona otra variable regresora y se repite el proceso hasta agotar todas las variables regresoras. Finalmente, se ponderan los resultados y se obtiene el resultado final. Es decir, el algoritmo genera múltiples modelos de regresión y pondera sus estimaciones (Breiman, 2001;Reis et al. 2018; G. Zhang et al. 2018).

- ***Refuerzo de gradiente extremo* (*XGBoost*)**

  XGBoost es un método que consiste en optimizar un modelo ya implementado. Es decir, se puede tomar alguno de los modelos de interpolación espacial o un RF y tomar alguna función de pérdida (Schapire y Freund, 2013). Por lo general, la función de pérdida utilizada es el error cuadrático medio. Por ejemplo, sea $\hat{y}_i = F(x_i)$ el valor predicho, y_i el valor observado con $i = 1, 2, ..., n$ con n el número de observaciones. Entonces la función de pérdida está dada por

$$L_{MSE} = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - F(x_i) \right)^2 , \tag{4-1}$$

  y el gradient boosting a realizar

$$-\frac{\partial L_{MSE}}{\partial F(x_i)} = \frac{2}{n} \left(y_i - F(x_i) \right) . \tag{4-2}$$

Donde la idea es encontrar un nuevo estimador $h(x_i)$ que contribuya a mejorar el modelo. Es decir,

$$y_i = F(x_i) + h(x_i), \tag{4-3}$$

por lo tanto

$$L_{MSE} = \frac{2}{n} h(x_i). \tag{4-4}$$

4.2. Mortalidad y análisis de supervivencia

La mortalidad, o tasa de mortalidad, se refiere al número de muertes ocurridas en un periodo de tiempo respecto al total de un grupo poblacional específico . Se suele expresar en el número de muertes por cada 1000 habitantes en un año (Bureau, 2007). Puede ser calculada para toda la población en un período de tiempo determinado, desglosada por género, edades y otros criterios. Esto permite comprender el comportamiento de los fallecimientos en una población particular. También se suele expresar en términos del número de años vividos por persona en lugar del número de muertes (Wachter, 2014).

Por otro lado, para establecer la relación entre el número de muertes y los niveles de $MP_{2,5}$, es necesario definir los tipos de muerte que podrían estar involucrados. Por ejemplo, los decesos relacionados con asesinatos, suicidios, accidentes de tránsito y otros eventos no están directamente vinculados a factores ambientales. Por lo tanto, siguiendo la propuesta de Yang et al. 2020, se consideran dos grupos de muertes según la 10ma Edición de la Clasificación Internacional de Enfermedades[1] dada por la Organización Mundial de la Salud (OMS):

- *No accidentales:* Aquí se incluyen las muertes que no fueron provocadas por factores externos, como envenenamiento o traumatismos, entre otros. Este grupo corresponde a las causas con códigos entre $A00$ y $R99$.
- *Cardio-metabólicas:* Se trata de un subconjunto de las muertes no accidentales que abarcan los códigos entre $I00$ y $I99$, así como del $E10$ al $E14$.

Asimismo, para estimar el número de muertes prematuras, Liang et al. 2020 consideran únicamente las muertes cardio metabólicas.

[1] La 10 edición estuvo vigente hasta el 31 de diciembre de 2021 la cual está disponible en *https://icd.who.int/browse10/2016/en*. Actualmente se está implementado la 11ra edición.

En cuanto al análisis de supervivencia, que consiste en analizar y modelar el tiempo transcurrido hasta la ocurrencia de un evento, como la muerte, es necesario tener en cuenta algunos conceptos. Sea T una variable aleatoria positiva que indica el tiempo transcurrido hasta la muerte del individuo.

Función de supervivencia: Probabilidad de un individuo de sobrevivir más allá del tiempo t. Es decir

$$S(t) = P(T > t), t > 0. \tag{4-5}$$

Función de riesgo: También llamada función de Hazard. Es la probabilidad de un individuo de morir en el siguiente instante de tiempo dado que sobrevivió hasta el tiempo t (Klein, Moeschberger et al. 2003). Es conocida también como la fuerza de la mortalidad y está dada por

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T \leq t + \Delta t | T \geq t)}{\Delta t}. \tag{4-6}$$

Y cumple con

$$h(t) = -\frac{dln[S(t)]}{dt} \tag{4-7}$$

Además, al comparar las curvas de supervivencia de dos grupos se define:

Riesgo relativo (RR): Probabilidad de un individuo de morir antes de un tiempo t en un grupo dividida entre la probabilidad de morir antes de un tiempo t en el otro grupo. Es decir, para un tiempo t el riesgo relativo es

$$RR(t) = \frac{P(T \leq t | X = 1)}{P(T \leq t | X = 2)}, \tag{4-8}$$

donde X es una variable que indica si el individuo pertenece al grupo 1 o 2 (Stare y Maucort-Boulch, 2016).

Cociente de riesgo: Es la razón entre las funciones de riesgo entre los dos grupos. Es decir, es la probabilidad de un individuo en el grupo 1 de morir en el siguiente instante de tiempo dado que sobrevivió hasta el tiempo t, dividida entre la probabilidad de un individuo del grupo 2. Así

$$HR(t) = \frac{h(t, 1)}{h(t, 2)}, \tag{4-9}$$

donde $h(t,1)$ y $h(t,2)$ son las funciones de riesgo para el grupo 1 y 2, respectivamente. Es usual encontrar en la literatura que estas dos cantidades son similares, pero hay diferencias en ellas. Pues el riesgo relativo se mide sobre la probabilidad total de sobrevivir mientras el cociente de riesgo sobre la probabilidad condicional (Klein, Moeschberger et al. 2003; Stare y Maucort-Boulch, 2016).

4.2.1. Modelo de Cox

Este modelo permite explicar la función de riesgo $h(t)$ en función del tiempo y de covariables asociadas a las características sociodemográficas de la población, de la siguiente forma

$$h(t) = h_0(t)exp\left(\vec{X}^T\vec{\beta}\right), \tag{4-10}$$

donde $h_0(t)$ se le denomina riesgo base, $\vec{X}$ es un vector con las covariables y $\vec{\beta}$ un vector con los parámetros del modelo. Las cantidades $exp(\beta_i)$ son conocidas como los cociente de riesgos para cada covariable en particular, y para ser significativas dentro del modelo su p-valor debe ser menor a un nivel de significancia α.

Por otro lado, el modelo plantea dos supuestos. El primero es conocido como el supuesto de riesgos proporcionales, y hace referencia a que las funciones de riesgo $h(t)$ para los grupos analizados no se deben intersecar. Es decir, se debe cumplir que el cociente de riesgo es independiente del tiempo. Por ejemplo, sean dos individuos k y k' entonces se debe cumplir

$$\frac{h_k(t)}{h_{k'}(t)} = \frac{h_0(t)exp\left(\vec{X}^T\vec{\beta}\right)}{h_0(t)exp\left(\vec{X'}^T\vec{\beta}\right)} = c, \tag{4-11}$$

donde $h_k(t)$ y $h_{k'}(t)$ son las funciones de riesgo para los individuos k y k', y c es una constante. Este supuesto se puede validar con la prueba de residuales escalados de Schoenfeld. Para esto se consideran n tiempos de observación durante el análisis, donde $t_{(j)}$ el j-ésimo tiempo de observación con $j = 1, ..., n$. Así, sea $R_{(j)}$ el conjunto de la población en riesgo para el tiempo $t_{(j)}$, y $\vec{X}_{(j)}$ el vector de covariables para un individuo. Con ello se definen los residuales de Schoenfeld $\hat{\vec{r}}_{(j)}$ como

$$\hat{\vec{r}}_{(j)} = \vec{X}_{(j)} - \frac{\sum_{k\in R_{(j)}} \vec{X}_k exp\left(\vec{X}_k^T\hat{\vec{\beta}}\right)}{\sum_{k\in R_{(j)}} exp\left(\vec{X}_k^T\hat{\vec{\beta}}\right)}. \tag{4-12}$$

Además, la matriz de varianza en este caso se definen como

$$V_{(j)} = \frac{\sum_{k\in R_{(j)}} \vec{X}_k \vec{X}_k^T exp\left(\vec{X}_k^T \hat{\vec{\beta}}\right)}{\sum_{k\in R_{(j)}} exp\left(\vec{X}_k^T \hat{\vec{\beta}}\right)} - \frac{\sum_{k\in R_{(j)}} \left[\vec{X}_k exp\left(\vec{X}_k^T \hat{\vec{\beta}}\right)\right]\left[\vec{X}_k exp\left(\vec{X}_k^T \hat{\vec{\beta}}\right)\right]^T}{\left(\sum_{k\in R_{(j)}} exp\left(\vec{X}_k^T \hat{\vec{\beta}}\right)\right)^2}. \tag{4-13}$$

Así, los residuales escalados de Schoenfeld $\hat{\vec{r}}_{s(j)}$ se definen como

$$\hat{\vec{r}}_{s(j)} = V_{(j)}^{-1} \hat{\vec{r}}_{(j)}, \tag{4-14}$$

para un individuo de la población en riesgo (Winnett y Sasieni, 2001). De esta manera, la prueba de los residuales escalados de Schoenfeld tiene como hipótesis nula que los residuales escalados no tienen una correlación con el tiempo, frente a la hipótesis alterna donde la tienen. Si el p-valor es mayor al nivel de significancia definido, que por lo usual es $\alpha = 0{,}05$, no se rechaza la hipótesis nula y por lo tanto se cumple el supuesto de los riesgos proporcionales (Schoenfeld, 1982). Esta prueba se encuentra en la mayoría de software y paquetes estadísticos, como en el caso de R.

El segundo supuesto se tiene para las variables continuas incluidas en el modelo, para las cuales se debe garantizar que no hay relaciones no lineales entre el $ln[h(t)]$ y las covariables. Para esto se usan métodos gráficos donde se ve busca que no exista no linealidad entre los residuales de Martingale y los valores de las covariables (David et al. 1972).

4.3. Curva concentración - respuesta (CR)

La curva concentración respuesta busca ajustar el cociente de riesgo o cociente de riesgo a los niveles de concentración de $MP_{2,5}$. Como se mencionó en el capítulo 3, esto se puede realizar con la metodología del GEMM, en el cual se propone la siguiente forma funcional

$$HR(z) = exp\left(\theta T(z) w(z)\right) \tag{4-15}$$

donde θ es un parámetro a ajustar, $z = C - a$ con C el nivel de concentración $MP_{2,5}$ y $a = min(C)$, $T(z) = ln(z+1)$ y $w(z)$ una función que sirve para controlar la forma de la curva. En el caso general se puede tomar

$$w(z) = \frac{1}{1 + exp\left(-\frac{(z-\mu)}{\nu}\right)}, \tag{4-16}$$

donde μ y ν son parámetros para ajustar. Lo anterior implica un alto costo computacional para la estimación de los parámetros. Sin embargo, para simplificar el modelo, se puede

abordar el caso donde μ toma valores negativos muy grandes. De esta manera, $w(z) \approx 1$ y el modelo se puede llevar a una forma lineal mucho menos costosa computacionalmente (Apte et al. 2015; R. Burnett et al. 2018).

Existe otra forma de encontrar la curva concentración - respuesta, pero ya no en términos del cociente de riesgo sino del riesgo relativo. La cual está dada por

$$RR(z)) = 1 + \alpha \left[1 - exp\left[-\gamma z^{\delta}\right]\right], z > a, \tag{4-17}$$

donde α, γ y δ son parámetros para ajustar, y $z = C - a$ con C el nivel de concentración $MP_{2,5}$ y $a = min(C)$ (R. T. Burnett et al. 2014).

4.4. Estimación muertes prematuras

La estimación de muertes prematuras relacionadas con la exposición de $MP_{2,5}$ corresponde a calcular el exceso de muertes que se presentan en un grupo respecto a otro de referencia, el cual está a un nivel de dado de $MP_{2,5}$. De esta manera, se puede realizar a través del cociente de riesgo o los riesgos relativos (RR). En el primer caso, se sigue la metodología utilizada por Liang et al. (2020), aplicando la siguiente ecuación:

$$M_{ij} = \frac{HR(C_{ij}) - 1}{HR(C_{ij})} CM_{ij}, \tag{4-18}$$

donde M_{ij} es el número de muertes prematuras en la localidad i en el año j; $HR(C_{ij})$ es el cociente de riesgo estimado para la concentración de $MP_{2,5}$ en la localidad i en el año j; y CM_{ij} es el número total de muertes cardio - metabólicas en la localidad i en el año j. Es importante notar que usando la ecuación 4-15 para $HR(C_{ij})$, el mínimo valor posible es 1 con lo cual M_{ij} es no negativo. En el segundo caso, usando riesgos relativos se tiene

$$\Delta M = M_0 P \left(1 - exp(-\beta \Delta MP)\right), \tag{4-19}$$

donde ΔM son las muertes prematura estimadas, M_0 es la mortalidad base, P es la población expuesta, ΔMP es el cambio anual promedio de la concentración de $MP_{2,5}$, y $\beta = ln(RR)/\Delta Q$ con $\Delta Q = 10 \mu g/m^3$ (Rodriguez-Villamizar et al. 2022).

5. Metodología

En este capítulo se presenta la metodología implementada para la estimación del número de muertes prematuras relacionadas con la exposición de material $MP_{2,5}$. Se abordan las fuentes utilizadas, el tratamiento de los datos para la construcción de la base de datos y la forma en que se desarrollaron cada uno de los modelos necesarios.

5.1. Fuentes de datos y tratamiento

Para la construcción de la base de datos, se utilizaron las fuentes mostradas en la tabla **5-1**, que comprenden cuatro fuentes primarias. La primera corresponde a los microdatos de defunciones no fetales del Departamento Administrativo Nacional de Estadística (DANE), a partir de los cuales se identificaron las defunciones por causas cardio-metabólicas en Bogotá durante el periodo de estudio. Es importante señalar que estos datos son consolidados y no incluyen información detallada por localidad. La segunda fuente son las defunciones por localidad en Bogotá, registradas por la Secretaría Distrital de Salud. La tercera fuente incluye los registros de las estaciones de monitoreo de la calidad del aire de la Secretaría Distrital de Ambiente (SDA), con información comprendida entre 2008 y 2021. Estas estaciones de monitoreo son: Carvajal, CDAR, Fontibón, Guaymaral, Kennedy, Las Ferias, MinAmbiente, Móvil 7ma, Puente Aranda, San Cristóbal, Suba, Tunal y Usaquén. Por último, la cuarta fuente corresponde a las estimaciones de la población por localidad en Bogotá, proporcionadas por la Secretaría Distrital de Salud, a partir de los censos realizados en 2005 y 2018.

De esta manera, se construyó una base de datos consolidada de defunciones por localidad, junto con las causas de muerte y características sociodemográficas utilizando las dos primeras fuentes. Para lograr esto, se realizaron cruces de información por campos como año del deceso, género, edad, causa básica de la muerte y estado de la seguridad social. Además, se utilizaron dos tablas maestras para estandarizar la información: una para el grupo de edad y otra para la identificación de las causas cardio-metabólicas. Estos pasos fueron necesarios debido a que los datos abiertos del DANE no incluían la localidad para las defunciones en Bogotá, mientras que la información de la Secretaría Distrital de Salud no proporcionaba las características sociodemográficas. Gracias a este proceso, se obtuvo información completa de 129 622 defunciones cardio-metabólicas registradas en Bogotá para mayores de 20 años, de un total de 134 269 defunciones registradas durante el periodo comprendido entre 2008 y

2021.

Nombre de la fuentes	Descripción	Periodo de tiempo	Propietario de la fuente
Defunciones no fetales	Microdatos con las estadísticas de defunciones no fetales en el todo el territorio nacional	2008 al 2021	Departamento Administrativo Nacional de Estadística - DANE.
osb-demografia-causasmortalidad	Microdatos con las estadísticas de las defunciones en Bogotá por localidad	2008 al 2021	Secretaría Distrital de Salud
DatosMetereologicos	Información detallada registrada por las estaciones de monitoreo de la calidad del aire en Bogotá.	2008 al 2021	Secretaría Distrital de Ambiente.
osb-demografia-poblacion	Estimaciones de la población por localidad en Bogotá de la Secretaría Distrital de Salud	2008 al 2021	Secretaría Distrital de Salud.

Tabla 5-1.: Resumen de fuentes usadas para la base de datos

Seguido a esto, se tomó la información meteorológica registrada por las estaciones de monitoreo y se resumió de forma mensual. Es decir, se calcularon los niveles mensuales promedios para las siguientes variables

- MP_{10}: Densidad de material particulado con diámetros menores $10\mu m$. Se mide en $\mu g/m^3$
- $MP_{2,5}$: Densidad de material particulado con diámetros son menores a $2{,}5\mu m$.
- O_3: Concentración de ozono. Medido en partes por mil millones.
- CO: Concentración de monóxido de carbono. Medido en partes por millón.
- NO: Concentración de monóxido de nitrógeno. Medido en partes por mil millones.
- NO_2: Concentración de dióxido de nitrógeno. Medido en partes por mil millones.

- NO_X: Concentración de óxidos de nitrógeno. Medido en partes por mil millones.
- $t2$: Temperatura medida a 2 metros del suelo. Medido en $°C$.
- $t8$: Temperatura medida a 8 metros del suelo. Medido en $°C$.
- $t20$: Temperatura medida a 20 metros del suelo. Medido en $°C$.
- ws: Velocidad del viento. Medida en m/s.
- wd: Dirección del viento. Medida en grados
- rn: Precipitación. Medida en mm.
- rh: Porcentaje de humedad relativa
- p: Presión, medida en mmHg.
- sr: Radiación solar, medida en W/m^2.

Posteriormente, como se muestra en la figura **5-1**, se diseñó una cuadrícula de 1 km por 1 km para el mapa de Bogotá y se ubicaron las estaciones de monitoreo en ella. Luego, se asignaron las coordenadas correspondientes a cada estación dentro de la cuadrícula. Este paso es fundamental para utilizarlo como insumo en los modelos de estimación de los niveles mensuales de $MP_{2,5}$ por localidad.

5.2. Estimación del nivel promedio mensual de $MP_{2,5}$

A partir de la cuadrícula con resolución de 1 km por 1 km, mencionada anteriormente, se seleccionaron únicamente los registros con información mensual de los niveles de $MP_{2,5}$, que constituye la variable objetivo. Sin embargo, debido al reducido número de estaciones en comparación con el tamaño de la ciudad, fue necesario disminuir la resolución de la cuadrícula a 2km x 2km y excluir la localidad de Sumapaz. Esto se hizo con el propósito de aumentar el número de registros sin valores nulos y mejorar la calidad de los datos para el desarrollo de los modelos. Además, se tomaron las variables con una completitud mayor al 70 % para entrenar los modelos, que incluyen la temperatura a 2 metros del suelo, la concentración de CO, O_3 y MP_{10}, así como la precipitación, velocidad y dirección del viento. Por último, se agregaron como variables la población y el número de muertes cardio - metabólicas de las localidades donde se ubican las estaciones de monitoreo.

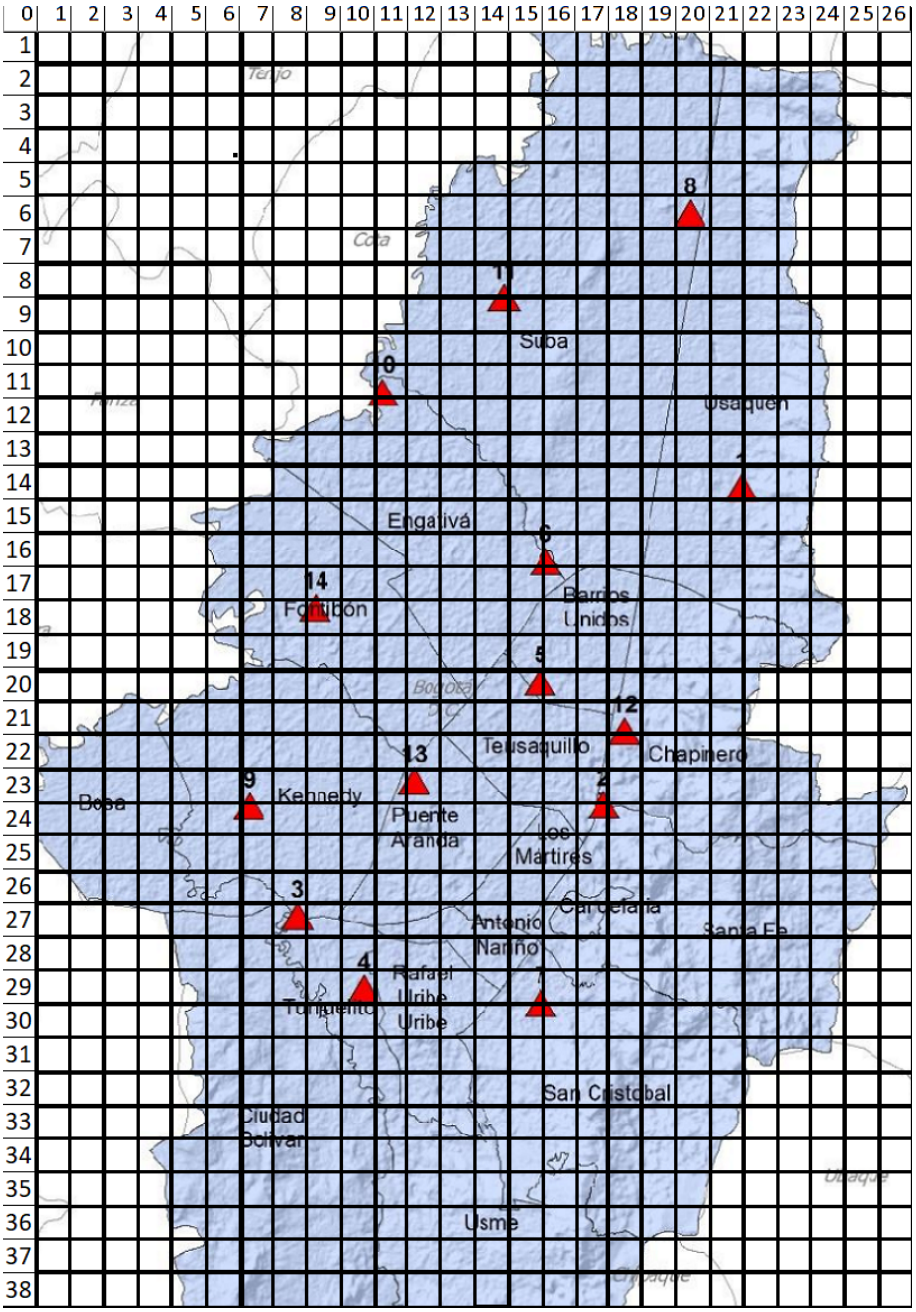


Figura 5-1.: Cuadrícula de 1km por 1km diseñada para Bogotá junto con la localización de las estaciones de monitoreo de calidad del aire de Bogotá. Fuente: Secretaría Distrital de Ambiente - Esquema propio.

De esta manera, se realizó un proceso de limpieza de datos, excluyendo los niveles atípicos de $MP_{2,5}$, que corresponden a valores superiores a $100\mu g/m^3$ presentados antes de 2010 en la estación de Guaymaral. También, se aplicó un proceso de imputación para los datos faltantes utilizando el método de *bosque desaparecido*. Este método consiste en implementar un *bosque aleatorio* para predecir los datos faltantes, utilizando como datos de entrenamiento los datos

reales y minimizando el error cuadrático medio. Para lograr imputar los datos faltantes, se requirieron 5 iteraciones. Con este preprocesamiento, se generó la base para el desarrollo de los modelos de *bosque aleatorio* (RF) y *Refuerzo de gradiente extremo* (*XGBoost*).

Bosque aleatorio (RF)

Para esta parte, se seleccionaron las celdas con estaciones de monitoreo y se desarrolló un modelo de *bosque aleatorio* de regresión con 10, 20, 50 y 100 árboles de decisión, utilizando la función de costo de error absoluto medio (MAE), y se empleó una división de datos del 80 % para entrenamiento y el 20 % para pruebas. Luego, se seleccionó el mejor modelo en función del menor error absoluto medio (MAE) y el mayor coeficiente de determinación (R^2) al ajustar una regresión lineal entre los valores mensuales estimados y observados de $MP_{2,5}$. A continuación, se llevó a cabo una optimización de hiperparámetros con el objetivo de encontrar el número óptimo de árboles y variables por regresión que minimizaran el MAE y maximizaran el mencionado R^2.

Posteriormente, utilizando el modelo de *bosque aleatorio* para estimar los niveles mensuales de $MP_{2,5}$, se aplicó el método de *bosque desaparecido* en las celdas donde no se encontraban las estaciones de monitoreo. Esto permitió imputar los datos de las variables distintas a los niveles de $MP_{2,5}$, con el fin de aplicar el modelo de *bosque aleatorio* también en esas celdas.

Refuerzo de gradiente extremo (*XGBoost*)

Para esta parte, se seleccionaron las celdas con estaciones de monitoreo y se desarrolló un modelo de *XGBoost* con un rango de hiperparámetros entre 0 y 20 para la profundidad máxima de los árboles, una tasa de aprendizaje entre 0.001 y 0.1, un máximo de 1000 árboles de decisión, y utilizando la función de costo error absoluto medio (MAE), con una división de datos del 80 % para entrenamiento y el 20 % para pruebas. Siguiendo un procedimiento similar al realizado para el *bosque aleatorio*, se imputaron los datos faltantes en las celdas donde no se encontraban las estaciones de monitoreo mediante el método de *bosque desaparecido*, y posteriormente se aplicó el modelo de *XGBoost* a esas celdas.

Finalmente, se realizó la estimación final de los niveles mensuales de $MP_{2,5}$ como el valor promedio obtenido a partir de los resultados de ambos modelos (*bosque aleatorio* y *XGBoost*), para el periodo de tiempo comprendido entre los años 2008 y 2021. Este resultado sirve como insumo para calcular el cociente de riesgo y la curva de concentración - respuesta.

5.3. Modelamiento del cociente de riesgo

Para el modelamiento del cociente de riesgo se tienen en cuenta las muertes cardio-metabólicas. Como se mencionó en la sección de construcción de la base de datos, de las 134 269 muertes cardio-metabólicas registradas en Bogotá para mayores de 20 años durante el periodo de 2008 al 2021, se tiene información completa para 129 622 muertes gracias a los datos proporcionados por el DANE y la Secretaría Distrital de Salud. A partir de estos registros, se identificó el nivel mensual promedio de $MP_{2,5}$ del lugar de residencia para 129 584 muertes.

Luego, la función de riesgo por un aumento de $10\mu g/m^3$ en los niveles anuales de $MP_{2,5}$ se calculó mediante un modelo proporcional de Cox como el mostrado en la ecuación 4-10, tomando como covariables: los niveles de $MP_{2,5}$ C, el género, grupo etario, estado civil, estado de afiliación a la seguridad social, máximo nivel educativo alcanzado y grupo étnico. En el caso de las variables categóricas, para cada categoría se creó una variable indicadora. Sin embargo, algunas de las variables mencionadas no fueron incluidas en el modelo seleccionado debido a que no resultaron significativas o no cumplieron con los supuestos necesarios. El nivel de significancia fue de 0,05 tanto para los cociente de riesgos de las covariables, como para la prueba de los residuales escalados de Schoenfeld.

Seguido a esto, para verificar la estabilidad del modelo, siguiendo la metodología propuesta por Yang et al. (2020), se excluyeron los registros de las personas que fallecieron en los tres primeros años y se comparó con el primer modelo ajustado para seleccionar el modelo final. Finalmente, se identificó la localidad con los niveles más bajos de concentración y se calculó el cociente de riesgo con respecto a este grupo poblacional.

5.4. Curva de concentración - respuesta

A partir del resultado de la parte anterior, se realizó el ajuste del cociente de riesgo en función de los niveles estimados de $MP_{2,5}$. Para ello, se siguió la metodología del GEMM en el cual se propone la siguiente forma funcional

$$HR(z) = exp\left(\theta T(z) w(z)\right) \tag{5-1}$$

donde θ es un parámetro a ajustar, $z = C - a$ con C el nivel de concentración $MP_{2,5}$ y $a = min(C)$, $T(z) = ln(z+1)$ y $w(z)$ una función que sirve para controlar la forma de la curva (Apte et al. 2015; R. Burnett et al. 2018). En el presente caso de estudio se tomó $w(z) = 1$, que corresponde al caso más sencillo. Con ello el modelo ajustado se llevó a una regresión lineal con los supuestos usuales de la forma

$$ln(HR(z)) = \theta T(z) + \beta_0 + \epsilon, \tag{5-2}$$

y un nivel de significancia 0,05. Finalmente, se calculó el número de muertes prematuras relacionado a la exposición de $MP_{2,5}$ usando la ecuación 4-18, lo cual corresponde al exceso de muertes respecto a la población de referencia identificada en la sección 5.3.

6. Resultados y Discusión

En este capítulo se presentan los resultados obtenidos de acuerdo con la metodología explicada en el capítulo anterior, junto con la discusión de estos.

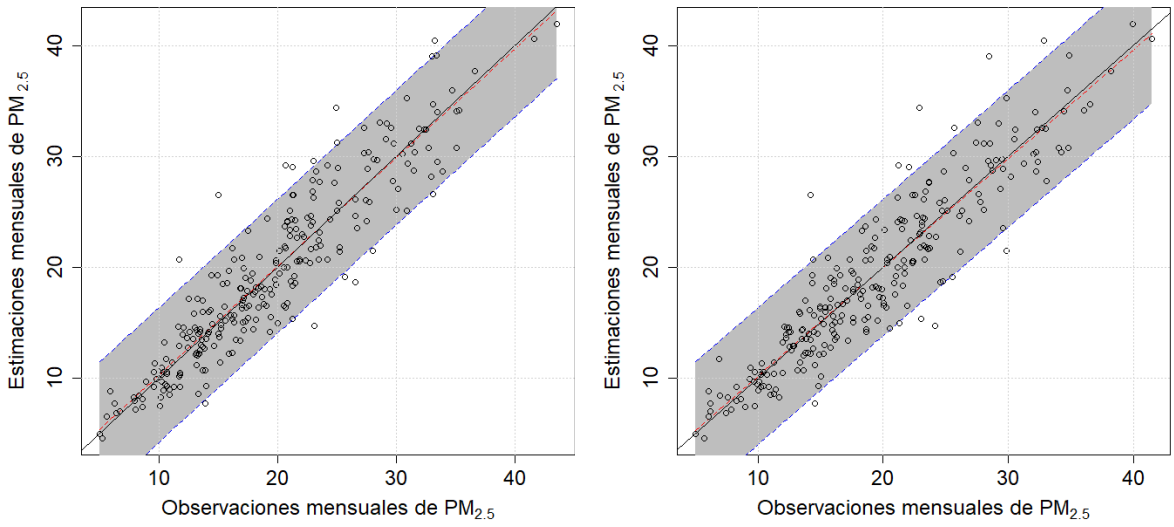
6.1. Estimación de niveles mensuales de $MP_{2,5}$ en Bogotá

***Bosque aleatorio* (RF)**

En este caso, el mejor ajuste inicial, aplicado a las celdas donde se ubican las estaciones de monitoreo, corresponde a un RF de 10 árboles con un error de predicción de $2{,}42 \mu g/m^3$ y un $R^2 = 0{,}83$. Este valor de R^2 se obtiene al realizar el ajuste lineal entre los valores mensuales estimados y observados de $MP_{2,5}$, como se muestra en la figura **6-1**a. En la figura, la línea negra indica que las estimaciones y los valores observados son iguales, la línea roja representa el ajuste y el área sombreada corresponde al intervalo de confianza.

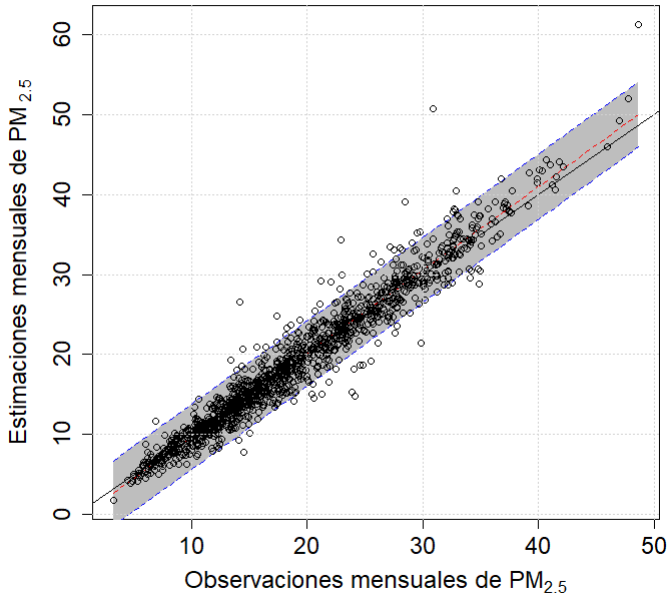
A partir de este modelo, se realizó una optimización de hiperparámetros donde se tomaron 12 árboles y cada regresión con 8 variables, lo que redujo el error de predicción a $2{,}34 \mu g/m^3$ y aumentó el R^2 a 0,85, como se muestra en la figura **6-1**b. Luego, al aplicar el modelo a todas las celdas de la cuadrícula, el R^2 aumentó a 0,93, lo cual se muestra en la figura **6-1**c.

El resultado obtenido es coherente con los resultados reportados por Liang et al. (2020), ya que en este caso la incorporación de las celdas sin mediciones produce un aumento de 0,11 en el R^2, mientras que en el estudio de Liang et al. (2020) el ajuste provocó un aumento de 0,31, pasando de un R^2 de 0,66 a 0,97. Es decir, el aumento en este estudio es cerca de un tercio del aumento reportado en estudios similares, lo cual se puede explicar porque este estudio contempla un área más pequeña y no se hace necesario el uso de datos de satélites.



(a) Ajuste lineal con modelo inicial para las estaciones $R^2 = 0{,}83$

(b) Ajuste lineal con modelo optimizado para las estaciones $R^2 = 0{,}85$



(c) Ajuste lineal con modelo optimizado para todas las celdas de la cuadrícula. $R^2 = 0{,}93$

Figura 6-1.: Ajustes lineales entre los valores mensuales estimados y observados de $MP_{2,5}$ para las estaciones y todas las celdas de la cuadrícula con el modelo RF.

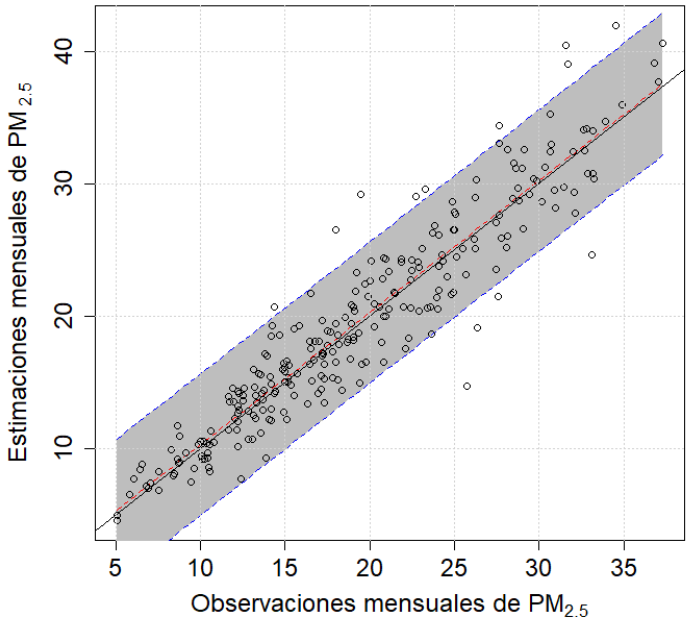
Refuerzo de gradiente extremo (XGBoost)

En este caso los mejores hiperparámetros encontrados fueron una profundidad máxima de 20, una tasa de aprendizaje de 0.01 y 535 árboles de decisión. Con ello se tuvo un error de predicción de $1{,}97\mu g/m^3$ y un $R^2 = 0{,}88$; lo que se muestra en la figura **6-2**a, donde la línea negra indica que las estimaciones y los valores observados son iguales, la roja el ajuste y el área sombreada el intervalo de confianza. Similar al procedimiento realizado para el *bosque aleatorio*, se aplicó el modelo para todas las celdas y se obtuvo un $R^2 = 0{,}96$; lo cual se muestra en la figura **6-2**b.

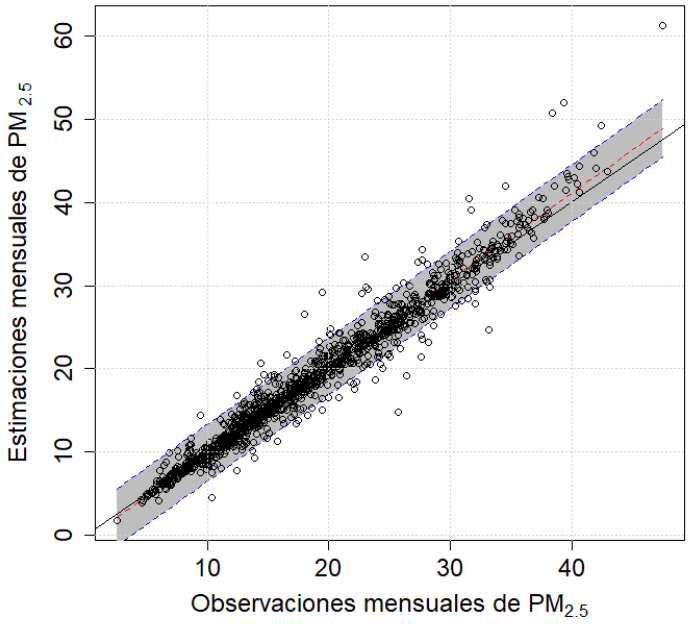
Estimación de los niveles anuales de $MP_{2,5}$ por localidad

Como se mencionó en la metodología, para obtener la estimación final de los niveles mensuales de $MP_{2,5}$, se tomó el valor promedio de los dos modelos. A partir de estas estimaciones mensuales, se calculó el nivel anual para resumir los resultados. En la figura **6-3**, se presentan los niveles anuales promedios de $MP_{2,5}$ por localidad en el periodo de 2008 al 2021. Los niveles más altos se muestran en rojo, los niveles medios en amarillo, y los niveles más bajos en verde. En general, se observa que las localidades de Bosa, Ciudad Bolívar y Kennedy presentan los niveles anuales más altos de $MP_{2,5}$, mientras que Usaquén, San Cristóbal y Suba son las localidades menos afectadas.

Por su parte, la Secretaría Distrital de Ambiente señala que en 2021 los niveles más altos se presentaron para el suroccidente de la ciudad, en las localidades de Kennedy, Bosa, Ciudad Bolívar y Tunjuelito con valores superiores a $21{,}0\mu g/m^3$ (de Ambiente, 2022). Al comparar con los resultados presentados, se encuentra el mismo comportamiento para las localidades mencionadas y también para Fontibón. Por otro lado, Rodriguez-Villamizar et al. (2022) presentan las estimaciones de los niveles de $MP_{2,5}$ a nivel nacional para 2019; donde señalan que en el caso de Bogotá se tienen valores superiores a $20{,}0\mu g/m^3$. Este límite es superior al estimado de $17{,}2\mu g/m^3$. Sin embargo, lo anterior contrasta con lo presentado por Farrow et al. (2022) para el 2021, pues señala el nivel anual en $13,7\mu g/m^3$; lo cual está por debajo del valor estimado de $19{,}2\mu g/m^3$. De esta manera, los resultados presentados para las estimaciones de los niveles anuales de $MP_{2,5}$ son coherentes con los valores reportados en la literatura.



(a) Ajuste lineal con modelo inicial para las estaciones $R^2 = 0{,}88$



(b) Ajuste lineal con modelo optimizado para todas las celdas de la cuadricula. $R^2 = 0{,}96$

Figura 6-2.: Ajustes lineales entre los valores mensuales estimados y observados de $MP_{2,5}$ para las estaciones y todas las celdas de la cuadrícula con el modelo *XGBoost*.

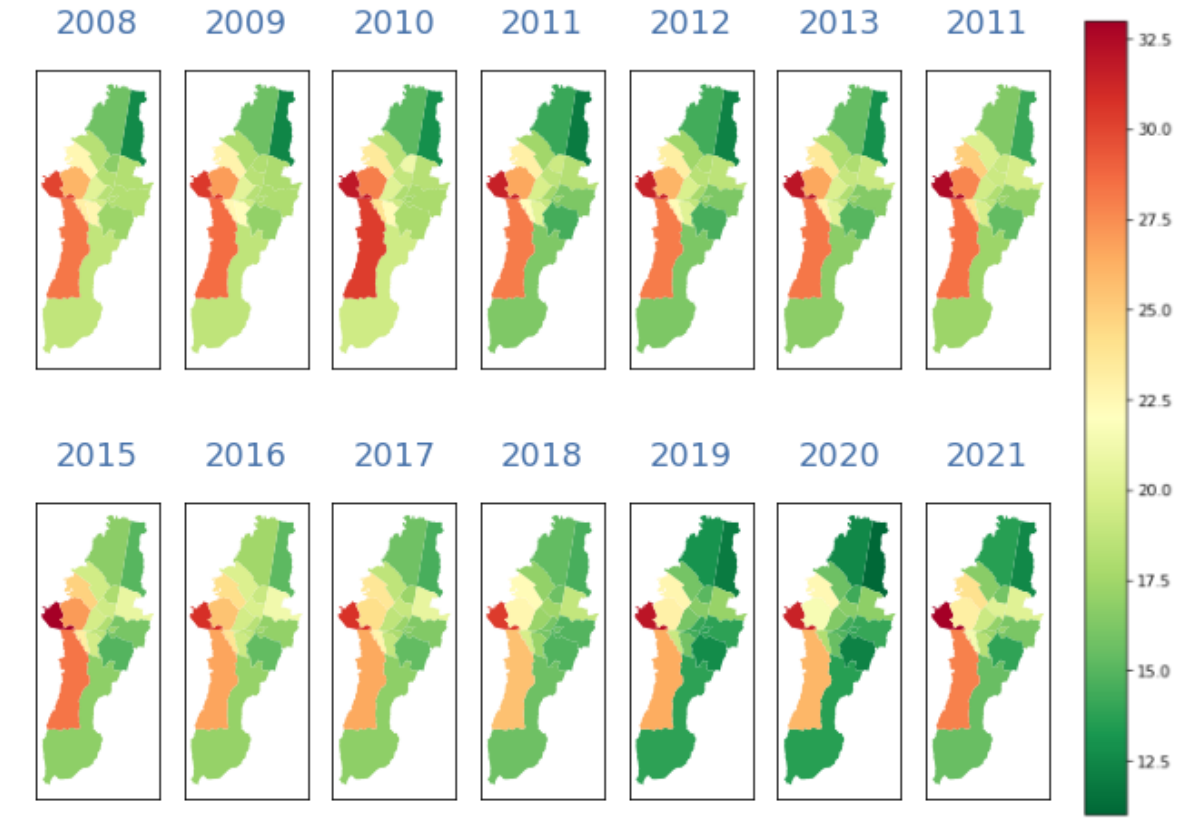


Figura 6-3.: Estimaciones de los niveles anuales de $MP_{2,5}$ en $\mu g/m^3$ por localidad para Bogotá desde 2008 a 2021.

6.2. Modelamiento del cociente de riesgo

En el tabla **6-1** se muestran los resultados para el modelo proporcional de Cox, teniendo en cuenta toda la base de defunciones entre 2008 y 2021. Para ello se presenta el nombre de la covariable, sus cociente de riesgos[1], su intervalo de confianza del 95 % y los p-valores, junto con el p-valor para la prueba de los residuales escalados de Schoenfeld [2]. Así, se tiene que ser hombre aumenta el riesgo respecto a ser mujer con un HR de 1,022 ($95\,\%IC, 1{,}011 - 1{,}033$); mientras no estar casado aumenta el riesgo respecto a estar casado con un HR de 1,205 ($95\,\%IC, 1{,}117 - 1{,}299$); por su parte no tener seguridad social subsidiada podría aumentar el riesgo con un HR de 1,034 ($95\,\%IC, 1{,}021 - 1{,}046$); y haber alcanzado estudios de doctorado aumentar el riesgo con un HR de 1,551 ($95\,\%IC, 1{,}311 - 1{,}835$). Estas covariables

[1]Los cociente de riesgos de las covariables son diferentes al cociente de riesgo que se define para comparar las poblaciones expuestas a altas y bajas concentraciones de $MP_{2,5}$.

[2]En las variables incluidas en el modelo no se incluyeron variables relacionadas con la edad porque no fueron significativas o no cumplieron los supuestos. Esto se puede dar por la calidad de los datos, porque se encontraban en grupos quinquenales. En el anexo 1 se muestran los resultados incluyendo la edad

podrían incluirse en el modelo, por ello es necesario ver la estabilidad de este.

Posteriormente, siguiendo la metodología de Yang et al. (2020) para ver la estabilidad del modelo, se excluyeron los registros de las personas que fallecieron en los tres primeros años, con ello se obtuvieron que las cuatro covariables mencionadas son significativas y cumplen los supuestos como se muestra en el tabla **6-2**. Además, los valores obtenidos son muy similares a los presentados en el tabla **6-1**.

Covariable	HR	Intervalo 95 %	p-valor Modelo de Cox	p-valor riesgos proporcionales
Hombre	1,022	(1,011 – 1,033)	$< 0{,}001^{***}$	0,117
No Casado	1,205	(1,117 – 1,299)	$< 0{,}001^{***}$	0,327
No Subsidiado	1,034	(1,021 – 1,046)	$< 0{,}001^{***}$	0,015
Nivel Educativo: Doctorado	1,551	(1,311 – 1,835)	$< 0{,}001^{***}$	0,313

Tabla 6-1.: Resumen del modelo de Cox para todas la base de defunciones

Covariable	HR	Intervalo 95 %	p-valor Modelo de Cox	p-valor riesgos proporcionales
Hombre	1,020	(1,008 – 1,032)	$0{,}001^{**}$	0,20
No Casado	1,122	(1,028 – 1,226)	$0{,}010^{*}$	0,46
No Subsidiado	1,019	(1,005 – 1,032)	$0{,}001^{**}$	0,18
Nivel Educativo: Doctorado	1,346	(1,086 – 1,670)	$0{,}001^{**}$	0,40

Tabla 6-2.: Resumen del modelo de Cox excluyendo las defunciones de los tres primeros años.

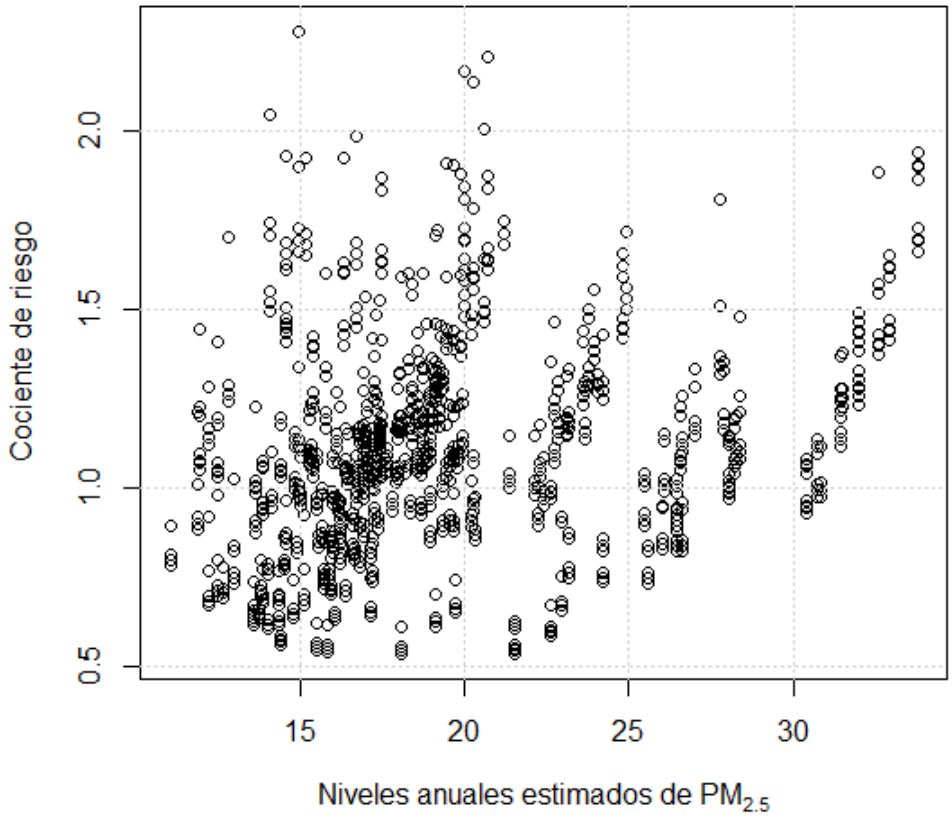


Figura 6-4.: Cociente de riesgo contra niveles anuales estimados de $MP_{2,5}$ con los resultados del modelo presentado en el tabla **6-2**

Debido a lo mencionado anteriormente, en este proyecto se consideró toda la población disponible para realizar el estudio y se tomaron los resultados del modelo presentado en el tabla anterior para calcular el cociente de riesgo. De esta manera, se identificó la localidad de Usaquén como la localidad con más bajos niveles de concentración de $MP_{2,5}$; pues en promedio anual son menores a $15{,}15\mu g/m^3$. Por esta razón se toma la población expuesta a niveles de $MP_{2,5}$ menores o iguales a $15{,}15\mu g/m^3$ como población de referencia, es decir $HR = 1$. Los resultados del cociente de riesgo en función de los niveles anuales estimados de $MP_{2,5}$ para toda la población se muestran en la figura **6-4**. Se puede observar que los valores se concentran para niveles de $MP_{2,5}$ por debajo de $25\mu g/m^3$, pero también se evidencia la presencia de valores atípicos de HR por encima de 1,6.

Por lo tanto, se llevó a cabo un proceso de limpieza y suavizado de los datos para ajustar la curva de concentración-respuesta. Dicho proceso consistió en eliminar los valores atípicos

para el cociente de riesgo, ordenar los registros, agruparlos en bloques de mil registros y tomar la mediana del cociente de riesgo y de $MP_{2,5}$ para cada bloque. Con ello, se obtiene el resultado final de la estimación del cociente de riesgo mostrado en la figura **6-5**.

Es importante tener en cuenta que en el cálculo del cociente de riesgo sólo se tuvo en cuenta la información de las muertes cardio - metabólicas. Debido a que la información de individuos vivos no estaba disponible. En el caso que se incluyan estos datos se puede tener un resultado diferente en el ajuste, porque las funciones de supervivencia cambian.

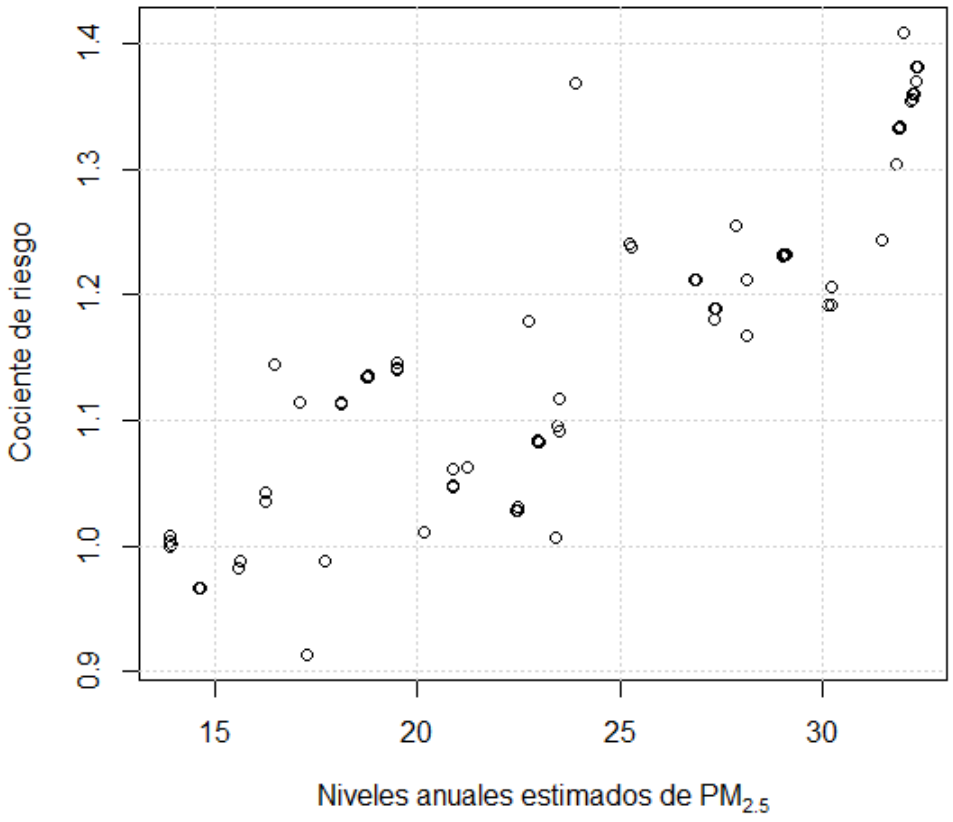


Figura 6-5.: Cociente de riesgo contra niveles anuales estimados de $MP_{2,5}$ con los resultados del modelo presentado en el tabla **6-2**, después del proceso de limpieza.

6.3. Ajuste curva CR

A partir del resultado obtenido en la sección anterior, se realizó el ajuste del cociente de riesgo en función de los niveles estimados de $MP_{2,5}$ utilizando la ecuación 5-2. De esta manera,

se obtuvo un valor estimado de $\hat{\theta} = 0{,}105$, $R^2 = 0{,}64$ y estadísticamente significativa (p-valor menor a 0,001. La curva ajustada se muestra en la figura **6-6**, donde la línea roja punteada corresponde al ajuste y el área sombreada representa el intervalo de confianza del 95 % para la predicción. Esta curva es consistente con lo reportado en la literatura para otras curvas concentración-respuesta, como se muestra en la figura **6-7**.

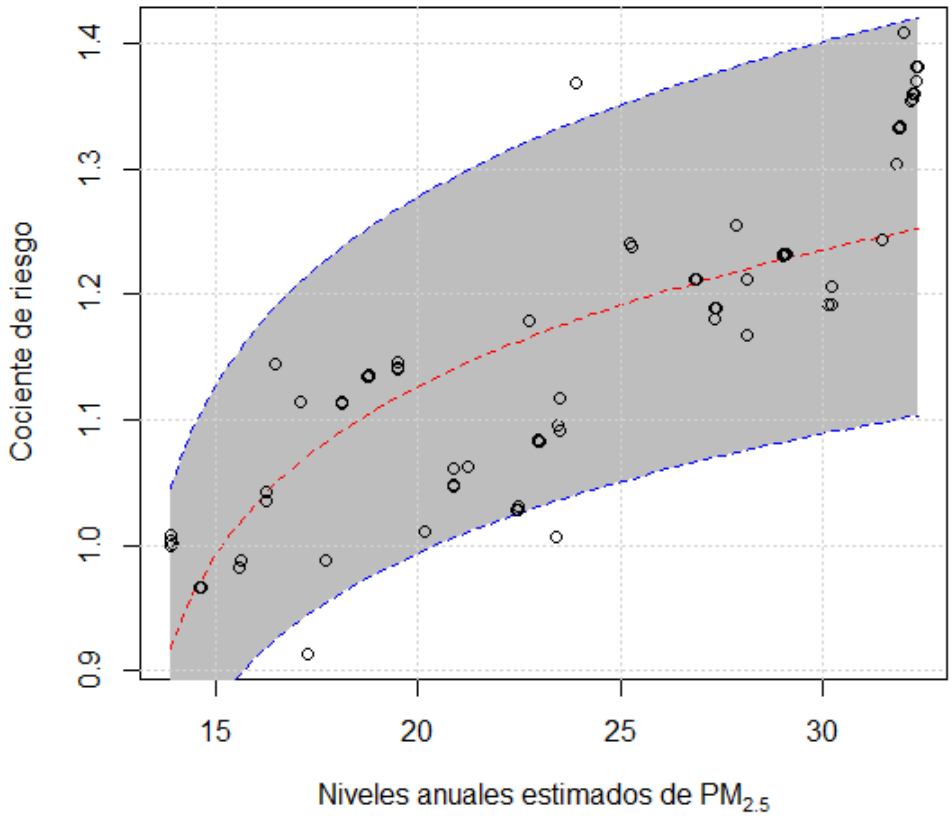


Figura 6-6.: Curva concentración - respuesta hallada para Bogotá durante los años 2008 al 2021. La línea roja corresponde al ajuste y el área sombreada al intervalo de predicción.

Por otro lado, en la figura **6-8** se muestra un resumen de la validación de los supuestos del modelo, que corresponden a los supuestos usuales de una regresión lineal. Así, en la gráfica de residuales vs valores ajustados se observa que los residuales no presentan una dispersión grande; con lo cual se tiene una varianza constante. Por otro lado, el Q-Q plot se ajusta bien a una línea y al aplicar la prueba de Shapiro - Wilk se concluye que no se rechaza el supuesto de normalidad, pues el p-valor es 0,29 es mayor al nivel de significancia definido de

0,05. Además, al ver la gráfica de los Leverage Points se tiene que no hay puntos influyentes. De esta manera, la curva de concentración - respuesta presentada cumple con los supuestos.

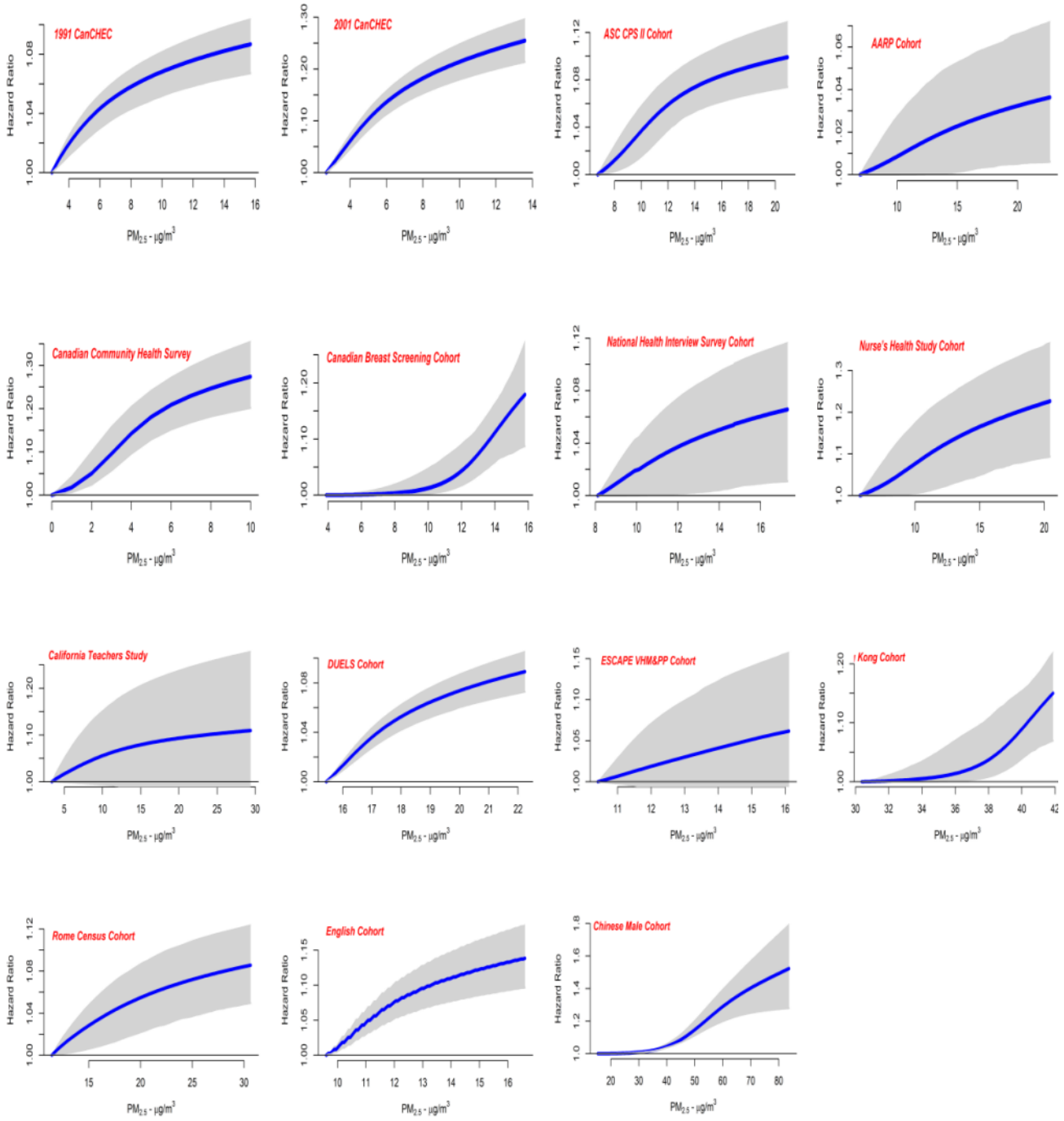


Figura 6-7.: Tipos de curva concentración - respuesta para ejercicios similares. En azul los ajustes y las áreas en gris corresponden a los intervalos de confianza. Fuente: Tomado de R. Burnett et al. (2018).

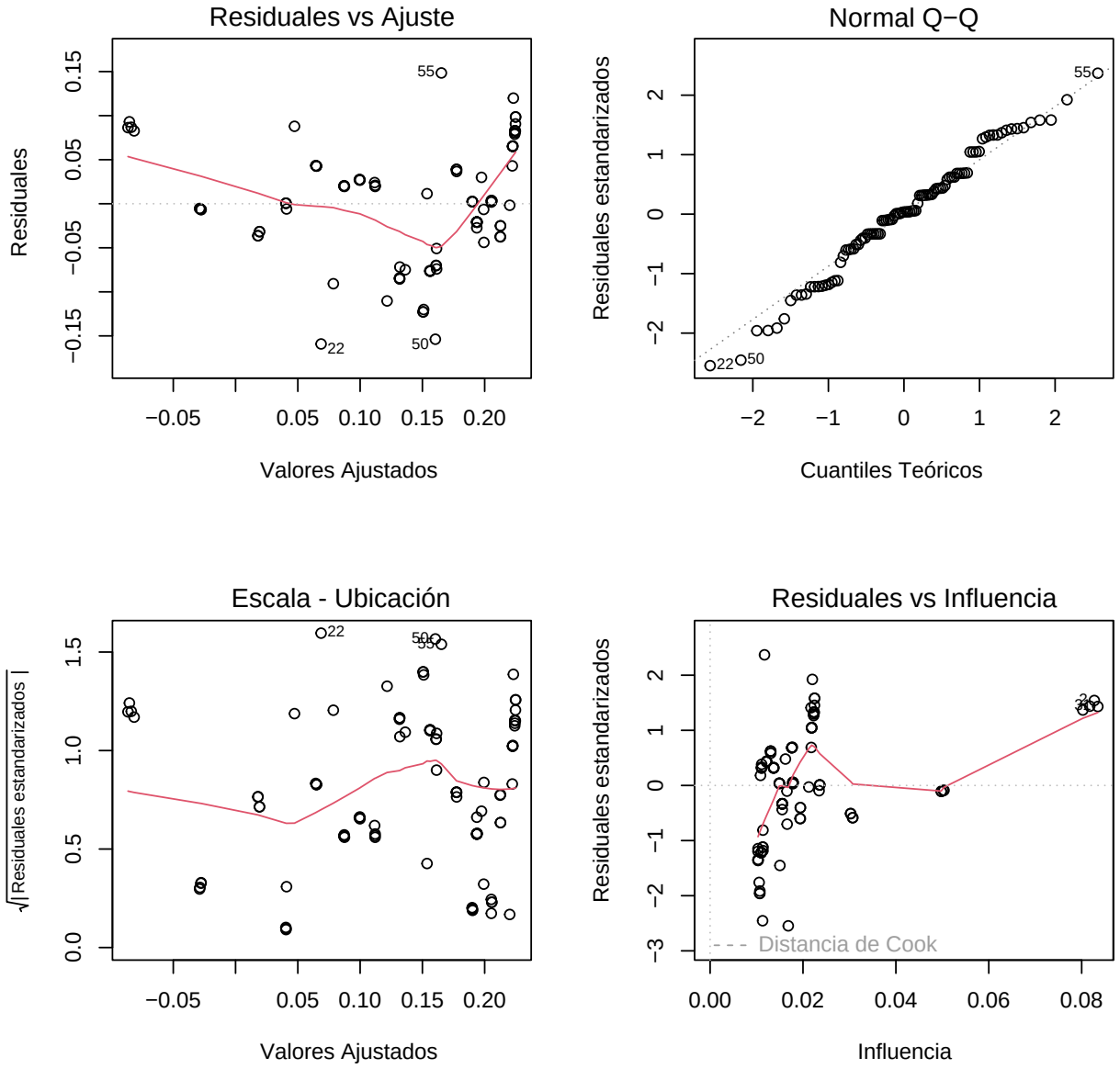


Figura 6-8.: Validación de supuestos para el ajuste de la curva concentración - respuesta hallada para Bogotá durante los años 2008 al 2021.

6.4. Estimación del número de muertes prematuras relacionadas con la exposición de $MP_{2,5}$

En cuanto a la estimación del número de muertes prematuras relacionadas con la exposición de $MP_{2,5}$ se utilizó la ecuación 4-18. De esta manera en la figura **6-9** se muestra el exceso de

muertes relacionado con la prolongada exposición a material $MP_{2,5}$, tomando como referencia la población de la localidad de Usaquén. En general se observa que al aumentar los niveles de $MP_{2,5}$ aumentan el número de muertes estimadas. Además, los años con mayor cantidad de muertes prematuras son 2016 y 2021, mientras el menor año corresponde al 2020. Esto último se puede explicar por el impacto de los aislamientos en los niveles de contaminación en el aire.

Por su parte, Rodriguez-Villamizar et al. (2022) estiman que entre los años 2014 y 2019 se presentaron en Bogotá más de 500 muertes anuales prematuras, debidas a la exposición prolongada a niveles superiores a $15\mu g/m^3$ de $MP_{2,5}$. Al comparar esto con los resultados encontrados se tiene que son coherentes, pues las muertes anuales para este periodo están en el rango de 869 a 1430. Por otro lado, de acuerdo con Farrow et al. (2022) el número de muertes prematuras para 2021 fue de 3400; mientras las encontradas para ese año fueron 1444, lo que representa el 42 % de las 3400. Esta diferencia se puede presentar por el nivel de referencia para calcular las muertes prematuras, pues Farrow et al. (2022) usaron un valor de $13,7\mu g/m^3$ mientras en el presente trabajo $15{,}2\mu g/m^3$.

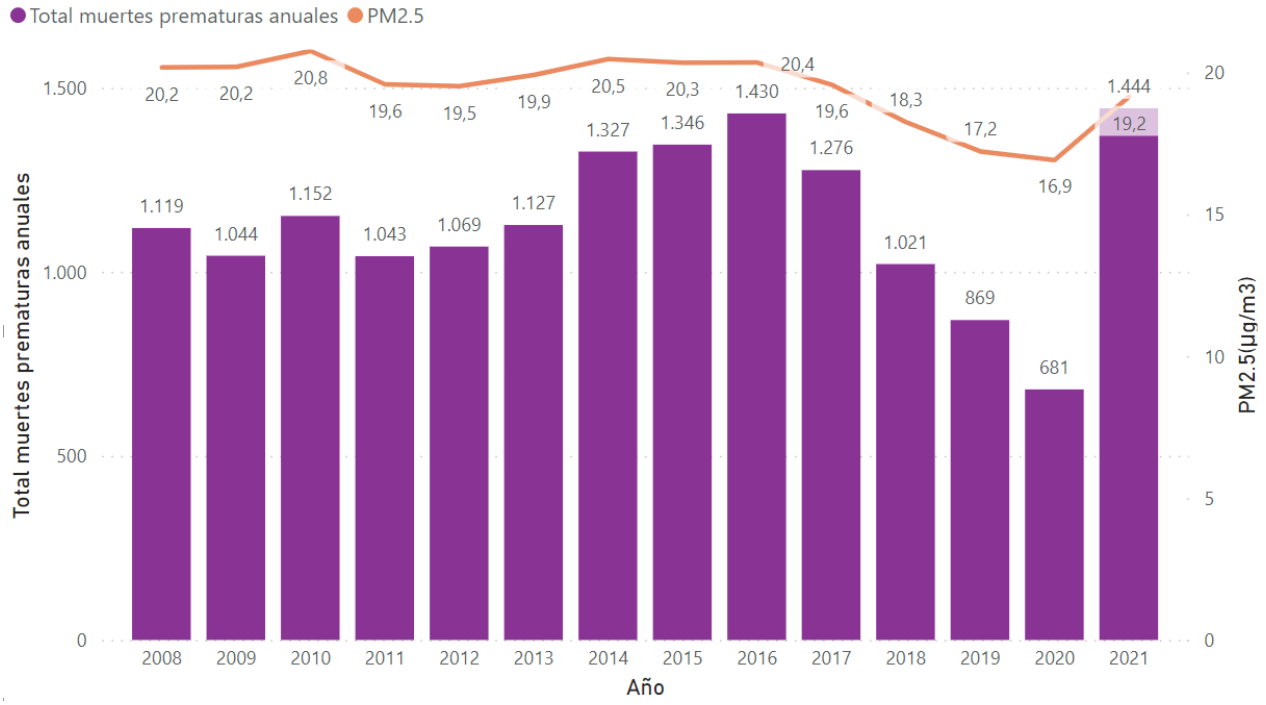


Figura 6-9.: Muertes prematuras anuales estimadas relacionadas con la larga exposición a material particulado $MP_{2,5}$ en Bogotá durante los años 2008 y 2021, para niveles por encima de $15{,}15\mu g/m^3$.

Además, en la figura **6-10** se observa el número total de muertes prematuras relacionadas con la prolongada exposición a material $MP_{2,5}$ para cada una de las localidades. En el caso de

la localidad de Usaquén el número de muertes es cero porque esta localidad se encuentra por debajo del nivel de $MP_{2,5} = 15{,}15\mu g/m^3$, y sirve como población para calcular el cociente de riesgo. A partir de los resultados obtenidos se observa que las localidades más afectadas son las localidades del suroccidente de la ciudad. Estas son Kennedy, Bosa y Ciudad Bolívar. Esto es coherente con los resultados presentados por Bonilla et al. (2021) donde señala que el suroccidente de la ciudad es el más afectado por la calidad del aire, y con lo ya discutido en la estimación de los niveles de $MP_{2,5}$.

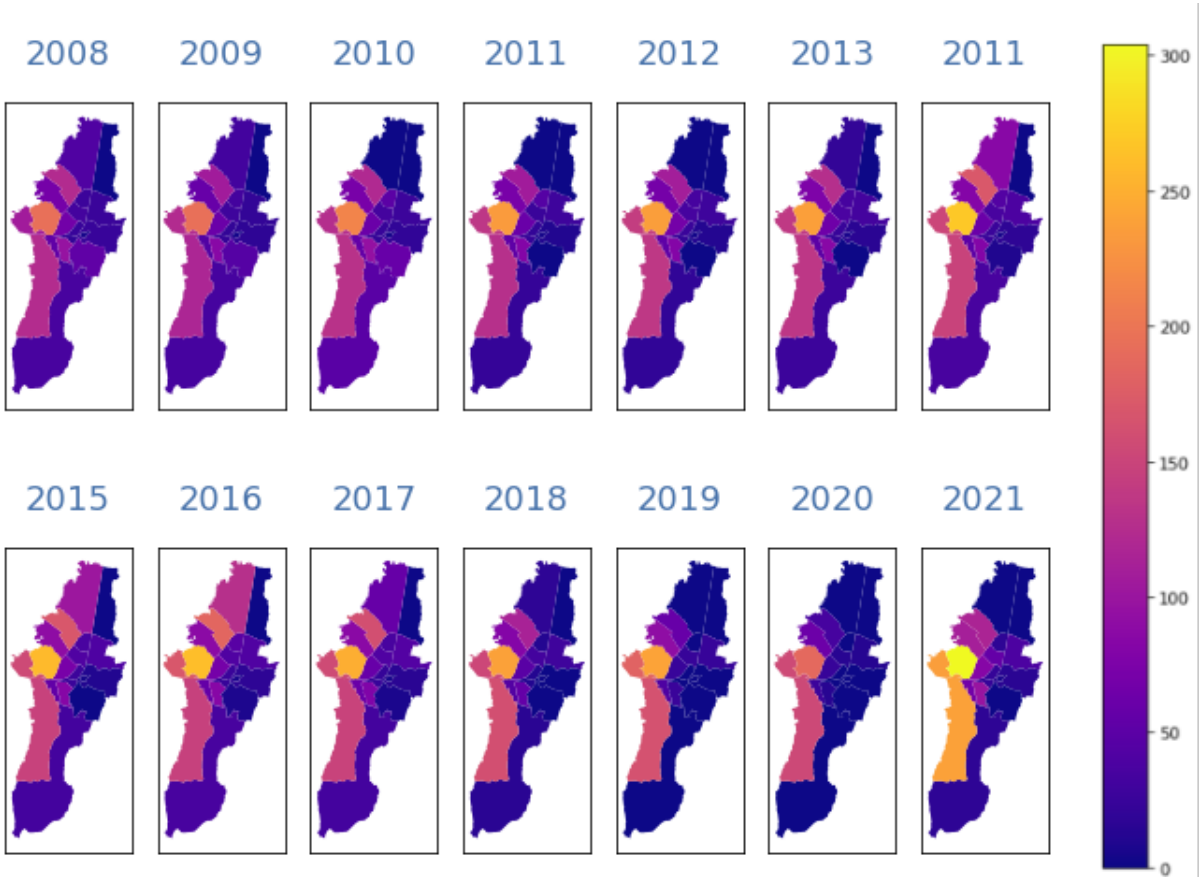


Figura 6-10.: Muertes prematuras anuales relacionadas a la larga exposición a material particulado $MP_{2,5}$ en Bogotá por localidades durante los años 2008 y 2021. Para niveles por encima de $15{,}15\mu g/m^3$

Por otro lado, en la figura **6-11** se muestra a que porcentaje del número de muertes cardio metabólicas corresponden las muertes prematuras estimadas por exposición prolongada a $MP_{2,5}$, por localidad y año. De esta manera, se puede observar que las localidades de Bosa, Kennedy y Ciudad Bolívar son las más afectadas, con el 24,4 %; 22,3 % y 20,7 % respectivamente en todo el periodo de análisis. Mientras para Bogotá se tiene que el porcentaje anual varía entre 8,2 % y 14,2 % de las muertes cardio metabólicas corresponde a muertes prema-

turas por exposición a $MP_{2,5}$. En particular, para 2018 se tiene un valor de 9,7 %, el cual está por encima del valor dado por el Instituto Nacional de Salud de 8 % (Instituto Nacional de Salud, 2018). Además, para los años 2019 y 2020 se ve un descenso de los porcentajes para la mayoría de localidades, lo cual se puede explicar por la reducción de los niveles de $MP_{2,5}$ para esos años, lo cual se acentúa en 2020 con los aislamientos.

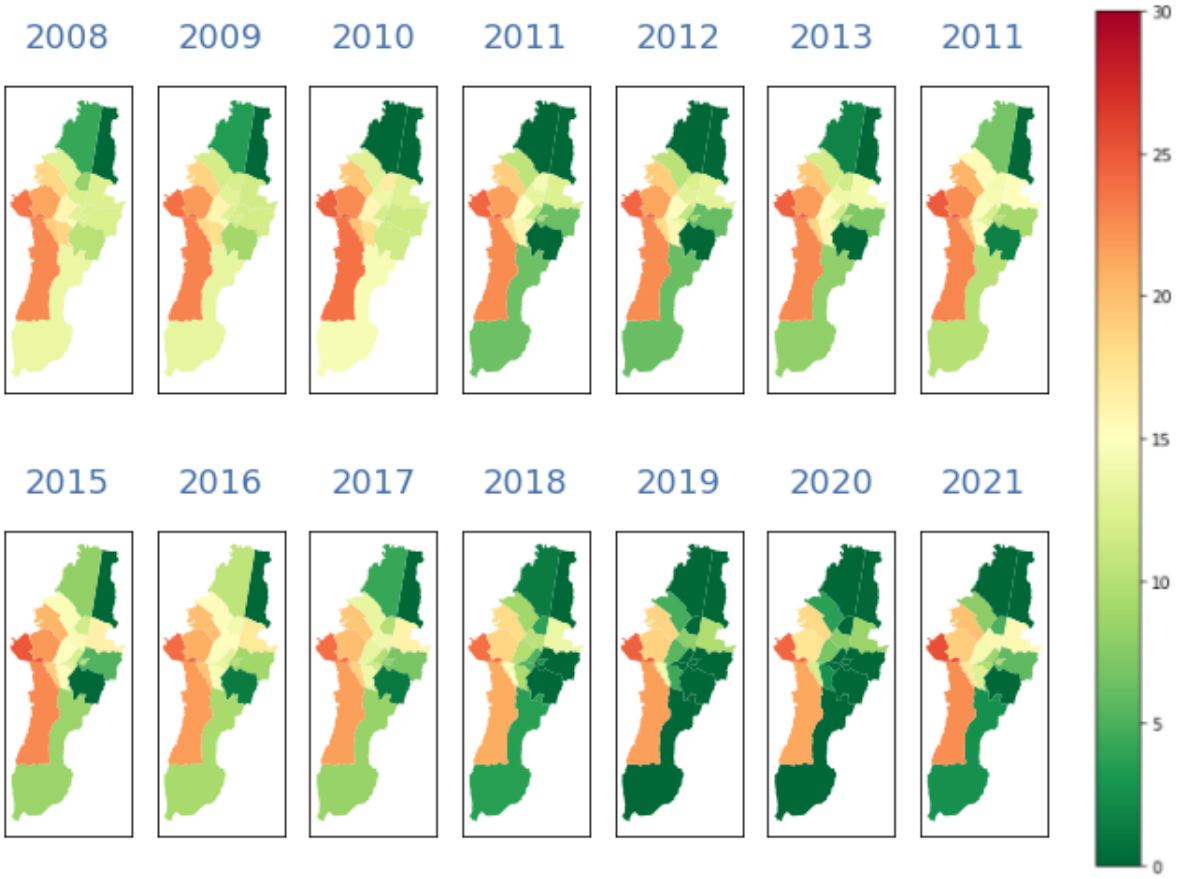


Figura 6-11.: Porcentaje de las muertes prematuras estimadas por exposición a $MP_{2,5}$ respecto a las muertes cardio metabólicas en Bogotá por localidad del 2008 al 2021.

Por último, es crucial considerar que los años 2020 y 2021 estuvieron marcados por el impacto de la pandemia de COVID-19. En este contexto, la Organización Mundial de la Salud proporcionó directrices para el manejo de clasificaciones de causas de muerte distintas a las cardio-metabólicas. Sin embargo, los efectos de la pandemia aún están siendo objeto de estudio, incluyendo las muertes cardio-metabólicas asociadas con la situación pandémica y no necesariamente vinculadas a los niveles de contaminación.

7. Conclusiones

En este trabajo se realizó la estimación del número de muertes prematuras relacionadas con la exposición prolongada a material particulado $MP_{2,5}$ en la ciudad de Bogotá, durante el período comprendido entre 2008 y 2021. Para llevar a cabo este proceso, se construyó una base de datos mensual que recopila información sobre las muertes debidas a causas cardio metabólicas entre los habitantes mayores de 20 años de cada localidad de la ciudad. Esta base de datos contiene detalles como la causa de la muerte y las características socioeconómicas de cada individuo. La consolidación de esta información representó un esfuerzo significativo, ya que implicó la integración de datos provenientes de distintas fuentes. Fue necesario unificar los datos publicados por el Departamento Administrativo Nacional de Estadística (DANE) relacionados con las estadísticas vitales, así como las bases de datos de causas de muerte proporcionadas por la Secretaría Distrital de Salud. En este proceso, se logró recopilar información de un total de 129 622 registros, a partir de un conjunto inicial de 134 269 muertes registradas por el DANE, que se utilizaron en el análisis y estudio posterior.

En cuanto a la tendencia espacio temporal de los niveles de $MP_{2,5}$ se desarrollaron os modelos cuyas estimaciones fueron promediadas para obtener el resultado final. Este enfoque permitió calcular los niveles de $MP_{2,5}$ con una resolución de 2km de forma mensual con un $R = 0{,}97$. Esto permitió ver el comportamiento de los niveles de $MP_{2,5}$ por localidad de forma anual, en donde se observaron valores consistentes con lo reportado en otros estudios; y una mayor afectación para las localidades del suroccidente de la ciudad. Así mismo, para el modelamiento del cociente de riesgo se ajustó un modelo de Cox con cuatro covariables socioeconómicas: el género, el estado civil, el tipo del seguro social y su nivel educativo. Esto permitió identificar a la población expuesta a niveles de $MP_{2,5}$ iguales o inferiores a $15,15 \mu g/m^3$ como la población de referencia para estimar las muertes prematuras relacionadas con la larga exposición a $MP_{2,5}$.

Por otro lado, para la curva concentración - respuesta se llevó el modelo a una forma lineal, donde se obtuvo un $R^2 = 0{,}64$. Con ello se puede identificar el comportamiento para cada localidad y estación de monitoreo de calidad del aire en la ciudad. El análisis reveló que las localidades de Bosa, Kennedy, Ciudad Bolívar y Fontibón han sido las más afectadas durante el período de 2008 a 2021 debido a la exposición prolongada a $MP_{2,5}$. Estos hallazgos se alinean con los datos presentes en la literatura y ofrecen una mayor granularidad, ya que permiten identificar el impacto de la contaminación del aire en cada localidad de la ciudad.

Por tanto, este estudio aporta un análisis que puede respaldar la formulación de políticas para mejorar la salud pública en Bogotá. Se recomienda dar prioridad a las localidades del suroccidente de la ciudad, como sugieren otros estudios, y también considerar a Fontibón en estas políticas.

Finalmente, para mejorar la precisión en los modelos que estiman los niveles de $MP_{2,5}$, es fundamental contar con la información que será registrada por las nuevas estaciones de monitoreo, lo que permitirá una mayor resolución espacial en futuros análisis. Asimismo, la incorporación de estudios de cohortes podría ayudar a identificar con mayor detalle la población más vulnerable a la contaminación del aire. En cuanto a futuras investigaciones, se propone aplicar una metodología similar para otras ciudades del país, como Medellín o Cali, con el fin de evaluar el impacto de la contaminación del aire en la salud en áreas más pequeñas y así obtener una visión más completa de la problemática a nivel nacional.

A. Anexo: Modelamiento del cociente de riesgo incluyendo grupos de edad

En los siguientes resultados se muestran los grupos de edad de forma quinquenal para mayores o iguales a 50 años.

Covariable	HR	Intervalo 95 %	p-valor Modelo de Cox	p-valor riesgos proporcionales
Hombre	1,046	(1,034 – 1,057)	$< 0{,}001^{***}$	0,174
No Casado	1,169	(1,084 – 1,261)	$< 0{,}001^{***}$	0,379
No Subsidiado	1,041	(1,029 – 1,055)	$< 0{,}001^{***}$	$0{,}024^{*}$
Nivel Educativo: Doctorado	1,565	(1,322 – 1,852)	$< 0{,}001^{***}$	0,313
Edad: 50 a 54 años	1,064	(1,023 – 1,107)	$0{,}002^{**}$	$< 0{,}001^{***}$
Edad: 55 a 59 años	0,956	(0,914 – 0,993)	$0{,}022^{*}$	$< 0{,}001^{***}$
Edad: 60 a 64 años	0,961	(0,930 – 0,992)	$0{,}014^{*}$	$< 0{,}001^{***}$
Edad: 65 a 69 años	1,010	(0,982 – 1,038)	0,474	$< 0{,}001^{***}$
Edad: 70 a 74 años	1,030	(1,005 – 1,056)	$0{,}016^{*}$	$< 0{,}001^{***}$
Edad: 75 a 79 años	1,003	(0,982 – 1,025)	0,798	$< 0{,}001^{***}$
Edad: 80 a 84 años	0,982	(0,963 – 1,002)	0,080	$< 0{,}001^{***}$
Edad: 85 a 89 años	0,928	(0,909 – 0,945)	$< 0{,}001^{***}$	$< 0{,}001^{***}$
Edad: Mayor a 90 años	0,898	(0,880 – 0,917)	$< 0{,}001^{***}$	$< 0{,}001^{***}$

Tabla A-1.: Resumen del modelo de Cox para todas la base de defunciones incluyendo los grupos de edad mayores o iguales a 50 años.

Covariable	HR	Intervalo 95 %	p-valor Modelo de Cox	p-valor riesgos proporcionales
Hombre	1,034	(1,021 – 1,046)	$< 0{,}001^{***}$	0,239
No Casado	1,112	(1,018 – 1,215)	$< 0{,}018^{*}$	0,418
No Subsidiado	1,024	(1,010 – 1,038)	$< 0{,}001^{***}$	0,155
Nivel Educativo: Doctorado	1,356	(1,094 – 1,682)	$< 0{,}001^{***}$	0,393
Edad: 50 a 54 años	1,072	(1,025 – 1,120)	$0{,}002^{**}$	$0{,}010^{*}$
Edad: 55 a 59 años	0,976	(0,935 – 1,020)	0,287	$0{,}010^{*}$
Edad: 60 a 64 años	0,960	(0,927 – 0,995)	$0{,}025^{*}$	$< 0{,}001^{***}$
Edad: 65 a 69 años	1,005	(0,974 – 1,036)	0,767	$0{,}014^{*}$
Edad: 70 a 74 años	1,034	(1,006 – 1,062)	$0{,}015^{*}$	$< 0{,}001^{***}$
Edad: 75 a 79 años	1,008	(0,983 – 1,032)	0,532	$< 0{,}001^{***}$
Edad: 80 a 84 años	1,001	(0,979 – 1,023)	0,921	$< 0{,}001^{***}$
Edad: 85 a 89 años	0,952	(0,931 – 0,972)	$< 0{,}001^{***}$	$< 0{,}001^{***}$
Edad: Mayor a 90 años	0,910	(0,890 – 0,930)	$< 0{,}001^{***}$	$< 0{,}001^{***}$

Tabla A-2.: Resumen del modelo de Cox excluyendo las defunciones de los tres primeros años, incluyendo los grupos de edad mayores o iguales a 50 años.

Bibliografía

Apte, J. S., Marshall, J. D., Cohen, A. J., & Brauer, M. (2015). Addressing global mortality from ambient PM2. 5. *Environmental science & technology*, *49*(13), 8057-8066.

Arregocés, H. A., Rojano, R., & Restrepo, G. (2023). Health risk assessment for particulate matter: application of AirQ+ model in the northern Caribbean region of Colombia. *Air Quality, Atmosphere & Health*, 1-16.

Blanco-Becerra, L. C., Miranda-Soberanis, V., Hernández-Cadena, L., Barraza-Villarreal, A., Junger, W., Hurtado-Dıaz, M., & Romieu, I. (2014). Effect of particulate matter less than 10μm (PM10) on mortality in Bogota, Colombia: a time-series analysis, 1998-2006. *salud pública de méxico*, *56*, 363-370.

Bonilla, J. A., Morales-Betancourt, R., & Aravena, C. (2021). Análisis de desigualdades múltiples y polıticas de reducción de la contaminación.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

Bureau, P. R. (2007). *Population: A Lively Introduction* (Vol. 62).

Burnett, R., Chen, H., Szyszkowicz, M., Fann, N., Hubbell, B., Pope III, C. A., Apte, J. S., Brauer, M., Cohen, A., Weichenthal, S., et al. (2018). Global estimates of mortality associated with long-term exposure to outdoor fine particulate matter. *Proceedings of the National Academy of Sciences*, *115*(38), 9592-9597.

Burnett, R. T., Pope III, C. A., Ezzati, M., Olives, C., Lim, S. S., Mehta, S., Shin, H. H., Singh, G., Hubbell, B., Brauer, M., et al. (2014). An integrated risk function for estimating the global burden of disease attributable to ambient fine particulate matter exposure. *Environmental health perspectives*, *122*(4), 397-403.

Casallas, A., Celis, N., Ferro, C., López Barrera, E., Peña, C., Corredor, J., & Ballen Segura, M. (2020). Validation of PM10 and PM2. 5 early alert in Bogotá, Colombia, through the modeling software WRF-CHEM. *Environmental Science and Pollution Research*, *27*(29), 35930-35940.

Casallas, A., Ferro, C., Celis, N., Guevara-Luna, M. A., Mogollón-Sotelo, C., Guevara-Luna, F. A., & Merchán, M. (2021). Long short-term memory artificial neural network approach to forecast meteorology and pm2. 5 local variables in bogotá, colombia. *Modeling Earth Systems and Environment*, 1-14.

Cheng, Q., Qu, C., Wang, Y., Wang, X., He, R., Cao, H., Liu, B., Zhang, H., Zhang, N., Lai, Z., et al. (2023). Global burden and its association with socioeconomic development status of meningitis caused by specific pathogens over the past 30 years: a population-based study. *Neuroepidemiology*, 1-1.

Chowdhury, S., Dey, S., & Smith, K. R. (2018). Ambient PM2. 5 exposure and expected premature mortality to 2100 in India under climate change scenarios. *Nature communications*, *9*(1), 318.

Cox, D. R. (1997). Some remarks on the analysis of survival data. *Proceedings of the First Seattle Symposium in Biostatistics*, 1-9.

David, C. R., et al. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society*, *34*(2), 187-220.

de Ambiente, S. D. (2022). Informe anual de calidad del aire de Bogotá Año 2021.

Dockery, D. W., Pope, C. A., Xu, X., Spengler, J. D., Ware, J. H., Fay, M. E., Ferris Jr, B. G., & Speizer, F. E. (1993). An association between air pollution and mortality in six US cities. *New England journal of medicine*, *329*(24), 1753-1759.

Farrow, A., Anhäuser, A., Chen, Y. J., & Cespedes, T. (2022). La carga de la contaminación del aire en Bogotá, Colombia 2021.

Gakidou, E., Afshin, A., Abajobir, A. A., Abate, K. H., Abbafati, C., Abbas, K. M., Abd-Allah, F., Abdulle, A. M., Abera, S. F., Aboyans, V., et al. (2017). Global, regional, and national comparative risk assessment of 84 behavioural, environmental and occupational, and metabolic risks or clusters of risks, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. *The Lancet*, *390*(10100), 1345-1422.

Grisales-Romero, H., Piñeros-Jiménez, J. G., Nieto, E., Porras-Cataño, S., Montealegre, N., González, D., & Ospina, D. (2021). Local attributable burden disease to PM 2.5 ambient air pollution in Medellın, Colombia, 2010–2016. *F1000Research*, *10*.

Han, C., Kim, S., Lim, Y.-H., Bae, H.-J., & Hong, Y.-C. (2018). Spatial and temporal trends of number of deaths attributable to ambient PM2. 5 in the Korea. *Journal of Korean medical science*, *33*(30).

Instituto Nacional de Salud, O. N. d. S. (2018). Carga de enfermedad ambiental en Colombia. Décimo informe técnico especial.

Johnston, F. H., Borchers-Arriagada, N., Morgan, G. G., Jalaludin, B., Palmer, A. J., Williamson, G. J., & Bowman, D. M. (2021). Unprecedented health costs of smoke-related PM2. 5 from the 2019–20 Australian megafires. *Nature Sustainability*, *4*(1), 42-47.

Klein, J. P., Moeschberger, M. L., et al. (2003). *Survival analysis: techniques for censored and truncated data* (Vol. 1230). Springer.

Liang, F., Xiao, Q., Huang, K., Yang, X., Liu, F., Li, J., Lu, X., Liu, Y., & Gu, D. (2020). The 17-y spatiotemporal trend of PM2. 5 and its mortality burden in China. *Proceedings of the National Academy of Sciences*, *117*(41), 25601-25608.

Lozano, N. (2004). Air pollution in Bogota, Colombia: A concentration-response approach. *Revista Desarrollo y Sociedad*, (54), 133-177.

Mudu, P., Gapp, C., & Dunbar, M. (2018). *AirQ+: example of calculations* (inf. téc.). World Health Organization. Regional Office for Europe.

Murray, C. J., Aravkin, A. Y., Zheng, P., Abbafati, C., Abbas, K. M., Abbasi-Kangevari, M., Abd-Allah, F., Abdelalim, A., Abdollahi, M., Abdollahpour, I., et al. (2020). Global burden of 87 risk factors in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. *The Lancet*, *396*(10258), 1223-1249.

Ortiz-Durán, E. Y., & Rojas-Roa, N. Y. (2013). Estimating air quality change-associated health benefits by reducing PM10 in Bogotá. *Revista de Salud Pública*, *15*(1), 90-102.

Pope III, C. A., & Dockery, D. W. (2006). Health effects of fine particulate air pollution: lines that connect. *Journal of the air & waste management association*, *56*(6), 709-742.

Reis, I., Baron, D., & Shahaf, S. (2018). Probabilistic random forest: A machine learning algorithm for noisy data sets. *The Astronomical Journal*, *157*(1), 16.

Rodriguez-Villamizar, L. A., Belalcazar-Ceron, L. C., Castillo, M. P., Sanchez, E. R., Herrera, V., & Agudelo-Castañeda, D. M. (2022). Avoidable mortality due to long-term exposure to PM2. 5 in Colombia 2014–2019. *Environmental Health*, *21*(1), 137.

Sampson, P. D., Richards, M., Szpiro, A. A., Bergen, S., Sheppard, L., Larson, T. V., & Kaufman, J. D. (2013). A regionalized national universal kriging model using Partial Least Squares regression for estimating annual PM2. 5 concentrations in epidemiology. *Atmospheric environment*, *75*, 383-392.

Schapire, R. E., & Freund, Y. (2013). Boosting: Foundations and algorithms. *Kybernetes*.

Schoenfeld, D. (1982). Partial residuals for the proportional hazards regression model. *Biometrika*, *69*(1), 239-241.

Southerland, V. A., Brauer, M., Mohegh, A., Hammer, M. S., Van Donkelaar, A., Martin, R. V., Apte, J. S., & Anenberg, S. C. (2022). Global urban temporal trends in fine particulate matter (PM2· 5) and attributable health burdens: estimates from global datasets. *The Lancet Planetary Health*, *6*(2), e139-e146.

Sram, R. J., BeneS, I., Binková, B., Dejmek, J., Horstman, D., Kotěsovec, F., Otto, D., Perreault, S. D., Rubes, J., Selevan, S. G., et al. (1996). Teplice program–the impact of air pollution on human health. *Environmental health perspectives*, *104*(suppl 4), 699-714.

Stare, J., & Maucort-Boulch, D. (2016). Odds ratio, hazard ratio and relative risk. *Advances in Methodology and Statistics*, *13*(1), 59-67.

Urbinato, D. (1994). London's historic"pea-soupers."(smog in London, England). *EPA journal*, *20*(1-2), 44-45.

Wachter, K. W. (2014). *Essential demographic methods*. Harvard University Press.

Winnett, A., & Sasieni, P. (2001). Miscellanea. A note on scaled Schoenfeld residuals for the proportional hazards model. *Biometrika*, *88*(2), 565-571.

Yang, X., Liang, F., Li, J., Chen, J., Liu, F., Huang, K., Cao, J., Chen, S., Xiao, Q., Liu, X., et al. (2020). Associations of long-term exposure to ambient PM2. 5 with mortality in Chinese adults: A pooled analysis of cohorts in the China-PAR project. *Environment international*, *138*, 105589.

Zafra-Mejía, C. A., Rodríguez-Miranda, J. P., & Rondón-Quintana, H. A. (2020). The relationship between atmospheric condition and human mortality associated with coarse material particulate in Bogotá (Colombia). *Revista Logos Ciencia & Tecnología*, *12*(3), 57-68.

Zhang, G., Rui, X., & Fan, Y. (2018). Critical review of methods to estimate PM2. 5 concentrations within specified research region. *ISPRS International Journal of Geo-Information*, *7*(9), 368.

Zhang, H., Wang, Z., & Zhang, W. (2016). Exploring spatiotemporal patterns of PM2. 5 in China based on ground-level observations for 190 cities. *Environmental Pollution*, *216*, 559-567.