Estimación de los efectos de la contaminación del aire en la salud pública en Bogotá utilizando la herramienta BenMAP-CE

Samuel Peña Garcia

Directora: Eddy Herrera Daza

Departamento de Matemáticas

Pontificia Universidad Javeriana

Facultad de Ciencias

Programa de Ciencia de Datos

Bogotá, D.C

2025

Agradecimientos

Agradezco especialmente a mi tutora, Eddy Herrera Daza, por poner a disposición todo su conocimiento y brindarme la confianza y el acompañamiento necesario para la elaboración de este trabajo de grado. A mis padres, por ser siempre ese apoyo fundamental en este proceso y a mis amigos: Tomás, Daniel, Mauricio y Juan David, gracias por ser unos increíbles coequiperos desde el día 1, su ayuda y acompañamiento ha sido invaluable, y a Dana Katalina, por su apoyo emocional y paciencia en los momentos difíciles de este camino.

Finalmente, quiero expresar mi gratitud a todos los profesores con los que tuve el privilegio de coincidir. Destaco a Lina Maria Acosta Avena, Edgar Enrique Ruiz Garcia, Julio Omar Palacio Niño y a John Jairo Corredor Franco, quien fuere un consejero en todo momento. En ellos destaco su humanidad, su habilidad para comprender a los estudiantes y el estar siempre dispuestos a ser guías en el arduo proceso de aprendizaje.

A todos, muchas gracias por ser partícipes y constructores de este proceso de aprendizaje y crecimiento académico y ante nada personal.

Tabla de Contenidos

Introducción 1

Objetivos 2

1. Objetivo General 2
2. Objetivos Específicos 2

Marco Teórico 2

3. Contaminación del Aire y sus Principales Componentes 2

Definición y fuentes de contaminación atmosférica 2

Tipos de contaminantes atmosféricos 3

4. Modelo conceptual del transporte y transformación del PM2.5 en entornos urbanos y su interacción con variables meteorológicas 3

Fuentes y emisión 4

Transporte atmosférico 4

Transformación química 5

Remoción y deposición 5

Interacciones con variables meteorológicas 5

Síntesis conceptual 7

Modelo Matemático de la dinámica del PM2.5 7

5. Escalas características espaciales y temporales de PM2.5 en Bogotá 8

Escalas espaciales de variabilidad del PM2.5 9

Escalas temporales de variabilidad del PM2.5 11

6. Efectos de la Contaminación del Aire en la Salud Pública 15

Relación entre contaminación del aire y enfermedades 15

Exposición a PM 2.5 16

Impacto en el sistema respiratorio 17

Efectos cardiovasculares de la contaminación del aire 17

7. Estudios de Caso sobre Contaminación del Aire y Salud Pública 18

Situación actual de la contaminación del aire en Bogotá 18

Comparación con otras ciudades latinoamericanas 20

8. Modelado y Simulación de Impactos en Salud por Contaminación del Aire 22

Fundamentos estadísticos del modelado epidemiológico aplicado a contaminación atmosférica. 22

Introducción al programa BenMAP-CE 24

Estudios previos realizados con BenMAP-CE 26
Validación estadística, sensibilidad y limitaciones en el modelado con BenMAP-CE 27
9. Algoritmo CatBoost 31
Funcionamiento Técnico 31
Base matemática 32
CatBoost en aplicaciones ambientales 35
Metodología 36
10. Diseño del estudio y alcance 36
11. Fuentes de datos y variables 37
Análisis exploratorio y espaciotemporal 38
12. Modelo de aprendizaje automático 38
13. Estimación de los efectos en salud con BenMAP-CE 40
Preparación de datos de entrada: 40
Datos de salud y población en BenMAP: 40
Selección de funciones concentración–respuesta: 41
Configuración de análisis y ejecución en BenMAP: 41
14. Supuestos a tener en cuenta 42
Supuestos asociados a la interpolación espacial 42
Supuestos de agregación temporal 44
Consistencia de las escalas de análisis con las escalas características del sistema 46
Resultados 47
15. Datos obtenidos 47
16. Análisis exploratorio espacial de datos (EDA) 48
Conjunto de datos 49
Análisis de Nulidad 50
Análisis Estadístico datos en Bruto PM2.5 51
Análisis Gráfico 53
Transformaciones a la base de datos de material particulado 57
17. Preparación de datos, validación temporal y modelado predictivo de PM2.5 en Bogotá 59
Ingeniería de variables 59
Partición temporal y validación cruzada 60
Modelo y entrenamiento 61

Métricas de evaluación 62
Resultados 64
Métricas 65
18. Metodología para la construcción de predictores meteorológicos (Bogotá) y proyección 2026 68
Selección de la fuente de datos: ERA5 (C3S/ECMWF) 68
Variables y preprocesamiento 68
Proyección para 2026 70
19. Generación de predicciones de concentraciones de PM2.5 para 2026 71
Generación Recursiva de Predicciones 71
Smoothing y Relleno de Datos Faltantes 71
Agregación por Localidad 72
Resultados y Almacenamiento 72
Impacto de la agregación temporal en las proyecciones 2025-2026 72
20. Proyección de Mortalidad por causas respiratorias o cardiopulmonares 73
Datos de Mortalidad Histórica 74
Datos de Población 74
Modelado de conteo con Offset Poblacional 74
Proyección a 2026 75
21. Generación de escenarios en BenMAP-CE 75
Función de Impacto en Salud 75
Mapa y detalle de incidencia. 77
Conclusiones 80
Recomendaciones 81
Referencias 84

Lista de Tablas

Tabla 1: *Tipos de contaminantes atmosféricos* 3
Tabla 2. Descripción hiperparámetros modelo CatBoost 39
Tabla 3. *Dimensionalidad bases de datos extraídas* 47
Tabla 4. *Detalle base de datos contaminación del aire* 47
Tabla 5. *Detalle base de datos climatología* 47
Tabla 6. *Detalle base de datos poblacional* 48
Tabla 7. Detalle base de datos salud - Mortalidad 48
Tabla 8. *Cantidad de estaciones por localidad de Bogotá D.C* 49
Tabla 9. *Estructura conjunto de datos contaminación del aire* 49
Tabla 10. *Análisis de nulidad para el conjunto de datos de contaminación del aire* 50
Tabla 11. Resumen estadístico material Particulado PM 2.5 por estación 52
Tabla 12. Porcentaje completitud de datos por estación según el umbral establecido 57
Tabla 13. Estructura base de datos predictores PM 2.5 59
Tabla 14. Estructura conjunto de datos de entrada para el modelo CatBoost 61
Tabla 15. Valores de los hiperparámetros utilizados para el modelo CatBoost 62
Tabla 16. Métricas por fold 65
Tabla 17. Métricas Hold-Out 67
Tabla 18. Variables seleccionadas del reanálisis ERA5 para la construcción del conjunto de proyección de climatología 69
Tabla 19. Detalle numérico para la Ilustración 10 78
Tabla 20. Resultados Ben-MAP 79

Lista de Ilustraciones

Ilustración 1. *Transición de ML a BenMAP-CE* 31

Ilustración 2. *Contaminantes y medidas meteorológicas con mayor cantidad de valores faltantes* 51

Ilustración 3. *Distribución de PM 2.5 en las 13 localidades de Bogotá presentes en el Conjunto de Datos* 53

Ilustración 4. Distribución de PM 2.5 en las 13 localidades de Bogotá presentes en el Conjunto de Datos 54

Ilustración 5. Evolución temporal del promedio mensual de PM2.5 en 13 localidades de Bogotá (2021-2024). 55

Ilustración 6. Mapa de la distribución promedio de PM 2.5 por localidad (2021-2024) 56

Ilustración 7. Proceso de entrenamiento y validación para la validación cruzada temporal..... 61

Ilustración 8. Importancia de variables del modelo CatBoost 64

Ilustración 9. Flujo de trabajo para la generación de mapa con puntos de estaciones y malla de ERA5 77

Ilustración 10. Impacto estimado en mortalidad para todas las causas en población de 30-99 años por PM 2.5 a 2026 basado en predicciones de contaminante a través de un modelo CatBoost y modelos de conteo para tasas de Mortalidad asociado a través de una función C-R. 78

Introducción

La contaminación del aire es uno de los problemas ambientales más críticos que enfrentan actualmente las grandes ciudades alrededor del mundo, pues esto tiene un efecto sobre la salud pública y el bienestar social. Bogotá, capital y centro urbano principal de Colombia, tiene desafíos significativos en relación con la contaminación del aire debido al crecimiento poblacional, la industrialización acelerada, el aumento del parque automotor y otros factores geográficos. Todo esto ha llevado la calidad del aire a niveles preocupantes, generando un riesgo elevado para la salud de sus habitantes, especialmente en la población urbana, que es la más expuesta a contaminantes provenientes del tráfico vehicular, la actividad industrial y eventos extremos como incendios forestales. Un reflejo de esto es un informe de Greenpeace en 2021, que destacaba que "En Bogotá la contaminación del aire por el PM 2.5 (Material Particulado) fue responsable de la pérdida estimada de 5400 vidas durante el pasado año."

De acuerdo con la Organización Mundial de la Salud: "Según estimaciones de 2019, la contaminación del aire ambiente (exterior) en las ciudades y zonas rurales de todo el mundo provoca cada año 4,2 millones de muertes prematuras; esta mortalidad se debe a la exposición a materia particulada fina, que causa enfermedades cardiovasculares y respiratorias, así como cánceres." En este contexto, la ciencia de datos y herramientas especializadas como BenMAP-CE emergen como soluciones clave para una comprensión más profunda y una gestión eficiente del problema. BenMAP-CE permite estimar impactos en la salud pública y su valoración económica derivados de cambios en la calidad del aire. Todo esto mediante datos poblacionales, tasas de incidencia de enfermedades, calidad del aire y funciones de concentración-respuesta, facilitando la evaluación de posibles escenarios futuros y la formulación de políticas ambientales y de salud pública efectivas.

Objetivos

1. Objetivo General

Diseñar y validar escenarios para estimar espacialmente la carga sanitaria y económica atribuible a la contaminación del aire a escala de localidad en Bogotá, integrando modelos de aprendizaje de máquina para proyección de exposición con la herramienta *BenMAP-CE*.

2. Objetivos Específicos

1.1 Integrar una base de datos de predictores geoespaciales de la calidad del aire y salud pública entre 2021 y 2024 en localidades de Bogotá.

1.2 Entrenar y validar el CatBoost de Aprendizaje de Máquina (ML) que proyecte las concentraciones de contaminantes del aire bajo diversos escenarios de emisión.

1.3 Integrar las series pronosticadas por el modelo de ML validado en BenMAP-CE con el fin de simular la carga sanitaria para 2026.

Marco Teórico

3. Contaminación del Aire y sus Principales Componentes

Definición y fuentes de contaminación atmosférica

Según la OMS (s.f.), “Air pollution is contamination of the indoor or outdoor environment by any chemical, physical or biological agent that modifies the natural characteristics of the atmosphere.”. Derivado de esto, tenemos que las fuentes de emisión se clasifican en dos tipos: antropogénicas, que incluyen emisiones provenientes de transporte, procesos industriales y actividades agrícolas como quema de residuos agrícolas y forestales junto con el uso excesivo de fertilizantes, y las fuentes naturales, como las erupciones volcánicas, tormentas de polvo y los ya

mencionados incendios forestales. En la tabla 2.2 se resumen los tipos de contaminantes atmosféricos y sus fuentes principales

Tipos de contaminantes atmosféricos

Tabla 1: *Tipos de contaminantes atmosféricos*

Contaminante	Definición	Fuentes
Material Particulado $PM_{2.5}$	Partículas sólidas y líquidas suspendidas en el aire con un diámetro menor a 2.5µm que por su tamaño pueden penetrar profundamente en los pulmones	Emisiones de vehículos, incendios forestales, industrias, y reacciones químicas
Material Particulado PM_{10}	Partículas sólidas y líquidas suspendidas en el aire con un diámetro menor a 10µm	Emisiones de vehículos, incendios forestales, industrias, y reacciones químicas
Ozono Troposférico (O_3)	Se forma a través de reacciones fotoquímicas entre óxidos de nitrógeno y compuestos orgánicos volátiles en presencia de luz solar intensa. Es un contaminante secundario cuya concentración varía diurnamente.	Las actividades antropogénicas incrementan la concentración de los precursores. Además, con el cambio climático, las temperaturas han aumentado, favoreciendo su formación y acumulación.
Dióxido de Azufre (SO_2)	Es un gas incoloro con olor penetrante en concentraciones elevades, puede transformarse en ácido sulfúrico, contribuyendo a la lluvia ácida	Se produce principalmente por la combustión de combustibles fósiles que contienen azufre como carbón y petróleo y procesos industriales
Monóxido de Carbono (CO)	Es un gas incoloro e inodoro, lo cual lo hace peligroso en tanto su presencia puede pasar desapercibida sin efectos de medición	Automóviles, estufas mal ventiladas y algunas industrias

4. Modelo conceptual del transporte y transformación del PM2.5 en entornos urbanos y su interacción con variables meteorológicas

El material particulado fino (PM2.5) es uno de los principales contaminantes atmosféricos en entornos urbanos, tanto por su prevalencia como por sus efectos en la salud pública. La dinámica de su comportamiento en la atmósfera está gobernada por un conjunto de procesos físicos y químicos interrelacionados, que pueden ser descritos mediante un modelo conceptual que integra fuentes de emisión, transporte, transformación, remoción y acoplamiento con variables meteorológicas. Entender esta dinámica resulta fundamental para interpretar patrones espaciales y temporales de concentración, evaluar riesgos en salud y establecer estrategias de gestión ambiental efectivas.

Fuentes y emisión

El PM2.5 tiene un origen mixto: puede ser emitido directamente como partícula (PM primario) o bien formarse por reacciones químicas atmosféricas a partir de precursores gaseosos (PM secundario), como óxidos de nitrógeno (NO), dióxidos de azufre (SO_2), compuestos orgánicos volátiles (COVs) y amoníaco (NH_3) (Zhang et al., 2007). En ciudades latinoamericanas como Bogotá, las principales fuentes incluyen el tráfico vehicular (especialmente motores diésel), procesos industriales, combustión residencial, y en menor medida, fuentes naturales como el polvo resuspendido (Mura et al., 2020).

Transporte atmosférico

Una vez emitido, el PM2.5 se desplaza en la atmósfera influenciado por el viento y la turbulencia. El transporte horizontal depende de la velocidad y dirección del viento en capas bajas de la atmósfera, mientras que el transporte vertical está condicionado por la estabilidad atmosférica y la altura de la capa límite planetaria (PBL, por sus siglas en inglés). En zonas como Bogotá, la complejidad topográfica y la presencia de cañones urbanos (calles rodeadas por edificaciones altas)

generan microclimas que afectan la dispersión local del contaminante (Díaz et al., 2021). Durante condiciones estables (e.g., inversión térmica), el aire frío queda atrapado cerca del suelo y limita la mezcla vertical, favoreciendo la acumulación de PM2.5 en la superficie.

Transformación química

El PM2.5 puede experimentar procesos de transformación por reacciones fotoquímicas, oxidación, condensación o nucleación. Por ejemplo, el dióxido de azufre puede oxidarse en fase gaseosa o acuosa para formar sulfatos, mientras que los NO_x contribuyen a la formación de nitratos (Sun et al., 2013). Estas reacciones están moduladas por la presencia de radiación solar, humedad relativa y temperatura, variables que controlan tanto la cinética como la ruta de formación de especies secundarias. Las altas temperaturas, por ejemplo, favorecen la volatilización de compuestos orgánicos semivolátiles, mientras que la humedad promueve reacciones acuosas que aumentan la formación de aerosoles secundarios.

Remoción y deposición

Los procesos de remoción del PM2.5 incluyen la deposición seca (por gravedad o contacto con superficies) y la deposición húmeda (lavado por lluvia). Ambos procesos están fuertemente influenciados por variables meteorológicas como la precipitación, el tamaño de partícula, la rugosidad del terreno, y la cobertura vegetal. En Bogotá, los episodios de lluvia intensa están asociados a caídas abruptas en las concentraciones de PM2.5 por efecto de lavado atmosférico.

Interacciones con variables meteorológicas

Cada variable meteorológica tiene un rol crítico y diferenciado en la dinámica del PM2.5. A continuación se resumen las principales interdependencias:

El viento controla la dispersión y el transporte de contaminantes. Vientos intensos dispersan el PM2.5, reduciendo su concentración local. Por el contrario, condiciones de calma favorecen la acumulación. Por su parte, la temperatura afecta la reactividad química de los precursores del PM2.5, la eficiencia de su formación secundaria y la estabilidad atmosférica. Además, temperaturas bajas favorecen inversiones térmicas que atrapan contaminantes (Seinfeld & Pandis, 2016).

La humedad relativa influye en el crecimiento higroscópico de partículas, lo que puede aumentar su masa y modificar su composición. También potencia reacciones acuosas en aerosoles. La radiación solar, por otro lado, activa reacciones fotoquímicas que transforman gases precursores en partículas secundarias. Días soleados, especialmente con alta carga de NO y COVs, pueden inducir aumentos en el PM2.5 secundario (Sun et al., 2020).

A esto se añaden variables como la precipitación, que constituye el mecanismo más eficiente de remoción por deposición húmeda. Episodios de lluvia intensa pueden reducir drásticamente las concentraciones ambientales de partículas (Mura et al., 2020). Por esa misma línea se encuentra la presión atmosférica, que, aunque su efecto directo es menor, se asocia a condiciones de estabilidad atmosférica. Altas presiones están ligadas a estancamiento del aire y pobre dispersión. Finalmente, la altura de la capa límite planetaria (PBL) controla el volumen atmosférico disponible para diluir contaminantes. Una PBL baja (por la mañana o en condiciones estables) implica concentraciones superficiales más altas de PM2.5 (Díaz et al., 2021).

Síntesis conceptual

El modelo conceptual resultante es sistémico: las concentraciones observadas de PM2.5 no dependen solo de las emisiones, sino de un balance dinámico entre transporte, transformación y remoción, mediado por condiciones meteorológicas locales y regionales. Este modelo explica, por ejemplo, por qué localidades en Bogotá como Kennedy y Fontibón presentan concentraciones persistentemente altas: a su carga emisora se suma la influencia de flujos de viento débiles, frecuentes episodios de inversión térmica y limitada ventilación urbana (Mura et al., 2020). Así, integrar la meteorología al análisis no solo mejora la comprensión del comportamiento del contaminante, sino que aporta fundamentos para decisiones de política pública orientadas a mitigar impactos en salud.

Modelo Matemático de la dinámica del PM2.5

Una forma de representar la dinámica del PM2.5 en la atmósfera es mediante una ecuación de balance de masa tipo advección-difusión-reacción:

$$\frac{\partial C(\mathbf{x},t)}{\partial t} = S(\mathbf{x},t) - \Lambda(\mathbf{x},t)\, C(\mathbf{x},t) - \mathbf{u}(\mathbf{x},t)\cdot\nabla C(\mathbf{x},t) + \nabla\cdot(K(\mathbf{x},t)\,\nabla C(\mathbf{x},t)),$$

Donde $C(\mathbf{x},t)$es la concentración del material particulado PM2.5, S agrupa las fuentes y la formación secundaria, Λ reúne los procesos de remoción (deposición seca y húmeda), $\mathbf{u}$ es el campo de viento y K representa la difusión turbulenta.

Al tener en cuenta la ecuación diferencial parcial, en el tiempo sobre un intervalo de 24 horas (1 día) y en el espacio sobre el entorno de una estación de monitoreo o localidad (como unidad de muestreo) se obtiene una aproximación de la dinámica en tiempo discreto para los promedios diarios:

$$\bar{C}_{d+1} \approx \bar{C}_d + \Delta t\, f(\bar{C}_d, \mathbf{M}_d),$$

Donde $\bar{C}_d$es la concentración del PM2.5 media diaria en el día d que además se desagrega como el día del mes y del año de la serie de tiempo asociada, M_d es el vector de condiciones meteorológicas promedio (temperatura, viento, humedad, precipitación, etc.) y la función $f$resume el efecto neto de emisión, transporte, transformación y remoción durante ese día. Hay que considerar, que en la práctica el modelo CatBoost entrenado con Lags (rezagos) de PM2.5 y predictores meteorológicos diarios aproxima de manera no paramétrica la función $f(\cdot)$ y el esquema recursivo que se implementará para la proyección 2025–2026 corresponde y considera lo anterior porque se aplicará iterativamente esta relación discreta día a día.

Bajo esta formulación, las operaciones de interpolación y agregación que se adopten en el procesamiento de datos del material particulado pueden interpretarse como formas de aproximar las integrales espaciales y temporales implícitas al pasar de la ecuación continua al modelo discreto, lo que permite el uso de promedios diarios y de climatologías mensuales cuando se trabaja a escala de localidad y en el marco de evaluaciones de impacto en salud.

5. Escalas características espaciales y temporales de PM2.5 en Bogotá

La concentración de material particulado fino (PM2.5) en entornos urbanos como Bogotá resulta de la interacción de procesos de transporte (advección y dispersión turbulenta), transformación química en la atmósfera y remoción por deposición. Estos procesos actúan en distintas escalas de espacio y tiempo y determinan la variabilidad observada en las concentraciones. A continuación, se analizan las escalas espaciales y temporales características del comportamiento del PM2.5 en Bogotá, de acuerdo con el modelo conceptual de transporte, transformación y remoción previamente descrito, sustentado en hallazgos de literatura académica.

Escalas espaciales de variabilidad del PM2.5

En Bogotá se observan fuertes gradientes espaciales de contaminación por PM2.5, con diferencias notables entre localidades. La heterogeneidad en la distribución de fuentes emisoras y las condiciones meteorológicas locales hace que ciertas zonas presenten concentraciones mucho más elevadas que otras. Estudios descriptivos de la calidad del aire en la ciudad han mostrado que la región suroccidental de Bogotá concentra los niveles más altos de PM, excediendo hasta 3–4 veces los valores guía anuales recomendados por la OMS.

Por ejemplo, estaciones de monitoreo ubicadas en localidades industrializadas o de alto tráfico (como Carvajal o Kennedy) registran concentraciones diarias de PM2.5 por encima del valor guía de la OMS (~25 μg/m³, 24 h) en más del 50% de los días, e incluso exceden la norma nacional de 50 μg/m³ en alrededor del 5% de los días muestreados. En contraste, estaciones en zonas residenciales con menos fuentes (p. ej. Usaquén al norte) reportan casi 100% de días por debajo del estándar diario nacional OMS.

Esto implica que, en un mismo periodo, las áreas más limpias de la ciudad pueden tener promedios diarios de PM2.5 que son solo una fracción de aquellos observados en las zonas más contaminadas. Consecuentemente, usar un solo valor medio para toda la ciudad resulta engañoso; es necesario un análisis espacialmente diferenciado, dado que existen diferencias significativas entre localidades incluso dentro del mismo mes (Mura et al., 2020).

Las causas de esta variabilidad espacial están ligadas tanto a la distribución de las emisiones como a factores geográficos. En Bogotá, las fuentes locales dominantes de PM2.5 incluyen el tráfico vehicular (especialmente diésel), la industria y actividades como la quema de residuos; la concentración de estas fuentes en ciertas zonas eleva significativamente las partículas en el aire circundante. Así, localidades con corredores viales de alto flujo o polígonos industriales

(suroccidente) presentan niveles de PM2.5 sustancialmente mayores que áreas suburbanas o residenciales con menos fuentes directas. Adicionalmente, la topografía urbana y los microclimas desempeñan un papel importante en la dispersión. Bogotá se sitúa en una sabana elevada (~2600 m s.n.m.) flanqueada por montañas al oriente, lo que crea condiciones propicias para microclimas locales diferenciados.

Por ejemplo, bajo ciertos escenarios meteorológicos, los contaminantes pueden quedar atrapados en la capa de aire cercana al suelo debido a inversiones térmicas frecuentes en la sabana cundiboyacense. Este fenómeno ocurre típicamente en las madrugadas de los primeros meses del año, cuando noches muy frías seguidas de días soleados generan una capa de inversión que impide la ventilación vertical.

El resultado es una "nata" de contaminación sobre la ciudad en horas de la mañana, con concentraciones más altas en sectores donde el terreno y las construcciones dificultan la dispersión. Adicionalmente, la dirección e intensidad de los vientos locales modulan la distribución: vientos predominantes pueden arrastrar las emisiones de las fuentes y concentrarlas sotavento de obstáculos topográficos (como los cerros orientales). Así, ciertas zonas bajas del suroccidente pueden actuar como receptores donde se acumula el material particulado transportado desde el centro y norte de la ciudad.

Así pues, la interacción de fuentes locales desigualmente distribuidas y la compleja geografía de Bogotá (orografía y urbanismo) da lugar a escalas espaciales de variación que van desde "puntos calientes" muy localizados (p. ej. en inmediaciones de vías congestionadas) hasta gradientes entre diferentes localidades urbanas (Secretaría Distrital de Ambiente, 2021).

Escalas temporales de variabilidad del PM2.5

Ciclos horarios y diarios

El PM2.5 en Bogotá exhibe marcados ciclos diurnos asociados a las rutinas humanas y a la dinámica de la atmósfera en 24 horas. En días típicos, las concentraciones comienzan a elevarse en la madrugada y alcanzan un pico pronunciado en las horas de la mañana (aproximadamente entre las 6:00 y 9:00 h). Este pico coincide con el inicio de la jornada laboral y escolar, cuando se incrementa el tráfico vehicular (particularmente de diésel y transporte público) y otras actividades emisoras. A esa hora el aire cercano al suelo suele estar poco mezclado: existe una capa límite muy baja debido al enfriamiento nocturno de la superficie, que limita la dispersión vertical de contaminantes. Bajo estas condiciones de inversión térmica matutina, las emisiones de las primeras horas quedan confinadas cerca del suelo, acumulando altas concentraciones de PM2.5 (Secretaría Distrital de Ambiente, 2021).

Estudios de la red de monitoreo indican que durante este rush hour las concentraciones horarias superan los valores de referencia de calidad del aire, alcanzando niveles considerados no saludables según estándares de 24 horas. Hacia el mediodía y primeras horas de la tarde, la intensa radiación solar calienta la superficie y rompe la inversión: la capa de mezcla atmosférica se expande cientos de metros, diluyendo los contaminantes. En consecuencia, las concentraciones de PM2.5 tienden a disminuir significativamente entre el final de la mañana y primeras horas de la tarde (Mura et al., 2020).

Más entrado el atardecer, conforme cae de nuevo la temperatura y se reduce la turbulencia, puede observarse un segundo pico de menor magnitud en horas de la noche (aprox. 18:00–22:00 h). Este comportamiento bimodal – pico mayor en la mañana y otro al final de la jornada – se atribuye al patrón de tráfico diario: una explosión concentrada de vehículos en la mañana, seguida de flujos vespertinos más dispersos. Durante la noche, la renovación del aire es escasa y las

emisiones residuales (p. ej. quema de biomasa para calentarse en zonas periurbanas, o la operación de algunas industrias nocturnas) pueden contribuir a mantener niveles moderadamente elevados hasta que un nuevo ciclo comience al día siguiente (Mura et al., 2020).

Además del ciclo dentro de cada día, existe una modulación día a día a lo largo de la semana. En términos generales, los días laborales tienden a registrar concentraciones mayores de PM2.5 que los fines de semana, reflejando la reducción de actividades emisoras durante sábados y especialmente domingos. Por ejemplo, históricamente se ha observado que el domingo (día con tráfico significativamente menor y menos actividad industrial) presenta los promedios diarios más bajos de PM, en contraste con los picos semanales que suelen ocurrir entre martes y viernes.

Esta variación inter-diaria sugiere que una porción importante del material particulado fino de Bogotá proviene de fuentes antropogénicas locales de carácter regular (tráfico rutinario, actividades económicas), las cuales disminuyen en días de descanso. No obstante, eventos atípicos (como festividades con fuegos artificiales, quemas esporádicas o episodios de contaminación regional) pueden romper esta regularidad semanal y causar alzas puntuales en días normalmente limpios (Mura et al., 2020).

Variaciones estacionales y factores meteorológicos

A escala estacional, la dinámica del PM2.5 en Bogotá está fuertemente controlada por las variaciones del clima local a lo largo del año. En la ciudad no se presentan estaciones térmicas marcadas (debido a su ubicación tropical), pero sí existen periodos diferenciados por el régimen de lluvias y vientos. Los estudios señalan un claro ciclo anual: los meses de enero, febrero y marzo (período seco inicial) suelen registrar las concentraciones más altas de PM2.5 del año, mientras que junio, julio y agosto presentan las más bajas. Esta diferencia estacional se ha asociado

principalmente a condiciones meteorológicas contrastantes. A inicios de año predominan cielos despejados, poca precipitación y frecuentemente situaciones atmosféricas estables (vientos débiles y capas de mezcla poco desarrolladas), lo que favorece la acumulación de contaminantes. En cambio, a mitad de año los vientos alisios y la incursión de masas de aire más frías traen mayor ventilación sobre la sabana, con brisas más intensas y una atmósfera más turbulenta que dispersa el material particulado (Mura et al., 2020).

Adicionalmente, los períodos de lluvia (principalmente abril–mayo y septiembre–noviembre) contribuyen a remover partículas del aire mediante lavado atmosférico. La precipitación es uno de los mecanismos de remoción más eficaces: durante aguaceros, buena parte del PM suspendido es arrastrado hacia el suelo. Sin embargo, estudios locales sugieren que la lluvia tiende a remover preferencialmente las partículas más gruesas, mientras que las finas (PM2.5) se reducen en menor proporción.

Por ejemplo, en la estación Kennedy se encontró que en horas lluviosas la razón $PM2.5/PM_{10}$ aumentaba (0,41 frente a 0,36 en horas secas), lo que indica que las gotas eliminan más eficientemente a la fracción gruesa, dejando relativamente más material fino. Esto implica que, aunque los eventos de lluvia disminuyen las concentraciones absolutas de PM2.5, no siempre logran restablecer condiciones totalmente limpias, especialmente si persisten emisiones continuas o si la lluvia es ligera/intermitente (Blanco et al., 2015).

Otro factor estacional crítico es la química atmosférica modulada por la radiación solar. En los meses secos con alta insolación, la intensa radiación UV y las mayores temperaturas diurnas potencian las reacciones fotoquímicas en la atmósfera urbana. Bajo estas condiciones se forma ozono troposférico y otros oxidantes que transforman gases precursores (como NO_x, SO_2 y compuestos orgánicos volátiles) en partículas secundarias (sulfatos, nitratos, aerosoles orgánicos).

De hecho, se estima que una fracción significativa del PM2.5 en Bogotá proviene de este origen secundario regional y urbano, especialmente en la temporada seca cuando la química fotooxidante es más activa.

La humedad relativa también influye en estos procesos: niveles altos de humedad pueden facilitar reacciones en fase acuosa dentro de las partículas o en las nubes, contribuyendo a la formación de sulfatos y nitratos secundarios, mientras que también propician el crecimiento higroscópico de las partículas existentes. Así, durante madrugadas húmedas o episodios de neblina, las partículas finas pueden absorber agua aumentando de tamaño, lo que puede modificar temporalmente las concentraciones medidas y su tasa de deposición seca.

En conjunto, temperatura, radiación solar y humedad afectan la tasa de transformación química de contaminantes gaseosos a PM2.5, modulando la composición del aerosol a lo largo del día y del año (p. ej., mayor fracción de aerosoles secundarios en épocas cálidas y soleadas).

Finalmente, es importante resaltar la interacción entre las escalas temporal y espacial mediante el transporte de contaminantes. En ciertos episodios críticos, las condiciones meteorológicas regionales favorecen el arribo de material particulado desde fuera de la ciudad (escala meso/regional). Por ejemplo, durante la temporada seca de comienzos de año coinciden a veces incendios forestales y quema de biomasa en los llanos orientales (Orinoquia) que emiten humo; masas de aire provenientes de esas regiones pueden ser transportadas cientos de kilómetros y llegar a la Sabana de Bogotá.

Cuando tal transporte de largo alcance coincide con una atmósfera local estable (inversión térmica), ocurre una acumulación exacerbada de PM2.5 sobre la ciudad – se han documentado eventos en los cuales el material particulado de fuentes regionales aportó alrededor de un 30–40%

del incremento estacional de PM2.5 en Bogotá (Mendez Espinosa, 2019). Estos episodios ilustran cómo las escalas temporales estacionales (sequía prolongada e incendios) y las escalas espaciales regionales se combinan para condicionar la dinámica local del PM2.5. En tales situaciones, reducciones locales de emisiones podrían no evitar niveles altos de contaminación si el fondo regional viene ya cargado.

Así pues, las concentraciones de PM2.5 en Bogotá emergen de un conjunto de procesos multiescala: a nivel espacial, varían desde microescala (entornos inmediatos a una fuente) hasta contrastes inter-barrios condicionados por la distribución de fuentes y la topografía urbana; y a nivel temporal, fluctúan en ciclos diarios marcados por la actividad humana y la evolución de la capa límite, así como en ciclos estacionales determinados por el clima andino tropical (regímenes de viento, lluvias y radiación).

Estas escalas características gobiernan la acumulación del PM2.5 (cuando hay emisiones sostenidas en atmósferas estables), su dispersión (favorecida por la mezcla turbulenta diurna y los vientos sinópticos) y su remoción (principalmente vía deposición húmeda en épocas lluviosas). Comprender estas escalas y sus interacciones es fundamental para interpretar la dinámica del aerosol fino en Bogotá y diseñar estrategias de gestión de la calidad del aire efectivas a corto y largo plazo, según la variabilidad esperada en cada contexto (espacial o estacional).

6. Efectos de la Contaminación del Aire en la Salud Pública

Relación entre contaminación del aire y enfermedades

La contaminación del aire es un riesgo que, pese a no ser fácil de percibir a simple vista, es un grave problema de salud pública a nivel global, donde las áreas densamente pobladas como

Bogotá son más vulnerables a sus efectos perjudiciales, dado que la exposición prolongada al aire contaminado afecta gravemente a la salud de la población. Según Pavón (2024), esta exposición genera afecciones cardiorrespiratorias, así como afectaciones al sistema nervioso y problemas metabólicos e incluso de salud mental. Otros estudios, afirman que "*la exposición prolongada a una atmósfera enrarecida suele ser especialmente perjudicial para las personas con enfermedades respiratorias y cardiovasculares, cáncer, mujeres embarazadas, niños y ancianos*" (Torres et. al, 2023).

Exposición a PM 2.5

La exposición poblacional a PM2.5 se estima mediante combinaciones de monitoreo ambiental y modelación. A nivel local, se usan redes de estaciones (estaciones fijas) complementadas con modelos de uso de suelo (LUR) o satelitales para asignar la concentración diaria o anual a cada sector poblacional. Por ejemplo, Marín et al. (2025) estimaron las exposiciones intraurbanas prediciendo concentraciones de PM2.5 con un modelo LUR ajustado a datos de 17 estaciones en cada ciudad.

En salud pública se distingue la exposición aguda (horas-días) de la crónica (meses-años): los estudios de series de tiempo analizan efectos de picos cortos de PM2.5, mientras que estudios de cohorte evalúan exposiciones promedio anuales. Las escalas espaciales van desde diferencias entre ciudades hasta variaciones dentro de áreas locales densamente pobladas. También se reconoce que ciertos grupos vulnerables (niños, ancianos, personas con comorbilidades respiratorias o cardiovasculares, gestantes, etc.) sufren exposiciones de mayor impacto, aunque no se hallaron estudios específicos de Bogotá sobre esto. A escala global casi toda la población excede los estándares recomendados de PM2.5, lo que implica una exposición generalizada al riesgo.

Impacto en el sistema respiratorio

La exposición a contaminantes como el material particulado $PM_{2.5}$, el O_3 y el dióxido de nitrógeno (NO_2) se relaciona de manera significativa con un incremento en la incidencia y la mortalidad por diversas enfermedades no transmisibles, incluyendo las del sistema respiratorio e incluso una serie de respuestas inflamatorias que impactan negativamente a todo el cuerpo. Por ejemplo, el material particulado "puede resultar en estrés oxidativo, aumento de la reactividad bronquial y resistencia de las vías respiratorias, así como inflamación de las vías aéreas, lo que afecta la función pulmonar." (Pavón, 2024). Asimismo, "La contaminación del aire doméstico y por la exposición a combustibles fósiles gasolina, diésel provoca enfermedades respiratorias tanto en niños como adultos en las que se incluyen asma, infecciones como tuberculosis, enfermedad pulmonar obstructiva crónica, cáncer pulmonar con disminución de la función pulmonar y bronquitis crónica." (Ramírez et. al, 2023).

Efectos cardiovasculares de la contaminación del aire

La exposición a partículas $PM_{2.5}$ se asocia con el desarrollo de aterosclerosis, un proceso que endurece y estrecha las arterias, incrementando significativamente el riesgo de enfermedades cardíacas como el infarto de miocardio y los accidentes cerebrovasculares (Pavón, 2024). Asimismo, diversos estudios han encontrado una fuerte asociación entre la exposición crónica a $PM_{2.5}$ y un mayor riesgo de desarrollar diabetes tipo 2, lo que subraya la interconexión entre la calidad del aire y la salud metabólica, que a su vez impacta la salud cardiovascular.

Por lo tanto, la exposición prolongada a contaminantes del aire en general se asocia con un aumento en la mortalidad por enfermedades cardiovasculares y respiratorias teniendo en cuenta que la inhalación de partículas finas puede desencadenar estrés oxidativo e inflamación, procesos que pueden propagarse desde los pulmones incluyendo el cardiovascular (Pavón, 2024).

Finalmente, se resalta que el material particulado, especialmente $PM_{2.5}$, es sumamente fino, lo que facilita su entrada al torrente sanguíneo, provocando así un aumento en los niveles de factores procoagulantes como el fibrinógeno, la trombina y el factor de Von Willebrand, así como un incremento en la actividad plaquetaria, lo que disminuye los tiempos de coagulación (Ramírez et. al, 2023).

7. Estudios de Caso sobre Contaminación del Aire y Salud Pública

Situación actual de la contaminación del aire en Bogotá

La ciudad de Bogotá presenta un escenario complejo en cuanto a calidad del aire, el cual se ve afectado por una combinación de factores orográficos, geográficos, sociales, meteorológicos y antrópicos. Al Bogotá ubicarse en una cuenca montañosa, frecuentemente se experimentan inversiones térmicas, especialmente durante la noche y la madrugada. Al tener estas condiciones, la dispersión de contaminantes se complica, provocando así que todos los agentes contaminantes se queden en la ciudad, agravando la acumulación de $PM_{2.5}$ y Ozono. Aunado a esto, el acelerado crecimiento urbano hace que cada día Bogotá sea una ciudad más densa, lo cual puede contribuir a tener un mayor nivel de concentración de agentes contaminantes, siendo así el factor demográfico crítico en este estudio.

La Red de Monitoreo de Calidad del Aire de Bogotá (RMCAB) registra continuamente información sobre las concentraciones de contaminantes y variables meteorológicas en diferentes puntos de la ciudad. Durante el año 2024, la RMCAB estuvo conformada por 19 estaciones en operación. Estas estaciones monitorean continuamente la presencia contaminante como material particulado PM_{10} y $PM_{2.5}$ y gases contaminantes SO_2, NO_2, CO y O_3.

A continuación, se presentan datos de la Secretaría de Ambiente en su informe anual de calidad del aire para el año 2023. En cuanto a los promedios anuales de PM_{10}, la estación Móvil

Fontibón, ubicada al occidente de la ciudad, excedió el valor de la norma anual (50 μg/m^3) con un promedio de 53.2 μg/m^3. Las concentraciones promedio más bajas se registraron en las estaciones Colina y Usaquén, en la zona norte, con 20.5 y 20.1 μg/m^3 respectivamente. En relación con las excedencias de PM_{10} en 24 horas, la estación Carvajal – Sevillana registró el mayor número con 54 μg/m^3. Las concentraciones más altas de PM_{10} se presentaron en el sur y occidente de la ciudad, en las estaciones de Ciudad Bolívar, Kennedy y Móvil Fontibón, con un rango de concentraciones entre 40 y 53 μg/m^3.

Para el caso de los promedios anuales de $PM_{2.5}$ registrados en 2023, la concentración más alta se observó en la estación Kennedy con 21.9 μg/m^3, hacia el suroccidente de la ciudad. La concentración anual más baja se evidenció en Colina con 10.4 μg/m3, en la zona noroccidental. Con relación a las excedencias diarias de PM2.5, la estación Carvajal-Sevillana registró el mayor número con 45 μg/m^3, mientras que las concentraciones más altas de PM2.5 se presentaron en la zona sur y occidente, en las estaciones Kennedy, Ciudad Bolívar y Fontibón, con concentraciones entre 17 y 22 μg/m^3. Es importante notar que, al comparar con el valor permisible de la Resolución 2254 de 2017 —correspondiente a 25 μg/m^3 para PM2.5—, ninguna estación lo superó en promedio anual. Sin embargo, al considerar el valor recomendado por la Guía de Calidad del Aire de la OMS (5 μg/m^3), ninguna estación cumplió con este valor.

El Informe Anual de Calidad del Aire también analiza el comportamiento de los gases contaminantes. Para el ozono (O^3), las concentraciones más altas se observaron en varias zonas, donde la estación Fontibón registró los valores más altos. En cuanto al NO^2, los promedios anuales más altos se registraron en las estaciones Móvil Fontibón y Colina, mientras que, para el SO^2, la estación Usme registró el promedio anual más alto. Finalmente, La estación Kennedy registró el promedio más alto de CO^5.

El informe de la RMCAB señala que las zonas con mayores registros de PM10 y PM2.5 se encuentran hacia el sur y occidente de la ciudad. Se asocia el incremento de las concentraciones en ciertas épocas del año con factores como el tiempo seco, incendios forestales regionales, inversiones térmicas y baja dispersión de contaminantes. El tráfico vehicular, especialmente de vehículos pesados que utilizan diésel, también se identifica como una influencia notable en las concentraciones de contaminantes, particularmente en vías principales.

Comparación con otras ciudades latinoamericanas

Existen estudios sobre la contaminación del aire en Medellín, Ciudad de México, Santiago de Chile, entre otros. Estas ciudades, al igual que Bogotá, son centros urbanos importantes en Latinoamérica con altos niveles de urbanización, actividad económica y emisiones atmosféricas, lo que las convierte en focos de vulnerabilidad ante el cambio climático y la contaminación del aire.

En cuanto a los niveles de contaminación, un estudio comparativo de Romero-Lankao et al. de 2013 ofrece datos anuales promedio previos al 2023 para PM_{10}, ozono y NO_2 en Bogotá, Ciudad de México y Santiago. Los datos históricos del estudio mostraban promedios anuales de 51.6 mg/m^3 en México City y 70.2 mg/m^3 en Santiago. Estos valores sugieren que estas ciudades históricamente han enfrentado desafíos significativos con la contaminación por partículas. El estudio también destacó que, a pesar de los esfuerzos por controlar la contaminación en estas ciudades (como el Programa PROAIRE en el Valle de México y el Plan de Descontaminación en Chile), niveles altos de contaminación persisten como un problema serio. Esto se alinea con lo visto anteriormente, donde se observó que Bogotá no ha reducido significativamente sus niveles de contaminación en los últimos años.

Medellín, al igual que Bogotá, es identificada en otro estudio realizado en 2020 por Parra-Sánchez, et al. como una ciudad colombiana que ha presentado episodios críticos de contaminación y se encuentra entre las zonas más contaminadas del país, junto con Bogotá. Este estudio se centró en analizar el impacto de la contaminación por $PM_{2.5}$ en la salud respiratoria en Medellín, encontrando una relación entre la concentración de $PM_{2.5}$ y las asistencias médicas por IRA[1], EPOC y asma. La metodología de este estudio, basada en analítica de datos y el seguimiento de retardos en la atención médica, difiere del enfoque del estudio comparativo de Bogotá, Ciudad de México y Santiago, que se centró en la diferenciación socioeconómica y espacial de los riesgos.

Finalmente, Cartagena en un estudio de 2023 realizado por Aldegunde, et al., es señalada como una de las más contaminadas del país por PM2.5, donde las enfermedades respiratorias son una de las principales causas de mortalidad. El estudio empleó la función de impacto en la salud, derivada de funciones concentración-respuesta (C-R), y un modelo logístico para el asma infantil. Se utilizó el software BenMAP-CE para estimar los casos atribuibles a la contaminación por PM2.5, cuya exposición se determinó mediante la interpolación Voronoi. Para la evaluación económica, se aplicaron los conceptos de costo de la enfermedad y valor de la vida estadística. El estudio reveló que Cartagena presenta niveles elevados de $PM_{2.5}$, con una media anual de 25 $\mu g/m^3$, superando los límites recomendados.

Las estimaciones indicaron que una reducción del 20% en el $PM_{2.5}$ antropogénico podría evitar 104 muertes anuales por todas las causas, además de reducir la morbilidad infantil con menos hospitalizaciones por causas respiratorias y visitas a urgencias por asma. El estudio también calculó un beneficio económico significativo de entre USD 50 y 100 millones anuales con esta reducción. Se concluyó que Cartagena enfrenta un grave problema de salud pública debido al $PM_{2.5}$,

[1] IRA: Infección Respiratoria Aguda

principalmente generado por el tráfico rodado, y se recomienda implementar políticas para mitigar la contaminación, incluyendo educación ambiental y el fomento del transporte sostenible.

8. Modelado y Simulación de Impactos en Salud por Contaminación del Aire

Fundamentos estadísticos del modelado epidemiológico aplicado a contaminación atmosférica.

En el contexto de la epidemiología ambiental, el riesgo en salubridad por contaminación del aire se cuantifica basado en relaciones estadísticas entre la exposición a contaminantes y las consecuencias en la salud. Para realizar esto, se utilizan herramientas que pueden incluir:

a) Funciones de concentración-respuesta (C-R).

b) Análisis de regresión epidemiológica.

c) Atribución de carga de la enfermedad.

d) Simulaciones de Monte Carlo.

Las funciones de concentración-respuesta —también llamadas funciones C-R— representan la relación entre la concentración de un agente contaminante y la respuesta en salud. Usualmente, se asume una forma log-lineal con una función de 4 componentes, donde su forma básica es la siguiente:

$$\Delta y = y_0\left(e^{\beta \Delta x} - 1\right)Z$$

En este caso, y_0 es la tasa inicial de ocurrencia del efecto en salud, β indica el grado de asociación entre el contaminante y su efecto, Δx es el cambio estimado en los niveles de contaminación del aire, Z la población expuesta y Δy la variación esperada en los efectos sobre la salud como resultado del cambio en la exposición del contaminante. (Berman, et al, 2020).

Estas funciones se derivan de estudios epidemiológicos que estiman Riesgos Relativos[2] por unidad de concentración, teniendo que esta forma es la más utilizada. Por ejemplo, Chung et al postulaban en 2022 que "La manera más común de estimar el riesgo en salud atribuible a la exposición de la polución del aire en una población es a través de emplear una función de concentración-respuesta, que está generalmente basada en la estimación de un modelo de RR.". En este mismo artículo, se plantea cómo modelos recientes como el IER –Integrated Exposure-Response y GEMM – Global Exposure Mortality Model amplían los rangos de exposición de las funciones C-R, combinando datos de cohortes humanas y otras fuentes para escenarios altamente contaminados.

Respecto al análisis de regresión epidemiológica, vemos en un artículo de Berman, et al en 2020 que muestra cómo los coeficientes que alimentan a las funciones C-R vienen de métodos como la regresión de Poisson en series de tiempo para efectos agudos o modelos de regresión multinivel de Cox para efectos crónicos. Por ejemplo, se han utilizado series de tiempo diarias para estimar de qué manera un aumento de $10\mu g/m^3$ en el material particulado $PM_{2.5}$ eleva la mortalidad diaria en cierto porcentaje, o estudios de cohorte que comparan la mortalidad entre ciudades con distintas concentraciones promedio de agentes contaminantes. Así pues, los resultados de estos modelos son típicamente un modelo de Riesgos relativos o un coeficiente Beta que cuantifica el cambio porcentual en el riesgo en la salud por unidad de concentración.

Una vez determinado el RR o β, se calcula la fracción de casos atribuibles a la contaminación en la población. Esto es posible a través de fórmulas de riesgo atribuible, donde la tasa de casos en ausencia de contaminación adicional se combina con el exceso de riesgo debido a

[2] Un modelo de Riesgo Relativo (RR) estima el riesgo de un impacto adverso en la salud por una mayor exposición a la contaminación del aire, comparando poblaciones con diferentes niveles de exposición. Se usa para calcular la mortalidad atribuible y los costos de salud, basándose en funciones de concentración-respuesta.

la exposición. Así pues, recordemos el modelo log-lineal, donde β es el coeficiente de la función de concentración-respuesta y P es la población expuesta. Esta función asume un incremento porcentual constante por unidad de contaminante, lo que permite estimar cuántas muertes u hospitalizaciones se evitan al reducir estas unidades. En esencia, el producto $y_0 * Z$ da el número esperado de casos de afección respiratoria en la población, mientras que la expresión $e^{\beta \Delta x} - 1$ representa la fracción de esos casos atribuible al cambio en la contaminación.

Finalmente, dado que las estimaciones epidemiológicas conllevan un grado de incertidumbre[3], es común emplear métodos como simulaciones de Monte Carlo para propagar esa incertidumbre en el modelado de impactos. Este método consiste en asignar distribuciones de probabilidad a los parámetros clave —como lo puede ser β en la función C-R y realizar un muestreo aleatorio cientos o miles de veces para así obtener una distribución de resultados posibles. Por ejemplo, BenMAP-CE utiliza simulación Monte Carlo para construir intervalos de confianza alrededor de las estimaciones medias de impacto en salud, muestreando aleatoriamente una distribución de incertidumbre en torno a los coeficientes de efecto, todo esto mediante el error estándar de los efectos conjuntos. Así, se captura la variabilidad inherente a los parámetros epidemiológicos, ofreciendo una representación más robusta y transparente del rango de posibles resultados. (Coffman, et al, 2024).

Introducción al programa BenMAP-CE

BenMAP-CE (Environmental Benefits Mapping and Analysis program – Community Edition) es un software de código abierto código abierto basado en sistemas de información geográfica, desarrollado por la Agencia de Protección Ambiental de EE. UU. (EPA) para estimar

[3] Incertidumbre: Variabilidad y posibles errores en las estimaciones de parámetros derivados de limitaciones en los datos y metodologías. Esta incertidumbre se cuantifica asignando distribuciones de probabilidad a dichos parámetros y utilizando simulaciones para obtener intervalos de confianza en los resultados.

los impactos en salud atribuibles a cambios en la calidad del aire. En términos generales, BenMAP-CE implementa los cálculos epidemiológicos descritos en la sección anterior, permitiendo así realizar análisis de beneficios en salud por mejoras en los niveles de contaminación atmosférica en una ciudad o región. A continuación, se describen sus componentes y funcionamiento.

El software permite realizar evaluaciones del impacto en salud atribuido a la contaminación atmosférica mediante una aproximación poblacional. El modelo integra datos espaciales de concentración de agentes contaminantes y distribución poblacional, para después crear superficies de exposición que luego se utilizan para aplicar funciones C-R previamente seleccionadas. BenMAP-CE emplea funciones de impacto en salud de un tipo log-lineal similar al visto previamente, calculando primero la diferencia en concentración de contaminantes para cada ubicación y luego traduciendo esta diferencia en cambios específicos en resultados de salud.

En Estados Unidos, BenMAP-CE ha sido una herramienta clave para informar políticas regulatorias a nivel local y federal relacionadas con la exposición al ozono y al material particulado. Además, el software también se ha utilizado para simular los impactos en la salud bajo diversos escenarios de cambio climático. Una de las ventajas de BenMAP-CE es su flexibilidad. Por ejemplo, Coffman et. al en 2024 propusieron una manera para incluir varios factores contaminantes a modo de Betas, usando la siguiente fórmula:

$$\Delta Y = Y_0\left(1 - e^{-(\sum_{i=1}^{n} \beta_i \times \Delta \mathrm{AQ}_i}\right) \times Pop$$

Donde ΔY es el impacto conjunto estimado en la salud atribuible al cambio simultáneo en un conjunto de n contaminantes del aire. Por su parte, β_i y $\Delta \mathrm{AQ}_i$ son, respectivamente, la estimación del efecto y el cambio en la calidad del aire para un contaminante específico i dentro del conjunto de contaminantes n.

BenMAP-CE permite al usuario ingresar bases de datos propias y adaptar las funciones C-R y de valoración económica, lo que facilita hacer un estudio preciso en cuanto al contexto geográfico de estudio. Asimismo, la interfaz gráfica facilita la interacción con el usuario y posibilita obtener resultados a manera de mapas, tablas y datos, facilitando la reproductividad del análisis.

En cuanto a datos de calidad del aire, BenMAP-CE está diseñado para aceptar datos provenientes de actividades de monitoreo como lo podrían ser los datos de la Red de Monitoreo de Calidad del Aire de Bogotá, o RMCAB o de modelos predictivos de calidad del aire. Además, puede incorporar concentraciones de contaminantes en promedio predichas para los periodos donde se presentan y donde no se presentan con frecuencia estos eventos. Asimismo, facilitando la ejecución y el escalado espacial y temporal de los datos de calidad del aire para adaptarse a las necesidades del análisis.

Estudios previos realizados con BenMAP-CE

Diversos estudios que abordan la evaluación de los impactos en la salud y los beneficios económicos asociados con cambios en la calidad del aire se han abordado tanto en Estados Unidos como en otros lugares:

Para empezar, un estudio para Bogotá realizado por Ortiz y Rojas en el año 2013 se centró en estimar los beneficios en salud y económicos derivados de la reducción de la concentración promedio anual de material particulado PM_{10} a 50 $\mu g/m^3$ entre los años 2010 y 2020. Este análisis logró desagregar los resultados por Unidades de Planeamiento Zonal (UPZ) y por Localidades. Se estimó que se podrían evitar 21,000 muertes en mayores de 30 años y 900 muertes en menores de un año debido a la exposición a largo plazo al PM_{10}. Además, se proyectaron importantes reducciones en hospitalizaciones por causas respiratorias y cardiovasculares, así como en visitas a salas de urgencias y unidades de cuidados intensivos en diferentes grupos de edad.

En términos económicos, se calcularon beneficios de alrededor de 180 mil millones de pesos en costos de enfermedad y más de 21 billones de pesos en mortalidad para el período estudiado. La interpolación Voronoi fue utilizada para determinar los niveles de inmisión de PM_{10} en la ciudad a partir de los datos de la red de monitoreo. Las funciones de concentración-respuesta empleadas se basaron en estudios realizados en poblaciones de Estados Unidos y Canadá, asumiendo una reacción similar en la población de Bogotá.

Se observó que las estimaciones de los modelos que incluían varios agentes al mismo tiempo eran generalmente más pequeñas que la suma de las estimaciones de los modelos de un solo contaminante, lo que sugiere la posible existencia de confusión por contaminantes no medidos en los modelos de un solo contaminante. Este trabajo resalta la viabilidad técnica, aunque con complejidad computacional y requerimientos de datos detallados, de realizar evaluaciones de impacto en la salud considerando la exposición a múltiples contaminantes.

Validación estadística, sensibilidad y limitaciones en el modelado con BenMAP-CE

El utilizar BenMAP-CE conlleva la necesidad de evaluar la robustez y validez de sus resultados, así como entender sus limitaciones inherentes. Una forma de validar resultados es comparar las predicciones de BenMAP-CE con observaciones epidemiológicas reales en casos donde la contaminación ha cambiado. Por ejemplo, estudios de tipo antes y después han observado que reducciones reales de PM2.5 se asocian con mejoras en la salud poblacional (mayor esperanza de vida, menor mortalidad) coherentes con las estimaciones modeladas (Pope et. al, 2005).

Para probar la robustez de los resultados de BenMAP-CE ante distintas suposiciones, se llevan a cabo análisis de sensibilidad variando parámetros de entrada y opciones del modelo. Un enfoque común es utilizar diferentes funciones C-R de fuentes alternas y comparar los resultados.

De igual forma, puede evaluarse la sensibilidad a la elección de umbral contra fáctico (concentración de referencia por debajo de la cual no se cuenta efecto y así comparar estos casos brinda un rango de posibles impactos.

Por otro lado, BenMAP-CE incorpora explícitamente la incertidumbre estadística de las funciones C-R mediante sus simulaciones Monte Carlo, pero es importante entender el alcance de lo que se incluye en dicha incertidumbre. De forma predeterminada, la herramienta asume que la única fuente de incertidumbre aleatoria proviene del error estándar del coeficiente epidemiológico (β) (Sacks et. al, 2020). Es decir, al ejecutar el análisis probabilístico, BenMAP-CE muestrea valores de β según una distribución (normal, log-normal, etc. según se especifique) centrada en el estimador puntual y con la desviación estándar reportada por el estudio epidemiológico.

Por tanto, los intervalos de confianza de BenMAP-CE tienden a subestimar la incertidumbre total del análisis, ya que dejan por fuera incertidumbres *epistemológicas* o sistémicas (relacionadas a las elecciones de modelo y datos). Para abordar esto, lo mejor es realizar escenarios alternativos (como parte del análisis de sensibilidad descrito) o utilizar la opción avanzada de BenMAP-CE que permite asignar distribuciones de probabilidad a otros parámetros si se cuenta con información para ello. En consecuencia, la comunicación de resultados debe aclarar que la incertidumbre cuantificada es parcial. Por ejemplo, en un estudio se mencionaba que *"el intervalo calculado refleja la incertidumbre estadística del estudio epidemiológico, pero no otras fuentes de incertidumbre asociadas a los insumos restantes"* (Sacks et. al, 2020).

A pesar de que BenMAP-CE es una herramienta robusta y ampliamente usada, existen varias limitaciones a considerar al interpretar sus estimaciones. Para empezar, tenemos los supuestos de transferencia de funciones C-R. BenMAP-CE generalmente aplica coeficientes de riesgo derivados de estudios internacionales (frecuentemente de Europa o Norteamérica) a

poblaciones locales. Esto asume que la relación exposición-respuesta es igual en todos los contextos, lo cual puede no ser preciso. Por ejemplo, según un estudio de Chung et. al en 2022, se ha encontrado que el modelo integrado IER (basado mayormente en datos de EE.UU./Europa) podría subestimar el riesgo real en regiones con contaminación muy alta, como China, especialmente para ciertas enfermedades. Asimismo, las poblaciones pueden diferir en susceptibilidad (perfil de comorbilidades, estructura etaria) o la mezcla de contaminantes puede modificar la respuesta. Por ello, aplicar funciones C-R fuera del rango de exposición original añade incertidumbre. En ausencia de funciones locales, se acepta esta transferencia con cautela, pero se reconoce que es una limitación estructural. En consecuencia, los resultados deben interpretarse como aproximaciones mientras no haya evidencia epidemiológica local más específica.

Asimismo, BenMAP-CE permite calcular contaminantes por separado, pero la suma de impactos de, por ejemplo, PM2.5 y ozono sobre la mortalidad, debe considerar que un mismo deceso podría ser atribuido a ambos en estudios separados. (Coffman et. al, 2024) En la práctica, para evitar este problema, los análisis de BenMAP-CE suelen enfocarse en el contaminante considerado más importante o no suman desenlaces idénticos de múltiples contaminantes. La necesidad de un enfoque con múltiples contaminantes es reconocida en la comunidad científica, pero las herramientas y variaciones desarrolladas todavía no son 100% estables.

La calidad del resultado de BenMAP-CE es tan buena como la de sus insumos, esto implica que, en muchos países en desarrollo, donde los datos de monitoreo de aire son escasos o poco representativos espacialmente, las estadísticas de salud pueden tener subregistro y la resolución poblacional detallada puede faltar. Estas carencias de datos introducen error en las estimaciones que el modelo no puede corregir. Por ejemplo, si la concentración de PM2.5 en una zona se subestima por falta de monitores, BenMAP-CE calculará menos casos de los reales.

5.5 Integración de los modelos ML con BenMAP-CE

Dado un conjunto de entrenamiento $D = \{(x_i, y_i)\}_{i=1}^{n}$, un modelo f_θ minimiza el riesgo empírico regularizado:

$$J(\theta) = \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i)) \; + \; \lambda \Omega(\theta)$$

$$J(\theta) = \frac{1}{n} \sum_{i=1}^{n} L(y_{\,i}, f_{\,\theta}(x_{\,i})) \; + \; \lambda \Omega(\theta)$$

$$\hat{y}(x) = \frac{1}{B} \sum_{b=1}^{B} T_{\,b}(x)$$

$$I(S) = \frac{1}{|S|} \sum_{i \in S} (y_{\,i} - \bar{y}_{\,S})^2$$

$$\Delta I \; = \; I(S) \; - \; [\frac{|S}{|S|} I(S_{\,L}) \; + \; \frac{|S}{|S|} I(S_{\,R})]$$

$$f(x) = w_{\,q(x)}$$

$$\Omega(f) = \gamma T \; + \; \frac{\lambda}{2} \|w\|^{\,2}$$

$$\mathcal{L}^{\,(t)} \approx \sum_{i} g_{\,i} f_{\,t}(x_{\,i}) \; + \; \frac{1}{2} h_{\,i} f_{\,t}(x_{\,i})^2 \; + \; \Omega(f_{\,t}) \; + \; const.$$

Luego, la integración de modelos ML con BenMAP-CE está dada por:

$$RR_{\,e}(\Delta C_{\,g}) = exp(\beta_{\,e} \cdot \Delta C_{\,g})$$

$$\Delta Y_{\,g,e} = y_{\,0\,g,e} \cdot P_{\,g} \cdot (1 - exp(-\beta_{\,e} \cdot \Delta C_{\,g}))$$

Donde la conversión de salidas ML en superficies de concentración por escenario

$$\hat{C}_{\,g,t} \approx \sum_{i} w_{\,g,i} \cdot \hat{y}_{\,i,t}$$

$$\sum_{i} w_{\,g,i} = 1$$

Lo anterior se puede resumir en la Figura 1.

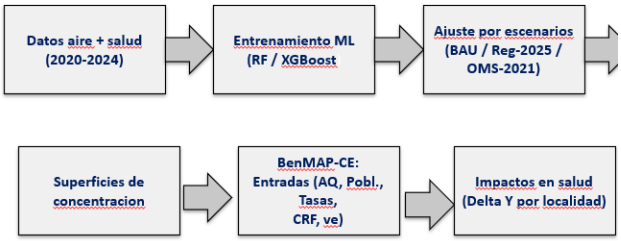


Ilustración 1. *Transición de ML a BenMAP-CE*

9. Algoritmo CatBoost

Funcionamiento Técnico

CatBoost es un algoritmo de aprendizaje automático basado en la técnica de boosting por gradiente. Su nombre proviene de 'Categorical Boosting', lo que refleja su principal característica: el manejo eficiente y nativo de variables categóricas. Este modelo pertenece a la familia de métodos ensemble, donde múltiples árboles de decisión se combinan para mejorar la capacidad predictiva y reducir errores individuales. A diferencia de otros algoritmos de boosting, como XGBoost o LightGBM, CatBoost incorpora mecanismos avanzados que corrigen sesgos y previenen el sobreajuste durante el entrenamiento (Dorogush, Ershov & Gulin, 2018).

El funcionamiento de CatBoost parte de un modelo base, usualmente una predicción promedio o constante, sobre el cual se van construyendo secuencialmente árboles de decisión. Cada nuevo árbol busca corregir los errores residuales del modelo anterior, de manera que las predicciones finales son el resultado de la suma ponderada de todos los árboles entrenados. En cada iteración, el algoritmo optimiza una función de pérdida específica, como el error cuadrático medio o la entropía cruzada, mediante gradientes. De esta forma, cada árbol adicional se centra en los casos mal clasificados o estimados, mejorando gradualmente la precisión global del modelo.

Una de las principales innovaciones de CatBoost es su tratamiento de las variables categóricas mediante un esquema denominado target encoding ordenado. En lugar de convertir las categorías en variables binarias o numéricas simples, el algoritmo las transforma utilizando promedios del valor objetivo calculados en orden aleatorio, evitando así la fuga de información. Además, el modelo emplea una estructura de árboles simétricos, donde todos los nodos de un mismo nivel usan la misma característica y umbral de división. Esta característica reduce la complejidad del modelo y mejora la estabilidad computacional, permitiendo un entrenamiento más rápido y con menor riesgo de sobreajuste (Prokhorenkova et al., 2018).

Base matemática

CatBoost se basa en el principio del boosting por gradiente, donde se construye un modelo aditivo compuesto por árboles de decisión. En cada paso, el modelo busca minimizar una función de pérdida mediante la adición de un nuevo árbol que aproxima el gradiente negativo del error. Esto significa que cada árbol intenta corregir las discrepancias entre las predicciones actuales y los valores observados. La contribución de cada árbol se pondera mediante una tasa de aprendizaje, que controla la magnitud de la corrección agregada al modelo.

En términos generales, el modelo puede representarse como una suma de árboles. Si se denota por $F_m(x)$ la predicción del modelo en la iteración m, entonces la actualización se expresa como:

$$F_m(x) = F_{m-1}(x) + vf_m(x)$$

donde v es la tasa de aprendizaje y $f_m(x)$ representa el nuevo árbol entrenado en esa iteración. Cada nuevo árbol se construye para minimizar el error residual entre las predicciones del modelo anterior y los valores reales.

La función de pérdida mide qué tan alejadas están las predicciones de los valores observados. Para determinar el nuevo árbol, se busca la función y f_m que minimice la suma de las pérdidas en todas las observaciones. Es decir, se selecciona el árbol que más reduce el error total. Los gradientes negativos, también llamados pseudo-residuales, indican la dirección en la que deben corregirse las predicciones. Cada residuo se calcula como la derivada parcial negativa de la función de pérdida con respecto a la predicción actual.

Por ejemplo, en el caso de la regresión, la pérdida utilizada suele ser el error cuadrático medio, expresado como:

$$L(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2$$

y su gradiente negativo equivale simplemente a la diferencia entre el valor real y la predicción actual, es decir,

$$r = (y - \hat{y})$$

En el caso de la clasificación binaria, la función de pérdida más común es la entropía cruzada o "log-loss", que se define como:

$$L(y, \hat{y}) = -[y * \log(\hat{y}) + (1 - y) * \log(1 - \hat{y})]$$

y el gradiente negativo correspondiente refleja el ajuste necesario en la probabilidad predicha.

CatBoost también incluye regularización L2 sobre los valores asignados a las hojas de cada árbol, con el fin de evitar el sobreajuste. Esta regularización penaliza valores extremos en las hojas, y el valor óptimo en cada una de ellas se obtiene como el promedio ponderado de los residuos en

esa región dividido por la suma de los pesos más el término de penalización. En términos sencillos, esto suaviza el ajuste y previene que el modelo memorice valores atípicos.

Una de las contribuciones distintivas de CatBoost es el tratamiento de las variables categóricas mediante estadísticas objetivo. Cada categoría se transforma en un valor numérico que representa el promedio del valor objetivo (por ejemplo, la concentración de un contaminante) para esa categoría. Para evitar sesgos, el promedio se calcula usando solo información disponible en otras observaciones, nunca la de la misma fila. La fórmula puede expresarse de manera conceptual como:

$$Valor\ codificado = \frac{\sum_{j:\, x_j = c} y_j + a * p}{\sum_{j:\, x_j = c} 1 + a}$$

Donde y_j son los valores del objetivo, p es un valor previo (prior), a es un parámetro de suavizado que evita divisiones por valores pequeños y $\sum_{j:\, x_j = c} 1$ es el número de observaciones con la categoría c.

En su versión "ordenada", CatBoost aplica este cálculo progresivamente a lo largo de una permutación de los datos, de forma que cada registro se codifica solo con información proveniente de observaciones anteriores. Esto garantiza independencia entre predictores y variable respuesta, evitando el problema de fuga de información (information leakage).

El proceso de entrenamiento de CatBoost incluye dos elementos técnicos distintivos: el ordered boosting y el manejo ordenado de categorías. En el ordered boosting, los árboles se entrenan sobre permutaciones parciales del conjunto de datos para evitar el sesgo por información futura. Por su parte, el manejo ordenado de categorías se basa en la estimación progresiva de promedios del valor objetivo, calculados a partir de observaciones previas y nunca del propio

registro, garantizando la independencia estadística entre características y respuesta. Ambos mecanismos hacen que CatBoost produzca modelos más estables y generalizables en comparación con otros métodos boosting tradicionales.

Formalmente, el modelo resultante puede entenderse como una combinación lineal de árboles de decisión, donde cada árbol aporta una fracción al resultado final. El modelo final se obtiene al sumar las predicciones ponderadas de todos los árboles entrenados. De este modo, la predicción de CatBoost para una observación es el promedio acumulado de las salidas de los árboles ajustados durante el proceso iterativo. Esta formulación asegura que el modelo capture tanto relaciones lineales como no lineales entre las variables predictoras y la variable objetivo.

CatBoost en aplicaciones ambientales

En los últimos años, el modelo CatBoost ha sido utilizado con éxito en diferentes áreas del modelamiento ambiental, especialmente en la predicción de contaminantes atmosféricos, índices de calidad del aire y estudios de salud pública relacionados con la contaminación. Su capacidad para manejar variables categóricas y numéricas de manera conjunta, junto con su robustez frente a datos ruidosos, lo convierten en una herramienta adecuada para estos contextos.

Wang et al. (2023) aplicaron CatBoost en la predicción simultánea de PM2.5 y ozono en la llanura de Fenwei, China, obteniendo valores de R^2 superiores a 0.7 y demostrando que el modelo supera a otros algoritmos como Random Forest y XGBoost. Mediante análisis de interpretabilidad con valores SHAP, identificaron que la concentración de PM_{10} y la radiación solar neta fueron las variables más influyentes en las predicciones. Estos resultados evidencian la utilidad de CatBoost para comprender la dinámica entre variables meteorológicas y contaminantes atmosféricos.

El éxito de CatBoost en estas aplicaciones puede explicarse por su arquitectura ordenada y regularizada, que permite capturar interacciones complejas entre variables ambientales sin perder

generalización. Además, su compatibilidad con métodos de interpretabilidad, como SHAP, facilita la comprensión de los factores determinantes de la contaminación. En consecuencia, CatBoost se presenta como una herramienta sólida para estudios ambientales, ya que combina precisión, eficiencia computacional y transparencia analítica en la interpretación de resultados.

Metodología

10. Diseño del estudio y alcance

El presente estudio se enmarca en un diseño observacional de tipo ecológico y correlacional, orientado a evaluar la asociación entre la contaminación atmosférica y los impactos en salud pública y la carga económica en la ciudad de Bogotá. La unidad de análisis puede ser espacial (localidades o estaciones de monitoreo) y temporal (series de tiempo diarias o anuales).

En resumen, la metodología de estudio está compuesta en las siguientes fases:

Fase 1:

i. Implementar la metodología CRISP-DM

Fase 2:

i. Selección de datos de entrenamiento y de validación
ii. Ajustes de hiperparámetros y entrenamiento CatBoost
iii. Evaluación de desempeño (RMSE, MAE, R^2 ajustado) y de cobertura de intervalos predictivos.
iv. Proyecciones bajo escenario BAU
v. Selección del mejor modelo por contaminante y horizonte; exportar predicciones.

Fase 3:

i. Integración de los datos en BenMAP-CE con: Funciones concentración-respuesta ajustadas y tasas de mortalidad/morbilidad 2026 (población proyectada).
ii. Simulación del impacto sanitario

11. Fuentes de datos y variables

Se integrarán diversas fuentes de datos oficiales y confiables para alimentar los análisis ambientales y de salud. Entre ellas se encuentran:

Contaminación del aire: Se emplearán los registros de la Red de Monitoreo de Calidad del Aire de Bogotá, administrada por la Secretaría Distrital de Ambiente. Esta red aporta datos de concentración de contaminantes criterio (PM2.5, PM_{10}, NO2, Ozono, etc.) medidos en estaciones distribuidas en la ciudad. En particular, se obtendrán las concentraciones horarias y/o diarias de PM2.5 como variable principal de exposición ambiental.

Climatología: Se complementarán los registros de la RMCAB con datos de climatología de la NASA, dado el grado de completitud de estos, estas variables pueden incluir temperatura, velocidad del viento, presión, humedad relativa, precipitación, entre otras. Por otro lado, para realizar las predicciones de las concentraciones del contaminante objetivo es necesario contar con datos de predicciones climatológicas, que están disponibles en el portal del programa Copernicus, de la Unión Europea.

Datos demográficos y de población: Para estimar adecuadamente los impactos en salud, se utilizarán las proyecciones y censos poblacionales del Departamento Administrativo Nacional de Estadística (DANE). En particular, se obtendrá la población expuesta por sexo en Bogotá durante el periodo de estudio a nivel de localidad. Estos datos permiten calcular las tasas base de mortalidad y morbilidad (baseline incidence) necesarias en las funciones de impacto en salud.

Indicadores de salud pública: La información de resultados en salud (variables dependientes) provendrá de registros oficiales de la Secretaría Distrital de Salud de Bogotá y otras fuentes nacionales. Se recopilarán datos sobre mortalidad (por causas cardiovasculares,

respiratorias y todas las causas, idealmente con detalle diario o mensual), morbilidad respiratoria asociados en la literatura con la contaminación. De no estar disponibles datos individualizados, se usarán datos agregados de epidemiología (por ejemplo, tasas de mortalidad anual por causa para Bogotá). Cada registro de salud será asociado temporalmente y/o geográficamente con los niveles de contaminación correspondientes.

Análisis exploratorio y espaciotemporal

Antes de probar hipótesis específicas, se realizará un análisis exploratorio de datos (EDA) exhaustivo. Esto incluye estadísticos descriptivos de las variables principales: distribuciones de concentraciones de PM2.5 (promedios, máximos, mínimos, percentiles) y tasas de eventos de salud. Un análisis espaciotemporal también será clave para el mapeo de las concentraciones promedio de PM2.5 por localidad de la ciudad para detectar heterogeneidad geográfica; paralelamente, se mapearán indicadores de salud (ej. mortalidad respiratoria por localidad) para ver correlaciones visuales entre zonas más contaminadas y mayor impacto en salud.

Se calcularán correlaciones entre los niveles de contaminación y las tasas de efectos en salud en diferentes escalas temporales. Dado que la literatura indica que ciertos grupos son más susceptibles a la contaminación (por ejemplo, niños, adultos mayores), se analizarán por separado las tendencias en subpoblaciones.

12. Modelo de aprendizaje automático

Se implementará un modelo basado en CatBoost realizado en Python, empleando librerías como CatBoost, scikit-learn y pandas para el procesamiento de datos y entrenamiento del modelo. Las variables predictoras consideradas incluyen la localidad, fecha y hora —con las variables que se pueden derivar de esto, como día de la semana— y variables climatológicas. La variable objetivo

del modelo será la concentración de material particulado $PM_{2.5}$, que es el principal contaminante en este estudio.

Asimismo, se realizará un proceso sistemático de optimización de los hiperparámetros del modelo mediante técnicas de búsqueda aleatoria. Entre los parámetros ajustados se incluyen los observados en la Tabla 2.

Tabla 2. *Descripción hiperparámetros modelo CatBoost*

Parámetro	Funcionalidad
Depth	Controla la profundidad máxima de los árboles, regulando la complejidad del modelo y su capacidad para capturar relaciones no lineales.
Learning Rate	Determina la magnitud de los pasos de actualización del gradiente.
L2 Leaf Reg	Penaliza la complejidad de los árboles para evitar el sobreajuste.
Bagging Temperature	Controla el grado de aleatoriedad en el muestreo de datos
Random Strength	Define la intensidad del ruido aleatorio añadido a las divisiones de los árboles
Iterations	Indica el número total de árboles que conforman el ensamble
Bootstrap Type	Especifica el método de muestreo de observaciones. (Para esta aplicación, el valor siempre será Bayesian, pues este enfoque introduce pesos probabilísticos para diversificar el entrenamiento, lo que resulta más práctico de cara a este estudio).
mamaEarly Stopping Rounds	Implemente un criterio de detención temprana. (Para esta aplicación, el valor siempre será 100, lo que implica que el entrenamiento se interrumpirá cuando no se mejore el error de validación durante 100 rondas).

Estos ajustes permitirán alcanzar un equilibrio adecuado entre sesgo y varianza, lo que optimiza la capacidad de generalización del modelo.

Una ventaja relevante del uso del modelo CatBoost es la posibilidad de evaluar la importancia relativa de cada variable predictora mediante medidas como Feature Importance. Esto permitirá identificar qué factores climatológicos y ambientales tienen mayor incidencia en las concentraciones de contaminantes, aportando información valiosa para la estimación de niveles de exposición en áreas donde no existen estaciones de monitoreo en sitio.

13. Estimación de los efectos en salud con BenMAP-CE

Para cuantificar de forma integral los efectos en salud atribuibles a la contaminación bajo distintos escenarios, se implementará la ya mencionada herramienta BenMAP-CE. Así pues, la implementación en este estudio seguirá los pasos estándar recomendados por la EPA (2015) y trabajos previos.

Preparación de datos de entrada:

Se formatearán los datos de concentración de PM2.5 de Bogotá en el formato requerido por BenMAP. Dado que BenMAP puede trabajar con superficies interpoladas de contaminación, se generará una grilla espacial de concentraciones anuales promedio de PM2.5 sobre Bogotá, usando técnicas de interpolación a partir de las estaciones de la Red de Monitoreo.

Datos de salud y población en BenMAP:

Se cargarán las tasas base de los desenlaces de salud de interés. Por ejemplo, la tasa de mortalidad anual por causas cardiovasculares en Bogotá (por 100,000 habitantes) y la tasa de ingresos hospitalarios por asma en niños (por 100,000) se ingresan para cada subgrupo poblacional correspondiente. Asimismo, se ingresarán los datos de población georreferenciada; típicamente, BenMAP permite suministrar la población por celdas de la grilla o por áreas para hacer el cálculo agregado de casos. En este estudio, se usarán las localidades de Bogotá, asignando cada área a las concentraciones de la grilla interpolada.

Selección de funciones concentración–respuesta:

Basados en la revisión de literatura, se elegirá para cada desenlace de salud una función C-R apropiada dentro de BenMAP. El software BenMAP-CE ya incluye por defecto varias funciones epidemiológicas derivadas de estudios (especialmente para PM2.5 y Ozono, enfocadas en contextos de EE. UU.), pero se pueden ingresar funciones personalizadas. En este caso, se optará por funciones adaptadas a Latinoamérica si están disponibles, o se utilizarán las de OMS/HRAPIE Europe ajustando si es necesario. Por ejemplo, para mortalidad total en adultos mayores de 30 años, se podría usar la función de Hoek et al. (2013) con RR 1.062 por 10 μg/m³, para mortalidad cardiovascular la función de Aldegunde et. al (2023) que reportó ~11% por 10 μg/m³, etc. Cada función se introduce en BenMAP con su coeficiente Beta y el intervalo de confianza para propagar la incertidumbre. Este paso es crítico ya que define la sensibilidad del resultado: se realizará con base en publicaciones revisadas por pares, citando la fuente de cada coeficiente en la documentación del estudio.

Configuración de análisis y ejecución en BenMAP:

Una vez cargados los datos de contaminación (escenario base y alternativo), población y funciones C-R, se configura BenMAP-CE para ejecutar un análisis de impacto en salud. Esto implicará definir el ámbito geográfico (Bogotá, con su extensión y división en celdas o áreas), el contaminante (PM2.5) y los endpoints de salud para evaluar. Se ejecutará entonces la simulación, donde el software calcula la variación esperada en los efectos sobre la salud para cada combinación de celda poblacional y endpoint, comparando el escenario alternativo con el base. Por ejemplo, en un escenario de reducción, BenMAP estimará cuántos casos de mortalidad se evitarían al bajar la contaminación, donde los valores negativos de Δy indican casos evitados, es decir, beneficios en salud. En un escenario de aumento de contaminación, Δy será positivo, indicando casos adicionales

atribuibles al incremento de emisiones. BenMAP también permite realizar análisis de incertidumbre como las ya mencionadas simulaciones de Monte Carlo. Tomando en cuenta la variabilidad en las funciones C-R, se implementarán iteraciones de Monte Carlo (p. ej., 1,000 iteraciones) para obtener intervalos de confianza del estimado de casos evitados/atribuibles, confiriendo significancia estadística a los resultados.

14. Supuestos a tener en cuenta

El procesamiento de datos realizado en este estudio, incluyendo la interpolación espacial de variables meteorológicas y la agregación temporal de series de PM2.5 y predictores, incorpora una serie de supuestos metodológicos que es necesario examinar a la luz del modelo conceptual de transporte y transformación del material particulado fino, así como de las escalas espaciales y temporales características descritas previamente en el marco teórico. El análisis crítico de estos supuestos permite no solo delimitar el alcance y las fuentes potenciales de incertidumbre asociadas al modelado, sino también justificar la validez de las decisiones adoptadas dadas las restricciones inherentes a la disponibilidad y resolución de los datos.

Supuestos asociados a la interpolación espacial

Interpolación bilineal de campos meteorológicos y climatológicos

La interpolación bilineal utilizada para derivar variables desde la grilla de ERA5 hacia los puntos de interés supone que dichas variables presentan una variación espacial suave y continua dentro de cada celda. Este supuesto resulta razonable para variables de gran escala, como la temperatura del aire, la presión o la humedad relativa, cuyo comportamiento está dominado por campos sinópticos cuya variabilidad horizontal suele ser menor que la resolución del reanálisis.

No obstante, a escala urbana, fenómenos relevantes para la dispersión del PM2.5 —como la configuración de los vientos locales, la turbulencia mecánica y térmica o el desarrollo de la capa

límite atmosférica— pueden presentar variabilidad significativa a distancias inferiores al tamaño de celda de ERA5. Tal discrepancia implica que la interpolación bilineal no puede capturar completamente microclimas urbanos ni gradientes asociados a la orografía local, especialmente aquellos que contribuyen a zonas de acumulación o estancamiento identificadas en la literatura para Bogotá.

Sin embargo, dado que el objetivo del estudio es caracterizar tendencias integradas a escala diaria, y no reconstruir campos atmosféricos de alta resolución, el uso de ERA5 y su interpolación es metodológicamente adecuado y consistente con prácticas establecidas en estudios epidemiológicos y de modelación de calidad del aire en contextos con limitaciones de datos.

Desde la perspectiva del modelo conceptual explicado anteriormente, la interpolación bilineal de ERA5 a las estaciones asume que las variables meteorológicas de gran escala como temperatura, presión, humedad, viento a 10 m se comportan de forma lo suficientemente suave como para que los gradientes a escala de subcelda no cambien de manera abrupta.

Esto es consistente con el hecho de que procesos que controlan el PM2.5 como los patrones sinópticos de viento, los regímenes de lluvia, entre otros, operan en escalas espaciales de decenas a cientos de kilómetros y en las escalas espaciales de variabilidad. Sin embargo, hay que tener en cuenta que el modelo conceptual subyacente reconoce procesos de micro y mesoescala (vientos canalizados a través de huecos urbanos, turbulencia mecánica generada por edificaciones) que pueden cambiar fuertemente en distancias de pocos cientos de metros.

Al trabajar con campos ERA5 interpolados bilinealmente se adopta el supuesto de que esta variabilidad sub-grilla se integra en términos de su efecto neto sobre las concentraciones diarias de PM2.5, es decir, que lo importante para el análisis es el “forzamiento meteorológico promedio”

sobre la ciudad y no el detalle fino de la circulación local en cada calle. Este supuesto es razonable para un estudio cuyo objetivo es estimar impactos en salud a escala de localidad y no a nivel de calle o manzana, asumiendo configuraciones locales de transporte y acumulación no serán explícitamente representadas o consideradas.

Representatividad espacial de las estaciones de monitoreo

El análisis presupone que las concentraciones observadas en cada estación son representativas de su área circundante y pueden ser utilizadas sin correcciones adicionales para el modelado estadístico. Si bien se reconoce que la Red de Monitoreo de Calidad del Aire de Bogotá presenta una cobertura heterogénea, la literatura respalda el uso de registros puntuales como aproximación válida para modelar variaciones temporales de PM2.5, especialmente cuando el interés no recae en la generación de mapas continuos, sino en establecer relaciones funcionales entre predictores y concentraciones. Asimismo, la integración de predictores meteorológicos y de transporte a escala de ciudad complementa la información puntual, mitigando parcialmente los efectos de la heterogeneidad espacial.

Supuestos de agregación temporal

Expansión diaria de climatologías mensuales

La asignación de valores mensuales de climatologías a cada día del mes constituye un supuesto de estacionariedad intramensual. Este tipo de aproximación es habitual cuando se integran predictores de gran escala en estudios de calidad del aire, ya que estas climatologías representan condiciones promedio que influyen en el comportamiento atmosférico regional. Sin embargo, dicha aproximación no captura la variabilidad intramensual que podría incidir en fenómenos de interés, como episodios de inversión térmica o cambios en la altura de la capa de mezcla.

A pesar de ello, esta decisión metodológica es razonable y metodológicamente defendible, pues el objetivo del estudio no es la reconstrucción detallada de la meteorología diaria, sino la

generación de predictores robustos que preserven patrones sinópticos amplios y permitan estimar escenarios de exposición compatibles con los requerimientos de BenMAP-CE. La literatura epidemiológica respalda ampliamente el uso de promedios mensuales o climatologías en modelos de exposición cuando la variabilidad subdiaria o intramensual no es el foco principal del análisis.

La expansión diaria de climatologías mensuales de ERA5 implica un promedio fuerte sobre la variabilidad intramensual de las condiciones meteorológicas. En términos del modelo conceptual, esto supone que las escalas rápidas (por ejemplo., días concretos con episodios de frentes fríos o lluvias extremas) tienen un impacto limitado sobre los promedios diarios de exposición en lo proyectado 2026, en comparación con el patrón estacional dominante capturado por la climatología anterior.

El escenario 2026 no busca reconstruir un año meteorológico específico, sino un año compatible con las condiciones climáticas de referencia. Esto hace que las series de predictores sean suaves y reproducibles, pero también que no se represente la posible ocurrencia de episodios extremos que según el modelo conceptual amplifican notablemente las concentraciones de PM2.5 en periodos secos.

Agregación diaria de PM2.5 a través de promedios

El empleo de valores diarios de PM2.5 implica suponer que el ciclo diurno —caracterizado por picos en horas de la mañana y la noche— puede ser representado adecuadamente mediante un valor promedio. Aunque este supuesto elimina la posibilidad de analizar dinámicas horarias asociadas a la capa límite o a la movilidad urbana, es coherente con los requerimientos metodológicos de BenMAP-CE, que opera fundamentalmente sobre promedios diarios. La agregación temporal mantiene la coherencia interna del estudio y evita incompatibilidades entre resoluciones de los distintos predictores.

Además, los efectos agudos asociados a incrementos horarios del PM2.5, si bien relevantes en estudios de riesgo inmediato, tienen menor impacto en análisis de carga de enfermedad basados en exposiciones diarias o anuales. Por lo tanto, aunque la agregación diaria atenúa la expresión de los picos subdiarios, no compromete la validez de las estimaciones de impacto en salud desarrolladas en este estudio.

Consistencia de las escalas de análisis con las escalas características del sistema

El modelo conceptual establece que el comportamiento del PM2.5 en Bogotá está gobernado por procesos que operan de forma simultánea en múltiples escalas temporales (horaria, diaria, semanal, estacional, interanual) y espaciales (microescala de calle, escala de localidad, escala regional).

Luego, las decisiones de procesamiento realizadas en este trabajo, como el uso de promedios diarios, climatologías mensuales expandidas, interpolación bilineal de ERA5 y agregación por localidad sitúan explícitamente el análisis en la escala diaria-localidad, que es coherente tanto con el tipo de desenlace de salud modelado (mortalidad anual proyectada) como con la resolución operativa de BenMAP-CE y la disponibilidad de datos.

Por lo tanto, se puede considerar que la agregación diaria actúa como un filtro de paso bajo sobre la señal de PM2.5, eliminando las oscilaciones de alta frecuencia asociadas a procesos subdiarios (picos en hora pico, variaciones transitorias de la capa límite) y reteniendo las variaciones más lentas asociadas a episodios de varios días y a cambios estacionales.

A su vez, la agregación por localidad integra la información de varios sitios de medición en un único valor representativo del nivel de exposición para la población residente. El resultado es un conjunto de series que reflejan patrones de exposición que son consistentes con las escalas de tiempo y espacio a las que responden las funciones concentración-respuesta utilizadas en BenMAP-

CE, las cuales se derivan de estudios que trabajan con métricas agregadas de concentración (diaria, mensual o anual) como lo realiza BenMAP-CE.

Finalmente, se asume que, aunque las variables de gran escala (p. ej. ERA5 y productos NASA) están definidas sobre celdas de decenas de kilómetros, sus valores son razonablemente representativos del entorno de cada estación dentro de cada localidad, tras la interpolación a las coordenadas puntuales.

Resultados

15. Datos obtenidos

Tras la extracción de los datos de Contaminación, Meteorología, Salud y Población, se obtuvieron finalmente los registros presentados en la tabla XX, junto con su diccionario de datos.

Tabla 3. *Dimensionalidad bases de datos extraídas*

Base de datos	Filas	Columnas
Contaminación	658939	8
Climatología	21915	8
Población	80	44
Salud - Mortalidad	3532	7

Tabla 4. *Detalle base de datos contaminación del aire*

Columna	Descripción
DateTime	Fecha y hora del registro tomado
Estacion	Nombre de la estación de monitoreo
PM25	Concentración de material particulado PM2.5 en microgramos por metro cúbico
PM10	Concentración de material particulado PM10 en microgramos por metro cúbico
NO2	Concentración de NO2 en partes por billón
OZONO	Concentración de NO2 en partes por billón
SO2	Concentración de NO2 en partes por billón
CO	Concentración de NO2 en partes por millón

Tabla 5. *Detalle base de datos climatología*

Columna	Descripción
PS	Presión atmosférica a nivel de la superficie del terreno (kPa).
PRECTOTCORR	Cantidad total de precipitación diaria acumulada en la superficie terrestre (mm/día), corregida por sesgos del modelo de reanálisis.
RH2M	Porcentaje de humedad relativa a 2m de altura
T2M	Temperatura del aire a 2 m sobre la superficie (°C)
WS2M	Velocidad promedio del viento a 2 m sobre la superficie (m/s).

Tabla 6. *Detalle base de datos poblacional*

Columna	Descripción
NOM_LOC	Nombre de la localidad
Año	Año de la proyección
Hombres 0-100+	Grupo de 21 variables que hace el conteo estimado de mujeres por edad
Mujeres 0-100+	Grupo de 21 variables que hace el conteo estimado de hombres por edad

Tabla 7. Detalle base de datos salud - Mortalidad

Columna	Descripción
Año	Año de la fatalidad
Localidad	Localidad de la fatalidad
Sexo	Sexo de los fallecidos
Rango Edad	Rango de edad de los fallecidos (de 5 en 5)
Grupo causa	Grupo de causa de muerte
Causa	Causa de muerte
Casos	Número de casos

16. Análisis exploratorio espacial de datos (EDA)

Para empezar, se realizó un análisis exploratorio de datos referente al conjunto de datos de contaminación del aire en la ciudad de Bogotá en el periodo temporal de 2021-2024, con el fin de lograr un entendimiento correcto de los datos para así lograr el mejor pre-procesado de cara a los modelos de aprendizaje de máquina y al modelamiento en la herramienta BenMAP-CE. Dicho conjunto de datos proviene de la Red de Monitoreo de Calidad del Aire de Bogotá (RMCAB), la cual recopila y publica información oficial sobre concentraciones de contaminantes atmosféricos

en diferentes estaciones distribuidas en la ciudad, garantizando su disponibilidad pública a través de su portal web institucional (RMCAB, s. f.).

Es importante aclarar que dadas las limitaciones que presenta la RMCAB en cuanto a que no todas las localidades de la ciudad tienen una estación de monitoreo de calidad del aire, el análisis a realizar estará sujeto a las 13 localidades que se presentan en el conjunto de datos donde hay 19 estaciones de monitoreo repartidas de la siguiente forma:

Tabla 8. *Cantidad de estaciones por localidad de Bogotá D.C*

Localidad	Número de estaciones
Suba	3
Engativá	2
Puente Aranda	2
Kennedy	2
Fontibón	2
Chapinero	1
Ciudad Bolívar	1
Barrios Unidos	1
Santa Fe	1
San Cristóbal	1
Tunjuelito	1
Usaquén	1
Usme	1

Conjunto de datos

El conjunto de datos de contaminación tiene la siguiente estructura:

Tabla 9. *Estructura conjunto de datos contaminación del aire*

Campo	Descripción	Unidad / Formato
DateTime	Fecha y hora de la medición	dd/mm/yyyy HH:MM
Estacion	Nombre de la estación de monitoreo de calidad del aire	Texto libre
PM25	Concentración de partículas finas con diámetro ≤ 2.5 µm	µg/m³
PM10	Concentración de partículas con diámetro ≤ 10 µm	µg/m³
NO2	Concentración de dióxido de nitrógeno	ppb

OZONO	Concentración de ozono (O3)	ppb
SO2	Concentración de dióxido de azufre	ppb
CO	Concentración de monóxido de carbono	ppm
Localidad	Localidad de Bogotá donde está ubicada la estación	Texto sin tildes

Análisis de Nulidad

Para tener un correcto entendimiento de los datos, es necesario realizar un análisis de nulidad del conjunto de datos. Tras realizar este análisis en Python los resultados fueron los siguientes:

Tabla 10. *Análisis de nulidad para el conjunto de datos de contaminación del aire*

Variable	Nulos	Ceros	% Nulos	% Ceros
PM2.5	51073	31517	7.75%	4.78%
PM10	67298	4042	10.21%	0.61%
NO2	122285	697	18.56%	0.11%
Ozono	174918	4619	26.55%	0.70%
SO_2	342954	2522	52.05%	0.38%
CO	98503	809	14.95%	0.12%
Temperatura	167237	12	25.38%	0.00%
Precipitación	128873	451577	19.56%	68.53%
Vel. Viento	205680	689	31.21%	0.11%

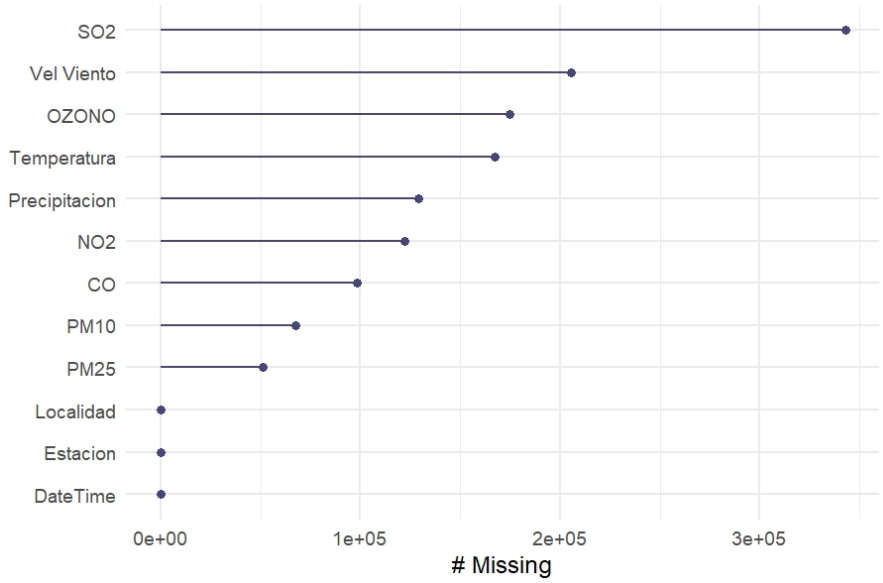


Ilustración 2. *Contaminantes y medidas meteorológicas con mayor cantidad de valores faltantes*

En el análisis de datos de calidad del aire, se observa que PM 2.5 (7,75 % de nulos) y PM 10 (10,21 %) son las variables con mejor cobertura, mientras que SO 2 presenta una cobertura muy baja (52,05 %), lo cual es preocupante de cara a la fiabilidad de estos datos, más aun teniendo en cuenta que, según las normas de la EPA, un año debe tener al menos 75 % de capturas válidas cada trimestre para ser considerado completo en PM 2.5

Análisis Estadístico datos en Bruto PM2.5

Para el análisis exploratorio de datos se decidió priorizar el contaminante PM2.5 frente a los demás (PM10, NO2, ozono, SO_2 y CO), debido a que las partículas finas representan un indicador más sensible y preciso de la calidad del aire en entornos urbanos. A diferencia del PM10, el PM2.5 puede penetrar profundamente en el sistema respiratorio e incluso llegar al torrente sanguíneo, lo que lo convierte en un parámetro clave para evaluar los riesgos en salud pública asociados a la contaminación atmosférica.

Además, el PM2.5 refleja de manera más directa los efectos de fuentes típicas en Bogotá, como el transporte vehicular y la quema de combustibles, factores determinantes en el deterioro ambiental de la ciudad. Desde la perspectiva estadística, aunque el conjunto de datos presenta valores nulos y ceros en todas las variables, el PM2.5 mantiene una proporción relativamente baja de datos faltantes en comparación con otros contaminantes, lo que garantiza una mayor robustez en los análisis exploratorios y posteriores fases de modelamiento.

Tabla 11. *Resumen estadístico material Particulado PM 2.5 por estación*

Localidad	Media	Mediana	Desv	Var	RIQ	Skewness	Kurtosis	Min	P25	P50	P75	Max
Barrios Unidos	13.6	11	10.2	104.6	14.4	1.1	1.4	0	5.3	11	19.7	84
Chapinero	16.2	13.5	12	144.2	16	1.2	1.9	0	6.9	13.5	22.9	134
Ciudad Bolívar	18	15.4	12.3	150.8	17.1	1.2	3	0	8	15.4	25.1	139.3
Engativá	14.7	12.8	9.8	96.8	13.2	1.1	1.9	0	7	12.8	20.2	106
Fontibón	19.8	17.3	12.4	154.1	16	1.2	2.6	0	10.4	17.3	26.4	162.4
Kennedy	20.7	19	16.5	271.2	22.1	1	2.3	0	8	19	30.1	188
Puente Aranda	14	11.5	12.2	147.9	16.7	1.2	1.8	0	4.3	11.5	21	101.8
San Cristóbal	12.4	9	11.1	123.8	12	1.6	3.8	0	5	9	17	137
Santa Fe	13.9	11	11.1	123.1	14	1.5	3.6	0	6	11	20	146
Suba	13.2	11	9.7	93.8	12	1.4	2.7	0	6	11	18	118
Tunjuelito	16.7	14.1	12	143.7	15	1.3	3.6	0	8	14.1	23	171
Usaquén	7.9	5	9	81.8	12	1.4	1.8	0	0	5	12	60
Usme	12.6	9.7	10.2	103.6	13.4	1.5	3.3	0	4.6	9.7	18	107.5

En los estadísticos descriptivos por localidad, Kennedy y Fontibón exhiben los promedios de PM 2.5 más altos (20,7 y 19,8 µg/m³, respectivamente), seguidos por Ciudad Bolívar (18,0 µg/m³) y Tunjuelito (16,7 µg/m³), mientras Usaquén muestra el promedio más bajo (7,9 µg/m³).

Las medianas guardan un patrón similar (p. ej., 19,0 µg/m³ en Kennedy frente a 5,0 µg/m³ en Usaquén), y la dispersión es mayor en zonas con mayor carga, destacando la desviación estándar de Kennedy (16,5) y Fontibón (12,4), así como sus amplios RIQ (22,1 y 16,0). La presencia de asimetría positiva en todas las localidades (skewness > 1) y colas en varias de ellas (kurtosis elevada en San Cristóbal, Santa Fe, Tunjuelito y Ciudad Bolívar) indica eventos puntuales de alta concentración —también visibles en los máximos (p. ej., 188 µg/m³ en Kennedy y 162,4 µg/m³ en Fontibón).

Análisis Gráfico

Distribución de PM 2.5 en las 13 localidades de Bogotá presentes en el Conjunto de Datos

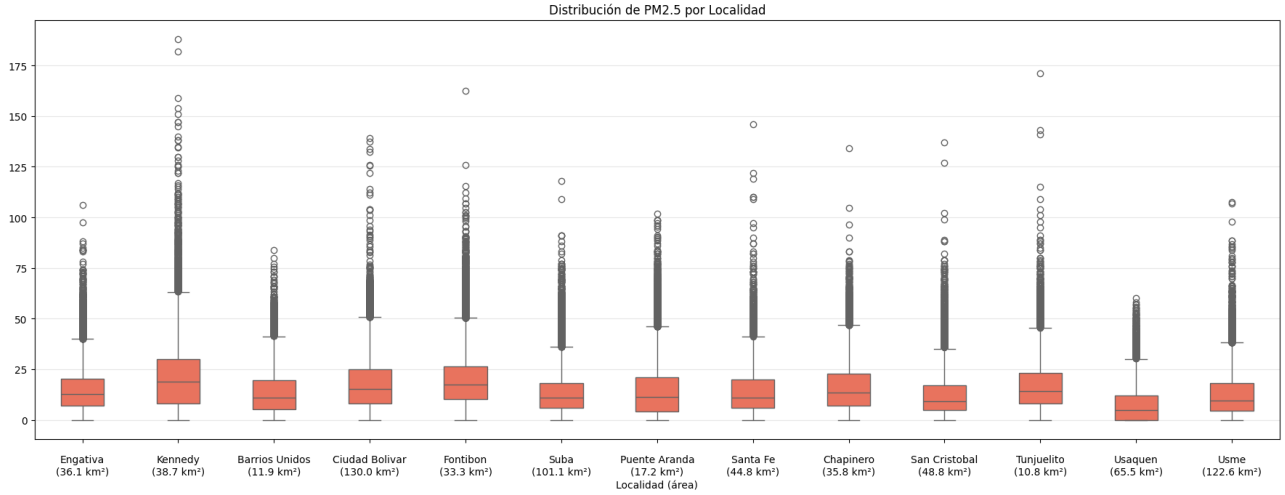


Ilustración 3. *Distribución de PM 2.5 en las 13 localidades de Bogotá presentes en el Conjunto de Datos*

En este caso, se aprecia una heterogeneidad marcada: Kennedy y Fontibón concentran las cajas más altas y anchas, con medianas y rangos intercuartílicos superiores y una cola derecha larga con numerosos valores extremos (episodios que superan ampliamente 100–150 µg/m³). Ciudad Bolivar y Tunjuelito muestran un patrón similar, aunque algo menor. En contraste, Usaquén, Suba y San Cristobal presentan niveles centrales más bajos y menor dispersión.

El boxplot revela una asimetría positiva generalizada y confirma que el comportamiento de PM 2.5 está dominado por eventos puntuales de alta concentración, más que por aumentos uniformes. Esto es consistente con fuentes locales de emisión y condiciones meteorológicas que favorecen la acumulación. Si se toma como referencia la guía de la OMS para PM 2.5 de 15 µg/m³ como promedio diario, varias localidades —en especial Kennedy y Fontibón— exhiben medianas o cuartiles superiores próximos o por encima de ese valor, lo que sugiere frecuencia elevada de días con riesgo sanitario y potencial desigualdad espacial en la exposición (OMS, 2021).

Distribución del parámetro PM 2.5 en Bogotá

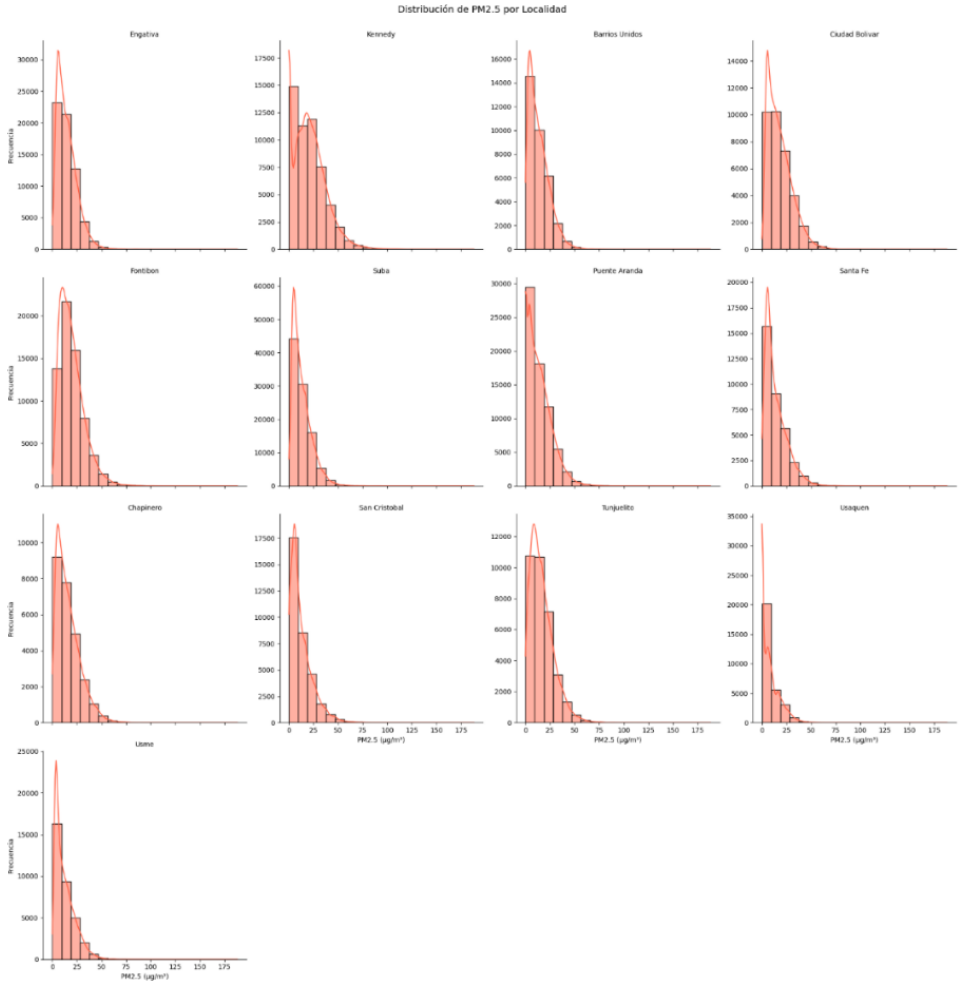


Ilustración 4. *Distribución de PM 2.5 en las 13 localidades de Bogotá presentes en el Conjunto de Datos*

Los histogramas por localidad confirman un patrón fuertemente asimétrico a la derecha en PM 2.5: la mayor parte de las horas se concentra en valores bajos–moderados y una cola de pocos eventos empuja los máximos. Se distinguen perfiles con mayor dispersión en Kennedy y Fontibón, seguidos por Ciudad Bolivar y Tunjuelito. En contraste, Usaquén, Suba y San Cristobal muestran distribuciones más comprimidas y picos en rangos bajos.

PM 2.5 Promedio por Mes en 13 localidades de Bogotá (2021-2024)

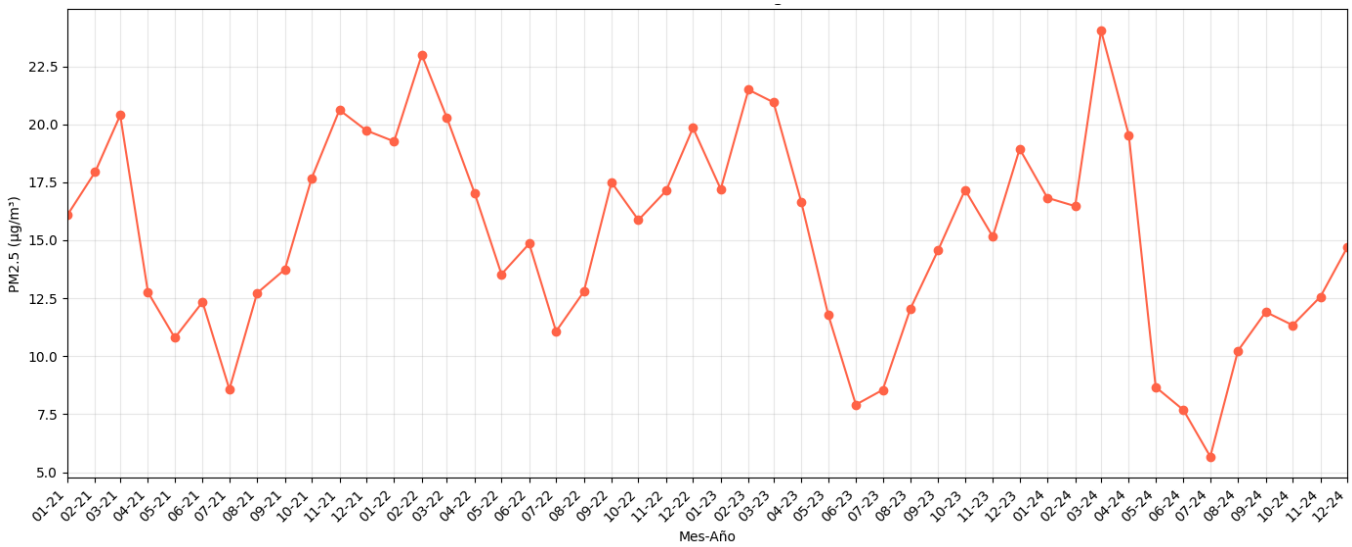


Ilustración 5. *Evolución temporal del promedio mensual de PM2.5 en 13 localidades de Bogotá (2021-2024).*

La serie mensual de PM 2.5 muestra un patrón estacional claro: los valores tienden a subir entre diciembre y marzo (picos cercanos o superiores a ~20–23 µg/m³) y a bajar entre junio y agosto (mínimos de ~7–10 µg/m³), con repuntes intermedios hacia finales de cada año. Destacan picos a inicios de 2022 y 2024 y un descenso pronunciado a mitad de 2024, consistente con periodos más lluviosos y mayor dispersión de contaminantes. Varios meses quedan por encima o muy cerca del referente sanitario de la OMS para PM 2.5 (15 µg/m³ como promedio diario), lo que muestra que existen temporadas de mayor riesgo de eventos climáticos, especialmente al inicio del año.

Distribución espacial de PM2.5 en las 13 localidades de Bogotá presentes en el Dataset

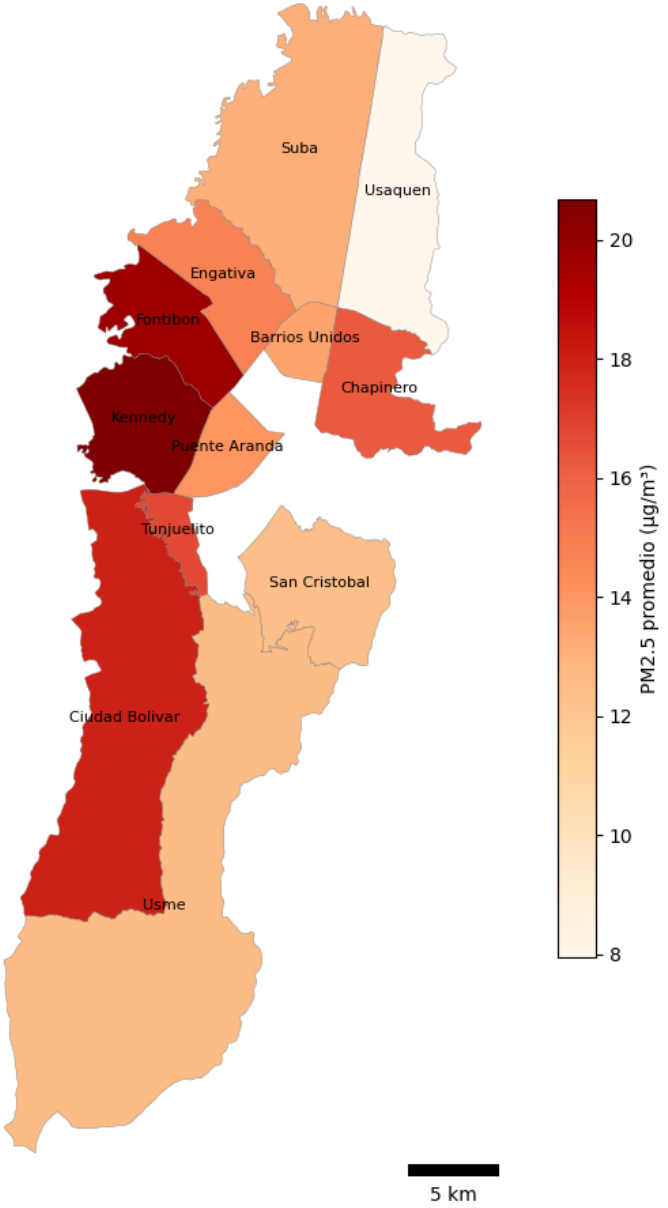


Ilustración 6. *Mapa de la distribución promedio de PM 2.5 por localidad (2021-2024)*

El mapa de promedios de PM 2.5 por localidad muestra una tendencia espacial nítida: las cargas más altas se concentran en el suroccidente-occidente de la ciudad, con máximos en Kennedy y Fontibón y una zona de concentración elevada que incluye Puente Aranda y Ciudad Bolívar. Asimismo, niveles intermedios aparecen en Barrios Unidos, Chapinero y Tunjuelito, mientras que los valores más bajos se observan hacia el nororiente, especialmente en Usaquén, y en menor

medida en Suba. San Cristóbal y Usme presentan promedios moderados a bajos, consistentes con una menor densidad de fuentes puntuales. Este patrón sugiere la acción combinada de fuentes móviles e industriales, corredores logísticos y condiciones de dispersión menos favorables en el occidente.

Transformaciones a la base de datos de material particulado

Teniendo en cuenta que no todas las estaciones tienen una cobertura buena (+75%) de todos los contaminantes, especialmente SO2, OZONO y NO, se decidió no tener en cuenta estos contaminantes, pues la completitud y calidad de los datos se ve muy comprometida dada la alta cantidad de nulos en estos contaminantes. Asimismo, se decidió crear un umbral de 18 horas por día, con el fin de excluir aquellas estaciones que no tuvieran por lo menos el 90% de la totalidad de sus días con un cumplimiento del umbral respecto a tener datos, todo esto, tomando como referencia el contaminante PM2.5, principal en este estudio.

Tabla 12. *Porcentaje completitud de datos por estación según el umbral establecido*

Estación	PM25	PM10	CO
Bolivia	86.52%	89.80%	92.06%
Carvajal - Sevillana	59.00%	57.08%	50.10%
Centro de Alto Rendimiento	96.69%	91.05%	98.14%
Ciudad Bolivar	98.42%	95.96%	98.08%
Colina	90.93%	89.47%	90.51%
Fontibon	95.82%	96.10%	91.23%
Guaymaral	96.39%	95.42%	85.51%
Jazmin	96.78%	96.10%	97.40%
Kennedy	94.07%	91.95%	87.36%
Las Ferias	94.69%	92.83%	88.83%
MinAmbiente	97.74%	89.92%	96.44%
Movil 7ma	73.15%	71.99%	20.41%
Movil Fontibon	95.76%	98.10%	66.93%
Puente Aranda	96.37%	77.93%	88.07%
San Cristobal	96.58%	98.15%	88.63%
Suba	93.72%	95.10%	97.24%

Tunal	96.30%	95.47%	83.33%
Usaquen	85.68%	86.58%	97.88%
Usme	95.48%	96.71%	93.63%

Teniendo en cuenta esto, se optó por eliminar del estudio las estaciones Bolivia, Carvajal – Sevillana, Móvil 7ma y Usaquén.

Para facilitar el preprocesado y los procesos de entrenamiento y validación del modelo CatBoost se decidió hacer una agregación al conjunto de datos, pasando de medidas horarias a medidas diarias, con el fin de aumentar la completitud de los datos y eliminar la volatilidad inherente a tener registros horarios, lo que puede provocar un grado sumamente alto de valores atípicos. Tras esta agregación, se sumaron variables de longitud, latitud y altura por estación, provenientes de la secretaría de medio ambiente, así como la adición de las nuevas variables climatológicas, extraídas por la NASA con una agregación diaria también. Todo esto, con el fin de mejorar el proceso de imputación de los datos, para aquellos valores que preservan nulidad.

Finalmente, se realizó un procedimiento que consistió en imputar los valores faltantes para los contaminantes mediante modelos de regresión multisalida, entrenados sobre las variables ambientales y climatológicas de cada estación. Este enfoque permite capturar la correlación entre contaminantes y las condiciones atmosféricas locales, preservando la coherencia temporal. Asimismo, se realizó una interpolación espacial complementaria tipo Inverse Distance Weighting (IDW) para suavizar los resultados entre estaciones cercanas, considerando la distancia geográfica como factor de ponderación.

Finalmente, se validó la coherencia de las imputaciones asegurando que los valores estimados respetaran los rangos físicos y la continuidad temporal de las series. De esta forma, el conjunto de datos final conserva la integridad estadística.

Tabla 13. *Estructura base de datos predictores PM 2.5*

Columna	Descripción
Estacion	Nombre de la estación de monitoreo de calidad del aire donde se registran los datos.
Date	Fecha exacta de la medición de los contaminantes y variables meteorológicas.
PM25	Concentración promedio diaria de PM2.5, expresada en µg/m³.
lat	Latitud geográfica de la estación de monitoreo, en grados decimales.
lon	Longitud geográfica de la estación de monitoreo, en grados decimales.
Altitud	Altitud de la estación sobre el nivel del mar, expresada en metros (m).
Temp	Temperatura promedio diaria del aire, registrada en grados Celsius (°C).
Hum	Humedad relativa promedio diaria del aire, expresada en porcentaje (%).
WindSpeed	Velocidad promedio diaria del viento, medida en metros por segundo (m/s).
Precip	Precipitación acumulada diaria, registrada en milímetros (mm).
Pres	Presión atmosférica promedio diaria, medida en kilopascales (kPa).
Localidad	Localidad de Bogotá donde se ubica la estación de monitoreo.

17. Preparación de datos, validación temporal y modelado predictivo de PM2.5 en Bogotá

Ingeniería de variables

Variables de calendario

A partir de la fecha t se construyeron $year_t$, $month_t$, día del año doy_t, día de la semana dow_t y un indicador de fin de semana $is_weekend_t$. Además, con el propósito de suavizar la representación de la estacionalidad anual y evitar discontinuidades entre el final e inicio de l año, se aplicó una codificación trigonométrica basada en el día del año. Esta transformación permite modelar los patrones cíclicos —como variaciones estacionales en la concentración de contaminantes— dentro de un espacio continuo, donde el 31 de diciembre y el 1 de enero son valores vecinos en el ciclo anual. Matemáticamente, se definieron las siguientes variables:

$$sin_doy_t = \sin\left(\frac{2\pi * doy_t}{365}\right), \qquad cos_doy_t = \cos\left(\frac{2\pi * doy_t}{365}\right)$$

De esta forma, la estacionalidad se representa en un plano circular continuo, preservando la periodicidad natural del fenómeno.

Lags y medias móviles de la variable objetivo (PM2.5)

Para capturar inercia temporal se generaron retardos (lags) y promedios móviles (solo con información del pasado):

$$PM25_lagL_t = PM25_{t-L}\,, \qquad L \in \{1, 3, 7\},$$

$$PM25_rollmeanW_t = \frac{1}{W}\sum_{k=1}^{W} PM25_{t-k}\,, \qquad W \in \{3, 7\},$$

Estas variables se calculan por localidad y luego se desplazan para evitar la fuga de información.

Lags de variables meteorológicas

De forma análoga, para cada variable climatológica

$$X \in \{Temp, Hum, WindSpeed, Precip, Pres\}:$$

$$X_lagL_t = \mathrm{X}_{t-L}, \qquad \mathrm{L} \in \{Temp, Hum, WindSpeed, Precip, Pres\}$$

Estas ayudan a capturar rezagos físicos entre condiciones climatológicas y concentraciones.

Codificación categórica

La variable categórica Localidad se trató como categórica nominal (one-hot en pipelines clásicos; en CatBoost se pasa el índice de la columna para tratamiento interno). Se decidió modelar por localidad (no por estación) por su mayor robustez espacial.

Partición temporal y validación cruzada

Hold-out

Se usó hold-out por año: entrenamiento en 2021–2023 y prueba en 2024. Esta separación respeta la causalidad temporal.

Validación cruzada temporal (ventana expansiva)

Se definieron $K = 4$ folds con ventana expansiva: si τ es el conjunto de días ordenados y B el tamaño del bloque de validación, en el fold k:

$$Train_k = \{t_1, \dots, t_{kB}\}, \qquad Val_k = \{t_{kB+1}, \dots, t_{(k+1)B}\}$$

Esto evalúa estabilidad a medida que aumenta la información histórica, lo que se resume en la figura 7.

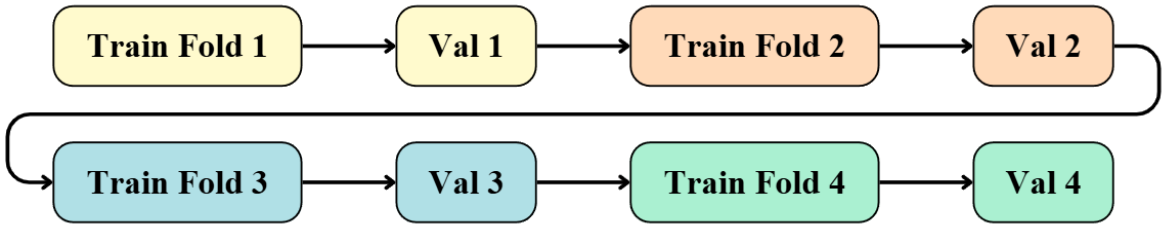


Ilustración 7. Proceso de entrenamiento y validación para la validación cruzada temporal.

Modelo y entrenamiento

Algoritmo, variables de entrada (X) y objetivo (Y)

Tabla 14. *Estructura conjunto de datos de entrada para el modelo CatBoost*

Variable	Descripción	Tipo
Localidad	Zona administrativa de Bogotá	Categórica
Estacion	Punto de monitoreo del contaminante	Categórica
lat	Latitud de la estación	Geoespacial
lon	Longitud de la estación	Geoespacial
Altitud	Altura sobre el nivel del mar (m)	Geoespacial
year	Año de la observación	Temporal
month	Mes de la observación	Estacional
dayofyear	Día del año (1–365)	Estacional
dow	Día de la semana (0=Lunes, 6=Domingo)	Estacional
is_weekend	Indicador binario de fin de semana	Estacional
sin_doy	Codificación trigonométrica seno del día del año	Estacional
cos_doy	Codificación trigonométrica coseno del día del año	Estacional
Temp_lag1	Temperatura del día anterior	Lag Climatológico
Temp_lag3	Temperatura tres días atrás	Lag Climatológico
Hum_lag1	Humedad relativa del día anterior	Lag Climatológico

Hum_lag3	Humedad relativa tres días atrás	Lag Climatológico
WindSpeed_lag1	Velocidad del viento del día anterior	Lag Climatológico
WindSpeed_lag3	Velocidad del viento tres días atrás	Lag Climatológico
Precip_lag1	Precipitación del día anterior	Lag Climatológico
Precip_lag3	Precipitación tres días atrás	Lag Climatológico
Pres_lag1	Presión atmosférica del día anterior	Lag Climatológico
Pres_lag3	Presión atmosférica tres días atrás	Lag Climatológico
PM25_lag1	Concentración de PM2.5 del día anterior	Lag Contaminación
PM25_lag3	Concentración de PM2.5 tres días atrás	Lag Contaminación
PM25_lag7	Concentración de PM2.5 siete días atrás	Lag Contaminación
PM25_rollmean3	Promedio móvil de PM2.5 en los tres días previos	Media Móvil
PM25_rollmean7	Promedio móvil de PM2.5 en los siete días previos	Media Móvil
PM25	Concentración diaria de material particulado fino (µg/m³)	Objetivo

Búsqueda de hiperparámetros y configuración

Tras la ejecución de la búsqueda aleatoria con validación temporal; la mejor configuración promedio en CV fue:

Tabla 15. *Valores de los hiperparámetros utilizados para el modelo CatBoost*

Hiperparámetro	Valor
Depth	5
Learning Rate	0.06
L2 Leaf Reg	5.0
Bagging Temperature	3.0
Random Strength	0.5
Iterations	2500
Bootstrap type	Bayesian
Early Stopping Rounds	100

Esta configuración se empleó para reentrenar sobre 2021–2023. Para el fit final se utilizó early stopping con un pequeño conjunto de validación interno (p. ej., diciembre de 2023).

Métricas de evaluación

Para un conjunto de evaluación $\{(y_i, \hat{y}_i)\}_{i=1}^{n}$:

MAE (Error absoluto medio):

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} | \, y_i - \hat{y}_i \, |$$

La MAE mide el error absoluto promedio entre los valores reales y las predicciones, dando una idea de la magnitud promedio del error sin considerar su dirección.

RMSE (Raíz del error cuadrático medio):

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

El RMSE penaliza más los errores grandes, ya que los eleva al cuadrado. Es útil cuando los errores grandes son especialmente indeseables en el modelo.

R^2 (Coeficiente de determinación):

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Donde $\bar{y}$ es la media de los valores reales y_i. El R^2 indica qué proporción de la varianza total de los datos es explicada por el modelo. Un valor de $R^2 = 1$ indica una predicción perfecta, mientras que $R^2 = 0$ sugiere que el modelo no mejora la predicción respecto a la media.

MDAE (Error absoluto mediano):

$$\text{MDAE} = \text{mediana}(| \, y_i - \hat{y}_i \, |)$$

El MDAE mide el error absoluto mediano, lo que lo hace menos sensible a valores atípicos o errores grandes que puedan afectar a la MAE y RMSE. Esto es útil cuando se quiere tener una medida robusta del rendimiento del modelo.

Resultados

Importancia de variables

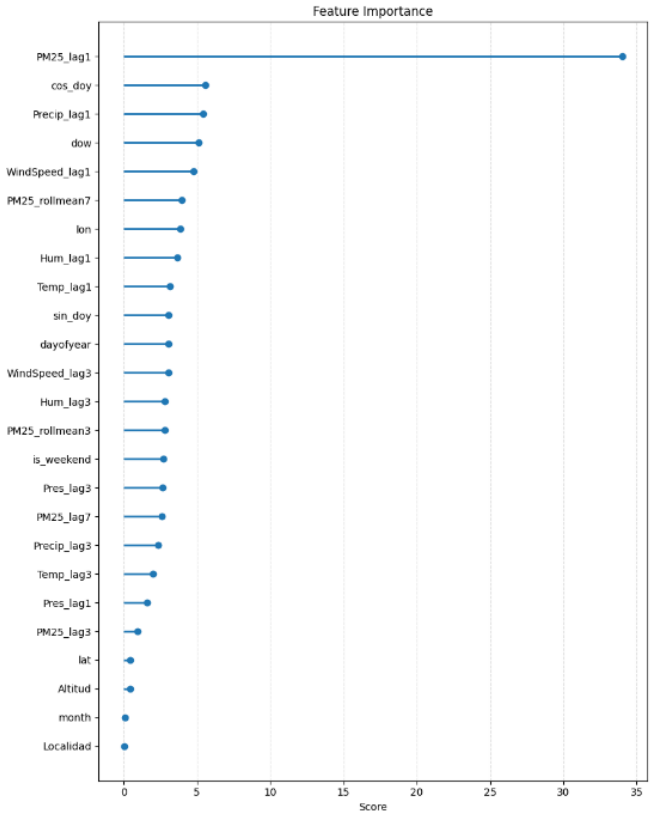


Ilustración 8. *Importancia de variables del modelo CatBoost*

En la gráfica de importancia de variables, se observa que la característica PM25_lag1 tiene una contribución significativamente mayor que el resto de las variables, lo que indica que el valor de la concentración de PM2.5 del día anterior es el principal factor explicativo en la predicción de las concentraciones actuales. Este comportamiento es esperado en modelos de series temporales, donde los valores rezagados de la variable objetivo muestran una fuerte autocorrelación temporal, un fenómeno común en la predicción de contaminantes atmosféricos. Además, otras variables

como cos_doy, Precip_lag1, dow y WindSpeed_lag1 también tienen una relevancia considerable en el modelo. Estas variables están relacionadas con factores meteorológicos y estacionales que influyen directamente en la variabilidad de la concentración de PM2.5, ya sea a través de su efecto en la dispersión de contaminantes o en los patrones climáticos que afectan la calidad del aire.

Por otro lado, variables como Localidad, Altitud y month muestran una importancia mucho menor, lo que sugiere que, en este caso específico, la ubicación geográfica y las fluctuaciones mensuales no aportan tanto a la predicción del PM2.5 en comparación con factores más inmediatos, como los lags del contaminante y las condiciones meteorológicas recientes. Esta distribución de importancia de variables resalta la relevancia de las relaciones temporales y meteorológicas en la predicción de la calidad del aire, en lugar de diferencias geográficas o estacionales a nivel macro. Estos resultados sugieren que el modelo podría beneficiarse de una mayor atención a las variables de autocorrelación temporal y factores meteorológicos inmediatos, lo que podría optimizar su capacidad predictiva y proporcionar pronósticos más precisos en contextos similares.

Métricas

Folds

Tabla 16. *Métricas por fold*

Fold	R^2	RMSE	MAE	MDAE	n	n_pct_used
1	0.521	5.315	4.114	3.321	3737	3737
2	0.476	5.717	4.416	3.483	3743	3743
3	0.63	5.414	4.141	3.281	3800	3742
4	0.724	5.065	3.882	3.1	3621	3490

En términos de R^2, el fold 4 presenta el mejor rendimiento con un valor de 0.724, lo que indica que el modelo explica aproximadamente el 72.4% de la variabilidad en los datos. Sin embargo, el fold 2 muestra un R^2 de 0.476, lo que sugiere que el modelo tiene una capacidad de explicación limitada en este conjunto de validación. Este comportamiento es indicativo de que,

aunque el modelo tiene un rendimiento aceptable en la mayoría de los folds, hay ciertos períodos o localizaciones en los que su capacidad predictiva es reducida.

Por otro lado, el RMSE, que mide la magnitud promedio del error, muestra que el mejor rendimiento se observa en el fold 4, con un RMSE de 5.065. Este valor sugiere que las predicciones en este conjunto son relativamente precisas. En contraste, el fold 2 presenta un RMSE de 5.717, lo que implica que el modelo genera mayores errores en este conjunto, lo que podría estar relacionado con características específicas de los datos de ese fold, como la presencia de outliers o patrones complejos no capturados adecuadamente por el modelo.

En cuanto al MAE, que calcula el error absoluto promedio, el fold 4 nuevamente muestra el mejor desempeño, con un valor de 3.882. Esto indica que, en promedio, el modelo tiene un error absoluto bajo en este conjunto de validación. Sin embargo, el fold 2 registra el MAE más alto con un valor de 4.416, lo que sugiere que los errores en este conjunto son más significativos en comparación con los otros folds.

Finalmente, el análisis del MDAE, que proporciona una medida robusta del error mediano, refuerza los hallazgos previos, con el fold 4 destacando nuevamente como el mejor (MDAE de 3.100). Esto indica que el 50% de los errores se encuentran por debajo de este valor, reflejando un buen desempeño general del modelo. El fold 2, por su parte, muestra un MDAE de 3.483, lo que implica que, a pesar de ser una métrica más robusta frente a los outliers, el modelo aún presenta un desempeño algo inferior en este fold.

Así pues, las métricas globales y desagregadas por fold sugieren que el modelo tiene un rendimiento generalmente bueno, pero con variabilidad entre los diferentes conjuntos de validación. El fold 4 se muestra como el mejor en todas las métricas, mientras que el fold 2 presenta

un desempeño inferior, lo que señala que ciertos factores en los datos de ese fold, como características no observadas o variabilidad en las condiciones, podrían estar afectando la capacidad predictiva del modelo.

Hold-Out

Tabla 17. *Métricas Hold-Out*

Métrica	Valor
R^2	0.684
RMSE	5.602
MAE	4.357
MDAE	3.489
MAPE	1254.186
MPE	1236.247
n_test	4571

El R^2 de 0.684 sugiere que el modelo explica aproximadamente el 68.4% de la variabilidad en los valores de PM2.5 del conjunto de prueba. Esto es indicativo de un buen ajuste, ya que más de dos tercios de la variabilidad del contaminante se pueden predecir utilizando las variables del modelo. Sin embargo, aunque este valor es aceptable, aún hay un margen significativo para mejorar la capacidad explicativa del modelo.

El RMSE de 5.602 indica que las predicciones del modelo, en promedio, están a una distancia de 5.6 unidades del valor real de PM2.5. Este valor es relativamente alto, lo que sugiere que el modelo sigue cometiendo errores considerables, especialmente en predicciones de valores extremos. Sin embargo, al compararlo con los resultados de la validación cruzada, este RMSE es consistente con los valores obtenidos en los folds, lo que implica que el modelo mantiene un rendimiento razonable en datos no vistos.

El MAE de 4.357 muestra que el error promedio absoluto entre las predicciones y los valores reales es de 4.357 unidades de PM2.5. Aunque este valor es relativamente bajo, indica que el modelo aún comete errores en una cantidad considerable de las predicciones. Comparado con el RMSE, el MAE está en una gama similar, lo que sugiere que los errores grandes no son tan predominantes, pero sí existen. Por otro lado, El MDAE de 3.489 refleja el error absoluto mediano, es decir, que el 50% de los errores de las predicciones están por debajo de este valor. Este valor más bajo en comparación con el MAE sugiere que la distribución de los errores está más concentrada cerca de los valores predichos, y los errores extremos son relativamente pocos, lo que es un buen indicio de la robustez del modelo frente a valores atípicos.

18. Metodología para la construcción de predictores meteorológicos (Bogotá) y proyección 2026

Selección de la fuente de datos: ERA5 (C3S/ECMWF)

Se utilizó ERA5, la reanálisis global de quinta generación del ECMWF distribuido por el Copernicus Climate Data Store (C3S). ERA5 ofrece cobertura global, consistencia temporal y disponibilidad de variables cercanas a superficie (T2M, D2M, U10, V10, SP, MSLP, TP), con resolución horaria y un mallado nominal ~31 km (T639), además de documentación y control de calidad extensos. Estas características lo hacen idóneo para estudios urbanos donde se requieren series completas y reproducibles y, a su vez, capas físicas coherentes para forzamiento/explicativas de modelos de calidad del aire.

Variables y preprocesamiento

Se trabajó con las siguientes magnitudes estándar de ERA5:

Tabla 18. *Variables seleccionadas del reanálisis ERA5 para la construcción del conjunto de proyección de climatología*

Variable	Descripción	Unidades
(T2M)	Temperatura promedio del aire a 2 metros	K (Kelvin)
(D2M)	Temperatura de rocío a 2 metros	K (Kelvin)
(U10)	Componente zonal del viento a 10 metros	m/s
(V10)	Componente meridional del viento a 10 metros	m/s
(SP)	Presión atmosférica a nivel del suelo	Pa (Pascales)
(TP)	Precipitación acumulada total durante el período	m (metros)

Interpolación espacial

Las variables reanalizadas sobre la grilla de ERA5 se interpolaron a las coordenadas de las 18 estaciones del SDA/Secretaría de Ambiente (interpolación bilineal sobre lat/lon). Esta etapa produce, por cada mes, una estimación representativa del entorno de cada sitio.

Conversión de unidades

Se aplicaron transformaciones estándar.

$$Temperatura\ Celsius = T_K - 273.15$$

$$Presión\ a\ kPa = \frac{P_Pa}{1000}$$

$$Precipitación\ a\ mm = TP_m * 1000$$

Velocidad del viento

A partir de componentes:

$$WS_{10} = \sqrt{U10^2 + V10^2}$$

En caso de requerir viento a 2m como aproximación neutral puede aplicarse el perfil logarítmico:

$$WS_2 = WS_z \frac{\ln\left(\frac{2}{z_0}\right)}{\ln\left(\frac{z}{z_0}\right)}, \qquad (aquí\ z = 10)$$

Humedad relativa

Se derivó con una formulación tipo Magnus (Alduchov & Eskridge,1996):

$$RH(\%) = 100\frac{e_s(T_d)}{e_s(T)}, \qquad e_s(T) \approx 6.112\exp\left(\frac{17.67T}{T + 243.5}\right)$$

Con T y T_d en °C.

Proyección para 2026

Para 2026 se toma la climatología 1991–2020 (promedios mensuales) como proyección "esperable" de condiciones típicas (sin perturbaciones extremas exógenas conocidas). El procedimiento fue:

1. Descarga de ERA5 mensual 1991–2020 para el rectángulo de Colombia (evita bordes sobre Bogotá).
2. Cálculo de climatología mensual: promedio sobre 1991–2020 y agrupación por mes ($m = 1, \dots, 12$).
3. Interpolación puntual a cada estación (bilineal).
4. Derivación de magnitudes:
5. Desagregación temporal a diario: para cada mes m de 2026, se replican valores mensuales a cada día d del mes. Esto produce una serie diaria 2026 consistente con la climatología mensual local de ERA5.

Este enfoque es reproducible, transparente y estable, especialmente cuando no es viable (o metodológicamente deseable) usar pronósticos estacionales operativos que pudieran no estar disponibles para todos los meses de 2026 o introducir cambios de sistema/modelo.

19. Generación de predicciones de concentraciones de PM2.5 para 2026

Tras entrenar y validar el modelo CatBoost, se generan las predicciones, teniendo como targets climatológicos la base de datos extraída de Copernicus, como se mencionó previamente. Estos datos permiten predecir cómo las condiciones meteorológicas influirán en la concentración de PM2.5 en los días futuros.

Generación Recursiva de Predicciones

El proceso de predicción recursiva es esencial para la proyección a futuro. Comienza con el buffer histórico de PM2.5 de los últimos 7 días de 2024 para cada estación de monitoreo, desde el cual se generan los primeros lags de PM2.5 para el inicio de 2026. A partir de ahí, el modelo realiza la predicción para el primer día de 2026, luego utiliza esa predicción como un nuevo valor para alimentar los lags en la predicción del siguiente día.

Este proceso se repite de forma iterativa hasta generar todas las predicciones diarias para el año 2026. Cada día predicho depende de las predicciones previas y de las condiciones meteorológicas, lo que genera un ciclo de retroalimentación en el que las predicciones de PM2.5 se actualizan recursivamente.

Smoothing y Relleno de Datos Faltantes

Es común que en los primeros días de la proyección (especialmente a principios de 2025) los buffers de PM2.5 no estén completos, ya que el modelo no tiene datos previos de PM2.5 para esos días. En este caso, se utilizan estrategias de relleno suave (por ejemplo, mediante

interpolación) para completar los datos faltantes y permitir una transición suave entre los valores predichos y las condiciones climáticas.

Agregación por Localidad

Finalmente, dado que se genera una predicción diaria por estación, es posible que algunas estaciones tengan pocos datos disponibles o tengan un comportamiento errático en algunos periodos. Para suavizar estas variaciones, las predicciones se agregan por localidad utilizando la mediana de las estaciones dentro de cada localidad. Esto da como resultado una serie temporal de PM2.5 a nivel de localidad, que refleja las condiciones promedio de la zona.

Resultados y Almacenamiento

Una vez que las predicciones han sido generadas para todo el periodo de 2025–2026, se almacenan en archivos CSV separados por estación y por localidad. Estos resultados pueden ser utilizados para alimentar modelos de impacto en la salud pública, como BenMAP-CE, o para realizar un análisis de las tendencias futuras en la calidad del aire de Bogotá.

Impacto de la agregación temporal en las proyecciones 2025-2026

La agregación de PM2.5 horaria a promedios diarios en la base de entrenamiento y validación del modelo de Ml (CatBoost) supone que en los impactos en salud considerados (mortalidad diaria por localidades) es el nivel medio de exposición en 24 horas y no la distribución intradiaria de picos. Este supuesto es coherente con la literatura de estudios de series de tiempo en epidemiología ambiental, que trabajan típicamente con promedios diarios de concentración y conteos diarios de desenlaces de salud, para los cuales se ha demostrado asociación robusta aún bajo errores de medición moderados por agregación (Dominici, 2004).

No obstante, al aplicar un filtro de media diaria se elimina la información sobre los picos horarios (por ejemplo, durante la hora pico de la mañana) que el modelo conceptual reconoce como

momentos de alta acumulación. En ese sentido, el modelo estadístico captura adecuadamente la dinámica de episodios de varios días (acumulaciones que se extienden durante periodos secos), pero es menos sensible a picos agudos muy breves que podrían ser relevantes para desenlaces agudos en la carga sanitaria reportada en los RIPS en servicios como urgencias u hospitalizaciones no modelados explícitamente en este trabajo.

Así pues, el uso de un esquema recursivo de predicciones diarias basado en lags de 7 días de PM2.5 implica asumir que la memoria del sistema respecto a la acumulación y disipación del material particulado se concentra en una ventana móvil de una semana, y que el resto de la dinámica de transporte y química atmosférica está suficientemente codificada en los predictores meteorológicos y de calendario.

Esto es consistente con la evidencia conceptual de que los episodios prolongados de mala calidad del aire en Bogotá suelen extenderse entre varios días y pocas semanas, modulados por secuencias de estabilidad/ventilación, pero introduce la posibilidad de acumulación de error en los horizontes de predicción más alejados, especialmente si se presentan patrones meteorológicos atípicos no representados en el periodo de entrenamiento.

20. Proyección de Mortalidad por causas respiratorias o cardiopulmonares

Dado que la incidencia base proyectada para el año 2026 no está directamente disponible, se diseñó un procedimiento para modelar y proyectar las tasas de mortalidad por todas las causas respiratorias o cardiopulmonares sexo y localidad, basándose en datos históricos (2021–2024). Este enfoque aprovecha modelos de regresión de conteo con offset logarítmico de población, lo cual es metodológicamente consistente con la literatura epidemiológica (Cameron & Trivedi, 1998).

Datos de Mortalidad Histórica

Se utilizó un conjunto de datos con registros anuales de mortalidad en Bogotá entre 2021 y 2024, clasificados por: Localidad (asignada a coordenadas de celda Row–Column), Sexo, Rango etario y Número de defunciones. Los datos fueron normalizados semánticamente y se agregaron para obtener los conteos de defunciones por año, celda, sexo y grupo de edad.

Datos de Población

Se utilizaron dos conjuntos de población:

1. Histórica (2021–2024): para estimar tasas ajustadas.
2. Proyectada para 2026: para obtener predicciones futuras.

Modelado de conteo con Offset Poblacional

Dado que las defunciones son variables de conteo y dependen del tamaño poblacional expuesto, se empleó un modelo de regresión de conteo con offset logarítmico. En particular, se aplicó un modelo de Poisson con estructura:

$$\log(u_{ijk}) = \beta_0 + \beta_1 * Año_i + \gamma_j + \delta_k + \phi_{rc} + \log\left(Población_{ijk}\right)$$

Donde:

- μ_{ijk}es la esperanza del número de muertes para el grupo i, j, k.
- $\beta_1$captura la tendencia temporal (año).
- γ_j, δ_k, ϕ_{rc}son efectos fijos para sexo j, grupo de edad ky celda geográfica rc.
- El término log (Población)es el offset logarítmico, lo cual convierte el modelo a una tasa.

Este enfoque es ventajoso porque permite controlar la exposición poblacional y modelar adecuadamente las tasas de mortalidad, lo cual es crucial en entornos urbanos con fuerte heterogeneidad demográfica como Bogotá.

Proyección a 2026

Se filtraron las combinaciones (celda, sexo, edad) que estaban presentes tanto en los datos históricos como en la proyección de población 2026. Luego, se predijo el número de muertes esperadas para 2026:

$$\hat{y}_{ijk}^{2026} = \exp\left(X_{ijk}^{2026} \cdot \hat{\beta} + \log\left(\text{Población}_{ijk}^{2026}\right)\right)$$

Donde X_{ijk} es el vector de variables categóricas del modelo.

La tasa de incidencia estimada (usada por BenMAP) se calculó como:

$$\text{Incidencia}_{ijk}^{2026} = \frac{\hat{y}_{ijk}^{2026}}{\text{Población}_{ijk}^{2026}}$$

Estas tasas son específicas por Localidad (row–column), sexo y edad, y se exportaron en el formato requerido por BenMAP-CE como tipo Incidence.

21. Generación de escenarios en BenMAP-CE

Una vez consolidadas todas las bases de datos, al ser cargados los escenarios, se obtuvieron los siguientes resultados:

Función de Impacto en Salud

Para este modelo se seleccionó el estudio de Hoek et al. (2013), una revisión sistemática con metaanálisis sobre exposición crónica a PM2.5 y mortalidad general, ampliamente aceptado por la comunidad científica y por organismos internacionales (OMS, EPA). β (coeficiente log-

lineal): derivado del riesgo relativo (RR) reportado en el estudio. Para una estimación de RR = 1.06 por cada 10 µg/m³, este valor fue el utilizado como parámetro base para la función de impacto.

Este valor de riesgo fue escogido por cumplir con los siguientes criterios:

1. Robustez metodológica: metaanálisis de múltiples cohortes.
2. Aplicabilidad internacional: aunque los estudios incluidos son de Europa y Norteamérica, se considera aplicable a contextos urbanos latinoamericanos por similitudes en la exposición.
3. Adopción institucional: la EPA y la OMS recomiendan o aceptan este coeficiente en análisis costo-beneficio (U.S. EPA, 2021; OMS, 2016).

Justificación del Modelo y Aplicación

El modelo log-lineal es adecuado para exposiciones ambientales crónicas porque asume una relación continua, sin umbral, entre la concentración del contaminante y el riesgo (Pope et al., 2002). Este enfoque permite capturar incrementos marginales en riesgo que se acumulan con el tiempo.

Adaptación Demográfica

El modelo se aplicó únicamente a personas de 30 años o más, conforme a los criterios del estudio base. Este límite inferior de edad evita sesgos relacionados con mortalidad infantil o juvenil no atribuible al PM2.5.

Limitaciones y Consideraciones

Aunque se emplea un coeficiente global, se reconoce que las condiciones de salud de fondo pueden variar entre Bogotá y las regiones estudiadas originalmente. También, la exposición puede ser subestimada si no se consideran contaminantes co-ocurrentes (e.g., NO2, O3). No obstante, el

uso de Hoek et al. (2013) sigue siendo defendible, dada la ausencia de estudios longitudinales equivalentes en Colombia, y la necesidad de mantener coherencia con estudios previos.

Mapa y detalle de incidencia.

Tras añadir todos los datos a BenMAP y correr el modelo correspondiente con la función C-R ya descrita, se exportaron los resultados a un archivo Shapefile que fue utilizado en la herramienta QGis para generar un nuevo mapa que mostrara, adicional a la mortalidad proyectada, datos como la representación gráfica de la malla del conjunto de datos de climatología extraído para predicción de material particulado y la ubicación de las estaciones utilizadas con sus respectivos nombres, como se evidencia en la figura 10.

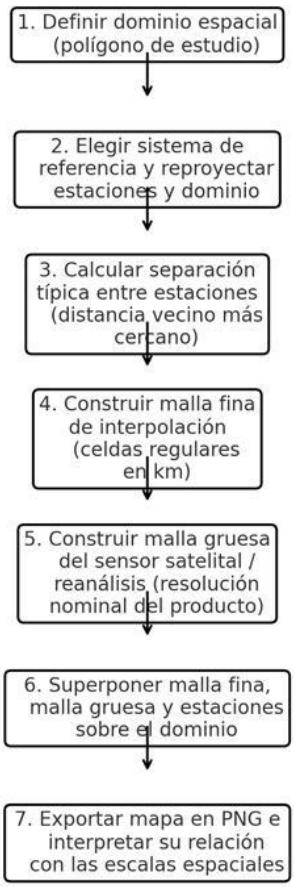


Ilustración 9. Flujo de trabajo para la generación de mapa con puntos de estaciones y malla de ERA5

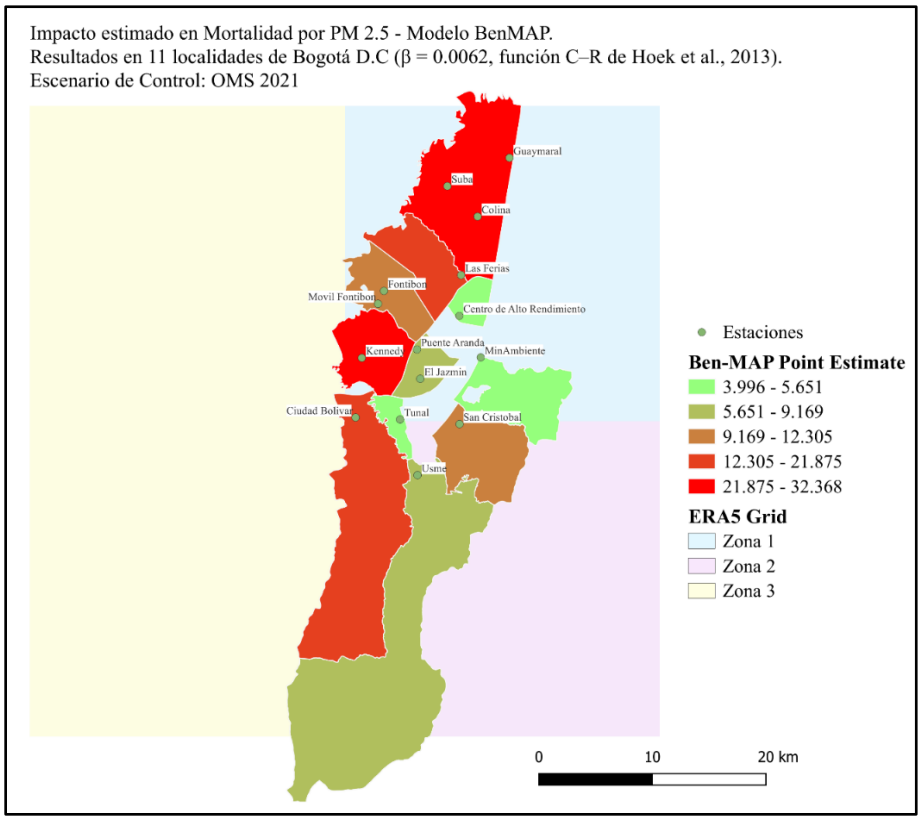


Ilustración 10. *Impacto estimado en mortalidad para todas las causas en población de 30-99 años por PM 2.5 a 2026 basado en predicciones de contaminante a través de un modelo CatBoost y modelos de conteo para tasas de Mortalidad asociado a través de una función C-R.*

Tabla 19. *Detalle numérico para la Ilustración 10*

Rango de Mortalidad Proyectada (muertes/100,000 habitantes)	Descripción
<= 5.651	Zonas con menor mortalidad proyectada por todas las causas, indicando bajo impacto por PM2.5.
5.651 - 9.169	Zonas con impacto moderado-bajo, donde la mortalidad es algo mayor, pero sigue siendo baja.
9.169 - 12.305	Áreas con mortalidad algo mayor, con una exposición moderada a la contaminación por PM2.5.
12.305 - 21.875	Incremento notable en la mortalidad, relacionado con mayores niveles de exposición y población.
21.875 - 32.368	Zonas con los valores más altos de mortalidad proyectada, asociadas a niveles críticos de contaminación.

En el mapa se puede observar cómo las localidades en el oriente de Bogotá, como lo son Usme, Chapinero y Santa Fe no presentan un escenario tan preocupante en contraste al observado en localidades del occidente como Kennedy o Suba, que son localidades que, o bien tienen una gran cantidad de fábricas o vías con un alto nivel de congestión vehicular (como Kennedy), o bien tienen una densidad poblacional alta (como es el caso de Suba).

Los resultados obtenidos concuerdan con las localidades con unos niveles más altos de contaminación, como se observaba en la figura 6, lo que muestra congruencia. No obstante, es fundamental comprender que el mapa no refleja la verdad absoluta, en tanto las estimaciones tienen ciertas limitaciones como el amplio tamaño de la malla del reanálisis ERA5 o la cobertura de las estaciones de monitoreo de calidad del aire respecto a las localidades, pues si bien es cierto que casos como el de Suba presentan 3 estaciones de monitoreo, localidades como Ciudad Bolívar o Usme solamente cuentan con 1 punto de monitoreo de calidad del aire cada una, lo que es una evidente limitación, en tanto la cobertura espacial no es la mejor.

Tabla 20. *Resultados Ben-MAP*

Localidad	Point Estimate	Population	Delta	Baseline	% of Baseline	Percentile 2,5	Percentile 97,5
Barrios Unidos	5494	105450	16050	57998	9,45%	4209	6702
Ciudad Bolívar	18357	359178	17681	176801	10,36%	14080	22371
Engativá	21875	536683	14189	259758	8,41%	16735	26723
Fontibón	12305	262732	17359	120595	10,18%	9436	14999
Kennedy	32368	647352	17901	308120	10,48%	24831	39441
Puente Aranda	9169	168129	15435	100465	9,11%	7021	11191
San Cristóbal	10337	230959	14318	121692	8,48%	7909	12627
Santa Fé	3996	59257	16050	42191	9,45%	3062	4876
Suba	27527	850421	13872	334007	8,23%	21054	33636
Tunjuelito	5651	106230	16303	58779	9,60%	4330	6893
Usme	8600	223955	14948	97162	8,83%	6583	10501

Conclusiones

El presente estudio integró modelos de aprendizaje automático, análisis multiescala del material particulado fino (PM2.5) y herramientas epidemiológicas basadas en funciones concentración-respuesta para estimar los impactos en salud atribuibles a la calidad del aire en Bogotá. Los resultados permiten concluir que la articulación entre dinámica atmosférica, exposición diferenciada y carga sanitaria proyectada constituye un enfoque sólido para la comprensión del riesgo ambiental en contextos urbanos complejos.

En primer lugar, el análisis conceptual de transporte, transformación y remoción del PM2.5, junto con la caracterización de sus escalas espaciales y temporales, evidenció que Bogotá presenta heterogeneidad marcada en la distribución del contaminante, con persistencia de concentraciones elevadas en el suroccidente de la ciudad. Esto confirma que la exposición poblacional no es homogénea y que los riesgos sanitarios deben interpretarse bajo una lógica espacialmente diferenciada, coherente con la estructura emisiva, la topografía de sabana y las condiciones meteorológicas que regulan la dispersión.

En segundo lugar, el estudio mostró que los supuestos metodológicos utilizados, incluyendo interpolación espacial, agregación temporal, construcción de predictores climáticos a partir de ERA5 y validación cruzada estrictamente temporal, introducen incertidumbres inevitables pero controladas. La evaluación de estos supuestos, realizada en coherencia con el modelo conceptual y las escalas características, indica que las transformaciones aplicadas no alteran de manera sustancial la estructura de variabilidad del PM2.5 y son consistentes con prácticas aceptadas en modelación atmosférica y epidemiológica. No obstante, se reconoce que la resolución espacial de los datos y la dependencia de funciones concentración-respuesta internacionales implican

márgenes de incertidumbre que deben ser considerados al interpretar los valores absolutos de carga sanitaria.

En tercer lugar, la integración del modelo predictivo basado en CatBoost con el sistema BenMAP-CE permitió estimar la mortalidad atribuible al PM2.5 para 2026 bajo escenarios específicos. Los resultados sugieren que, aun con reducciones moderadas en los niveles proyectados del contaminante, persisten impactos considerables en salud, especialmente en localidades con mayores niveles de exposición y alta vulnerabilidad demográfica. La consistencia interna entre la superficie de exposición modelada, la estructura poblacional y las funciones concentración-respuesta aporta validez al procedimiento y demuestra la utilidad del enfoque combinado para la evaluación de riesgo sanitario.

Finalmente, se concluye que el estudio proporciona una aproximación metodológicamente robusta, transparente y alineada con estándares internacionales para la estimación de impactos sanitarios por contaminación del aire. Si bien existen limitaciones inherentes al uso de modelos estadísticos y epidemiológicos, particularmente en cuanto a la transferencia de funciones C-R y la dependencia de insumos secundarios, los resultados obtenidos ofrecen información relevante para la formulación de políticas públicas y estrategias locales de control de emisiones en Bogotá. El marco conceptual y analítico desarrollado constituye un punto de partida sólido para futuros estudios de mayor resolución y para el fortalecimiento de capacidades institucionales en evaluación de riesgos ambientales.

Recomendaciones

Desde una perspectiva científica y metodológica, es recomendable que futuros estudios avancen hacia una mayor resolución espacial en la caracterización del PM2.5, incorporando

tecnologías emergentes como sensores de bajo costo o productos satelitales de alta granularidad que permitan reducir la incertidumbre asociada a los esquemas de interpolación utilizados. Asimismo, resulta pertinente promover el desarrollo de funciones concentración–respuesta locales que reflejen de manera más fidedigna las condiciones epidemiológicas de Bogotá, disminuyendo la dependencia de coeficientes internacionales. Complementariamente, sería valioso implementar análisis de sensibilidad ampliados que exploren variaciones en los supuestos de interpolación, agregación temporal y condiciones meteorológicas, con el fin de cuantificar con mayor rigor la incertidumbre estructural que subyace a las estimaciones de exposición y riesgo.

En cuanto a implicaciones para la política pública, los resultados obtenidos sugieren la necesidad de focalizar las intervenciones de control de emisiones en las zonas de la ciudad donde convergen altas concentraciones de PM2.5 y vulnerabilidad poblacional, particularmente en el suroccidente de Bogotá. La evidencia epidemiológica integrada en este estudio respalda la urgencia de fortalecer estrategias dirigidas a la reducción de emisiones móviles, especialmente las asociadas al transporte pesado diésel, principal contribuyente a los niveles urbanos de material particulado fino.

Igualmente, se recomienda ampliar la cobertura y densidad de la red de monitoreo de calidad del aire, con el propósito de mejorar la representatividad espacial de las mediciones y, por ende, la precisión de las estimaciones de exposición. Incorporar herramientas de evaluación sanitaria como BenMAP-CE en los procesos de formulación y evaluación de políticas públicas permitiría, además, que las decisiones en materia de calidad del aire integren explícitamente estimaciones de carga de enfermedad, beneficios evitables y análisis costo–beneficio.

Se debe considerar y reconocer que fenómenos urbanos relevantes para la dispersión del PM2.5 (vientos locales, turbulencia y topografía urbana) pueden variar a escalas menores que el

tamaño de celda del reanálisis, lo cual introduce una limitación estructural en la representación espacial del campo meteorológico. Es por esto que, en los supuestos adoptados, se pierde detalle de microescala y de variaciones subdiarias, pero se gana en estabilidad, completitud y coherencia con el marco. Esta elección es consistente con el objetivo central del trabajo, aunque debe ser tenida en cuenta al interpretar los resultados: las estimaciones de casos atribuibles o evitados representan el impacto de cambios en la exposición promedio diaria a nivel de localidad, y no el efecto de exposiciones agudas en puntos específicos de la ciudad.

Finalmente, desde el ámbito académico y de investigación, se considera relevante profundizar en la modelación de escenarios que integren los posibles efectos del cambio climático sobre la dinámica atmosférica y la formación de PM2.5, dado que modificaciones en la temperatura, la humedad o la altura de la capa límite podrían alterar significativamente los patrones de exposición futura. También sería pertinente explorar modelos multivariados que incorporen otros contaminantes atmosféricos de importancia sanitaria —como NO2 u O3— para analizar posibles efectos combinados o sinérgicos. Asimismo, se recomienda ampliar la investigación hacia la dimensión socioespacial del riesgo, integrando factores como estructura demográfica, condiciones socioeconómicas y desigualdades territoriales, con el fin de construir una visión más integral y equitativa del impacto de la contaminación del aire en la salud pública.

Referencias

Agencia de Protección Ambiental de los Estados Unidos. (2024). *How BenMAP-CE estimates the health and economic effects of air pollution*. U.S. Environmental Protection Agency. Recuperado de https://www.epa.gov/benmap/how-benmap-ce-estimates-health-and-economic-effects-air-pollution

Alcaldía de Bogotá. (2025). *Incendios en otras regiones aumentan material particulado en el aire de Bogotá*. Recuperado de https://bogota.gov.co/mi-ciudad/ambiente/calidad-del-aire-en-bogota-incendios-en-la-region-aumentan-particulas

Alduchov, O. A., & Eskridge, R. E. (1996). *Improved Magnus form approximation of saturation vapor pressure*. *Journal of Applied Meteorology*, 35(4), 601–609. https://doi.org/10.1175/1520-0450(1996)035<0601:IMFAOS>2.0.CO;2

AQICN. (2025). *Mapa en tiempo real de la calidad del aire en Bogotá* [Imagen]. AQI Bogotá. Recuperado de https://aqicn.org/map/bogota/es/

A. Colin Cameron, & Pravin K. Trivedi. (1998). *Regression Analysis of Count Data* (Vol. 00030). Cambridge University Press.

U.S. Global Change Research Program. (2023). *Fifth National Climate Assessment: Air Quality and Climate Change*. U.S. Global Change Research Program. Recuperado de https://nca2023.globalchange.gov/chapter/14/#:~:text=Climate%20change%20can%20worsen%20air,to%20reach%20air%20quality%20goals

Organización Mundial de la Salud (OMS). (s.f.). *Calidad del aire ambiente (exterior) y salud*. Recuperado de https://www.who.int/es/news-room/fact-sheets/detail/ambient-%28outdoor%29-air-quality-and-health

Aldegunde, J. A. Á., Bolaños, E. Q., Fernández-Sánchez, A., Saba, M., & Caraballo, L. (2023). Environmental and Health Benefits Assessment of Reducing PM2.5 Concentrations in Urban Areas in Developing Countries: Case Study Cartagena de Indias. *Environments*, *10*(3), 42. https://doi.org/10.3390/environments10030042

Berman, J. D., Fann, N., Hollingsworth, J. W., Pinkerton, K. E., Rom, W. N., Szema, A. M., Breysse, P. N., White, R. H., & Curriero, F. C. (2012). Health Benefits from Large-Scale Ozone Reduction in the United States. *Environmental Health Perspectives*, *120*(10), 1404-1410. https://doi.org/10.1289/ehp.1104851

Chung, C. Y., Yang, J., Yang, X., & He, J. (2022). Mathematical modeling in the health risk assessment of air pollution-related disease burden in China: A review. *Frontiers In Public Health*, *10*. https://doi.org/10.3389/fpubh.2022.1060153

Coffman, E., Rappold, A. G., Nethery, R. C., Anderton, J., Amend, M., Jackson, M. A., Roman, H., Fann, N., Baker, K. R., & Sacks, J. D. (2024). Quantifying Multipollutant Health Impacts Using the Environmental Benefits Mapping and Analysis Program–Community Edition

(BenMAP-CE): A Case Study in Atlanta, Georgia. *Environmental Health Perspectives*, *132*(3). https://doi.org/10.1289/ehp12969

Greenpeace Colombia. (2021, 18 de febrero). *La contaminación atmosférica causó 160,000 muertes en las 5 ciudades más grandes del mundo en 2020*. https://www.greenpeace.org/colombia/noticia/uncategorized/la-contaminacion-atmosferica-causo-160000-muertes-en-las-5-ciudades-mas-grandes-del-mundo-en-2020/

Hoek, G., Krishnan, R. M., Beelen, R., Peters, A., Ostro, B., Brunekreef, B., & Kaufman, J. D. (2013). Long-term air pollution exposure and cardio- respiratory mortality: a review. *Environmental Health*, *12*(1). https://doi.org/10.1186/1476-069x-12-43

Hurtubia, J. (2019). Breve examen al cambio climático, contaminación del aire y salud en Chile. *Cuadernos Médico Sociales*, *59*(1), 7-16.

Dorogush, A.V., Ershov, V., & Gulin, A. (2018). CatBoost: gradient boosting with categorical features support. *ArXiv,* https://doi.org/10.48550/arXiv.1810.11363.

Leão, M. L. P., Zhang, L., & Da Silva Júnior, F. M. R. (2022). Effect of particulate matter (PM2.5 and PM10) on health indicators: climate change scenarios in a Brazilian metropolis. *Environmental Geochemistry And Health*, *45*(5), 2229-2240. https://doi.org/10.1007/s10653-022-01331-8

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: unbiased boosting with categorical features. *Advances in neural information processing systems*, *31*.

Nawaz, M. O., Henze, D. K., Huneeus, N. J., Osses, M., Álamos, N., Opazo, M. A., & Gallardo, L. (2023). Sources of Air Pollution Health Impacts and Co-Benefits of Carbon Neutrality in Santiago, Chile. *Journal Of Geophysical Research Atmospheres*, *128*(19). https://doi.org/10.1029/2023jd038808

Naciones Unidas. (s.f.). ¿Qué es el cambio climático? Naciones Unidas. Recuperado de https://www.un.org/es/climatechange/what-is-climate-change

Ortiz Durán, E. Y., & Rojas Roa, N. Y. (2013). Estimación de los beneficios económicos en salud asociados a la reducción de PM10 en Bogotá. *Revista de Salud Pública*, *15*(1), 90-102.

Organización Mundial de la Salud. (s.f.). Contaminación del aire. Organización Mundial de la Salud. Recuperado de x

Pavón, D. J. (2024). Cambio climático, contaminación del aire y su impacto en la salud: El rol de la actividad física. *Revista Andaluza de Medicina del Deporte*, *17*(1-2), 137-142. https://doi.org/10.33155/ramd.v17i1-2.1173

Ramírez Mendoza, A. A., Mendoza Magaña, M. L., Ramírez Herrera, M. A., Hernandez Nazará, Z. H., Domínguez Rosales, J. A., & Romero Prado, M. M. J. (2023). Efectos de la contaminación del aire en la salud humana. *Congreso Internacional de Investigación Academia Journals*, *15*(10), 1-4.

Romero-Lankao, P., Qin, H., & Borbor-Cordova, M. (2013). Exploration of health risks related to air pollution and temperature in three Latin American cities. *Social Science & Medicine*, *83*, 110-118. https://doi.org/10.1016/j.socscimed.2013.01.009

Pope, C. A., Burnett, R. T., Thun, M. J., Calle, E. E., Krewski, D., Ito, K., & Thurston, G. D. (2002). Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. JAMA, 287(9), 1132–1141. https://doi.org/10.1001/jama.287.9.1132

Sánchez, J. S. P., Carrascal, A. I. O., & Fernández, F. o. A. (2020). Analítica de datos: incidencia de la contaminación ambiental en la salud pública en Medellín (Colombia). *Revista de Salud Pública*, *22*(6), 1-9. https://doi.org/10.15446/rsap.v22n6.78985

Secretaría Distrital de Planeación. (2020). *Tomo 3: Población POT 2020 (Versión 2).* Alcaldía Mayor de Bogotá. Recuperado de https://www.sdp.gov.co/sites/default/files/tomo_3._poblacion_pot_2020_version_2.pdf

U.S. Environmental Protection Agency (EPA). (2021). BenMAP-CE User Manual. Office of Air Quality Planning and Standards. https://www.epa.gov/benmap

Secretaría de Ambiente de Bogotá. (2021). Inversión térmica: ¿qué es y cuál es su relación con la calidad del aire? Secretaría de Ambiente de Bogotá. Recuperado de https://www.ambientebogota.gov.co/historial-de-noticias/-/asset_publisher/VqEYxdh9mhVF/content/inversion-termica-que-es-y-cual-es-su-relacion-con-la-calidad-del-aire

Torres Vásquez, J. D., Garrido Guerra, L. M., & Garrido Guerra, E. (2023). Calidad del aire. Principales riesgos de contaminación para la salud. *Opuntia Brava*, *15*, 17-32.

Yang, X., Li, Y., Liu, L., & Zang, Z. (2025). Prediction of respiratory diseases based on random forest model. *Frontiers In Public Health*, *13*. https://doi.org/10.3389/fpubh.2025.1537238

Wang, S., Ren, Y., & Xia, B. (2023). PM2.5 and O3 concentration estimation based on interpretable machine learning. *Atmospheric Pollution Research*, *14*(9), 101866. https://doi.org/10.1016/j.apr.2023.101866

World Health Organization (WHO). (2016). Health risks of air pollution in Europe – HRAPIE project: Recommendations for concentration-response functions for cost-benefit analysis of particulate matter, ozone and nitrogen dioxide. https://www.euro.who.int/__data/assets/pdf_file/0006/238956/Health-risks-of-air-pollution-in-Europe-HRAPIE-project.pdf

World Health Organization. (2021). WHO global air quality guidelines: Particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide. Geneva: World Health Organization. https://aqmx.org/sites/default/files/resources/WHO2021_AQG.pdf

Díaz, J. J., Mura, I., Franco, J. F., & Akhavan-Tabatabaei, R. (2021). aiRe – a web-based R application for simple, accessible and repeatable analysis of urban air quality data. Environmental Modelling & Software, 138, 104976. https://doi.org/10.1016/j.envsoft.2021.104976

Mura, I., Franco, J. F., Bernal, L., Melo, N., Díaz, J. J., & Akhavan-Tabatabaei, R. (2020). A Decade of Air Quality in Bogotá: A Descriptive Analysis. Frontiers in Environmental Science, 8, 65. https://doi.org/10.3389/fenvs.2020.00065

Seinfeld, J. H., & Pandis, S. N. (2016). Atmospheric Chemistry and Physics: From Air Pollution to Climate Change (3rd ed.). Wiley.

Sun, Y., Wang, Z., Wild, O., Xu, W., Chen, C., Fu, P., ... & Worsnop, D. R. (2013). Aerosol composition, sources and processes during wintertime in Beijing, China. Atmospheric Chemistry and Physics, 20(1), 275–289. https://doi.org/10.5194/acp-13-4577-2013

Zhang, Q., Jimenez, J. L., Canagaratna, M. R., Allan, J. D., Coe, H., Ulbrich, I. M., ... & Worsnop, D. R. (2007). Ubiquity and dominance of oxygenated species in organic aerosols in anthropogenically-influenced Northern Hemisphere midlatitudes. Geophysical Research Letters, 34(13), 1718–1726. https://doi.org/10.1029/2007GL029979

Inversión térmica: qué es y cuál es su relación con la calidad del aire - busqueda - Secretaría Distrital de Ambiente. (2021). Secretaría Distrital de Ambiente. https://www.ambientebogota.gov.co/search?p_p_id=101&p_p_lifecycle=0&p_p_state=maximized&_101_struts_action=%2Fasset_publisher%2Fview_content&_101_assetEntryId=1755532&_101_type=content&_101_urlTitle=inversion-termica-que-es-y-cual-es-su-relacion-con-la-calidad-del-aire#:~:text=salga%2C%20formando%20la%20conocida%20popularmente,local%20y%20regional%20de%20part%C3%ADculas

Blanco-Becerra, L. C., Gáfaro-Rojas, A. I., & Rojas-Roa, N. Y. (2015). Influence of precipitation scavenging on the PM2.5PM10 ratio at the Kennedy locality of Bogotá, Colombia. Revista Facultad de Ingeniería Universidad de Antioquia, 76. https://doi.org/10.17533/udea.redin.n76a07

Mendez-Espinosa, J., Belalcazar, L., & Betancourt, R. M. (2019). Regional air quality impact of northern South America biomass burning emissions. Atmospheric Environment, 203, 131-140. https://doi.org/10.1016/j.atmosenv.2019.01.042

Marín, D., Herrera, V., Piñeros-Jiménez, J. G., Rojas-Sánchez, O. A., Mangones, S. C., Rojas, Y., Cáceres, J., Agudelo-Castañeda, D. M., Rojas, N. Y., Belalcazar-Ceron, L. C., Ochoa-Villegas, J., Montes-Mejía, M. L., Lopera-Velasquez, V. M., Castillo-Navarro, S. M., Torres-Prieto, A., Baumgartner, J., & Rodríguez-Villamizar, L. A. (2025). Long-term exposure to PM2.5 and cardiorespiratory mortality: an ecological small-area study in five cities in Colombia. Cadernos de saude publica, 41(4), e00071024. https://doi.org/10.1590/0102-311XEN071024

Dominici, F. (2004). Time-Series Analysis of Air Pollution and Mortality: A Statistical Review (Research Report 123). Health Effects Institute.

Korhonen, A., Sofiev, M., et al. (2019). Influence of Spatial Resolution on Population PM2.5 Exposure and Health Impact Assessment. Environmental Science and Pollution Research, 26(22), 22459–22472.