UNIVERSIDAD
NACIONAL
DE COLOMBIA

Predicción de ozono troposférico en Bogotá: un enfoque de Machine Learning

Diana Rocío García Millán

Universidad Nacional de Colombia

Facultad de Ingeniería, Departamento de Ingeniería Química y Ambiental

Bogotá D.C., Colombia

2024

Predicción de ozono troposférico en Bogotá: un enfoque de Machine Learning

Diana Rocío García Millán

Tesis o trabajo de investigación presentada(o) como requisito parcial para optar al título de:

Magister en Ingeniería Ambiental

Director:

Ph.D. Néstor Yesid Rojas Roa

Codirector:

PhDc. Alejandro Casallas García

Línea de Investigación:

Calidad del aire

Grupo de Investigación:

Grupo de Investigación de Calidad de Aire, GICA

Universidad Nacional de Colombia

Facultad de Ingeniería, Departamento de Ingeniería Química y Ambiental

Bogotá D.C., Colombia

2024

Satisfecha de hasta donde he llegado, disfrutando de mi presente y mi actual yo, y entusiasmada de lo que vendrá a mi vida.

Agradecimientos

Gracias a Dios por darme una hermosa familia y novio, los cuales me apoyaron en cada decisión tomada en este camino. Aunque no ha sido sencillo, todo este sacrificio ha sido gratificante por el conocimiento adquirido y como me aporta tanta profesional como personalmente.

Agradezco profundamente el apoyo continuo y paciencia del Dr. Néstor Yezid Rojas Roa y al Dr. Alejandro Casallas García, director y codirector de este trabajo respectivamente, que no solo fueron excelentes guías, sino también excelentes personas que acompañaron todo este proceso, compartiendo su conocimiento conmigo. Y, por último, pero no menos importante, a la ingeniera Janaai Daniela Bustos Quevedo, que desde el Grupo de Investigación de Calidad del Aire - GICA, con su paciencia y disposición, me permitió conocer diferentes herramientas para obtener los resultados plasmados en el documento.

Resumen

Predicción de ozono troposférico en Bogotá: un enfoque de Machine Learning

El ozono troposférico es una preocupación ambiental que tiene repercusiones tanto en la salud humana como en los ecosistemas, aún más, cuando su producción depende directamente de factores tanto ambientales como antropogénicos. En los últimos años, en Bogotá, el ozono ha tenido una tendencia de aumento en su concentración. Por ello, es imperativo estudiar la razón por la cual este incremento se está presentando, así como realizar avances en modelos que permitan realizar alertas tempranas y reducir el riesgo. La predicción es crucial para concientizar sobre la salud pública y la implementación de estrategias de gestión de la calidad del aire. Se realiza un estudio general del comportamiento y evolución del ozono (O_3) en las diferentes zonas de Bogotá. Este proyecto tiene como objetivo diseñar un modelo predictivo de aprendizaje automático (*machine learning*) de los niveles de O_3 en Bogotá utilizando datos de la Red de Monitoreo de Calidad del Aire de Bogotá (RMCAB), incluyendo mediciones de precursores de ozono y variables meteorológicas. Los modelos de aprendizaje automático utilizados fueron redes convolucionales y capas de memoria bireccional a largo plazo (LSTM) de la biblioteca Tensor Flow Keras y el paquete Python Sklearn, para facilitar la categorización de técnicas de inteligencia artificial. Esto da como resultado un modelo que destaca por su capacidad de ofrecer pronósticos altamente precisos al considerar y cuantificar la influencia de los precursores. Los resultados del modelo determinaron una correlación de Spearman mayor a 0.6, la raíz del error cuadrático medio se mantuvo por debajo de 10 $\mu g/m^3$ en todos los casos y el Índice de Ajuste superó el valor de 0.5 en todos los casos, categorizados como bueno, excelente y bueno respectivamente, lo cual sugiere que el modelo replica con precisión y exactitud el comportamiento y la tendencia del ozono. Se espera que la metodología aplicada sirva como referencia para continuar con la predicción de otros contaminantes atmosféricos de

interés y de esta forma dar apoyo a la toma de decisiones que permitan minimizar los impactos ambientales enfocados a la calidad del aire.

Palabras clave: Ozono troposférico, machine learning, calidad del aire, predicción, precursores.

Abstract

Tropospheric ozone prediction in Bogotá: A Machine Learning Approach

Tropospheric ozone is an environmental concern has repercussions on both human health and ecosystems, even more so when its production depends directly on both environmental and anthropogenic factors. In recent years, in Bogotá, ozone has had a trend of increasing concentration. Therefore, it is imperative to study the reason why this increase is occurring, as well as make advances in models that allow early warnings and risk reduction. The prediction is crucial for public health awareness and implementation of air quality management strategies. A general study of the behavior and evolution of ozone (O_3) in the different areas of Bogotá is carried out. This project aims to create a predictive machine learning model for ozone levels in Bogotá using data from the air quality monitoring network, including measurements of O_3 precursors and meteorological variables. Machine learning models used were Convolutional Networks and Birectional Long Short-Term Memory (LSTM) layers from the Tensor Flow Keras library, and the Python package Sklearn, to facilitate the categorization of artificial intelligence techniques. This results in a model that stands out for its ability to offer highly accurate forecasts by considering and quantifying the influence of precursors. The results of the model determined a Spearman correlation greater than 0.6, the root mean square error remained below 10 µg/m³ in all cases and the Fit Index exceeded the value of 0.5 in all cases, categorized as good , excellent and good respectively, which suggests that the model accurately and precisely replicates the behavior and trend of ozone. It is expected that the applied methodology will serve as a reference to continue with the prediction of other atmospheric pollutants of interest and in this way provide support for decision-making that allows minimizing environmental impacts focused on air quality.

Keywords: Tropospheric ozone, Machine Learning, Convolutional, Air quality.

Contenido

1. Marco de referencia 5
1.1 Marco teórico 6
1.1.1 Formación de O_3 6
1.1.2 Cálculo de tendencia de los contaminantes 7
1.1.2.1 Descomposición de Tendencia Estacional 7
1.1.2.2 Regresión de Mínimos Cuadrados Generalizados 8
1.1.3 Aprendizaje automático (Machine Learning) 9
1.1.3.1 Redes neuronales artificiales 12
1.1.4 Aplicaciones de inteligencia artificial en calidad del aire 15
1.1.5 Riesgo poblacional 18
1.1.6 Marco legal 22

2. Metodología 24
2.1 Área de estudio 25
2.2 Fuente de datos 27
2.3 Estadística descriptiva 29
2.4 Estructura y diseño del modelo de Machine learning 31
2.4.1 Datos de entrada y selección de variables 31
2.4.2 Matriz de entrada y salida 36
2.4.3 Estructura y aplicación del modelo de Machine Learning 36
2.5 Validación y evaluación del modelo 39
2.6 Cálculo del riesgo poblacional y afectación a la 4salud 41

3. Resultados y discusión 45
3.1 Estadística Descriptiva 45
3.1.1 Análisis espacial 45
3.1.2 Análisis temporal 48
3.1.2.1 Descripción general de datos 48
3.1.2.2 Densidad de datos medidos de ozono 50
3.1.2.3 Comportamiento mensual del ozono 51
3.1.2.4 Tendencia multianual mensual del ozono 53
3.2 Modelo de predicción *machine learning* 56
3.3 Validación del Modelo 62
3.3.1 Distribución 62
3.3.2 Errores estadísticos 64
3.4 Riesgo poblacional 67
3.4.1 Impactos a la salud - Mortalidad 72

4. Conclusiones y recomendaciones 77

Lista de figuras

Figura 1-1: Redes neuronales y sus respectivas capas. 12
Figura 1-2: Ejemplo de una red neuronal CNN 1-D con 3 CNN y 2 layers. 13
Figura 1-3: Estructura de una red neuronal convolucional aplicando tensor 2-D. .. 14
Figura 1-4: Estructura de la red neuronal recurrente LSTM con unidades cerradas. 15
Figura 2-1: Metodología general dividida por fases para el diseño del modelo predictivo de O_3 24
Figura 2-2: Distribución de empresas en Bogotá entre enero - junio de 2023 por localidad. 26
Figura 2-3: Ubicación de las estaciones de RMCAB..... 27
Figura 2-4: Comportamiento mensual de O_3, NO y NO_2 32
Figura 2-5: Correlación Spearman entre los precursores del O_3..... 34
Figura 2-6: Estructura del modelo ML para predicción de ozono 37
Figura 2-7: Arquitectura del modelo ML para predicción de ozono 38
Figura 2-8: Metodología adaptada para el cálculo de riesgo poblacional y afectación a la salud. 41
Figura 3-1: Promedio multianual de O_3 en Bogotá, calculado a partir de datos diarios del máximo promedio de 8 horas. 46
Figura 3-2: Boxplot de contaminantes O_3, NO, NO_2, PM_{10}, $PM_{2.5}$ y CO. La línea naranja representa la media de los datos, la caja tiene los datos desde 25 hasta 75 percentil, y los bigotes van de 5 a 95 percentil. 48
Figura 3-3: Distribución y densidad de datos medidos de O_3. La línea negra dentro de la caja azul es la media de los valores, así como también la línea roja dentro de la distribución normal de la gráfica de abajo..... 50
Figura 3-4: Comportamiento multianual mensual de los niveles de ozono por estación. Las líneas continuas representan la media y la sombra es la desviación estándar. 51
Figura 3-5: Comportamiento multianual mensual de los niveles de ozono por estación. Las líneas continuas representan la media y la sombra es la desviación estándar. (Continuación Figura 3-4). 52
Figura 3-6: Descomposición STL y regresión GLS de la media mensual de los niveles de O_3 en Bogotá. Se tuvieron en cuenta la información disponible por cada estación. La línea violeta es el SLT y el componente estacional. La línea negra

representa el componente de tendencia STL y la línea roja es la pendiente ajustada GLS que representa la tendencia lineal del O_3. 54
Figura 3-7: Descomposición STL y regresión GLS de la media mensual de los niveles de O3 en Bogotá. Se tuvieron en cuenta la información disponible por cada estación. La línea violeta es el SLT y el componente estacional. La línea negra representa el componente de tendencia STL y la línea roja es la pendiente ajustada GLS que representa la tendencia lineal del O3. (Continuación 55
Figura 3-8: Valores de observación vs predicción de O_3 para las estaciones que realizan mediciones de las variables necesarias como entradas en el modelo. La hora 0 y la hora 100 representa desde el primer valor predicho hasta el valor número 100. 58
Figura 3-9: Diagrama de dispersión de observación vs predicción de ozono de estaciones que realizan mediciones de las variables necesarias como entradas en el modelo. La línea negra corresponde a la regresión lineal 1:1, las líneas punteadas azul y rojas corresponden a 1:2 y 1:0,5 respectivamente. 61
Figura 3-10: Dispersión de datos respecto al O_3 superficial de Bogotá a partir de las medias del conjunto ML para un subconjunto de 2000 puntos de datos submuestreados aleatoriamente del conjunto de datos de validación. El O_3 se representa gráficamente en función de la radiación, NO_2, NO, CO, PMC, $PM_{2.5}$, componentes del viento (u y v) y velocidad del viento. El panel inferior derecho representa la observación del ozono frente a la media del conjunto ML. 63
Figura 3-11: Resultado de los errores estadísticos para la validación del modelo, graficando los valores sombreados en rosa de los parámetros NMB, RMSE, Rho, FAC2, IOA, MB, HIT, FAR, POC de las estaciones que se modelaron de Bogotá. La línea roja punteada indica el valor de referencia según la Tabla 2-4. 65

NOTA: Se hace la aclaración que este trabajo la separación de decímales se realizó con el signo “.” mientras que los miles con el signo “,”

Lista de tablas

Tabla 1-1: Estimación de mortalidad según cada contaminante. 21
Tabla 1-2: Normatividad aplicable vigente para el sector ambiente. 22
Tabla 2-1: Contaminantes y variables meteorológicas medidas por cada estación de la ciudad 28
Tabla 2-2: Variance Inflation Factor (VIF) para las variables de entrada del modelo de predicción de O_3. 35
Tabla 2-3: Clasificación de parámetros para la validación del modelo. 39
Tabla 2-4: Valores de referencia de los parámetros de validación FAR, POC, HIT, NMB, MB y FAC2. 40
Tabla 2-5: Niveles límites permisibles de O_3 por tiempo de exposición 40
Tabla 2-6: Porcentaje de la población de cada una de las localidades por grupo de referencia. 42
Tabla 3-1: Valor media y curtosis de las variables..... 49
Tabla 3-2: Resultados obtenidos de correlación para cada estación. 66
Tabla 3-3: Concentraciones promedio anual $PM_{2.5}$ y promedio anual calculado del promedio diario máximo de 8 horas de O_3 para las estaciones de Bogotá entre los años 2013 y 2022. 68
Tabla 3-4: Indicador de exposición para $PM_{2.5}$ y O_3 68
Tabla 3-5: Clasificación del riesgo para exposición. 69
Tabla 3-6: Tabla riesgo poblacional por rango de edad para cada estación..... 70
Tabla 3-7: Resultados del cálculo de factor de exposición y clasificación de riesgo por exposición para cada estación de los contaminantes $PM_{2.5}$ y O_3. 71
Tabla 3-8: Casos confirmados de mortalidad en Bogotá. 72
Tabla 3-9: Datos de Incidencia inicial (por cien mil personas) de mortalidad por todas las causas, cardiovascular y respiratoria, y Coeficientes β para estimar la asociación por cada aumento en la concentración de contaminantes. 73
Tabla 3-10: Total de casos de mortalidad excedentes al año 2021 por cada estación. .. 74

Introducción

La contaminación atmosférica es originada por actividades naturales y antropogénicas, estas últimas han aumentado significativamente los niveles de contaminación, produciendo efectos adversos a la salud y al medio ambiente (IARC, 2016). Por ende, es un problema ambiental a nivel global al ser el cuarto factor de riesgo de mortalidad en 2019 (WHO, 2018), pues cerca de 4.2 millones de muertes en todo el mundo están relacionadas con la contaminación en el aire, a causa de enfermedades cardíacas, cerebrovasculares, pulmonares e infecciones respiratorias. Las principales fuentes antropogénicas de emisión son: el consumo de combustible fósil para generación de electricidad, transporte, industria, y procesos industriales (Agencia Europea de Medio Ambiente, 2020).

La mayoría de las emisiones contaminantes se producen en la superficie de la Tierra, en la capa más baja de la atmósfera denominada capa de mezcla, que, dependiendo de las condiciones meteorológicas, los contaminantes pueden acumularse en ciertas zonas dando lugar a la aparición de "*smog*" (European Environment Agency [EEA], 2008). Son seis contaminantes del aire comunes conocidos como "contaminantes criterio", material particulado (PM), ozono a nivel del suelo (O_3), monóxido de carbono (CO), óxidos de azufre (SO_x), óxidos de nitrógeno (NO_x) y plomo (Pb), donde el O_3 y el PM son las amenazas para la salud más diseminadas (EPA, 2007). A nivel nacional se consideran 5 de los 6 contaminantes como criterio, pues no se incluye el plomo. Estos contaminantes están establecidos en la Resolución 2254 de 2017 del Ministerio de Ambiente y Desarrollo Sostenible (MADS), la cual establece el nivel máximo permisible de 100 $\mu g/m^3$ para O_3 en un tiempo de exposición de 8 horas promedio.

Siendo uno de los contaminantes criterio por sus diversos efectos negativos en la salud humana y medio ambiente (EPA, 2007), el ozono ha sido monitoreado y estudiado satelitalmente, así como también, en estaciones en superficie de calidad de aire, pues el

aumento de sus concentraciones puede traer problemas a la salud humana (Monks et al. 2015). En Colombia, por ejemplo, se realiza monitoreo y seguimientos a los contaminantes atmosféricos por medio de Sistemas de Vigilancia de la Calidad de Aire SVCA, creado por la Resolución 601 de 2006 de MADS. Estos sistemas han permitido realizar diversos estudios sobre el comportamiento de ozono, los cuales han concluido que las medidas efectivas para reducir NO_x y PM no funcionarán para reducir O_3, ya que se requiere más estudios de su dinámica química (Henao et al., 2021). Esta compleja formación fue estudiada por Sokhi et al. (2021) y por Casallas et al (2023), quienes encontraron que el proceso de titulación es esencial para entender el comportamiento del ozono en Bogotá. Otro factor esencial es que existe una relación entre el incremento en los niveles de O_3 y las emisiones de la quema de biomasa especialmente en las temporadas de sequía (Ballesteros-González et al., 2020, Casallas et al. 2023), pues vientos meridionales transportan masas de aire contaminadas desde la frontera Colombo-Venezolana hasta Bogotá (Casallas et al. 2022a).

Según el Informe Anual de Calidad del Aire año 2020 de la Red de Monitoreo de Calidad del Aire de Bogotá - RMCAB y Casallas et al. (2024), el mes con más alta concentración de Ozono (O_3) fue marzo, seguido del mes de abril, posiblemente por la temporada de quemas de biomasa. Esto aunado con la complejidad topográfica de los Andes del Norte, donde se encuentra Colombia, induce el transporte químico regional de materia orgánica en aerosol, que representa aproximadamente el 39% de la masa total de $PM_{2.5}$ debido a la quema de biomasa, y se estima que puede ser responsable de enfermedades respiratorias en áreas densamente pobladas (Ballesteros-Gonzalez et al., 2020). Se corrobora, entonces, que las variables meteorológicas como la dirección y rapidez del viento, impactan en la concentración y dispersión de los contaminantes atmosféricos (Bolaños-Diaz et al., 2022), así como, la correlación entre éstos y el número de incendios (Mendez-Espinosa et al., 2022). Aparte de las quemas, se ha evidenciado en los últimos estudios la resuspensión de material particulado por el tráfico rodado (Ramírez et al., 2019), aportando de igual manera a que los niveles de O_3 se eleven. Aun cuando los valores no superan el nivel máximo permisible establecido por la Resolución 2254 del 2017 de MADS, ha sido notable el aumento en las concentraciones de ozono troposférico lo cual genera preocupación (Casallas et al. 2024).

Adicionalmente, en el estudio realizado por Sokhi et al. (2021), en algunas ciudades de Colombia se vio un aumento en concentraciones de O_3 relacionadas con la reducción de NO por la reacción del NOx (debido a restricciones en la movilidad en pandemia por el Covid), con las concentraciones de compuestos orgánicos volátiles (VOCs), y CH_2O particularmente con los incendios forestales (Ballesteros-Gonzalez et al. 2020; Casallas et al. 2024), lo cual considera un desafío formular estrategias para minimizar los niveles de O_3. Dicho de otra manera, es importante prevenir escenarios que pueden ser positivos para la reducción de los contaminantes atmosféricos, pero que originen un impacto adverso en los niveles de O_3, por lo que este estudio se encamina a diseñar un modelo para la predicción de O_3 y de esta forma aportar una herramienta para el seguimiento y toma de decisiones.

Este estudio constituirá una valiosa contribución para la predicción de contaminantes criterio en Latinoamérica, dada la escasez de herramientas similares en la región. Su objetivo general es predecir el comportamiento del ozono troposférico en Bogotá a corto plazo. Para lograrlo, se plantean los siguientes objetivos específicos: identificar las variables meteorológicas y de contaminación del aire relacionadas con el ozono troposférico que no presenten multicolinealidad, diseñar un modelo de Machine Learning para predecir las concentraciones de ozono troposférico y, finalmente, validar el modelo seleccionado mediante parámetros estadísticos y categóricos. Como aporte adicional de este estudio, se hace un estimado del riesgo poblacional por exposición que se relaciona directamente con salud pública, el cual podría encaminar problemas futuros en las ciudades con altas concentraciones de los contaminantes criterio.

1. Marco de referencia

La contaminación atmosférica ha sido ampliamente estudiada a nivel mundial debido a sus implicaciones en la salud humana y las repercusiones que tiene en el medio ambiente. Por ejemplo, las concentraciones elevadas de O_3 pueden afectar la salud humana, irritando las vías respiratorias. En cuanto al medio ambiente, este se puede ver afectado cuánto niveles altos de O_3 generan una disminución en la capacidad de las plantas para realizar fotosíntesis (Jin et al., 2023).

En los últimos años, el ozono troposférico ha venido presentando un aumento considerable de sus concentraciones en diferentes zonas. Esto indica que los controles que se han implementado para reducir o evitar el incremento del O_3 troposférico no han dado los mejores resultados, por lo que debe re-evaluarse. Lo anterior, sumando con los resultados de Sokhi et al (2021) entre otros estudios (Barten et al., 2020; Deroubaix et al., 2021; Henao et al., 2018), reportaron que la reducción del NO puede dar como resultado un incremento de O_3 debido a que el NO es el encargado de consumir el ozono y reducir su concentración por medio de la titulación.

Así mismo, el comportamiento de los contaminantes y la química de la atmósfera han sido estudiado con el modelo WRF-Chem, donde se encontró que el transporte a escala regional del O_3 sumado a la dinámica que hay entre PM_{10} y O_3 y la meteorología, puede afectar los niveles del O_3, según la complejidad orografía local como sucede en la Región Andina en Sur América (González et al., 2018). Este comportamiento dependerá de las emisiones y la química del lugar a estudiar (urbano, rural, suburbano), factores meteorológicos (e.g., temperatura, radiación para fotoquímica y altura de la Planetary Boundary Layer - PBL), precipitaciones, deposiciones húmedas, entre otros factores, los cuales pueden considerarse posibles fuentes de error para las predicciones de los contaminantes atmosféricos (Kumar, et al., 2015).

En diferentes ciudades del mundo, y más significativamente en Bogotá, durante el periodo de confinamiento, se vio un aumento en los niveles de O_3, principalmente por la reducción de NO, relacionado con la movilidad y con la reducción de ozono a causa de la titulación (Liu et al., 2021; Sokhi et al., 2021; Casallas et al. 2023). Debido a esto, en las distintas investigaciones se hace evidente e imperativo continuar con los estudios que determinen los factores para tener en cuenta en la titulación del ozono y su producción.

1.1 Marco teórico

1.1.1 Formación de O_3

Los niveles de concentración de ozono dependen de la intensidad lumínica, la concentración de óxidos de nitrógeno, los VOCs que son liberados en la combustión de gasolina por automóviles, refinerías de petróleo y plantas de fabricación (EPA, 2007) y también por las variables meteorológicas, especialmente la radiación que permite las reacciones químicas que producen ozono (Seinfeld & Pandis, 2006). Seinfeld & Pandis (2016) describen que la formación de ozono ocurre como resultado de la fotólisis del NO2 cuando NO2 y NO se encuentran en presencia de luz solar (hv < 424 nm) como se observa en la Ecuación 1-1 y Ecuación 1-2.

Ecuación 1-1: $$NO_2 + hv \rightarrow NO + O$$

Ecuación 1-2: $$O + O_2 + M \rightarrow O_3 + M$$

El oxígeno monoatómico reacciona con una molécula de oxígeno del aire para formar ozono, donde M puede ser O_2 o N_2 de la atmósfera. Adicional al NO_2, la presencia VOC también es determinante para la formación de O_3 por su participación en las reacciones de radicales OH, existiendo una relación óptima de NO_2:VOCs que favorece la formación de O_3, donde su aumento conduce al incremento en los niveles de O_3 (Seinfeld & Pandis, 2006).

El consumo de ozono se produce por un proceso llamado titulación, en el cual el ozono es consumido por el NO, produciendo NO_2 y O_2. Se genera, entonces, un ciclo continuo

entre estas 3 variables y la radiación solar, como se ve en la reacción química (ver Ecuación 1-3)

Ecuación 1-3: $$NO + O_3 \rightarrow NO_2 + O_2$$

Dado que el ozono troposférico puede causar daños graves a la salud, es importante diseñar modelos de predicción del ozono, pues estos permiten que las ciudades se preparen para eventos de alta concentración de este contaminante.

1.1.2 Cálculo de tendencia de los contaminantes

Con el fin de estudiar mejor el comportamiento de los datos del contaminante O_3 obtenidos por la RMCAB, se realizan los siguientes cálculos de tendencia:

1.1.2.1 Descomposición de Tendencia Estacional

La descomposición de tendencia estacional (SLT, por sus siglas en inglés) es el procedimiento desarrollado por Cleveland et al. (1990), basado en el Locally wighted scatterplot smoothing (LOESS o LOWESS), el cual es una regresión lineal para utilizado para suavizar las curvas de los datos, mediante la técnica de regresión lineal ponderada. El LOESS se puede usar para suavizar una función de cualquier número de variables independientes, a diferencia de SLT, que utiliza solo una variable independiente.

Según Cleveland et al. (1990), el cálculo de $g(x)$ se realiza teniendo en cuenta que q es un número entero, y suponiendo que $q \leq n$. Se seleccionan los valores q de x_i que son más cercanos a x y a cada uno se le asigna un “peso de vecindad” basado en su distancia a x. Sea $\lambda_q(x)$ la distancia del q-ésimo x_i más alejado de x y W la función de peso del tricubo. Por lo tanto, los x_i cercanos a x tienen los pesos más grandes y cuando x_i aumenta la distancia desde x, obteniendo cero en el q-ésimo más lejano (ver Ecuación 1-4, Ecuación 1-5 y Ecuación 1-6)

Ecuación 1-4: $$W(u) = (1 - u^3)^3 \quad for\ 0 \leq u < 1$$

Ecuación 1-5: $$W(u) = 0 \quad for\ u \geq 1$$

Ecuación 1-6: $$v_i(x) = W\left(\frac{|x_i - x|}{\lambda_q(x)}\right)$$

Donde *vi* es el peso de vecindad de cualquier *x*.

El siguiente paso es ajustar el polinomio de grado *d* a los datos con peso $v_i(x)$ en (x_i, y_i). El valor del polinomio ajustado localmente en *x* es *g(x)*, utilizando *d=1* o *2*, es decir, el ajuste es localmente lineal o localmente cuadrático. $\lambda_q(x)$ es la distancia desde *x* a la x_i más lejana, asumiendo que *q* > *n* (ver

Ecuación 1-7: $$\lambda_q(x) = \lambda_n(x)\frac{q}{n}$$

Posteriormente, se procede a utilizar la definición de pesos de vecindad usando los valores de $\lambda_q(x)$. Los valores de *d* y *q* son escogidos, a medida que *q* aumenta, *g(x)* se suaviza más, lo cual fue discutido por Cleveland et al. (1990).

1.1.2.2 Regresión de Mínimos Cuadrados Generalizados

La regresión de mínimos cuadrados generalizados (GLS, por sus siglas en inglés) se usa para estimar la relación lineal entre una serie de tiempo autocorrelacionada, el tiempo para obtener residuos independientes y la estimación correcta de la varianza de los coeficientes de regresión (Singh et al. 2021). Según Brockwell & Davis (2002), la estimación de GLS se realiza según el modelo (ver Ecuación 1-8 y Ecuación 1-9):

Ecuación 1-8: $$Y_t = x_t'\beta + W_t\ , \quad t = 1, \ldots, n$$

o en forma de matriz:

Ecuación 1-9: $$Y = X\beta + W$$

donde, $Y = (Y_1, \ldots, Y_n)'$ es el vector en observaciones de tiempo *t = 1*, ….., *n*, *X* es el diseño de la matriz donde cuya *t*-esima fila $x'_t = (x_{t1}, \ldots, x_{tk})$, que consta de los valores de

las variables explicativas *t*, *β (β1,...,βk)* es el vector de regresión de coeficientes, y los componentes de W (W1,...,Wn) son valores de un proceso de Modelo Autorregresivo de Media Móvil (ARMA por sus siglas en inglés) (*p*, *q*) de media cero causal. El estimador de *GLS* de *β*, es el *βGLS*, que minimiza la suma ponderada de los cuadrados (ver Ecuación 1-10 y Ecuación 1-11)

Ecuación 1-10: $$(Y - X\beta)'\Gamma_n^{-1}(Y - X\beta)$$

Donde:

Ecuación 1-11: $$\Gamma_n = E(WW'), es\ la\ matriz\ de\ covarianza\ de\ W$$

Posteriormente se deriva parcialmente con respecto a cada componente de *β* e iguala las derivadas a cero, obteniendo la Ecuación 1-12:

Ecuación 1-12: $$\beta_{GLS} = (X'\Gamma_n^{-1}X)^{-1}X'\Gamma_n^{-1}Y$$

1.1.3 Aprendizaje automático (Machine Learning)

En el campo de la inteligencia artificial, se han desarrollado modelos basados en la programación de algoritmos capaces de procesar datos y clasificarlos según sus características más importantes (Chaparro et al., 2022). De igual forma, es uno de los métodos que permite minimizar los impactos de estos contaminantes con la información existente, al realizar simulaciones de los niveles de contaminación atmosférica (Casallas et al., 2023; Yafous, et al., 2022). Por ello, el uso de algoritmos de *Machine Learning* (de ahora en adelante, ML) se ha hecho cotidiano en los sistemas con entornos cambiantes, pues tiene la capacidad de aprender y entrenarse continuamente (Alpaydin, 2016).

Como lo menciona Telikani et al. (2021), las técnicas de aprendizaje han estado inspiradas en la naturaleza, y están clasificadas de la siguiente manera:

- Bioinspirados: incluyen enfoques basados en inteligencia de enjambre y algoritmos inspirados en la evolución, que se originan en el comportamiento natural de los organismos.
- Inspirados en la física: residen en las reglas físicas/químicas.

- Inspirados en la geografía: generan soluciones aleatorias en materia geoespacial.
- De inspiración cultural: están inspirados en el comportamiento humano con respecto a las interacciones con otros.

Antes de iniciar con el manejo de algoritmos, se debe tener en cuenta la naturaleza de los datos, como lo describe Suthaharan (2015). Hay dos tipos de datos, los etiquetados y los no etiquetados. Los etiquetados tienen parámetros de entrada y salida en un patrón completamente legible por la máquina, pero requieren de una etiqueta previa por parte del desarrollador; mientras que los no etiquetados sólo tienen un o ningún parámetro en formato legible, requiriendo validaciones previas.

Como sucede con los diferentes métodos, existen clasificaciones para entrenar los algoritmos de ML. Según Aljarah et al. (2019), el aprendizaje de máquina se divide en tres clases de acuerdo con el uso requerido: Supervisado, no supervisado y de refuerzo:

a. Aprendizaje supervisado: hay un supervisor que da lineamiento de cuando una acción o decisión es buena o mala. Los más usados son los siguientes:

- Máquina de vectores soporte - SVM (por sus siglas en inglés, Support vector machine): el cual aprende mediante el ejemplo para asignar etiquetas a los objetos, maximizando una función matemática particular con relación a un conjunto de datos, y puede ser aplicado por ejemplo en reconocimiento de tarjetas de crédito fraudulentas, o en la clasificación automática de perfiles de expresión génica (Noble, 2006).
- Bosque aleatorio (Random Forest, RF): construido a partir de un conjunto de árboles que tienen características aleatorias en cada nodo, seleccionando aleatoriamente los datos de entrenamiento (Ali et al., 2012), siendo un método de clasificación y regresión de propósito general, y es aplicado para predicción en calidad del aire, reconocimiento de objetos 3D (Biau & Scornet, 2016).
- Red neuronal: Modelo formado por neuronas artificiales conectadas con coeficientes (pesos) que constituyen en la estructura neuronal, y se aplica principalmente en los procesos farmacéuticos (Agatonovic-Kustrin & Beresford, 2000). Se profundiza en el capítulo 1.1.3.1.

b. Aprendizaje no supervisado: es usado cuando el conjunto de datos no está etiquetado, por lo que el algoritmo debe encontrar las etiquetas y definirlas. Así mismo, deberá aprender la estructura del conjunto de datos y la relación entre las características. Los más usados son los siguientes:

- K-means clustering: es un modelo simple, que clasifica un conjunto de datos mediante un número de conglomerados (suponiendo k clusters) fijados previamente, y es utilizado en investigación de mercados, bioinformática, gestión de clientes, entre otras áreas (Kodinariya & Makwana, 2013).
- Red neuronal autoorganizada - SOFM (por sus siglas en inglés, Self-Organizing Feature Maps): Consta de dos capas neuronales, la primera, la capa de entrada de vectores de n dimensiones de números reales, y la segunda, la capa de unidad en cuadrícula bidimensional (Ultsch, 1993).

c. Aprendizaje de refuerzo: este algoritmo obtiene recompensas en caso de una acción correcta y se publica en caso de una acción incorrecta, simulando la forma en la que las criaturas aprenden a través de recompensas y castigos. Los más usados son los siguientes:

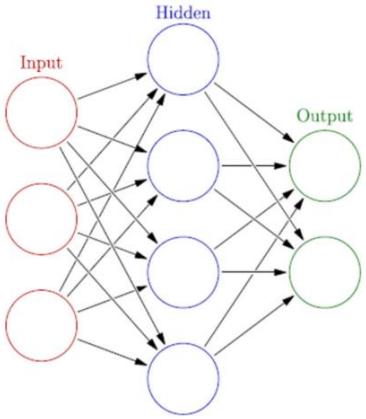
- Q learning: es una aproximación incremental o estocástica para estimar la función Q (Clifton & Laber, 2020).
- Deep Q Network (DQN): es una variación de Q learning con 3 mejoras: arquitectura de CNN profunda para la aproximación de la función Q; usar mini lotes de datos de entrenamiento aleatorios; usar parámetros de red más antiguos para estimar los valores Q del siguiente estado; este último proporciona un objetivo de entrenamiento estable (Roderick et al., 2017).
- Deep Deterministic Policy Gradient (DDPG): es un algoritmo para controles continuos a gran escala y manejo de grandes de entradas, un ejemplo en su aplicación es el control de la red de tráfico de una gran ciudad (Casas, 2017)

Los modelos de ML han empezado a ser utilizados en varias investigaciones (e.g., Albawi et al., 2017; Ali et al., 2012; Aljarha et al., 2019; AlOmar et al., Baena-Salazar et al., 2019; Casallas et al., 2021; Casallas et al., 2022b; Celis et al., 2022; Chaparro et al., 2022), por lo cual aplicarlos a diferentes ciudades y entornos es importante. Por ello, y porque esta investigación tiene la intención de usar este tipo de modelos, la siguiente sección describe algunas de las características de los modelos de ML.

1.1.3.1 Redes neuronales artificiales

Las redes neuronales artificiales - ANN son un sistema de procesamiento de conjuntos neuronales (o nodos, unidades) que se interconectan y entrelazan en una forma distribuida para aprender colectivamente de los datos de entrada, optimizando su salida (O'Shea & Nash, 2015). Un nodo es el lugar donde ocurren los cálculos, simulando la actividad de una neurona en el cerebro humano, que inicia cuando se encuentra el estímulo suficiente, combinando la entrada de datos con un conjunto de coeficientes o pesos que amplifican o amortiguan la entrada (Khalil et al., 2019). La ANN (ver Figura 1-1), consta de capas de entrada de neuronas, capas ocultas de neuronas y capas finales de neuronas que se conocen como capas de salida. Estas capas cuentan con una función de activación que es capaz de introducir elementos no lineales al aprendizaje de la red, y, por ende, la red neuronal será capaz de resolver problemas cada vez más complejos (Wang, 2003).

Figura 1-1: Redes neuronales y sus respectivas capas.



Fuente: (Wu & Feng, 2017)

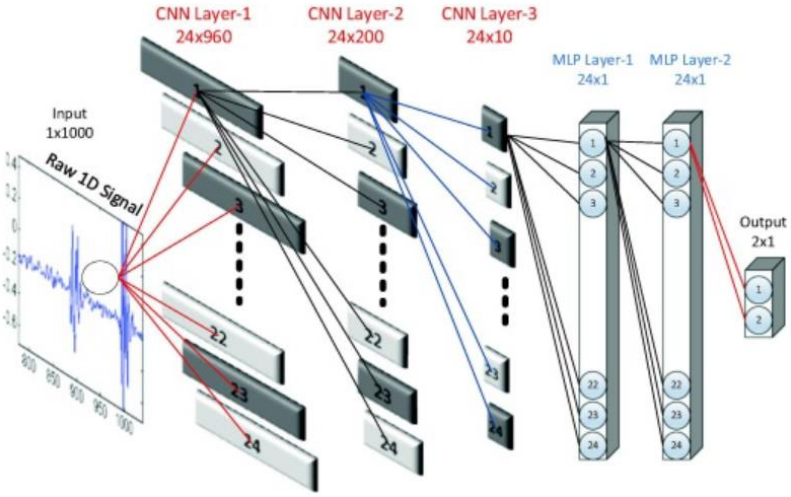
Según Wu & Feng (2017), las redes neuronales se caracterizan al ser no lineales, teniendo mejor rendimiento y mejoran la tolerancia a las fallas; siendo ilimitadas, al constar con múltiples neuronas conectadas entre sí, dependiendo de la interacción de las interconexiones entre estas; siendo no cualitativo, con capacidad de autoadaptación, autoorganización, y autoaprendizaje; y no tener conexidad, ya que la dirección evolutiva del sistema dependerá de una función de estado particular. Sus aplicaciones van más allá de patrones de reconocimiento, procesamiento de señales, optimización.

1.1.3.1.1 Redes neuronales convolucionales

Para reducir el número de parámetros en ANN, se aplican las redes neuronales convolucionales (CNN), que han sido desarrolladas para realizar tareas más complejas que no eran posibles con la ANN (Albawi et al. 2017). Según Goodfellow et al. (2016) en la CNN se aplica la operación matemática convolución sobre dos funciones de argumento de valor real. El primer argumento se denomina entrada y el segundo como núcleo o kernel en inglés, y la salida en ocasiones es denominada como mapa de características. En algunas aplicaciones de CNN, se utiliza la propiedad conmutativa, donde el kernel es invertido en relación con la entrada, o la correlación cruzada donde el kernel no es invertido. Normalmente se utiliza la convolución en combinación con otras funciones para que el ML aprenda los valores apropiados para el kernel.

La convolución discreta puede verse como una multiplicación por una matriz, en el caso de la convolución discreta univariante 1-D, cada fila de la matriz está restringida para ser igual a la fila anterior desplazada por un elemento (Figura 1-2). Los CNN 1-D pueden ser aplicados para aquellos casos en los que los datos de entrenamiento son escasos y se requiere la implementación en tiempo real y de bajo costo, por lo tanto, son relativamente fáciles de entrenar y ofrecen una complejidad computacional mínima, sin afectar el rendimiento (Kiranyaz et al., 2021).

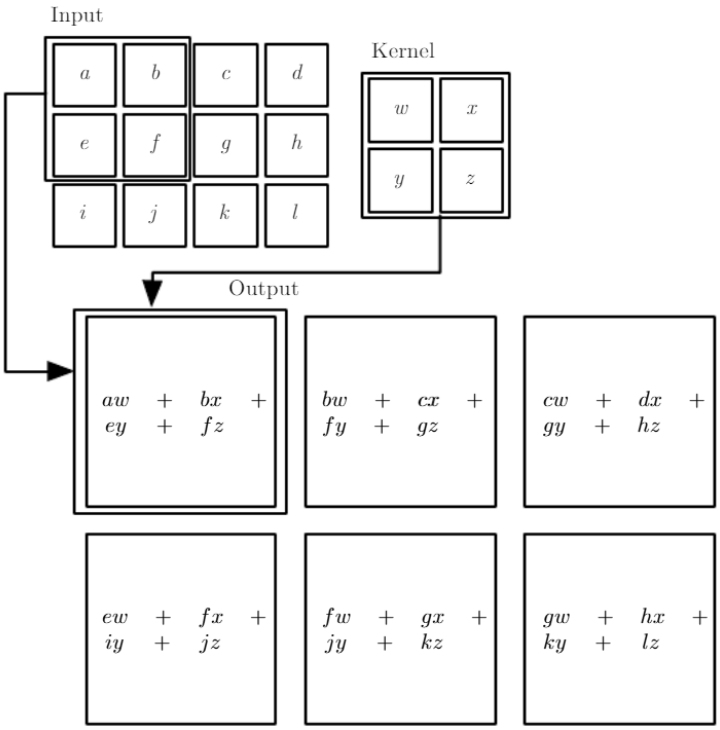
Figura 1-2: Ejemplo de una red neuronal CNN 1-D con 3 CNN y 2 layers.



Fuente: (Kiranyaz et al., 2021)

Para el caso de dos dimensiones 2-D (Figura 1-3), la matriz corresponde a la multiplicación de doble bloque. Las CNN 2-D son utilizadas principalmente para clasificación de imágenes o información con propiedades espaciales. Sin embargo**,** presenta inconvenientes, ya que requiere alta complejidad computacional especialmente para los entrenamientos, y que el conjunto de datos sea de gran tamaño (Kiranyaz et al., 2021).

Figura 1-3: Estructura de una red neuronal convolucional aplicando tensor 2-D.



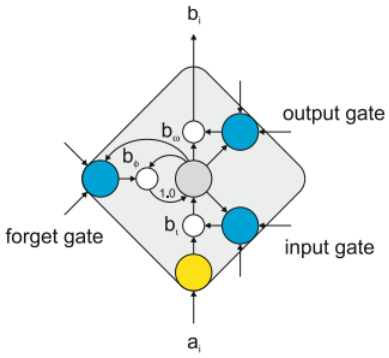
Fuente: (Goodfellow et al., 2016)

1.1.3.1.2 Redes neuronales recurrentes – LSTM

La red neuronal recurrente (RNN) consiste en una combinación de redes bucle, que permiten que la información persista pasando la información a la siguiente capa, lo cual tiene la desventaja de retrasar exponencialmente el aprendizaje a largo plazo (Kumar et al., 2018). Sin embargo, la red de memoria a largo plazo, o Long Short Term Memory (LSTM, por sus siglas en inglés), es un tipo de RNN, que no tiene algunas de las desventajas de la RNN común, ya que añade varios pasos intermedios como se muestra en la Figura 1-4**,** que reducen el costo computacional extra, pues después de aplicar la

función de activación a a_i, el resultado se multiplica por el factor *bl*, entre otros ajustes descritos en Sundermeyer et al. (2012).

Figura 1-4: Estructura de la red neuronal recurrente LSTM con unidades cerradas.



Fuente: (Sundermeyer et al., 2012).

La arquitectura de las celdas de memoria en LSTM permite el almacenamiento de información por un plazo más prolongado, usando en su formulación la función sigmoidea que se usa para formar tres puertas en la celda de memoria (Shewalkar et al., 2019), en otras palabras, una unidad de red neuronal estándar *i* solo consta de la entrada de activación a_i y la activación de salida b_i (ver Ecuación 1-13) que están relacionadas cuando se usa una función de activación tanh por (Sundermeyer et al., 2012):

Ecuación 1-13: $$b_i = tanh(a_i)$$

1.1.4 Aplicaciones de inteligencia artificial en calidad del aire

Los modelos predictivos han contribuido a fortalecer los sistemas de alertas tempranas en temas de calidad del aire (Casallas et al., 2020), especialmente los modelos ML como se observa a continuación en los estudios realizados.

Internacionalmente, se han realizado avances en el desarrollo de modelos de ML, demostrando que, con la combinación de varias técnicas para la predicción de O_3, es posible tener resultados más precisos y certeros. Por ejemplo: Yilmaz (2021) tuvo como objetivo reducir el riesgo a la contaminación, aplicando algoritmos genéticos y alcanzando una tasa de precisión del 94%; en India, encontrando buenos resultados en los valores de predicción utilizando 15 neuronas y 1 capa oculta en el modelo ANN en la

etapa de inicio (Kapadia & Jariwala, 2022). AlOmar et al. (2020) adoptó el robusto modelo híbrido W-ANN para pronosticar varias horas antes de la concentración de gas ozono, obteniendo pocos errores frente al modelo convencional. En China se han realizado varios estudios al ser uno de los países con mayor contaminación en el aire, como, por ejemplo: en el 2021, que comparó entre Community Multiscale Air Quality Model (CMAQ) y CNN en cuanto a precisión, siendo un sistema híbrido preciso, combinado con modelos numéricos y aprendizaje automático, e incluye en el modelo, los precursores meteorológicos y de calidad del aire (Sayeed et al. 2021).

Así mismo, se ha usado el modelo de regresión lineal para identificar las variables meteorológicas clave que impulsan la variabilidad del ozono superficial diario para cada celda de la cuadrícula en el estudio de Li et al. (2018). Sayeed et al. (2021) reportó que el modelo Convolutional Long Short Term Memory (ConvLSTM) logra un mejor rendimiento que los modelos Deep Convolutional Generative Adversarial Network (DCGAN) y Extra Loss Functional Term (LDCGAN), así como con el modelo físico Nested Air Quality Prediction Modelling System (NAQPMS), especialmente en los picos de concentraciones de O_3 en verano (Cheng et al., 2023). Chen et al. (2022) propuso un sistema híbrido de predicción de la concentración de ozono que consta de un modelo CNN y un modelo basado en transformadores, con la limitación que no logra capturar de manera precisa e instantánea los puntos de mutación. También se ha usado ML para estimar concentraciones de NO_2 superficial, como el modelo adaptado por Ghahremanloo et al. (2021), que utilizó la red neuronal convolucional profunda (Deep-CNN) y el Instrumento de Monitoreo TROPOsférico (TROPOMI).

El PM es un contaminante altamente estudiado, ya que está vinculado directamente con los problemas de salud y además es uno de los que presenta más altas concentraciones en varios países (EPA, 2021). Por lo anterior, muchos estudios se han centrado en predecir su comportamiento y magnitud con bastante precisión por medio de métodos de ML (e.g., Ghahremanloo et al., 2022; Sayeed et al., 2021a). A raíz de estos resultados, se han iniciado esfuerzos para aplicar estos modelos a otros contaminantes, como por ejemplo el ozono troposférico. En cuanto a los estudios centrados en el PM, Chiarvetto et al. (2008) aplicaron redes neuronales con el objetivo de predecir el material particulado PM_{10} en Argentina con alta precisión. En un estudio más reciente Zhang et al. (2021) desarrolló modelos refinados para elementos traza ($PM_{1.0}$) utilizando regresión de uso de

suelo de mínimos cuadrados (OLS-LUR) fueron validados en datos obtenidos en verano y en otoño (Zhang et al. 2021).

En Colombia, se han realizado estudios en materia de $PM_{2.5}$, en el área metropolitana de Bogotá, e.g, Casallas et al. (2021) desarrollaron una red neuronal recurrente para predecir $PM_{2.5}$ y diferentes variables meteorológicas, y fue el primer acercamiento para la construcción de un modelo de predicción enfocado a alertas de calidad del aire. A partir de este estudio, Celis et al. (2022), toman el modelo de Casallas et al. (2021) y lo comparan con otras dos herramientas ML con el fin de determinar su precisión y validarlos para ser utilizados en la generación de una alerta temprana. Celis et al. (2022) mostraron que una red neuronal con capas convolucionales y recurrentes de bi-dirección es el mejor modelo para predecir $PM_{2.5}$. Por lo que este modelo puede ser un buen punto de partida para la construcción de un modelo de predicción de ozono. El estudio de Celis et al. (2022) fue utilizado por Casallas et al. (2022b), en el cual se usó la red bidireccional para construir un modelo regional de $PM_{2.5}$ a partir de datos satelitales.

Por otro lado, Mogollón-Sotelo et al. (2021) construyeron y evaluaron una máquina de vectores de soporte soportes (SVM) para pronosticar $PM_{2.5}$ a nivel del suelo en Bogotá. Siguiendo la metodología propuesta por Mogollon-Sotelo et al. (2021), Chaparro et al. (2022) hicieron uso del modelo LSTM para la predicción del comportamiento del contaminante SOx, y utilizaron estos resultados para determinar el riesgo a la población en Bogotá. Así mismo, se ha hecho uso de ANN en Santa Marta para predecir la concentración de material particulado PM_{10} (Valencia et al., 2016), en el Valle de Aburrá para $PM_{2.5}$ (Baena-Salazar et al., 2019), y en Bogotá para la meteorología y $PM_{2.5}$ (Casallas et al., 2021) y PM_{10} (Franceschi et al., 2018). Estos modelos de ML han dado mejores resultados en comparación a los modelos numéricos para la química antiguamente utilizados en la calidad de aire como WRF-Chem, ya que en Colombia se encuentran algunas limitaciones por su orografía, clima tropical y especialmente por la falta un preciso inventario de emisiones (Kumar et al., 2015; Gonzalez et al., 2018; Barten et al., 2020; Celis et al. 2022).

Aun cuando se han generado esfuerzos y avances en el diseño de herramientas de ML en Colombia, la mayoría se han enfocado en $PM_{2.5}$ y solamente dos a la fecha en ozono (Henao et al. 2021). Henao et al. (2021) realizaron un estudio en Medellín y sus

alrededores, utilizando Random Forest y regresión lineal múltiple para comparar los efectos de la pandemia en la calidad del aire para todos los contaminantes criterio incluyendo O_3. Casallas et al. (2022a) construyeron un ensamble de modelos para identificar las contribuciones de los precursores y las variables meteorológicas en la producción de O_3. Los autores concluyeron que el ensamble es capaz de describir el comportamiento del O_3 en Bogotá y Medellín con gran precisión, sin embargo, su objetivo no es generar predicciones, por lo que el paso natural, es realizarlas.

El presente estudio se enfatiza en las herramientas de ML, y tiene sus bases en las conclusiones de los demás estudios realizados en Colombia con diferentes métodos de predicción del ciclo diurno del O_3. Con el fin de complementar la descripción del estado del arte, la siguiente sección incluye el marco legal, que cuenta con aquellas leyes que contemplan las concentraciones del ozono troposférico.

1.1.5 Riesgo poblacional

La presencia de O_3 en la atmósfera plantea un desafío significativo para la salud humana, consolidándose como el segundo contaminante más relevante después de las partículas suspendidas, con consecuencias sustanciales en la mortalidad y morbilidad (Acevedo S, 2021). Diversos estudios han abordado el cálculo de estimaciones globales del exceso de mortalidad asociado a la exposición prolongada al O_3, presentando cifras que varían desde alrededor de 1.3 millones/año (S. Chowdhury et al., 2020) para el exceso de muertes por todas las enfermedades respiratorias, hasta un mínimo estimado de 142 mil/año para la Enfermedad Pulmonar Obstructiva Crónica (EPOC) (J. Lelieveld et al., 2015).

La alta reactividad y baja solubilidad del ozono conducen a su absorción principalmente a través de la vía respiratoria, generando problemas de salud asociados con enfermedades pulmonares y cardiovasculares (Pope et al., 2009; Prabhakaran et al., 2020). La mortalidad atribuible a la contaminación del aire ambiente ya sea por $PM_{2.5}$ y O_3, se asocia con diversas enfermedades, como diabetes, trastornos neurológicos y cánceres (R. Burnett et al., 2018; Hahad et al., 2021; J. Lelieveld et al., 2020; Peters et al., 2019; Turner et al., 2020; Wong et al., 2016). La exposición prolongada a partículas finas $PM_{2.5}$ y O_3 se vincula con una disminución en la esperanza de vida y en años de vida saludable (RT Burnett et al., 2014; J. Lelieveld et al., 2015; Murray et al., 2020).

Este escenario resalta la complejidad y gravedad de los impactos del ozono en la salud humana, subrayando la imperiosa necesidad de abordar de manera efectiva y de conocer este problema crítico. De tal manera, en el estudio de Pozzer et al. (2022), se llevó a cabo una revisión de las estimaciones de exceso de mortalidad atribuible a la contaminación del aire a nivel mundial, centrándose en los contaminantes $PM_{2.5}$ y O_3. Al comparar diversos estudios disponibles, se identificó una tendencia de agrupamiento en dos categorías principales: un grupo con un exceso de muertes superior a 6 millones al año y otro con un rango de 2 a 4 millones al año. Este análisis reveló que la exposición a $PM_{2.5}$ genera aproximadamente un exceso de mortalidad 12 ± 8 veces mayor que la asociada al ozono.

Uno de los estudios más relevantes para el cálculo de la mortalidad asociado con la contaminación del aire, fue La aplicación del Modelo de Mortalidad por Exposición Global (GEMM) en China (Zheng, S, et al., 2021) donde se calculó una mortalidad prematura total de 1.55 millones de personas, con una tasa per cápita de 11.2 por cada 10,000 personas en el año 2016 para el $PM_{2.5}$.

La relación estrecha entre el $PM_{2.5}$ y ozono se observa en el estudio de (Dunner et al., 2018 en el que se observa un mayor impacto del $PM_{2.5}$ en días de O_3 alto. Además, un análisis epidemiológico en Madrid identificó una conexión directa entre las concentraciones diarias de ozono y la morbi-mortalidad. Un aumento de 25 µg/m³ sobre el nivel crítico se asoció con un incremento del 12% en la tasa de mortalidad y un aumento del 18 % en los ingresos hospitalarios de urgencia, abarcando patologías respiratorias y cardiovasculares, siendo seis veces mayores que los observados con otros contaminantes primarios como partículas en suspensión o dióxido de azufre (Cañada T, 2021).

Ciertos grupos de población se ven más afectados por este contaminante, ya sea por una mayor exposición, como los grupos de nivel social y económico más bajo, o por su mayor vulnerabilidad, como las personas mayores, embarazadas o niños (EEA, 2019). Un estudio en México señala que los adultos mayores de 60 años son la población más vulnerable al aumento de ozono troposférico, siendo la mortalidad por causas

respiratorias y cardiovasculares la más afectada por este contaminante (Río Camargo, 2020).

Otro de los estudios globales a destacar es la investigación sobre contaminación del aire ambiente por partículas y mortalidad diaria en 652 ciudades, que incluye a Colombia. Las estimaciones específicas de cada país sobre el cambio porcentual en la mortalidad diaria por todas las causas mostraron variaciones considerables, desde el 0.03% (para Colombia) hasta el 1.32% (para Australia) (Liu, C, et al., 2019).

En el estudio llevado a cabo por Acevedo (2020), se analiza la correlación de la calidad del aire con las enfermedades respiratorias en Colombia y su influencia en la morbilidad y mortalidad por COVID-19, el análisis se realizó tomando datos desde el año 2010 al 2019 para 11 localidades de la ciudad de Bogotá y de 15 ciudades a nivel nacional, mediante la metodología dosis respuesta teniendo en cuenta modelos previos utilizados por investigadores a nivel mundial y aplicándolos al caso colombiano, evidenciando que el O_3 es uno de los contaminantes con mayor influencia, en los casos de mortalidad.

De esta manera, la población de Bogotá, la capital de Colombia se ve afectada por la mala calidad del aire, principalmente debido a la presencia de $PM_{2.5}$ y O_3. En el año 2020, un grupo de investigadores llevó a cabo una evaluación cuantitativa basada en los valores máximos permisibles propuestos por la Organización Mundial de la Salud (OMS) para la concentración de $PM_{2.5}$ Los resultados revelaron un aumento en la correlación, que oscila entre el 1,2% y el 9%, en la tasa de morbilidad y mortalidad a corto plazo por enfermedades cardio-respiratorias, según las guías de calidad del aire y objetivos intermedios para $PM_{2.5}$ propuestos por la OMS (Rodríguez-Camargo et al., 2020).

Otro estudio realizado en Bogotá, enfocado en la estimación de la mortalidad asociada a la contaminación del aire, reveló que el ozono se encontraba asociado con 152 defunciones por causas respiratorias. Aunque las concentraciones de este contaminante en la ciudad rara vez superan los estándares normativos a lo largo del año del estudio, los resultados sugieren un impacto significativo en la salud de la población. Es importante destacar que estos hallazgos subrayan la complejidad de los efectos de la contaminación atmosférica, incluso cuando los niveles de ozono se mantienen dentro de los límites reglamentarios, evidenciando la necesidad de medidas preventivas y de monitoreo

continuo para salvaguardar la salud pública en Bogotá, algunos de los datos significativos como estudio del valor económico por calidad de aire está dada por la Secretaria Distrital de Ambiente – SDA (2019) con las cifras observadas en la Tabla 1-1:

Tabla 1-1: Estimación de mortalidad según cada contaminante.

Contaminante	Mortalidad	Edad (años)	Defunciones estimadas	Costo (106 dólares)	Porcentaje del PIB Bogotá
PM2.5	Todas las causas	>30	1746	1002	1.1
PM_{10}	Infantil	<1	330	190	1.2
O_3	Todas las causas	>30	152	88	0.1

Fuente: SDA, 2021.

En el contexto de los estudios que emplean cálculos de riesgo poblacional a partir de contaminantes atmosféricos y valores de predicción mediante ML o simulaciones del contaminante en Colombia, se tienen los casos de Celis et al (2020) y Bustos et al. (2021) que han adoptado el enfoque de evaluación de riesgo de la EPA de la siguiente manera:

- Identificación y caracterización del riesgo: En esta etapa, se busca identificar el peligro, considerado como el máximo permitido para el contaminante. No obstante, para la identificación del riesgo diario, se prefiere utilizar el Índice de Calidad de Aire (ICA) en lugar del Índice Bogotano de Calidad de Aire (IBOCA), ya que el AQI es reconocido internacionalmente.
- Evaluación de exposición: Se procede al cálculo de la dosis de exposición teniendo en cuenta los efectos del contaminante, clasificando a la población en cinco subgrupos según la zona de estudio: Bebés (< 3 años), Niños (3-12 años), Jóvenes (12-30 años), Adultos (30-65 años) y Adultos Mayores (> 65 años), con categorización por género. Este enfoque tiene como objetivo mejorar la calidad de vida y reducir los riesgos para los grupos más propensos a enfermedades y más probablemente expuestos al contaminante.

Por su parte, en la investigación realizada por Ballesteros et al (2020)., se utilizaron los resultados de simulaciones del modelo de transporte químico para determinar la fracción de $PM_{2.5}$ y ozono atribuible a las emisiones de biomasa abierta (BB) en Colombia. La magnitud de los efectos en la salud se estimó mediante funciones de respuesta a concentraciones logarítmicas (CRF), vinculando el número de casos en exceso con un aumento en la concentración de un contaminante atmosférico específico. Según sus hallazgos, la mortalidad por todas las causas relacionada con $PM_{2.5}$ atribuible a emisiones de biomasa (BB) durante el mes de febrero en el país asciende a 52 casos en exceso, con un aumento proyectado de 12 casos en exceso para la mortalidad cardiovascular y 7 casos en exceso para la mortalidad respiratoria durante el mismo período.

1.1.6 Marco legal

Actualmente, Colombia cuenta con normatividad vigente para el sector ambiente que define políticas ambientales para promover el manejo, uso y aprovechamiento de recursos naturales, las cuales se muestran en la Tabla 1-2.

Tabla 1-2: Normatividad aplicable vigente para el sector ambiente.

Normas	Emisor	Descripción
Ley 99 de 1993	Congreso de Colombia	Por la cual se crea el Ministerio del Medio Ambiente, se reordena el Sector Público encargado de la gestión y conservación del medio ambiente y los recursos naturales renovables, se organiza el Sistema Nacional Ambiental, SINA y se dictan otras disposiciones.
Decreto Ley 2811 de 1974	Presidencia de la República de Colombia	Por el cual se dicta el Código Nacional de Recursos Naturales Renovables y de Protección al Medio Ambiente.
Decreto 1600 de 1994	Presidencia de la República de Colombia	Por el cual se reglamenta el Sistema Nacional Ambiental – SINA, en lo relacionado con los sistemas nacionales de investigación ambiental y de información ambiental; o la que la modifique o sustituya.
Decreto 948 de 1995	Presidencia de la República de Colombia	Por el cual se reglamentan, parcialmente la Ley 23 de 1973, los artículos 33, 73, 74, 75 y 75 del Decreto-Ley 2811 de 1974; los artículos 41, 42, 4 3, 44, 45, 48 y 49 de la Ley 9 de 1979; y la Ley 99 de 1993, en relación con la prevención y control de la contaminación atmosférica y la protección de la calidad del aire.
Decreto 2107 de 1995	Presidencia de la República de Colombia	Por medio del cual se modifica parcialmente el Decreto 948 de 1995 que contiene el reglamento de protección y control de la calidad del aire.
Decreto 1697 de 1997	Ministerio de Ambiente (MADS)	Por medio del cual se modifica parcialmente el Decreto 948 de 1995 que contiene el reglamento de protección y control de la calidad del aire.
Decreto 1552 de 2000	Presidencia de la República de Colombia	Por el cual se modifica el artículo 38 del Decreto 948 de 1995, modificado por el artículo 3o del Decreto 2107 de 1995.
Decreto 2622 de 2000	Ministerio de Ambiente, Vivienda y Desarrollo	Por medio del cual se modifica el artículo 40 del Decreto 948 de 1995, modificado por el artículo 2o. del Decreto 1697 de 1997.

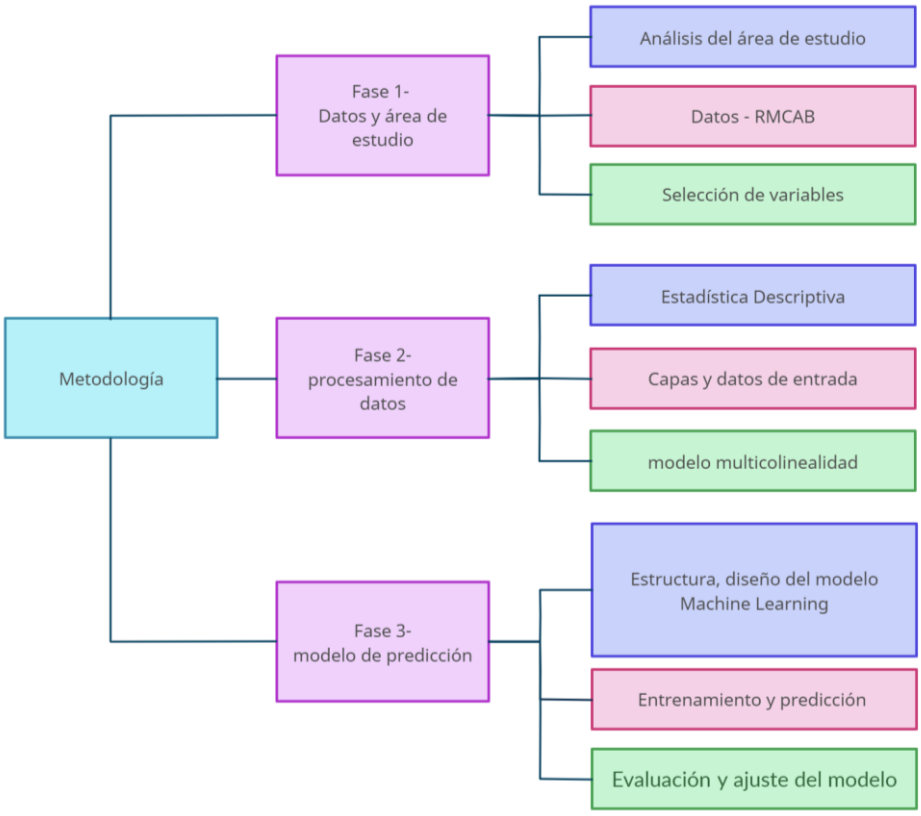
Normas	Emisor	Descripción
	Territorial (MADS)	
Decreto 979 de 2006	Ministerio de Ambiente, Vivienda y Desarrollo Territorial (MADS)	Por el cual se modifican los artículos 7, 10, 93, 94 y 108 del Decreto 948 de 1995; "Decreto por el cual se reglamentan, parcialmente la Ley 23 de 1973, los artículos 33, 73, 74, 75 y 75 del Decreto-Ley 2811 de 1974; los artículos 41, 42, 43, 44, 45, 48 y 49 de la Ley 9 de 1979; y la Ley 99 de 1993, en relación con la prevención y control de la contaminación atmosférica y la protección de la calidad del aire.
Decreto 2570 de 2006	Ministerio de Ambiente, Vivienda y Desarrollo Territorial (MADS)	Por el cual se adiciona el Decreto 1600 de 1994 y se dictan otras disposiciones; o la que la modifique o sustituya.
Decreto 1076 de 2015	Ministerio de Ambiente y Desarrollo Sostenible (MADS)	Por medio del cual se expide el Decreto Único Reglamentario del Sector Ambiente y Desarrollo Sostenible
Resolución 601 de 2006	Ministerio de Ambiente, Vivienda y Desarrollo Territorial (MADS)	Por el cual se dictan normas para simplificar, suprimir y reformar trámites, procesos y procedimientos innecesarios existentes en la administración pública.
Resolución 610 de 2010	Ministerio de Ambiente, Vivienda y Desarrollo Territorial (MADS)	Por la cual se establece la Norma de Calidad del Aire o Nivel de Inmisión, para todo el territorio nacional, en condiciones de referencia.
Resolución 650 de 2010	Ministerio de Ambiente, Vivienda y Desarrollo Territorial (MADS)	Por la cual se adopta el Protocolo para el Monitoreo y Seguimiento de la Calidad del Aire; o la que la modifique o sustituya.
Resolución 2154 de 2010	Ministerio de Ambiente, Vivienda y Desarrollo Territorial (MADS)	Por la cual se ajusta el Protocolo para el Monitoreo y Seguimiento de la Calidad del Aire adoptado a través de la Resolución 650 de 2010 y se adoptan otras disposiciones; o la que la modifique o sustituya.
Resolución 2254 de 2017	Ministerio de Ambiente y Desarrollo Sostenible (MADS)	Por la cual se adopta la norma de calidad del aire ambiente y se dictan otras disposiciones
Resolución 2410 de 2015	Secretaría Distrital de Ambiente - Secretaría Distrital de Salud	Por medio de la cual se establece el Índice Bogotano de Calidad del Aire –IBOCA– para la definición de niveles de prevención, alerta o emergencia por contaminación atmosférica en Bogotá D.C. y se toman otras determinaciones.

Fuente: Autor

2. Metodología

La metodología para la predicción de O_3 se realizó principalmente en 3 fases (ver Figura 2-1), iniciando por la recopilación de información de los datos de los contaminantes y variables a estudiar monitoreados por la RMCAB para el área de estudio durante los años 2013 al 2023.

Figura 2-1: Metodología general dividida por fases para el diseño del modelo predictivo de O_3



Fuente: Autor

En la segunda fase se aplicaron algoritmos de calidad de datos, como estadística descriptiva e información geoespacial para ordenar y clasificar los datos en conjuntos para su correcto procesamiento, finalmente en la tercera fase se usaron como variables de entrada para el modelo de ML en su entrenamiento y predicción y validando el modelo.

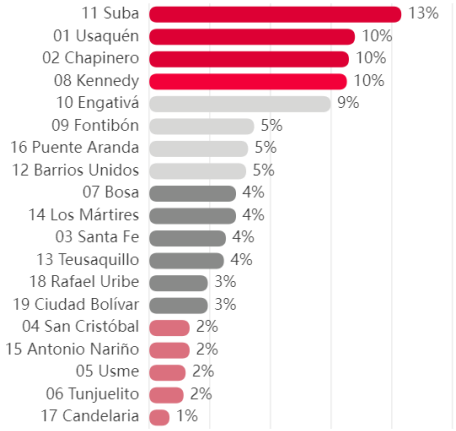
2.1 Área de estudio

Bogotá, capital de Colombia, está localizada en el centro del país, hace parte de la región Andina, se encuentra en una meseta a 2,600 msnm en la rama oriental de la Cordillera de los Andes y está bordeada por un sistema de cerros que llega a una altitud aproximada de 3,100 msnm. Además, la topografía extremadamente compleja de los Andes crea una variabilidad espacial particularmente alta en el comportamiento de la convección en toda región, con puntos críticos muy específicos que a menudo exhiben propiedades diversas (Hernández-Deckers et al., 2021). La ubicación estratégica de la ciudad hace que Bogotá no cuente con estaciones climáticas y tenga un clima homogéneo durante el año con temperatura media de 14°C (SDA, 2023). La ciudad tiene un régimen bimodal, característico del Nororiente de la Región Andina, con dos épocas secas y dos de altas precipitaciones, donde los meses más lluviosos son abril y mayo y octubre y noviembre (Guzmán et al., 2014).

Bogotá tiene una altitud promedio de 2,625 msnm, una longitud de 33 Km de norte a sur y de 16 Km de oriente a occidente, y está dividida en 20 localidades: Usaquén, Chapinero, Santa Fé, San Cristóbal, Usme, Tunjuelito, Bosa, Kennedy, Fontibón, Engativá, Suba, Barrios Unidos, Teusaquillo, Los Mártires, Antonio Nariño, Puente Aranda, La Candelaria, Rafael Uribe Uribe, Ciudad Bolívar y Sumapaz (Alcaldía Mayor de Bogotá D.C., 2021). Todas las localidades de Bogotá son urbanas, a excepción de Sumapaz, que es la única localidad completamente rural. Bogotá cuenta con más de 358,591 empresas registradas, de las cuales, aproximadamente 31% es comercial, 13% industrial, y 55% del sector servicios (Cámara de Comercio de Bogotá - CCB). Así mismo, el 51,5% de las empresas de Bogotá se localizan en Suba, Usaquén, Chapinero, Kennedy y Engativá, indicando una alta concentración empresarial en el norte y occidente de la ciudad, así como se observa en la Figura 2-2.

En cuanto al transporte, la ciudad cuenta con una malla vial urbana que está compuesta por la malla arterial, intermedia y local, con una extensión de 14,500 km (Instituto de Desarrollo Urbano - IDU). La movilidad de la ciudad principalmente se da en transporte público con 2,365 unidades troncales, 7,613 unidades zonales y 941 alimentadores (Transmilenio y SITP), 87,219 taxis, y 1.8 millones de vehículos privados, la mayor parte a gasolina, y una minoría a gas o híbridos, y 430.000 motos, aproximadamente, (Secretaría Distrital de Movilidad - SDM, 2023).

Figura 2-2: Distribución de empresas en Bogotá entre enero - junio de 2023 por localidad.



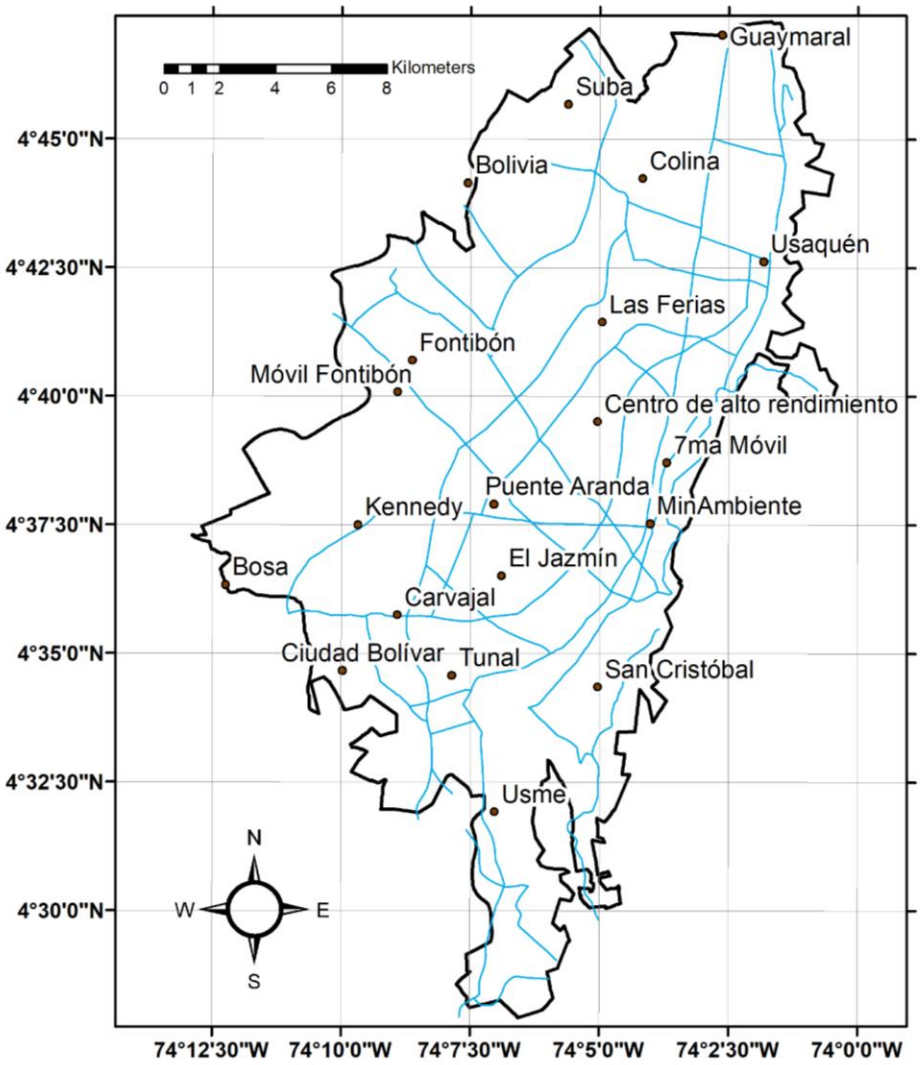
Fuente: Registro Mercantil, CCB, 2023.

La capital cuenta con la RMCAB, que hace parte de la Subdirección de Calidad del Aire, Auditiva y Visual de la Secretaría Distrital de Ambiente (SDA), como autoridad ambiental del Distrito Capital. La RMCAB inició en 1997 con el nombre DAMAIRE, con 12 estaciones distribuidas por la ciudad, y en 1999 se amplió a 2 estaciones ubicadas en Fontibón y Puente Aranda y eran operadas por IDEAM y empresas privadas. En 2005 se ampliaron a 15 estaciones y fueron operadas por la SDA, y desde el 2009 al 2015 se modifica la ubicación de las estaciones según necesidades de la ciudad. En la actualidad, la RMCAB cuenta con 20 estaciones de monitoreo, de las cuales 18 son fijas y 2 son móviles, y 18 miden ozono troposférico (SDA, 2023) como se observa en la Figura 2-3.

Se analizaron los datos de las variables influyentes en el comportamiento del O_3 en la última década. Así mismo, se observó la importancia en la calidad y representación de los datos, considerando que la RMCAB modernizó su sistema e instalaciones en el año

2008 y, sin embargo, contaba con tan solo 14 estaciones en 2013, sin monitorear localidades como es Usme, que fueron sumándose a través de los años.

Figura 2-3: Ubicación de las estaciones de RMCAB.



Fuente: Autor. Las líneas azules corresponden a las vías principales de transporte de la ciudad.

2.2 Fuente de datos

Para obtener la información medida por la RMCAB, se descargaron los datos directamente de la página de la Secretaría Distrital de Ambiente, donde se encuentran por fechas y variables a analizar de cada estación del Distrito Capital. Se obtuvo

información de las concentraciones registradas de los contaminantes precursores del ozono troposférico O_3, NO, NO_2, $PM_{2.5}$, PM_{10}, y los registros meteorológicos como radiación solar, temperatura, dirección y rapidez del viento, para el periodo de estudio.

Tabla 2-1: Contaminantes y variables meteorológicas medidas por cada estación de la ciudad

Estación	Contaminantes						Variables meteorológicas			
	O_3	NO	NO_2	PM_{10}	$PM_{2.5}$	CO	Vel. viento	T	Rad. solar	HR
Bolivia	✓	✓	✓	✓	✓	✓				
Bosa		✓	✓	✓	✓	✓		✓	✓	
Carvajal	✓	✓	✓	✓	✓	✓	✓	✓		
Centro de Alto Rendimiento (CDAR)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Colina	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Ciudad Bolívar	✓	✓	✓	✓	✓	✓	✓	✓		✓
Fontibón	✓	✓	✓	✓	✓	✓	✓	✓		✓
Guaymaral	✓	✓	✓	✓	✓	✓		✓	✓	✓
Jazmín	✓	✓	✓	✓	✓	✓	✓	✓	✓	
Kennedy	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Las Ferias	✓	✓	✓	✓	✓	✓	✓	✓		✓
Min Ambiente	✓			✓	✓	✓	✓			
Móvil Fontibón	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Puente Aranda	✓	✓	✓	✓	✓	✓		✓		
San Cristóbal	✓	✓	✓	✓	✓		✓	✓	✓	✓
Suba	✓	✓	✓	✓	✓	✓	✓	✓		
Tunal	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Usaquén	✓	✓	✓	✓	✓	✓	✓	✓		
Usme	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
7ma Móvil		✓		✓	✓	✓	✓	✓	✓	✓

Fuente: Autor.

Durante la recopilación inicial de información, se observó la ausencia de datos y su baja calidad en algunas estaciones, debido al largo periodo de observación (del 2013 al 2022) y los cambios a través del tiempo, donde en los primeros años la mayoría de las estaciones se limitaban al monitoreo del material particulado. Así mismo, se encontraron periodos vacíos para algunos años en particular, teniendo en cuenta que algunas estaciones como Carvajal fueron vandalizadas en el año 2021. Las estaciones de Colina y Móvil Fontibón fueron instaladas y puestas en marcha desde en el año 2021 (SDA, 2023), por lo que estas estaciones no cuentan con datos en años anteriores. Los datos que se encuentran por estación se pueden observar en la Tabla 2-1.

Además, es importante tener en cuenta que las estaciones de Monitoreo Ambiental del Ministerio de Ambiente y Desarrollo Sostenible (Minambiente) ubicadas en Puente Aranda y Fontibón, exceden la altura de 10 metros estipulada por el Protocolo para el Monitoreo y Seguimiento de la Calidad del Aire, adoptado mediante la Resolución 650 de 2010 y modificado por la Resolución 2154 de 2010.

Para abordar los datos faltantes, se aplicó un procedimiento de llenado utilizando el promedio multianual horario de los valores diarios máximos en un promedio móvil de 8 horas para el contaminante de interés. Respecto a los demás contaminantes, se llevó a cabo la imputación utilizando el promedio mensual, siguiendo la metodología propuesta por Celis et al. (2022).

Adicionalmente, se calcularon los componentes del viento (u y v) multiplicando la velocidad del viento por el seno de la dirección del viento para el componente u y por el coseno para el componente v.

2.3 Estadística descriptiva

Para el análisis espacial de los datos, se empleó la herramienta de mapas de calor en el software ArcGIS, junto con la técnica de interpolación kriging para generar mapas que representaran los valores promedio móvil de 8 horas del ozono durante los años 2019 al 2022 en distintas localidades de Bogotá. Este enfoque permitió visualizar de manera efectiva los cambios en las concentraciones de ozono a lo largo del tiempo en el área de estudio.

Los datos de superficie fueron tratados estadísticamente por medio del criterio de Chauvenet y el método de 3σ (Método de 3 desviaciones estándares) (Russo et al., 2018). Si bien este método es utilizado para distribuciones normales, es útil para dar una idea de los datos anómalos, que luego van a ser comparados con las estaciones cercanas, para definir si son datos reales o producidos por un error en el sensor. Luego de esto se tiene en cuenta que los datos pertenecientes a las estaciones automáticas deben tener al menos el 75% de cobertura de datos de control para el periodo analizado (USEPA, 2017). En el caso de los datos meteorológicos, también se realizará un filtro que elimine los datos que no pueden ser físicamente posibles (e.g., radiación negativa).

De esta manera, para el procesamiento de los datos con estadística descriptiva, se inició graficando la distribución y comportamiento de los datos a partir de diagramas de caja, con el fin de visualizar las medidas de posición y demás precursores del O_3. De igual forma, se calcularon los cuartiles para dividir la muestra de datos según la dispersión de las variables. Se procedió a visualizar con facilidad en las gráficas tipo histogramas donde se contó con la desviación estándar de los datos y la media. También, verificó que tipo de asimetría tiene, si es simétrica, o asimétrica negativa o positiva.

Por otro lado, se evalúo la curtosis, la cual indica que tan achatada o escarpada se encuentra la curva de distribución, teniendo en cuenta los criterios de Fernández et al. (2002):

- Leptocúrtica: la curtosis >0. Donde los datos se encuentran muy concentrados, siendo una curva muy apuntada
- Mesocúrtica: la curtosis = 0. Donde se presenta una distribución normal
- Platicúrtica: la curtosis < 0. Donde se presenta poca concentración de datos en la media, y la curva es achatada.

Se calcularon los promedios multianuales mensuales de los contaminantes NO y NO_2 de las principales estaciones de la ciudad para observar el comportamiento entre contaminantes y la estacionalidad. Para el caso de O_3 se calcularon los promedios diarios del máximo de 8 horas.

Así mismo, se calculó el promedio mensual utilizando el procedimiento de descomposición de tendencia estacional observado en Singh et al. (2021) y adoptado

para estudiar el comportamiento de $PM_{2.5}$ por Bigi & Ghermandi (2016), llamado Seasonal Trend (STL), el cual descompone series de tiempo en tres componentes: tendencia, estacional y residual, con el fin de suavizar la heterocedasticidad de los datos y normalizar la distribución. Posteriormente, se calculó la pendiente significativa en el componente de tendencia mensual utilizando la regresión de mínimos cuadrados generalizados (GLS, por sus siglas en inglés Generalized Least Squares) (Brockwell & Davis, 2002; Singh et al. 2021) en un intervalo de confianza del 95% (alfa = 0,05).

Finalmente, con el fin de identificar las variables meteorológicas y de contaminación del aire relacionadas con el ozono troposférico, se graficó la correlación estadística entre los datos de entrada, las tendencias y distribución de los diferentes contaminantes y variables meteorológicas seleccionadas a través de los años, analizando su comportamiento individual y la posible influencia en la predicción del modelo.

2.4 Estructura y diseño del modelo de Machine learning

2.4.1 Datos de entrada y selección de variables

Para comprender los factores que influyen en el comportamiento del ozono, se seleccionaron datos de entrada basados en diversos estudios que resaltan la contribución de las emisiones de ciertos precursores (contaminantes atmosféricos) en áreas urbanas, variables meteorológicas, que afectan el transporte de contaminantes entre regiones (Liu et al., 2021), y la química atmosférica y procesos biogénicos generados por ecosistemas terrestres (Wu et al., 2020), como factores que nos permiten prever cambios en la formación del ozono.

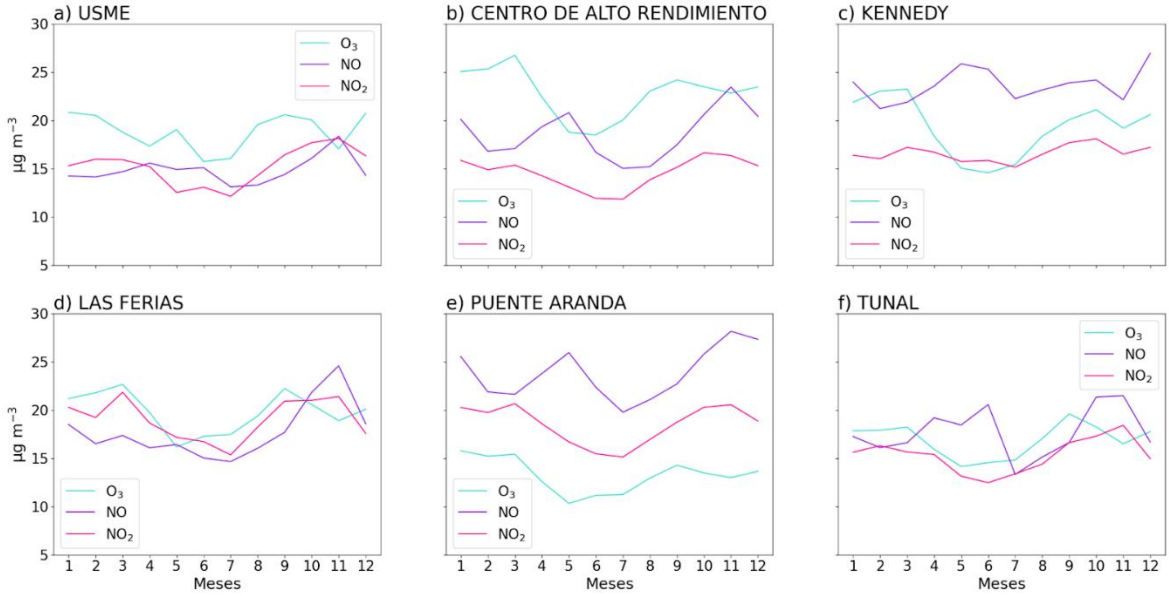
En este contexto, fueron seleccionadas 10 variables: O_3, CO, NO, NO_2, $PM_{2.5}$, PM_{10}, radiación solar, temperatura, dirección y velocidad de viento. Los cuales fueron tomados como datos de entrada debido a su importancia e influencia en la producción del Ozono troposférico.

Por su parte, CO y NO_x fueron elegidos como precursores del ozono debido a sus reacciones fotoquímicas, aunque su relación no es lineal, es esencial estudiar repercusiones como el NO, que puede consumir el ozono debido a la disminución de la

titulación de O_3 y los radicales (Liu, Wang 2020). Esto se evidencia en la gráfica de promedios multianuales mensuales de los contaminantes O_3, del NO y NO_2 de las principales estaciones de la ciudad para observar su comportamiento, la correlación y la estacionalidad, como se observa en la Figura 2-4.

En la Figura 2-4 se observa que las estaciones que registraron menores concentraciones de los contaminantes fueron Usme y Tunal, que no superaron 22.5 µg/m^3, contrario a la estación Puente Aranda que fue la que registró mayor concentración principalmente en NO, acercándose a concentraciones de 30 µg/m^3. Por lo tanto, se puede evidenciar la fuerte correlación entre O_3 y NO_2, y una relación inversa con NO en la mayoría de los meses, teniendo en cuenta que cuando hay bajas concentraciones de NO, se reduce la titulación del O_3 en condiciones anormales de temperatura superficial (Sokhi et al., 2021, Casallas et al. 2024).

Figura 2-4: Comportamiento mensual de O_3, NO y NO_2



Fuente: Autor.

Por otro lado, las partículas finas $PM_{2.5}$ fueron incluidas por indicar contaminación transportada y aerosoles orgánicos, derivados en gran medida de emisiones biogénicas de compuestos orgánicos volátiles (VOC), aunque su impacto es menor que las emisiones antropogénicas, ya que pueden afectar la formación de O_3 (Wu et al., 2020). En cuanto al PM_{10}, generado por la quema de biomasa, influye directamente en la

formación de ozono, reflejando la radiación solar y dilatando la reacción fotoquímica, por lo que, en ausencia de PM, la formación de O_3 aumenta (Casallas et al., 2024).

Asimismo, se analizaron las variables meteorológicas a considerar en la formación de O_3 tales como la humedad y temperatura atmosférica y radiación solar, precipitación, dirección y rapidez del viento, e intrusión de masas de aire estratosférico como la radiación solar (Monks et al., 2015). Los patrones meteorológicos coinciden con las anomalías que se presentan en los niveles de O_3, así como la correlación que hay entre estos, por ejemplo, mayor radiación, menor humedad, mayor presión, conduciendo a un aumento de O_3 (Deroubaix et al., 2021).

La correlación Spearman entre el O_3 y sus precursores y variables muestra relaciones importantes. Se destaca una correlación fuerte y positiva con la temperatura y la velocidad del viento, mientras que se observa una correlación fuerte pero negativa con las variables de NO, NO_2 y humedad relativa (ver Figura 2 5). Estas relaciones sugieren que el aumento de temperatura superficial y la baja humedad relativa pueden contribuir al incremento de O_3, mientras que la disminución de las concentraciones de NOx también puede influir, como se demostró en un estudio realizado por Sokhi et al. (2021).

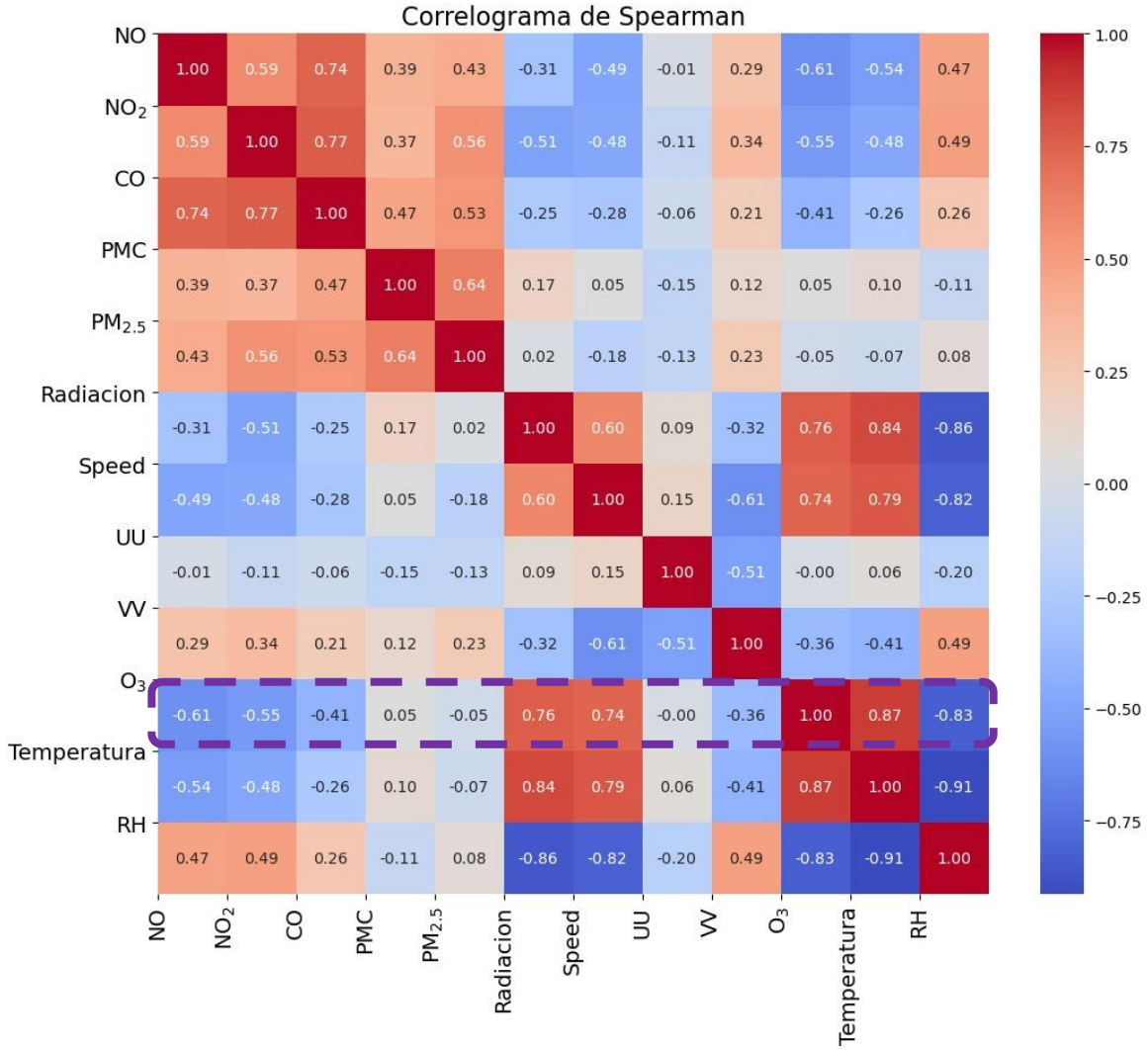
En contraste, se observa que las concentraciones de O3 tienden a reducirse en ambientes con baja radiación solar, alta humedad relativa y vientos débiles, lo que limita la advección del O_3 regional. En el caso específico de Bogotá, donde la velocidad del viento es baja y la presencia de nubes es alta, la radiación solar no ejerce una influencia significativa en la formación de O_3, como se menciona en el estudio de Casallas et al. (2023).

Además, se observa una correlación negativa entre el material particulado (PMC y $PM_{2.5}$) y el monóxido de carbono (CO), indicando que estas variables no tienen una contribución significativa a la formación positiva de O_3, lo cual coincide con los hallazgos del estudio de Casallas et al. (2023).

Por el contrario, en dicho estudio, se observó cómo se reducían las concentraciones de O_3 al encontrarse en un ambiente bajo en radiación solar, alta humedad relativa y débiles vientos, que limitaron la advección del O_3 regional. En el caso de Bogotá, la rapidez del

viento es baja, y no es suficiente para diluir y reducir las concentraciones de O_3, así como también, es una ciudad con una fracción alta de nubes, la radiación no tiene mayor influencia en la formación del O_3 (Casallas et al., 2023). A su vez, la correlación que hay entre el material particulado (PMC y $PM_{2.5}$) y CO son negativas y no tienen significancia para la contribución positiva de la formación del O_3, igualmente observado en el estudio de Casallas et al. (2023).

Figura 2-5: Correlación Spearman entre los precursores del O_3



Fuente: Autor

Una vez seleccionadas las variables mediante revisión bibliográfica, se llevó a cabo un análisis de estadística descriptiva utilizando el Correlograma de Spearman (ver Figura

2-5) para comprender la relación entre las variables reales y el ozono observado. Posteriormente, se aplicó el método del Factor de Inflación de la Varianza (VIF) para identificar y eliminar variables que contribuyeran significativamente a la multicolinealidad, la cual puede generar errores en el modelo (Ghahremanloo et al., 2021).

El cálculo del VIF se realizó utilizando la librería de código abierto Pandas en el lenguaje Python, dividiendo 1 entre la diferencia de 1 y el coeficiente de determinación al cuadrado de la regresión lineal de las variables al compararlas. Tanto el VIF como el correlograma se calcularon utilizando datos de todos los contaminantes obtenidos de la RMCAB desde el año 2013 hasta el 2022.

Tabla 2-2: Variance Inflation Factor (VIF) para las variables de entrada del modelo de predicción de O_3.

Variable	VIF
NO	3.12
NO_2	4.48
CO	4.54
PMC	2.08
$PM_{2.5}$	2.30
Radiación	2.24
Rapidez de viento	3.06
Dirección del viento	2.29
UU	1.48
VV	2.31
Humedad Relativa	6.75
Temperatura superficial	9.53

Fuente: Autor

Teniendo para el modelo los resultados de la Tabla 2-2, mostrando que los valores más altos fueron para las variables NO_2, CO, NO, velocidad, radiación, Material particulado, temperatura y humedad relativa. Esta expresión sugiere que cuanto mayor sea el valor

VIF, más problemática será la variable en cuanto a permitir que el método de ML se entrene correctamente (Curto & Pinto, 2011).

Por otro lado, la temperatura superficial tiene una alta correlación en la formación de O_3 como se observa en la Figura 2-5, se tuvo en cuenta como entrada dentro del modelo, aun cuando, tiene valores de multicolinealidad demasiado altas, lo que generaría mayor sesgo entre la predicción y lo observado.

2.4.2 Matriz de entrada y salida

Una vez obtenidos los resultados del VIF se seleccionaron en total 9 variables, como entradas al modelo, con un total de 70128 datos para el conjunto de entrenamiento. Para crear la matriz de entrada X por cada estación con las variables de NO, NO_2, CO, PM_{10}, $PM_{2.5}$, radiación, rapidez del viento, humedad relativa, temperatura y los valores calculados de los componentes del viento. Por otro lado, el vector Y de salida se realizó a partir del promedio horario de las observaciones del O_3 en cada estación. Así mismo se crean los vectores de entrenamiento y predicción, contando con el 85% del total de datos para el entrenamiento: 59,584 muestras, mientras que el 15% de los datos, se usaron para validación.

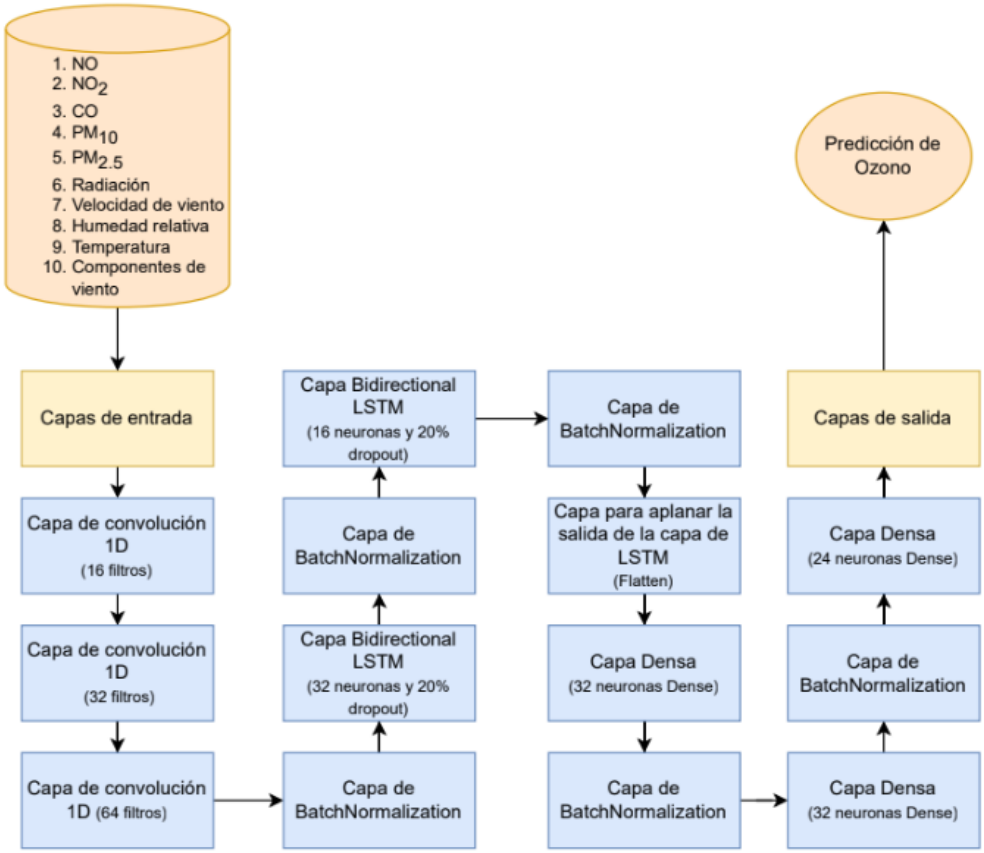
2.4.3 Estructura y aplicación del modelo de Machine Learning

Para realizar la predicción, se generó una red basada en el modelo ML propuesto por Celis et al (2022), modificando la estructura para adaptarla a las particularidades del contaminante O_3 (ver Figura 2-6).

Así las técnicas de ML utilizadas comprendieron las redes convolucionales y las capas bidireccionales de LSTM de las librerías de Python Tensor Flow (Abadi et al. 2015) y Keras (Chollet, 2015), y el paquete Sklearn de Python. Este modelo propuesto por Celis et al. (2022) fue seleccionado después de haber evaluado diversas alternativas. Entre los modelos probados, este destacó al proporcionar pronósticos de alta precisión, teniendo en cuenta y cuantificando la contribución de contaminantes atmosféricos y variables meteorológicas en la formación del ozono. Sin embargo, la elección se fundamentó en la realización de ajustes y modificaciones en el número de neuronas, activaciones,

funciones de pérdida y otros parámetros mediante una exhaustiva búsqueda en cuadrícula (Grid-search). Dando como resultado un modelo que destaca por su capacidad de ofrecer pronósticos altamente precisos al considerar y cuantificar la influencia de precursores.

Figura 2-6: Estructura del modelo ML para predicción de ozono

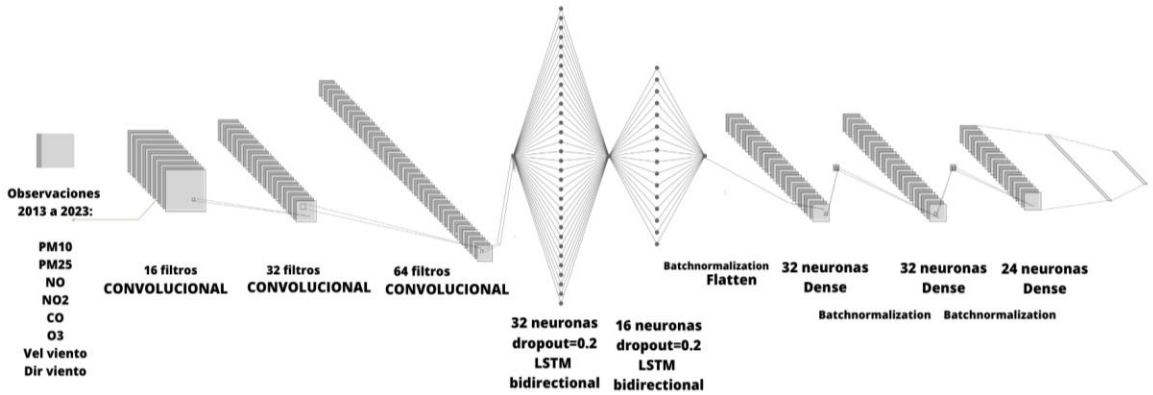


Fuente: Autor

Como se puede apreciar en la Figura 2-7 que muestra la arquitectura del modelo, se emplearon tres capas convolucionales 1D con 16, 32 y 64 filtros, respectivamente. Se realizaron ajustes del modelo en algunos parámetros, utilizando un tamaño de kernel de 3 y una función de activación "relu" (Fukushima, 1975) para introducir no linealidad en la secuencia de datos en cada una de las tres capas convolucionales. Posteriormente, se implementó una capa de batch normalization con el propósito de mantener los valores de activación en un rango óptimo durante el proceso de entrenamiento.

A continuación, se añadieron dos capas bidireccionales de LSTM: la primera con 32 neuronas y la segunda con 16, ambas incorporando un dropout del 20% para prevenir el sobreajuste en la red neuronal. Finalmente, se incluyeron tres capas Dense con 32, 32 y 24 neuronas, respectivamente, junto con su correspondiente función de activación "relu".

Figura 2-7: Arquitectura del modelo ML para predicción de ozono



Fuente: Autor

Para la compilación del modelo, se optó por usar el optimizador "Adam" (Kingma & Ba, 2014), ampliamente utilizado y altamente recomendado en el entrenamiento de redes neuronales. Se seleccionó una pérdida de tipo Maximum Likelihood Sequence Estimation (MLSE). Este enfoque ha demostrado resultados precisos en la predicción O_3, permitiendo alcanzar valores de hasta 32 µg/m³, lo cual coincide con el comportamiento y los valores máximos observados de O_3. Una de las diferencias más significativas en comparación con otras funciones de pérdida observadas radica en la capacidad del modelo para predecir y suavizar los picos pequeños a lo largo del tiempo.

Es importante destacar que la selección de estos optimizadores y parámetros se realizó después de probar diversas configuraciones, incluyendo el análisis de distintas funciones de pérdida como Mean Square Error (MSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE) y MLSE, siendo esta última la opción preferida debido a su mejor desempeño.

Para el entrenamiento del modelo, se utilizaron 25 épocas y un batch size de 64. Para el ajuste de la red, se empleó una época y un batch size de 10. Estos parámetros se configuraron utilizando el método fit, con un promedio de 64 segundos por época.

2.5 Validación y evaluación del modelo

Después de obtener el modelo de predicción junto con las rutinas de algoritmos de análisis estadístico, se procede a su implementación con datos reales y se lleva a cabo la validación para evaluar su aplicación y tasa de precisión. La validación se realiza siguiendo el enfoque propuesto por Celis et al. (2022), quienes validaron su modelo durante 182 días distribuidos según las categorías del Índice de Calidad del Aire (AQI) para asegurar que el modelo sea capaz de reproducir diferentes escenarios de contaminación.

La validación se realiza utilizando 9 parámetros estadísticos y 3 categóricos, los cuales se describen detalladamente en el material suplementario de Celis et al. (2022). Estos parámetros incluyen el coeficiente de correlación (R2) o el coeficiente de correlación de Spearman (Rho), el índice de concordancia (IOA) y el error cuadrático medio (RMSE). Se utilizan los criterios de evaluación propuestos por Mogollon-Sotelo et al. (2021) para clasificar la calidad de la predicción, como se describe en la Tabla 2-3 del documento mencionado.

Este enfoque de validación integral garantiza una evaluación completa del rendimiento del modelo en diferentes condiciones y escenarios de contaminación, proporcionando una base sólida para su aplicación en futuras predicciones de calidad del aire.

Tabla 2-3: Clasificación de parámetros para la validación del modelo.

Errores	Excelente	Bueno	Deficiente
Rho	≥0.9	(0-9-0.5)	<0.5
RMSE	<10	(10-15)	>15
IOA	(1-0.7)	(0.7-0.3)	(0.3-0)

Fuente: Mogollon-Sotelo et al. (2021)

Los demás parámetros fueron: False Alarm Rate (FAR) es la fracción de eventos predichos por encima del umbral que son incorrectos; Proporción de corrección o Proportion of Correctness (POC), siendo la fracción de predicciones que coinciden con el umbral superior/inferior con las observaciones; la tasa de aciertos o Hit Rate (HIT) que mide la fracción de eventos observados por encima del umbral que se predice correctamente (Chai et al., 2013); Normalized Mean Bias (NMB) y Mean Bias (MB); Factor of Two (FAC2) que representa la relación entre el error absoluto y un factor de dos veces la desviación estándar del valor observado. Los valores de referencia para dar aceptación al modelo se adoptaron siguiendo estudios previos (Chai et al., 2013; WHO, 2021) (ver Tabla 2-4).

Tabla 2-4: Valores de referencia de los parámetros de validación FAR, POC, HIT, NMB, MB y FAC2.

Parámetros	**FAR**	**POC**	**HIT**	**NMB**	**MB**	**FAC2**
Valores de referencia	< 40%	< 50%	> 60%	< 15%	< 0.5 µg/m^3	> 35%

Fuente: Chai et al. (2013); WHO (2021)

Como umbral se seleccionó el valor estándar de acuerdo con el tiempo de exposición de la ley colombiana, siendo 100 ppb el límite máximo en aire ambiente con un estándar de tiempo de 8 horas diarias en lugar de 24 horas (ver Tabla 2-5), debido a la fuerte variación diurna de las concentraciones del O_3.

Tabla 2-5: Niveles límites permisibles de O_3 por tiempo de exposición

Referencia	**Límite permitido por tiempo de exposición**	
	1 hora ppb	**8 horas ppb**
OMS	-	100
EPA	120	80
EU	90	60
Colombia Resolución 2254 de 2017	-	50

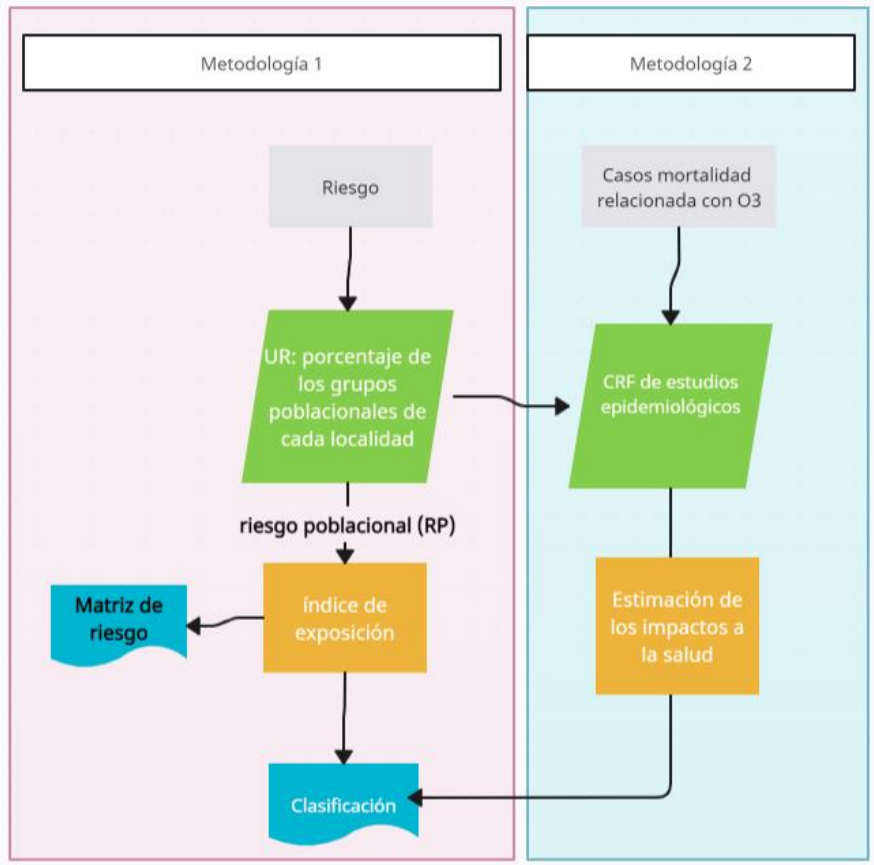
Fuente: Autor

2.6 Cálculo del riesgo poblacional y afectación a la salud

Con el fin de calcular el riesgo poblacional y su contribución a la salud humana, se diseñó la metodología que se puede observar en la

Figura **2-8:**, que combina la metodología utilizada por Ballesteros et al (2020) y la utilizada por Bustos et al. (2021), con el fin de adaptarlo al estudio del proyecto y robustecer el cálculo.

Figura 2-8: Metodología adaptada para el cálculo de riesgo poblacional y afectación a la salud.



Fuente: Autor

Con el fin de calcular el riesgo en la población de referencia derivado de las concentraciones de ozono de la predicción del modelo, se remite a la metodología propuesta por Bustos et al (2021) para la ciudad de Bogotá, donde el riesgo poblacional no cancerígeno, se encuentra en función de C (concentración) y de UR (porcentaje de cada grupo poblacional), para el porcentaje de los grupos poblacionales de cada

localidad donde se encuentra ubicada la red de monitoreo se utilizan los datos de la Secretaría Distrital de Planeación (SDP) con la colaboración del Departamento Administrativo Nacional de Estadística Año 2021 (Dane), dependiendo de la categoría de edad y género (ver Tabla 2-6).

Dentro de esta misma metodología, se propone la evaluación del índice exposición para cada una de las localidades representado con la Ecuación 2-1.

Ecuación 2-1: $$IE = \frac{CX}{NMPX}$$

IE: Índice de exposición
CX: Concentración del xenobiótico por estación de monitoreo (µg/m³)
NMPX: Nivel máximo permisible del xenobiótico según el indicador (µg/m³)

Tabla 2-6: Porcentaje de la población de cada una de las localidades por grupo de referencia.

Localidad	Estaciones de monitoreo	Grupo poblacional	Edad	Femenino (%)	Masculino (%)
Kennedy	Carvajal	Bebés	<3	2.8	2.9
		Niños	(3-12]	9.1	9.5
		Personas Jóvenes	(12-30]	20.2	20.7
		Adultos	(30-65]	26.5	23.5
		Adultos Mayores	>65	5.1	3.6
Barrios Unidos	Centro de alto rendimiento	Bebés	<3	2.5	2.7
		Niños	(3-12]	6.8	7.4
		Personas Jóvenes	(12-30]	18.5	19.4
		Adultos	(30-65]	27.8	26.0
		Adultos Mayores	>65	7.4	5.1
Fontibón	Fontibón	Bebés	<3	2.6	2.7
		Niños	(3-12]	8.1	8.4
		Personas Jóvenes	(12-30]	20.1	19.4

Localidad	Estaciones de monitoreo	Grupo poblacional	Edad	Femenino (%)	Masculino (%)
		Adultos	(30-65]	28.5	24.0
		Adultos Mayores	>65	5.7	3.8
Kennedy	Kennedy	Bebés	<3	2.8	2.9
		Niños	(3-12]	9.1	9.5
		Personas Jóvenes	(12-30]	20.2	20.7
		Adultos	(30-65]	26.5	23.5
		Adultos Mayores	>65	5.1	3.6
Puente Aranda	Puente Aranda	Bebés	<3	2.4	2.6
		Niños	(3-12]	7.5	8.0
		Personas Jóvenes	(12-30]	18.4	20.2
		Adultos	(30-65]	27.0	25.3
		Adultos Mayores	>65	6.9	4.6
Suba	Suba	Bebés	<3	2.8	3.0
		Niños	(3-12]	8.6	9.0
		Personas Jóvenes	(12-30]	19.8	19.5
		Adultos	(30-65]	27.9	23.3
		Adultos Mayores	>65	5.8	3.9
Tunjuelito	Tunal	Bebés	<3	3.6	3.8
		Niños	(3-12]	9.7	10.0
		Personas Jóvenes	(12-30]	19.9	21.2
		Adultos	(30-65]	25.0	23.5
		Adultos Mayores	>65	5.2	3.8
Usaquén	Usaquén	Bebés	<3	2.6	2.7
		Niños	(3-12]	7.7	7.8
		Personas Jóvenes	(12-30]	18.6	17.5
		Adultos	(30-65]	29.5	23.8
		Adultos Mayores	>65	7.9	5.3

Fuente: (SDP, 2021)

Con el resultado obtenido de la ecuación se realizará la debida clasificación y normalización de los datos en la matriz de riesgo. Estos se hacen con el fin de identificar el grado de riesgo que se presenta en cada una de las estaciones de monitoreo y así mismo proponer las posibles soluciones y actores que se pueden asociar a estas.

Para determinar los niveles máximos permisibles, se han empleado los valores proporcionados por la (OMS), según lo detallado en la Tabla 2-6 y las concentraciones estimadas para $PM_{2.5}$ y O_3 se derivan de nuestro modelo.

Asimismo, se llevó a cabo la metodología 2 propuesta por Ballesteros et al (2020) en la que destacaron las posibles afectaciones a la salud relacionadas con los resultados de la simulación de ozono. En este contexto, la extensión de los efectos sobre la salud se calculó utilizando funciones de respuesta a la concentración (CRF) logarítmicas lineales. Estas funciones relacionan el número de casos excedentes asociados con el criterio de valoración de salud específico para las afectaciones de ozono y el excedente en la concentración del contaminante (ver Ecuación 2-2).

Ecuación 2-2: $$\Delta y = y_0(1 - e^{-\beta \Delta x})$$

Donde, *y* representa el número de referencia de casos para el criterio de valoración sanitario, *x* indica el exceso en la concentración del contaminante con respecto al límite permisible de la OMS (2021) para el contaminante correspondiente, y β es el parámetro vinculado al riesgo relativo El parámetro β se analizará con base en los datos sobre afectaciones y defunciones por todas las causas, mortalidad cardiovascular y respiratoria, según lo informado en el análisis reciente (Liu et al., 2019).

La incidencia inicial de otros efectos sobre la salud se obtuvo del informe anual del Departamento de Estadística de Colombia (DANE, 2018) y se asumió que se tiene la misma toxicidad que las partículas $PM_{2.5}$ provenientes de otras fuentes.

3. Resultados y discusión

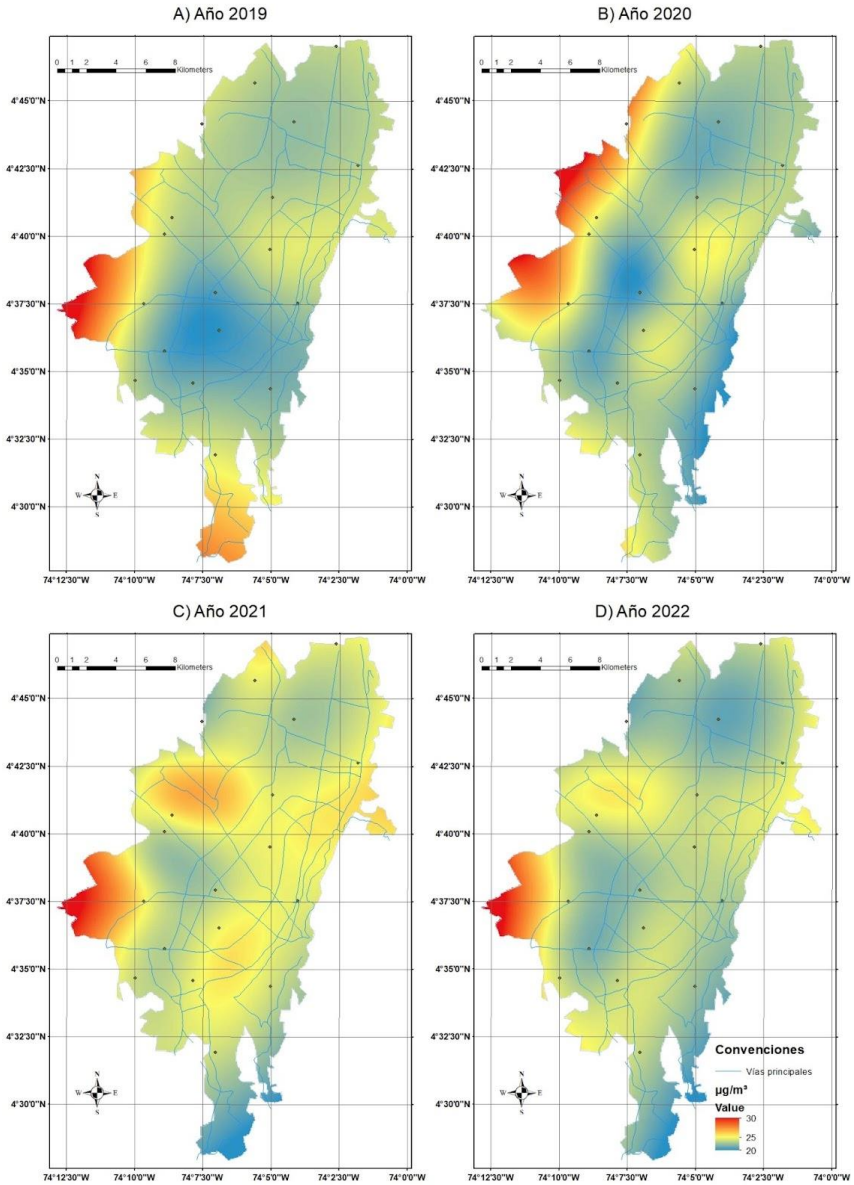
3.1 Estadística Descriptiva

3.1.1 Análisis espacial

Se realizó el análisis espacial teniendo en cuenta la comparación del comportamiento de las concentraciones de O_3 del año 2019 a 2022, usando mapas de calor para evidenciar las variaciones y distribución del O_3 entre las localidades en toda Bogotá, como se muestra en la Figura 3-1.

Se puede observar un comportamiento de los niveles de O_3 cercanos a 25 µg/m^3 en el año 2019, sin embargo, en el occidente las zonas aledañas a las estaciones Kennedy, Bosa y Usme se puede observar algunos niveles que superan los 25 µg/m^3. Las localidades ubicadas en el norte se encuentran con niveles cercanos a 25 µg/m^3, destacando la zona aledaña a la estación CDAR. En el occidente se destaca al tener mayores concentraciones de O_3, esto debido a la cantidad de industrias que existen en la zona y la movilidad de vehículos a diesel (Rojas, 2004; Castillo-Camacho et al., 2020) que emiten NO_x que influye en la producción de ozono, sumándose al transporte de contaminantes que hay en la ciudad, ya que el comportamiento del viento predomina en dirección sur a norte y del oriente a lo largo de los cerros orientales (SDA, 2023).

Figura 3-1: Promedio multianual de O_3 en Bogotá, calculado a partir de datos diarios del máximo promedio de 8 horas.



Fuente: Autor

En el año 2020, se puede observar altas concentraciones que superan 30 µg/m^3 en todo el occidente donde están ubicadas las estaciones Kennedy, Móvil Fontibón, Fontibón y Bolivia. Así mismo, se denota niveles de O_3 cercanos a 25 µg/m^3 en una franja en el oriente de norte a sur donde están ubicadas a las estaciones Guaymaral, Usaquén, Las Ferias, 7ma móvil, Jazmín, San Cristóbal, Tunal y Usme, destacando la zona aledaña a la estación CDAR. Se evidenció un pico en las concentraciones de O_3 principalmente en

los meses de junio y julio, debido a las restricciones de movilidad. Durante estos meses el ozono presentó concentraciones un 70% más altas con respecto al promedio de los años anteriores, principalmente a causa de una disminución en el NO, que disminuye la titulación y por ende no consume el ozono, permitiendo que este aumente (Sokhi et al., 2021; Casallas et al., 2023). Adicionalmente, durante el inicio de la cuarentena se presentaron incendios forestales al norte del país, los cuales contribuyeron con el incremento del contaminante (Ballesteros et al. 2020; Méndez-Espinosa et al., 2020; Casallas et al., 2023,2024). Asimismo, Bogotá presentó una anomalía positiva de radiación, la cual también contribuyó (aunque en menor medida) con el incremento del ozono durante la cuarentena y por ende en general, durante el año (Sokhi et al., 2021).

Para el año 2021, el comportamiento del O_3 se encontró cercano a 30 µg/m^3 en las zonas periféricas a las estaciones de Kennedy, Fontibón y Suba. La región oriental de norte a sur donde se encuentran las estaciones Usaquén, Las Ferias, CDAR, 7ma Móvil, Jazmín, San Cristóbal y Tunal alcanzó niveles superiores a 25 µg/m^3. Así mismo, la localidad de Usme presenta valores cercanos a 20 µg/m^3. Esto pudo deberse a la tendencia positiva que han presentado el NO y NO_2 en los últimos años (Casallas et al. 2024), en donde el crecimiento del NO_2 ha sido más acelerado que el del NO, lo que conlleva a que la relación NO/NO_2 favorezca al aumento de O_3, que se derivó de la reactivación económica tras la pandemia de COVID-19, por el con el ascendente consumo de gasolina de baja calidad impulsado, posiblemente por la inflación de precios del diésel y el aumento de uso de transporte aéreo (Casallas et al., 2024).

En cuanto al año 2022, se encuentra la zona aledaña a la estación Kennedy con valores cercanos a los 30 µg/m^3, y a la estación Fontibón con valores superiores a los 25 µg/m^3. De nuevo se ve el comportamiento en la franja oriental de norte a sur, se destaca niveles de O_3 por debajo de 25 µg/m^3 donde están ubicadas a las estaciones Guaymaral, Usaquén, Las Ferias, CDAR, 7ma móvil, Jazmín y Tunal. Esto puede deberse a que los niveles de NO y NO_2 siguen altos, aun cuando disminuyeron en comparación al año 2021, la relación NO/NO_2 continúa favoreciendo a los niveles de O_3 (Casallas et al., 2024).

Por lo anterior, se puede concluir que las fluctuaciones en las concentraciones de O_3 pueden ser afectadas por algunas situaciones adversas como se evidenció en las

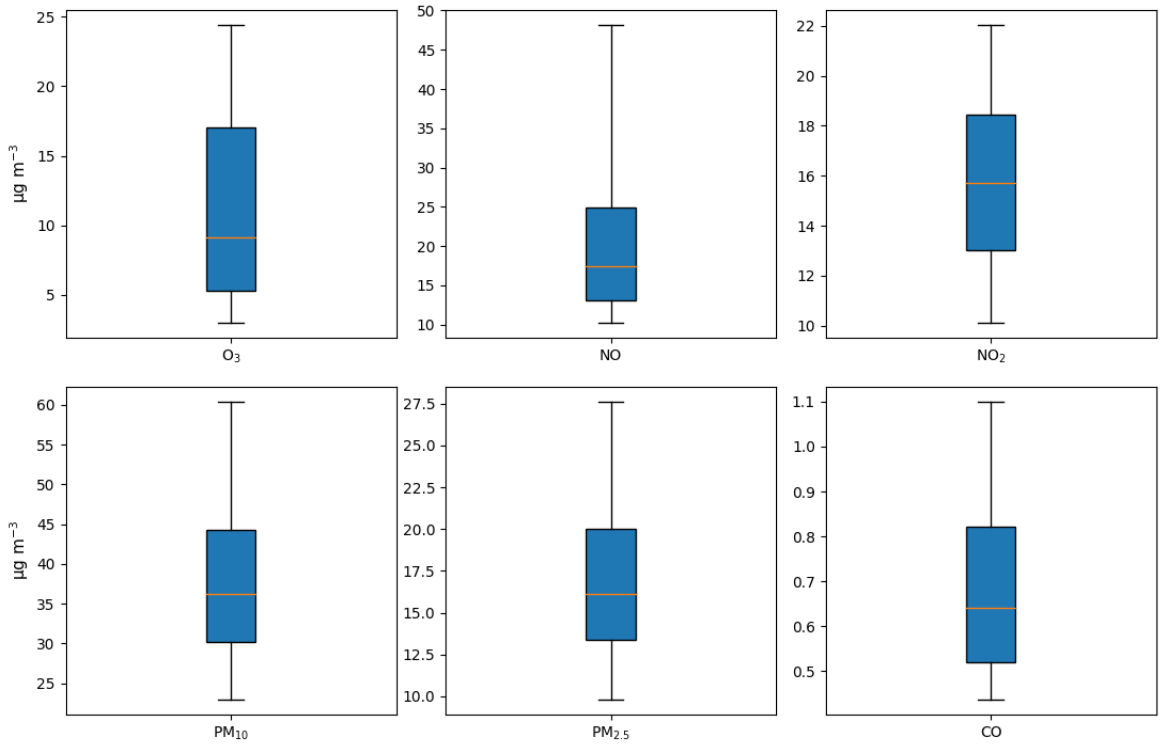
restricciones de movilidad, en la calidad de combustible mayoritariamente usada en la ciudad y por anomalías meteorológicas. De este modo, algunos contaminantes se vuelven más relevantes en el control y seguimiento de la calidad del aire en términos de O_3, como lo son el NO y el NO_2 y su relación, que pueden favorecer la titulación y producción del O_3.

3.1.2 Análisis temporal

3.1.2.1 Descripción general de datos

Se realizó un análisis de la distribución de los datos recopilados de la RMCAB en el periodo de estudio de los contaminantes, como se observa en la Figura 3-2. Los promedios obtenidos fueron: para O_3 de 5.9 µg/m^3, NO de 15.2 µg/m^3, NO_2 de 16.3 µg/m^3, CO de 0.7 µg/m^3, PM_{10} de 38.1 µg/m^3 y $PM_{2.5}$ de 16.8 µg/m^3.

Figura 3-2: Boxplot de contaminantes O_3, NO, NO_2, PM_{10}, $PM_{2.5}$ y CO. La línea naranja representa la media de los datos, la caja tiene los datos desde 25 hasta 75 percentil, y los bigotes van de 5 a 95 percentil.



Fuente: Autor

Para complementar el análisis de los datos, se llevó a cabo un análisis de distribución mediante el cálculo de la curtosis, la cual proporciona información sobre la forma de la distribución en comparación con una distribución normal o de campana de Gauss. Los resultados de la curtosis se presentan en la Tabla 3-1.

Se observa que para las concentraciones de O_3, NO, CO, PM_{10} y $PM_{2.5}$, la curtosis es positiva, lo que indica que la distribución tiene colas más pesadas y un pico más pronunciado en comparación con la distribución normal. Esto sugiere que los datos están más concentrados cerca de la media y que hay menos valores atípicos presentes en los conjuntos de datos.

Por otro lado, para el NO_2, la curtosis es negativa, lo que indica que la curva es más achatada y tiene colas más ligeras en comparación con la distribución normal. Esto sugiere que los valores de NO_2 están menos concentrados cerca de la media y que hay una mayor probabilidad de valores atípicos.

En resumen, los resultados sugieren que para los contaminantes O_3, NO, CO, PM_{10} y $PM_{2.5}$, hay pocos eventos de alta contaminación, mientras que, para el NO_2, la presencia de valores atípicos es más probable.

Tabla 3-1: Valor media y curtosis de las variables.

Variable	Media	Curtosis
Ozono	12.23	0.27
Monóxido de nitrógeno	19.49	4.31
Dióxido de nitrógeno	16.32	-0.07
Monóxido de carbono	0.76	1.77
PM_{10}	40.76	2.90
$PM_{2.5}$	18.00	2.39

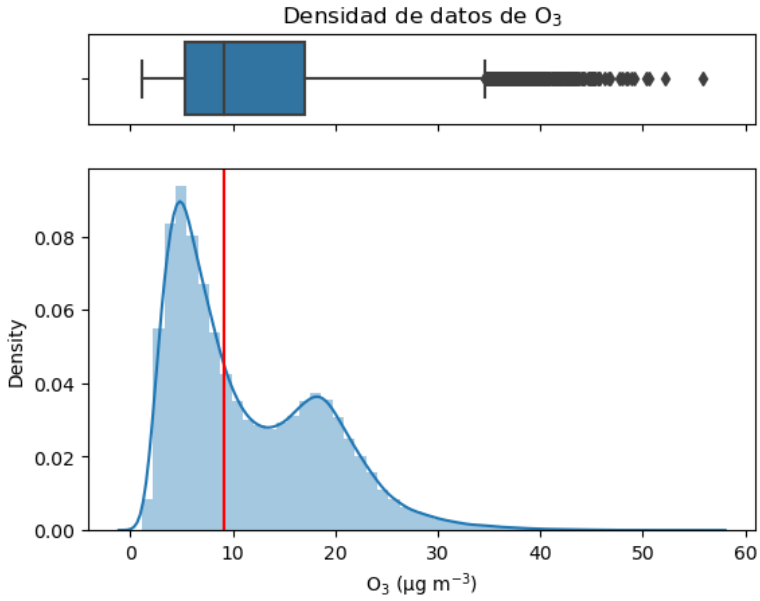
Fuente: Autor

Así mismo, se realiza la exploración del contaminante estudio, por medio de los análisis de densidad, tendencia, evolución mensual, distribución y por último la correlación entre otras variables. Esto con dos objetivos: (i) comprender mejor las variables que pueden ser utilizadas como entrada del modelo de predicción y (ii) describir la evolución del O_3, para explicar su comportamiento.

3.1.2.2 Densidad de datos medidos de ozono

Para continuar con el análisis del contaminante de estudio de O_3, se generó el gráfico de densidad, que permite suavizar la distribución y ruido, permitiendo facilitar el análisis de datos. Se graficó los datos recopilados de las estaciones de la ciudad en el periodo de estudio para observar la distribución y densidad de los datos, como se representa en la Figura 3-3.

Figura 3-3: Distribución y densidad de datos medidos de O_3. La línea negra dentro de la caja azul es la media de los valores, así como también la línea roja dentro de la distribución normal de la gráfica de abajo.



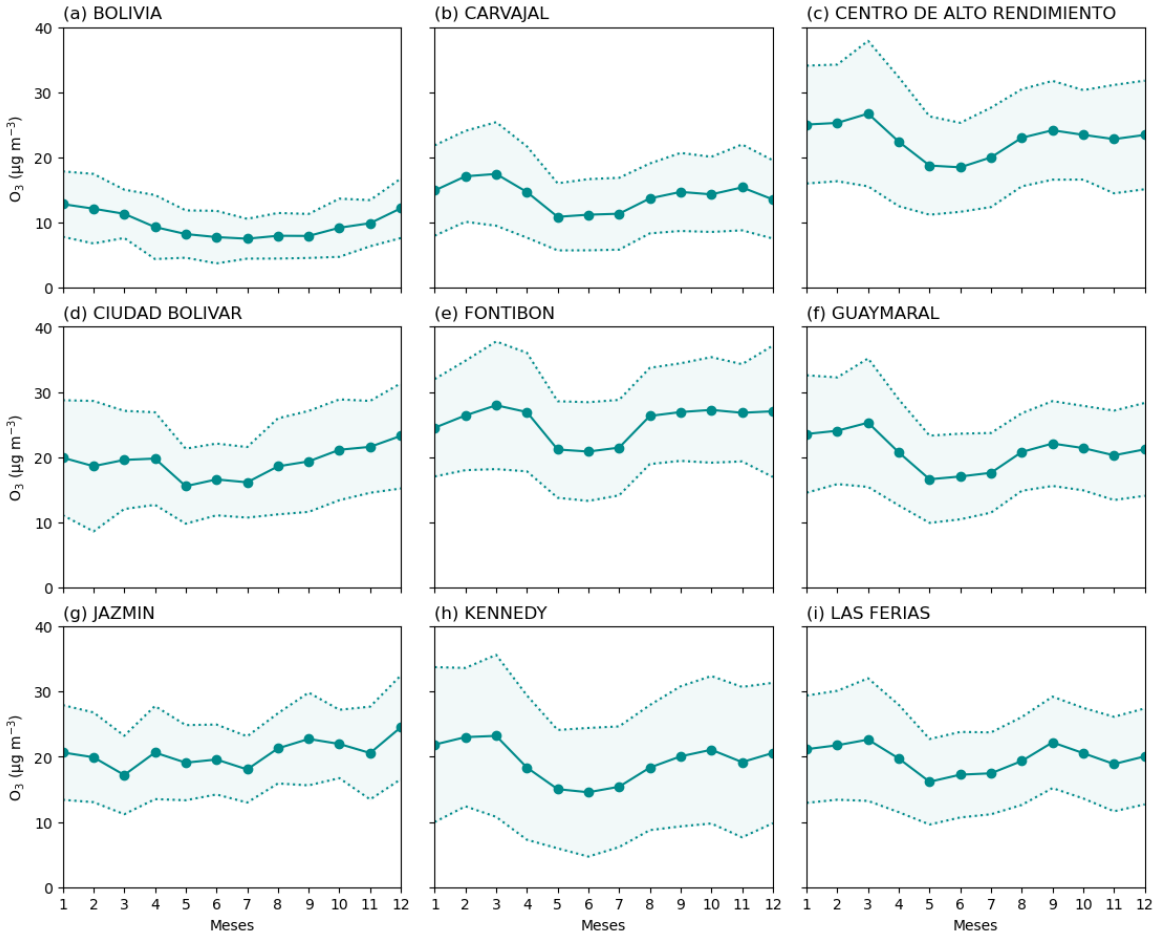
Fuente: Autor

Se puede observar en la Figura 3-3 que, para el caso de las concentraciones alcanzadas por el contaminante objeto de estudio, el 25% de los datos se encuentran debajo de 5.9 µg/m³, el promedio es de 9.9 µg/m³, así como también, el 75% de los datos se

encuentran sobre 18 µg/m³. Los datos aislados se encuentran sobre 34 µg/m³. La distribución se muestra bimodal al tener dos picos, donde se puede denotar una concentración de valores cerca de 5.9 µg/m³, encontrándose la mayor densidad de los datos de 0.09, seguido del segundo pico de datos que están concentrados alrededor de 18 µg/m³ con una densidad superior a 0.03. Por lo tanto, los datos mayormente se encuentran por debajo de 10 µg/m³, y los datos que se encuentran sobre 30 µg/m³ son datos aislados, evidenciando que es baja la probabilidad de eventos de alta concentración de O_3.

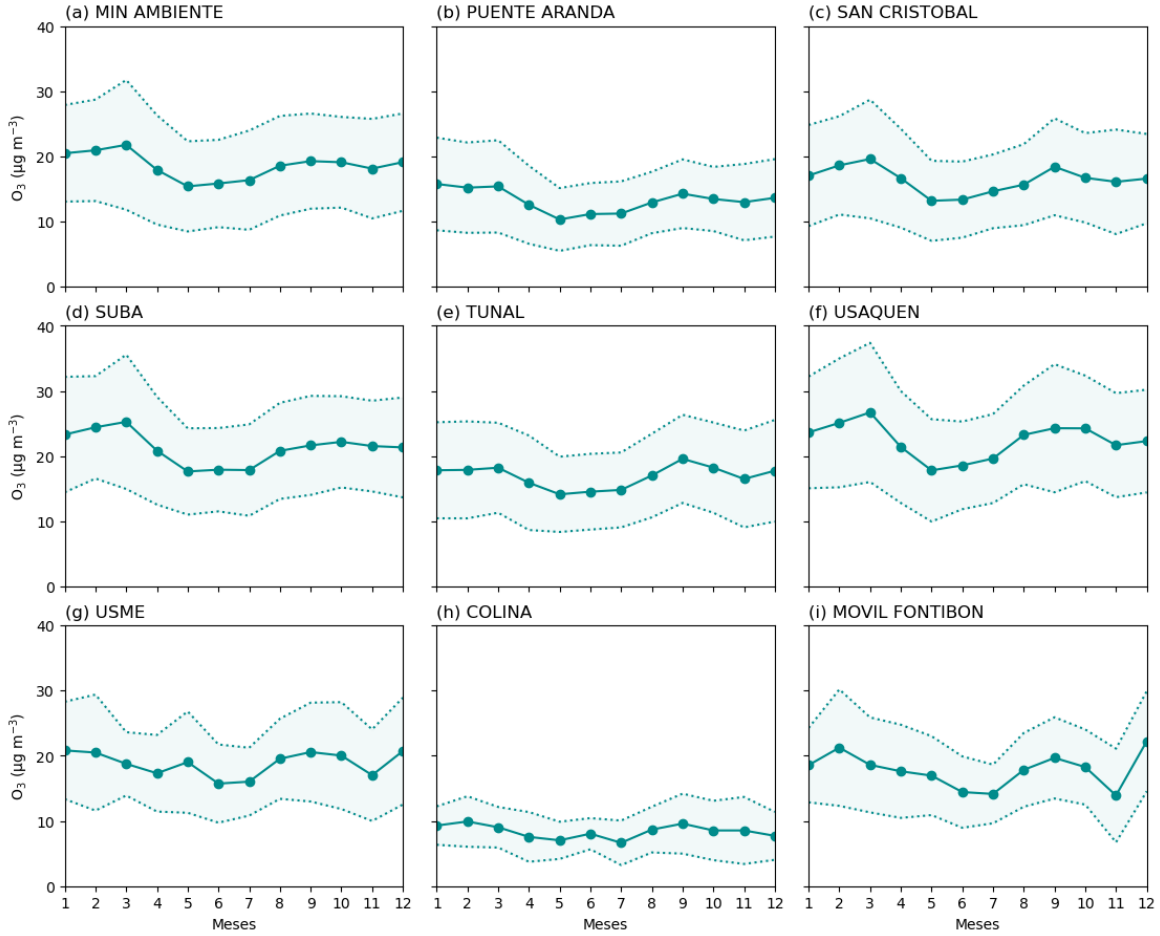
3.1.2.3 Comportamiento mensual del ozono

Figura 3-4: Comportamiento multianual mensual de los niveles de ozono por estación. Las líneas continuas representan la media y la sombra es la desviación estándar.



Fuente: Autor

Figura 3-5: Comportamiento multianual mensual de los niveles de ozono por estación. Las líneas continuas representan la media y la sombra es la desviación estándar. (Continuación Figura 3-4).



Fuente: Autor

Por otro lado, se analizaron los promedios mensuales de los valores diarios máximos en un promedio móvil de 8 horas para O_3, lo que resultó en la creación de la Figura 3-4 y la Figura 3-5. En estas figuras, se puede observar que las estaciones CDAR, Usme, Tunal, Fontibón, Kennedy, Las Ferias, Usaquén, San Cristóbal, Puente Aranda y Minambiente muestran un comportamiento de picos altos en los niveles de O_3 durante los meses de enero a marzo, seguido de una disminución desde abril hasta julio y luego un aumento en los meses siguientes, para finalmente disminuir en noviembre y aumentar nuevamente en diciembre.

Esta tendencia podría atribuirse al hecho de que, durante el primer y último trimestre, Bogotá se ve influenciada por los incendios forestales frecuentes, que contribuyen con más del 85% de los episodios de alta contaminación en la ciudad (Casallas et al., 2024), así como a la temporada seca y al aumento de los niveles de radiación debido a la baja nubosidad (Ballesteros-González et al., 2020; Casallas et al., 2023). Por otro lado, las estaciones de Bolivia y Colina muestran concentraciones más bajas de O_3 y presentan un comportamiento leve similar al del resto de las estaciones.

Sin embargo, en contraste con la mayoría de las estaciones, la estación de Jazmín no sigue este patrón en los meses de febrero y marzo, ya que experimenta una disminución en sus niveles de O3 en ese período.

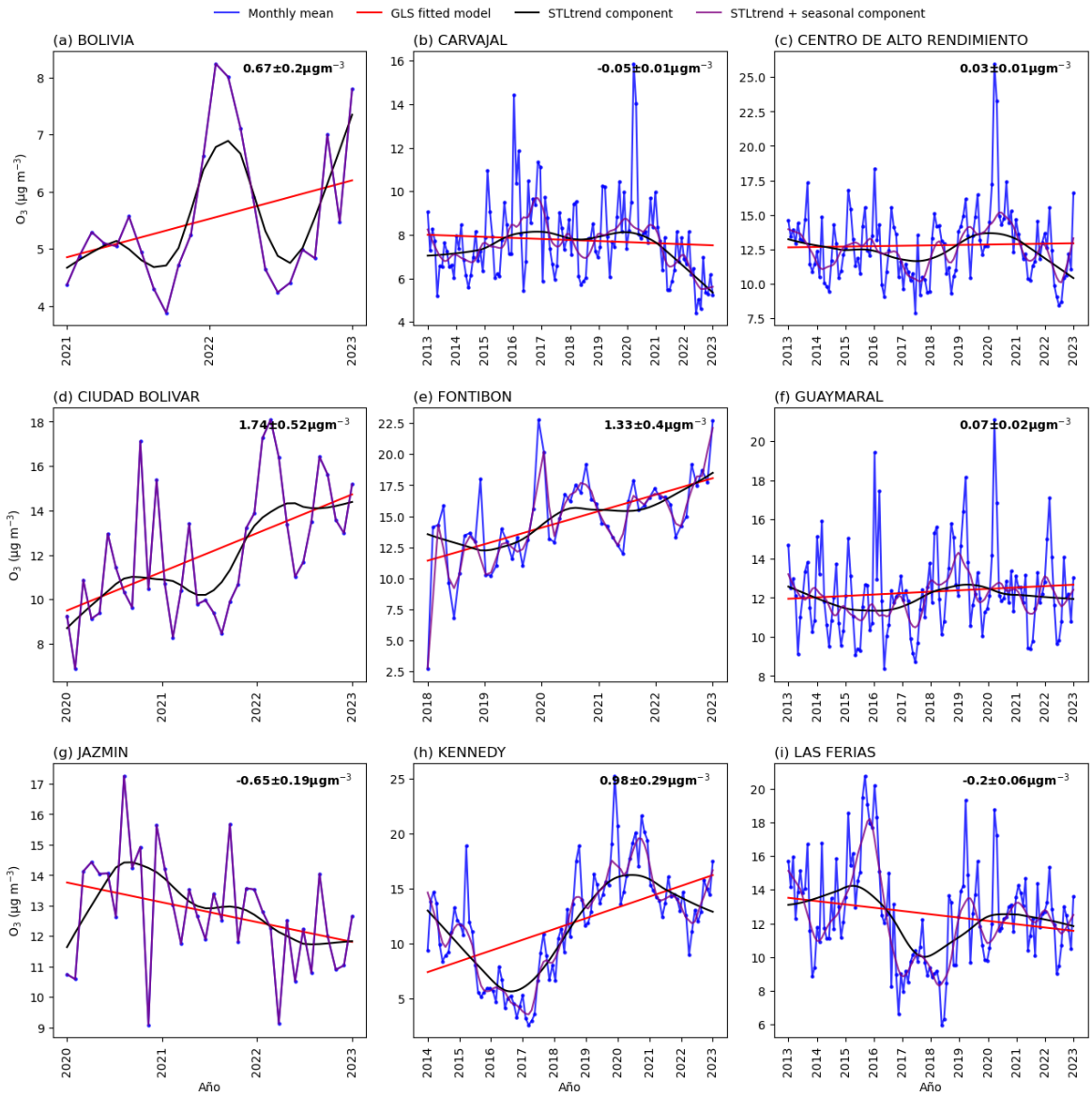
3.1.2.4 Tendencia multianual mensual del ozono

La Figura 3-6 y la Figura 3-7 presenta el GLS que estima la relación lineal entre una serie temporal autocorrelacionada y el tiempo para obtener residuos independientes y de esta forma obtener una estimación correcta de la varianza de los coeficientes de regresión. En estas figuras se observa que varias estaciones muestran una tendencia positiva en los niveles de O_3. Estas estaciones incluyen Bolivia, CDAR, Ciudad Bolívar, Fontibón, Guaymaral, Kennedy, Min Ambiente, Puente Aranda, Tunal, Usme y Móvil Fontibón, lo que representa 12 de las 18 estaciones que realizan mediciones de O_3 en la ciudad y muestran un aumento en sus niveles.

Este incremento podría atribuirse al período de confinamiento que se observó a nivel mundial, ya que las restricciones de movilidad contribuyeron a la disminución de los niveles de NOx, combinado con las condiciones meteorológicas y los incendios forestales que se registraron en la ciudad (Sokhi et al., 2021; Casallas et al., 2023).

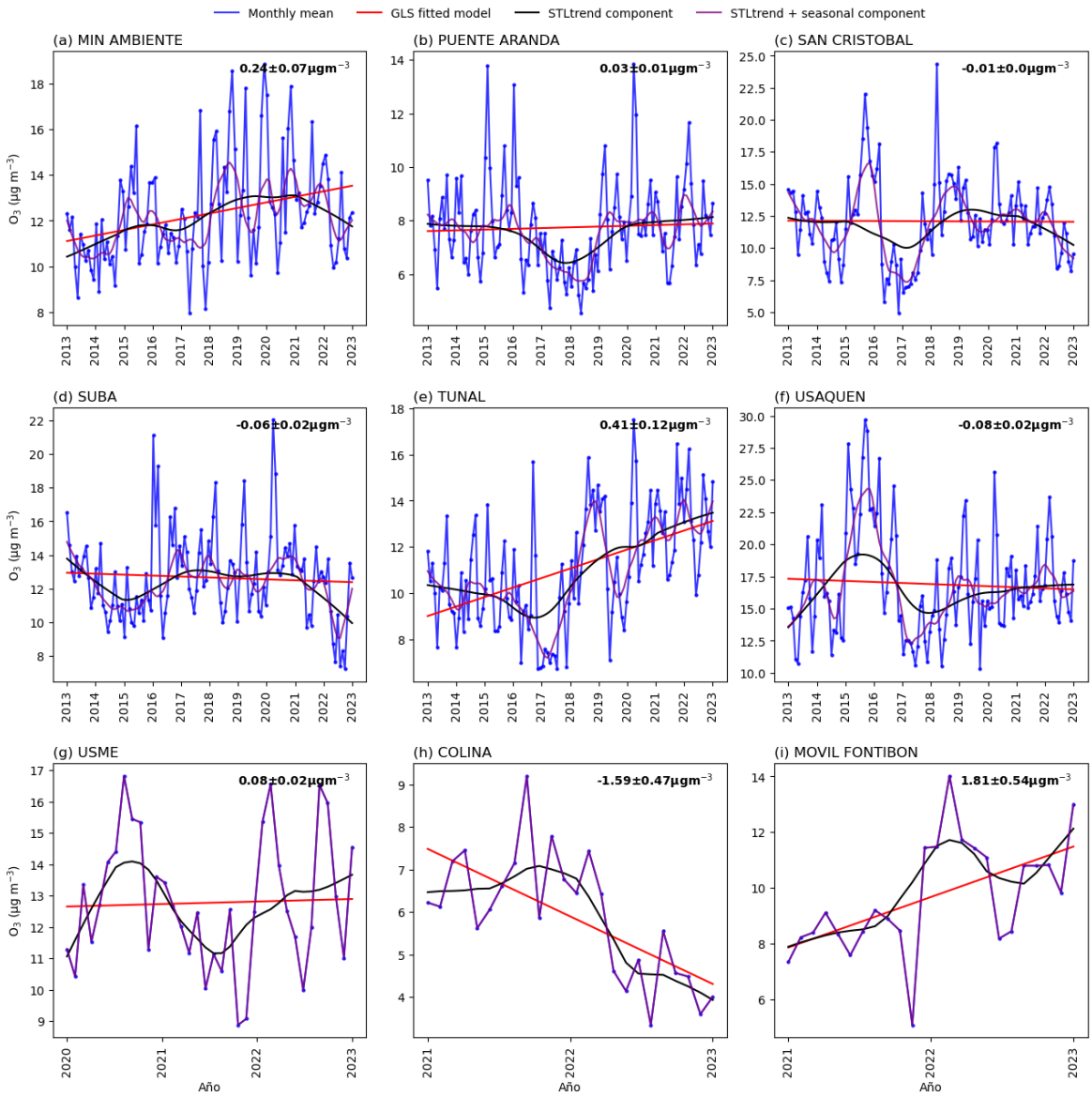
Por otro lado, algunas estaciones como Carvajal, CDAR, Ciudad Bolívar (aunque en menor medida que en 2022), Fontibón, Guaymaral, Jazmín, Kennedy, Las Ferias (aunque en menor medida que en 2019), Min Ambiente (cercanas a las presentadas en 2018), Puente Aranda, Suba, Tunal, Usaquén (aunque en menor medida que en 2015) y Usme mostraron un aumento significativo en los valores promedios mensuales durante el año 2020. Estas estaciones representan el 70% de las localidades de la ciudad.

Figura 3-6: Descomposición STL y regresión GLS de la media mensual de los niveles de O_3 en Bogotá. Se tuvieron en cuenta la información disponible por cada estación. La línea violeta es el SLT y el componente estacional. La línea negra representa el componente de tendencia STL y la línea roja es la pendiente ajustada GLS que representa la tendencia lineal del O_3.



Fuente: Autor

Figura 3-7: Descomposición STL y regresión GLS de la media mensual de los niveles de O3 en Bogotá. Se tuvieron en cuenta la información disponible por cada estación. La línea violeta es el SLT y el componente estacional. La línea negra representa el componente de tendencia STL y la línea roja es la pendiente ajustada GLS que representa la tendencia lineal del O3. (Continuación).



Fuente: Autor

Es de resaltar que, durante los años 2017 y 2018, se observó una disminución considerable en los niveles de O_3 en las estaciones de Kennedy, Las Ferias, Puente Aranda, San Cristóbal, Tunal y Usaquén, donde los valores aproximados fueron de 6, 10,

7, 10, 9 y 15 µg/m³ respectivamente, debiéndose a un comportamiento local de Bogotá, principalmente en las localidades centrales cercanas. Esto posiblemente fue debido a que en los años anteriores (se presentó en cuatro meses del 2010, de mayo del 2015 a mayo del 2016) y posteriores (1 mes del año 2018 y 3 meses en el 2019) a estas fechas, se presentó el fenómeno del niño, el cual es el aumento de temperatura superficial, reflejándose en el aumento de la capa de altura de mezcla y también se relaciona con aumento de NOx en estos años en algunas estaciones de la ciudad (Baquero, 2023).

Por otro lado, la estación de Carvajal no tiene datos representativos después de junio de 2022, debido al vandalismo presentado en la zona, por lo tanto, en la gráfica se denota una disminución luego de ese periodo al realizar el cálculo con los pocos datos existentes. No obstante, algunas estaciones presentaron valores mensuales promedio mayores o similares respecto al año 2020, como, por ejemplo: Fontibón, Ciudad Bolívar y Usme en 2022; San Cristóbal en el 2018; Suba en el 2016; Las Ferias, Usaquén y Puente Aranda en el 2015.

3.2 Modelo de predicción *machine learning*

El uso del modelo se llevó a cabo para las estaciones de la RMCAB, con excepción de las estaciones Bosa, Colina, Móvil Fontibón, Bolivia y Móvil 7ma, debido a diferentes razones como la falta de mediciones continuas o permanentes de las variables requeridas para el modelo o la disponibilidad limitada de datos. Por ejemplo, la estación Bosa solo tiene mediciones de O_3 durante 2 meses en 2020, mientras que las estaciones Colina, Móvil Fontibón y Bolivia comenzaron a medir O_3 en fechas posteriores. En el caso de la Estación Móvil 7ma, no se realizan mediciones.

Entre las estaciones restantes, CDAR, Jazmín, Kennedy, Puente Aranda, Tunal y Usme cuentan con mediciones de las variables de entrada requeridas para el modelo. Sin embargo, otras estaciones no realizaron mediciones continuas o permanentes de todas las variables necesarias. Por ejemplo, algunas estaciones carecen de mediciones de radiación, humedad relativa, velocidad del viento o CO.

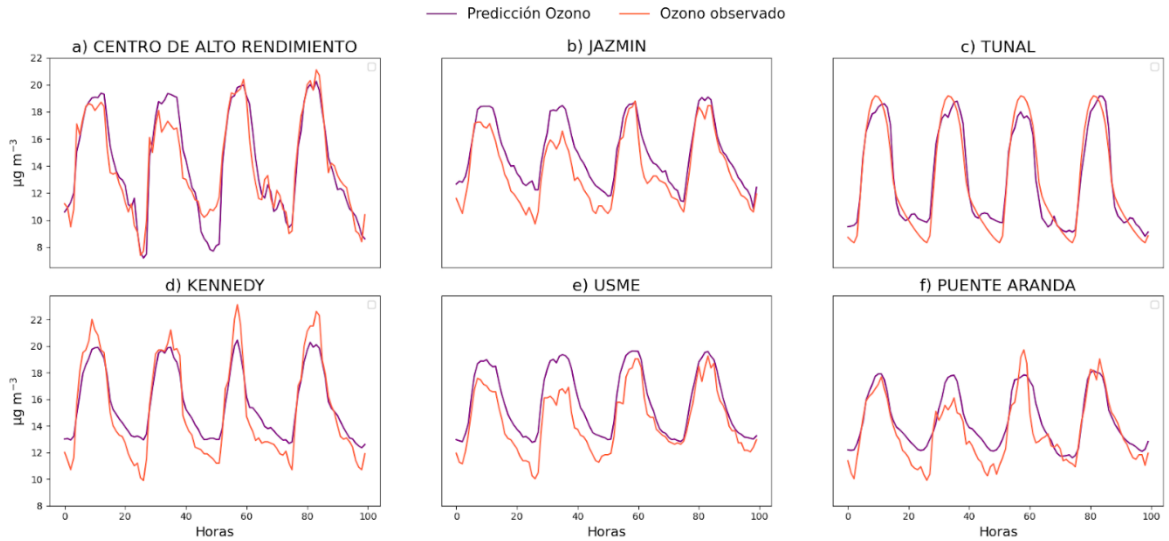
Para abordar la falta de datos en las estaciones donde no se disponía de información completa para las variables de entrada requeridas, se utilizó un enfoque de imputación. Se utilizaron las mediciones promedio imputadas de todas las estaciones a nivel distrital para calcular la predicción en estas estaciones.

Para visualizar la relación entre los valores predichos y observados, se generaron gráficos que muestran los primeros 100 resultados de predicción en comparación con los valores observados (ver Figura 3-8). Estos gráficos proporcionan una perspectiva inicial sobre la capacidad del modelo para predecir los niveles de O3 en las estaciones donde se aplicó el enfoque de imputación de datos.

En los resultados obtenidos en la Figura 3-8, se observan similitudes en la tendencia de los datos observados y los valores predichos para la estación CDAR. En el primer día estudiado, se inicia con valores cercanos, con una diferencia menor a 1 µg/m^3 entre el ozono predicho y el ozono observado, comenzando en 10.6 µg/m^3 y 11.2 µg/m^3, respectivamente. La tendencia de los datos comienza a descender de manera paralela en ambas situaciones.

De igual forma, se denota que, en los mínimos locales que se dan entre los días 26 al 29, por debajo de 9 µg/m^3 en el ozono observado y la predicción desciende de manera similar, con valores de 7.4 µg/m^3 y 7.6 µg/m^3 respectivamente para el punto más bajo. Sin embargo, en el segundo mínimo local se encuentra entre los días 47 y 53, donde la diferencia llega a ser de 3 µg/m^3, con valores de 7.7 µg/m^3 y 10.7 µg/m^3 respectivamente del punto más bajo. Pero a partir del día 58, los valores mantienen cercanía y una tendencia muy similar, donde el modelo reproduce la tendencia de elevación, máximos y mínimos luego del entrenamiento, alcanzando el máximo local donde el ozono observado y la predicción alcanzan 21.1 µg/m^3 y 20.3 µg/m^3 respectivamente, teniendo menos de 1 µg/m^3 de diferencia. El modelo para esta estación sobreestima los primeros 40 días aproximadamente, sin embargo, luego de esto, los valores permanecen cercanos a los observados.

Figura 3-8: Valores de observación vs predicción de O_3 para las estaciones que realizan mediciones de las variables necesarias como entradas en el modelo. La hora 0 y la hora 100 representa desde el primer valor predicho hasta el valor número 100.



Fuente: Autor

En el caso de la estación Jazmín, también inicia con una diferencia menor a 1 µg/m^3 entre el ozono observado y la predicción, con valores 12.6 µg/m^3 y 13.6 µg/m^3 respectivamente. Sin embargo, el modelo tiende a subestimar en los primeros dos máximos locales, con diferencias de más de 1 µg/m^3, donde el primer máximo tiene 17.6 µg/m^3 y 18.8 µg/m^3 respectivamente, y en el segundo 17.1 µg/m^3 y 18.8 µg/m^3 de valores observados y predichos respectivamente. A pesar de ello, el modelo logró replicar el comportamiento de los valores observados, y la diferencia se redujo en el tercer y cuarto máximo local, donde para este último, obtuvo valores de predicción de 19.4 µg/m^3 respecto al valor observado de 18.8 µg/m^3, siendo una diferencia muy baja.

Para la estación Tunal, se observa un patrón similar, con una representación adecuada de la tendencia del ozono a lo largo del tiempo, incluyendo fluctuaciones en los valores. No obstante, en los mínimos locales, el modelo tiende a subestimar los valores. Aunque los valores observados alcanzan un mínimo de 9.6 µg/m^3 en los días 28 y 52, los valores de ozono predicho alcanzan valores de hasta 11 µg/m^3, sobreestimando más de 1 µg/m^3. Esto mejora en el último mínimo y máximo local, donde las diferencias entre lo observado y la predicción es menor de 1 µg/m^3.

Caso contrario sucede en la estación Kennedy, donde aun cuando el modelo replica el comportamiento, los valores predichos se mantienen por debajo de los valores de ozono observados. Para los máximos locales, el modelo subestima el comportamiento, y en los mínimos locales, subestima los valores, alcanzando diferencias de más de 1 µg/m³. Es importante destacar que el mayor máximo local en esta estación alcanza 19.4 µg/m³, mientras que los valores de predicción llegan hasta 18.4 µg/m³.

En lo que se refiere a la estación Usme, el modelo arrojó resultados contrarios a la estación Kennedy, ya que subestimó los valores, destacando los primeros dos máximos y mínimos locales, superando en más de 2 µg/m³, donde el máximo local fue de 18.1 µg/m³ respecto al valor observado de 15.8 µg/m³, y el mínimo loca observado fue de 9.8 µg/m³ y la predicción de 12.3 µg/m³. Sin embargo, desde el tercer máximo local, el modelo mejora y reduce la diferencia de valores, donde se denota para el cuarto máximo local una diferencia menor a 1 µg/m³, alcanzando valores de predicción de 18.9 µg/m³ respecto a 18 µg/m³ O_3 observado.

Así mismo, la estación Puente Aranda tuvo un comportamiento variable, donde el modelo sobreestimó los mínimos y el segundo máximo local, pero subestimó el tercer máximo local. Sin embargo, el modelo logró corregir la predicción para el tercer mínimo local (11.3 µg/m³ respecto al observado de 10.6 µg/m³) y el cuarto máximo local (16.8 µg/m³ respecto al observado de 17.8 µg/m³), consiguiendo valores diferencias menores a 1 µg/m³, y reproduciendo la tendencia con mayor precisión.

Por otro lado, se obtuvo los resultados de las demás estaciones que no contenían alguna entrada (ver Anexo 1), donde el modelo consiguió en la mayoría de los casos, reproducir la tendencia y el comportamiento al menos en los últimos máximos locales, donde la diferencia es menor a 1 µg/m³. Para el caso de Las Ferias, MinAmbiente y San Cristóbal, el modelo sobreestimó en los dos primeros máximos locales, con diferencias de más de 1 µg/m³. Sin embargo, el modelo siguió el comportamiento de las demás estaciones, reduciendo esta diferencia. En cuanto a la estación Usaquén, sobreestimó los valores, pero en el último máximo local logra acercarse y reducir la diferencia entre lo observado y la predicción, cerca de 1 µg/m³, obteniendo 19.4 µg/m³ y 20.5 µg/m³ respectivamente.

En cuanto a la estación Ciudad Bolívar, obtuvo los mínimos locales con valores más bajos tanto en la predicción, como los observados, alcanzando datos máximos de 17 µg/m³, sin embargo, el modelo logró adaptarse a la tendencia, y desde el tercer máximo local, se asemejó significativamente. Las estaciones que alcanzaron valores de predicción sobre 20 µg/m³, fueron Carvajal, Suba y Usaquén, replicando a los valores observados en los máximos locales. Igualmente, la estación Suba, obtuvo valores de predicción por debajo de 9 µg/m³, en los días 75 a 77, reproduciendo el comportamiento de los valores observados.

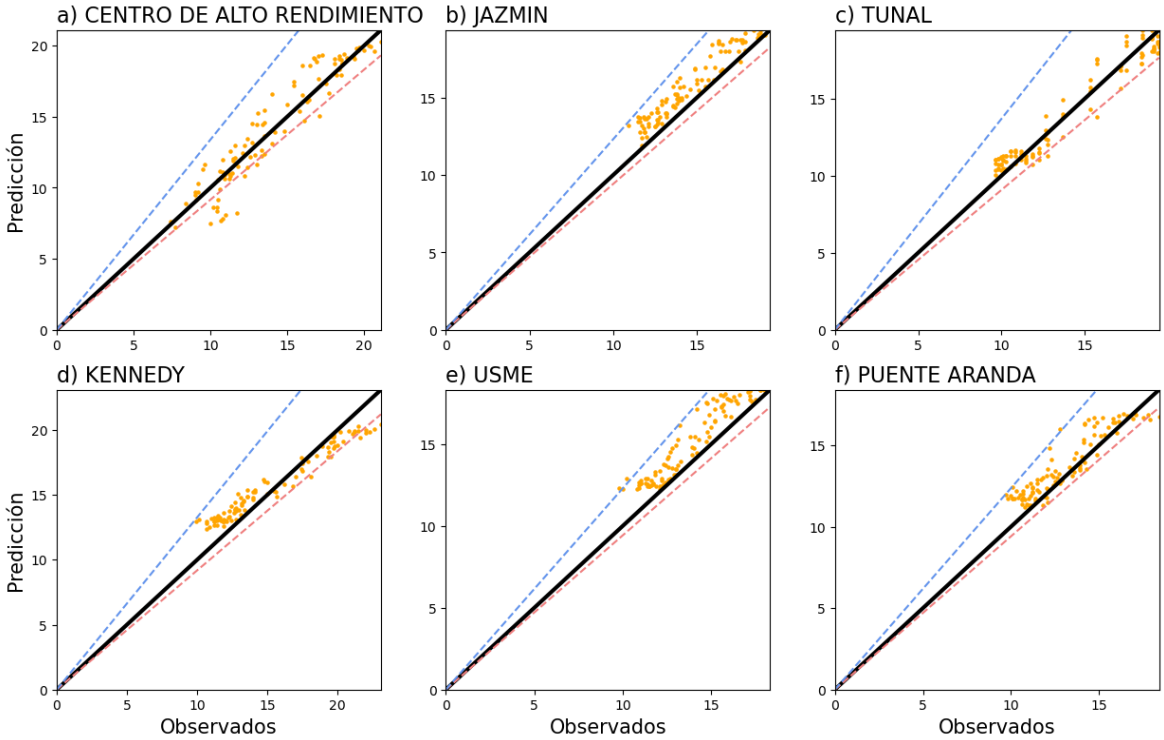
Los resultados obtenidos muestran que, aunque el modelo de la red convolucional y LSTM es capaz de aprender patrones complejos en los datos y en la tendencia del comportamiento, resulta desafiante predecir valores exactos para cambios rápidos y bruscos en el tiempo. Además, es posible que el modelo no esté generalizando correctamente situaciones que difieran de las condiciones presentes en los datos de entrenamiento. A diferencia de los resultados encontrados por Celis et al. (2022) y Casallas et al. (2020), donde los modelos tendían a sobreestimar los datos, en este modelo ajustado se observa en más casos una subestimación de los datos. Esto podría conducir a predicciones menos precisas en condiciones extremas o inusuales de picos de contaminación.

Una posible razón para estas diferencias podría ser el número de épocas utilizadas en el modelo. En comparación con modelos más robustos, como el estudio de Abdul-Wahab & Al-Alawi. (2002), donde se utilizaron más de 13 mil épocas, el número de épocas en este modelo puede no ser suficiente para minimizar el error entre los valores reales y las predicciones.

Además, es importante tener en cuenta que en estudios anteriores (Abdul-Wahab & Al-Alawi, 2002) se utilizaron variables adicionales como CH4 y polvo, las cuales no están disponibles para el análisis en Bogotá. La falta de estas variables podría influir en la precisión de las predicciones, ya que no se están considerando todos los factores relevantes para la formación de contaminantes atmosféricos en la ciudad.

Los resultados anteriores, se pueden contrastar con los diagramas de dispersión en la Figura 3-9 y el Anexo 1, utilizando la ecuación de regresión al modelo para describir mejor su comportamiento.

Figura 3-9: Diagrama de dispersión de observación vs predicción de ozono de estaciones que realizan mediciones de las variables necesarias como entradas en el modelo. La línea negra corresponde a la regresión lineal 1:1, las líneas punteadas azul y rojas corresponden a 1:2 y 1:0,5 respectivamente.



Fuente: Autor

La Figura 3-9 representa con mayor claridad la relación lineal positiva modelada con la línea ajustada de regresión, que existe entre los valores para la predicción y los observados. Los valores se mantienen constantes y agrupados dentro de las líneas punteadas, permaneciendo cerca de la línea central, a excepción de la estación CDAR, donde en valores observados cercanos a 10 µg/m³, el modelo subestima los valores.

Para los casos de Jazmín y Usme, los valores en su mayoría se encuentran sobre la línea negra, representando una tendencia en que el modelo sobreestima los valores, sin

embargo, en general, este comportamiento también se denota en las demás estaciones, ya que la línea baja roja se encuentra siempre más cercana a la línea negra, observándose que la dispersión siempre se acerca más hacía arriba que abajo. En la Figura 3-9 no se denotan valores atípicos evidentes, por lo que se puede decir que el modelo es preciso.

Por otro lado, la dispersión obtenida para las demás estaciones que no contenían alguna entrada requerida para el modelo, se puede ver un comportamiento similar a las primeras estaciones, donde con una relación lineal positiva. La dispersión de las estaciones Carvajal, Ciudad Bolívar, Las Ferias, Minambiente, San Cristóbal y Usaquén, se encuentra visiblemente sobre la línea negra, mostrando la tendencia a sobreestimar los valores de predicción al igual que las estaciones del Anexo 2. Se puede apreciar que, para la estación de Suba, hay datos que se alejan de la línea roja, pero no excesivamente como para aseverar que son valores atípicos, al igual que en la estación Usaquén, que contiene valores sobre la línea azul.

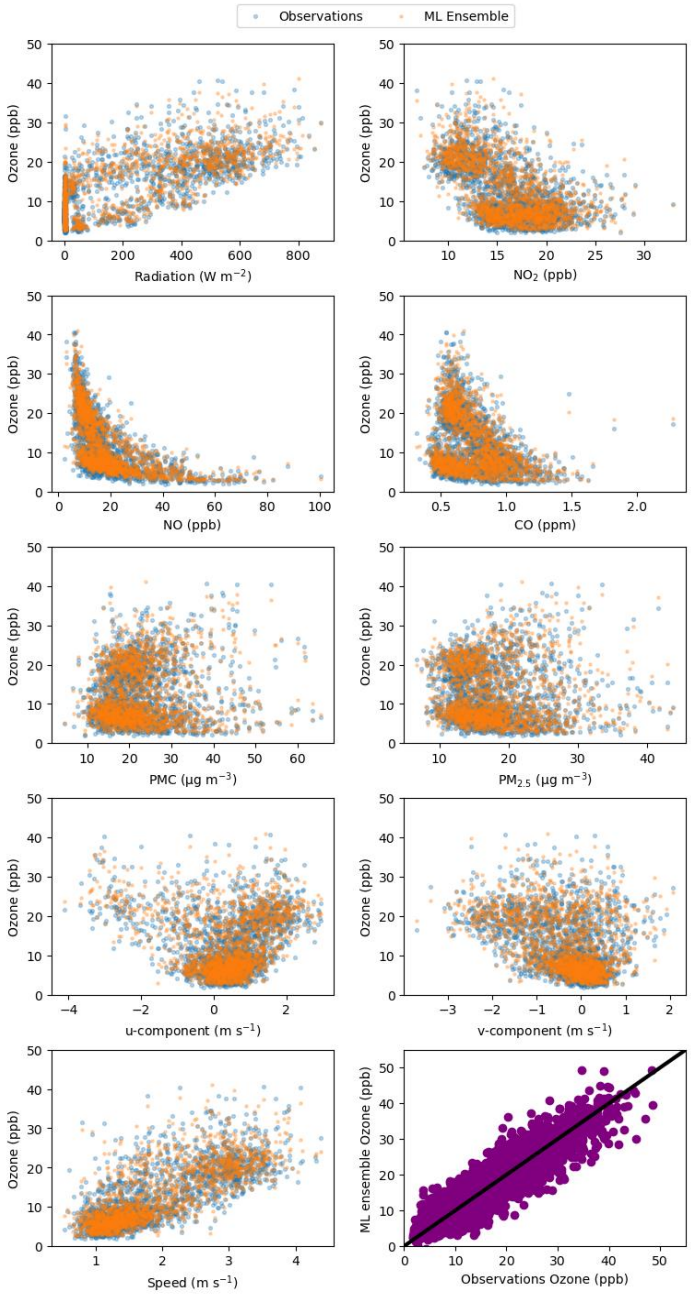
3.3 Validación del Modelo

3.3.1 Distribución

En la Figura 3-10 se observa un diagrama de dispersión para la validación del modelo, representado la concentración obtenida del modelo y de las observaciones de ozono (medida en µg/m³) en función de los diferentes contaminantes y características meteorológicas. Con el objetivo principal de evaluar la capacidad del modelo para predecir los niveles de ozono a partir de los diferentes precursores.

En relación con el diagrama de dispersión en la Figura 3-10, se evidencia una correlación positiva entre el ozono y la radiación, donde la tendencia suele ser de aumento, al igual que en el caso de la rapidez del viento. Por otro lado, se observa una correlación negativa con el NO, donde los valores de concentración de ozono tienden a disminuir en presencia de concentraciones elevadas de NO, este mismo caso se observa en el caso del CO, se encuentran valores similares en relación con el ozono y material particulado.

Figura 3-10: Dispersión de datos respecto al O_3 superficial de Bogotá a partir de las medias del conjunto ML para un subconjunto de 2000 puntos de datos submuestreados aleatoriamente del conjunto de datos de validación. El O_3 se representa gráficamente en función de la radiación, NO_2, NO, CO, PMC, $PM_{2.5}$, componentes del viento (u y v) y velocidad del viento. El panel inferior derecho representa la observación del ozono frente a la media del conjunto ML.



Fuente: Autor

Es importante señalar que, según estudios previos (Liu, Wang 2020), la relación entre el CO y NOx no sigue una dependencia lineal. Por lo que, hay repercusiones en el NO, que puede actuar como consumidor del ozono debido a la disminución de la titulación de O_3 y sus radicales. Esta complejidad en la relación entre los precursores resalta la necesidad de un análisis exhaustivo para entender completamente las interacciones y los impactos específicos de cada contaminante en la formación del ozono.

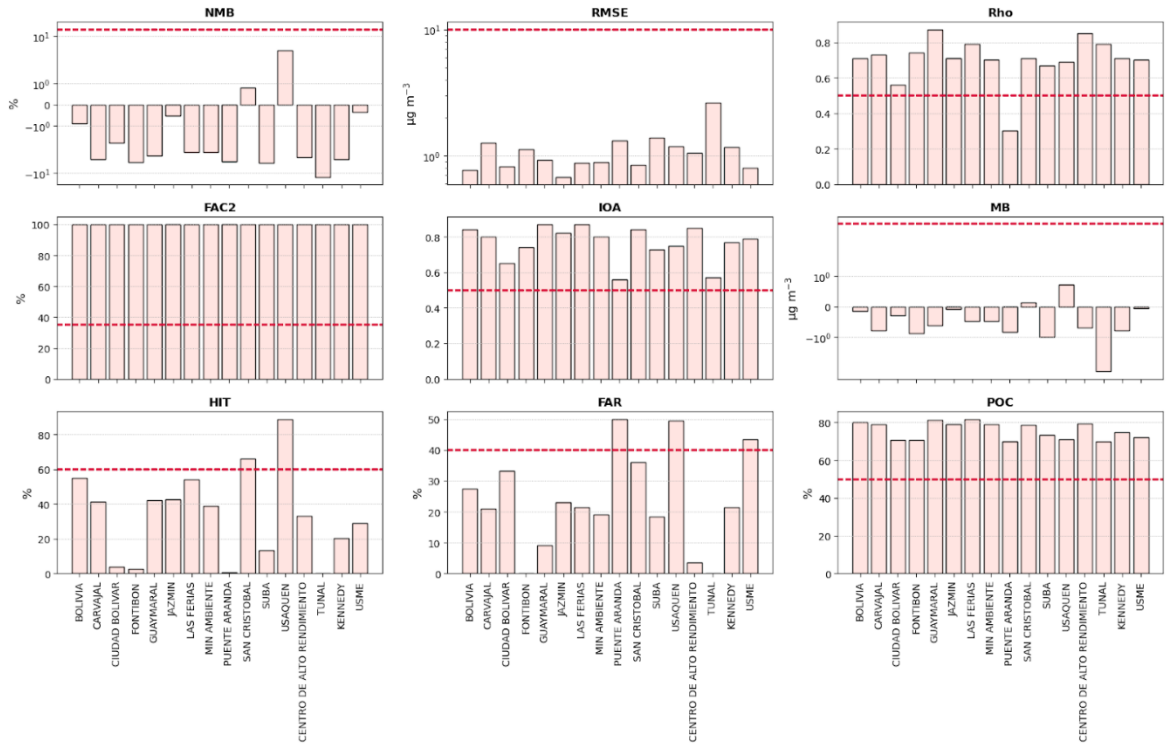
A partir de estas gráficas (ver Figura 3-10), como validación del modelo, se puede apreciar una buena calidad del modelo evaluado en función de su capacidad para hacer predicciones precisas con el ozono, considerando la correlación positiva hallada entre el método de ML y las observaciones.

3.3.2 Errores estadísticos

Con el propósito de evaluar la eficacia del modelo convolucional-LSTM, se ha establecido el resultado de la validación estadística a partir de los resultados de los criterios más comunes para validar la calidad y precisión del modelo. La comparación entre las concentraciones observadas de Ozono se ha llevado a cabo considerando, calculando los promedios diarios de ocho horas de los últimos 13125 datos simulados a partir de datos horarios y entradas de variables precursores de la generación de ozono, respecto a promedios diarios de ocho horas de los valores observados en las fechas mencionadas en la metodología, obteniendo la Figura 3-11.

Los resultados obtenidos para la validación del modelo son favorables, como se evidencia en la fuerte correlación positiva de Spearman Rho, que supera 0.6 entre los valores observados y los predichos para la mayoría de las estaciones, a excepción de Puente Aranda. Sin embargo, la correlación más destacada obtenida en el modelo fue de 0.87 para el caso de Guaymaral. En comparación con el modelo de Abdul-Wahab & Al-Alawi (2002), donde se alcanzó un resultado de 0.96, esta diferencia podría atribuirse al tiempo de entrenamiento más breve y a la exclusión de variables de entrada clave, como los hidrocarburos distintos al metano o NMHC (Non-Methane Hydrocarbons, por sus siglas en inglés), los cuales se consideran precursores químicos clave para la formación de ozono en presencia de luz solar (Abdul-Wahab & Al-Alawi, 2002).

Figura 3-11: Resultado de los errores estadísticos para la validación del modelo, graficando los valores sombreados en rosa de los parámetros NMB, RMSE, Rho, FAC2, IOA, MB, HIT, FAR, POC de las estaciones que se modelaron de Bogotá. La línea roja punteada indica el valor de referencia según la Tabla 2-4.



Fuente: Autor

Con resultados similares, el POC obtuvo valores superiores al 60% en todas las estaciones, sobresaliendo las estaciones Bolivia, Guaymaral, Las Ferias y CDAR que se encontraron superiores o iguales al 80%, evidenciando predicciones correctas realizadas por el modelo en relación con el conjunto total de valores observados. Igualmente, se obtuvieron resultados de HIT para las estaciones Usaquén y San Cristóbal superiores a 80% y 60% respectivamente como tasa de proporción de aciertos de las predicciones. Así mismo, los valores obtenidos de FAC2 se encuentran en todos los casos en 100%, debido a que, es la fracción de datos que satisfacen la relación que existe entre los valores predichos y los valores observados, que para todas las estaciones se encuentra entre 0.5 y 2.

Adicionalmente, se han aplicado los criterios estadísticos definidos por Mogollon-Sotelo et al. (2021) para evaluar los siguientes parámetros: IOA, RMSE, correlación. De los

cuales se pueden identificar los valores superiores de IOA 0.5 que se considera en clasificación “Excelente” en todas las estaciones, un RMSE por debajo de 10 µg/m³ categorizado como “Excelente” y una correlación superior de 0.7 en la mayoría de las estaciones categorizado como “Bueno” (ver Tabla 3-2), a excepción de las estaciones Ciudad Bolívar con 0.62 y Puente Aranda, que se categorizan en “Bueno” y “Deficiente” respectivamente.

Tabla 3-2: Resultados obtenidos de correlación para cada estación.

Estación	R	Estación	R
Usme	0.73	Fontibón	0.77
CDAR	0.85	Guaymaral	0.87
Jazmín	0.74	Las Ferias	0.82
Tunal	0.82	Minambiente	0.73
Kennedy	0.75	San Cristóbal	0.74
Puente Aranda	0.36	Suba	0.70
Carvajal	0.75	Usaquén	0.71
Ciudad Bolívar	0.62	***Promedio***	***0.73***

Fuente: Autor

Por su parte, se encuentra una subestimación para algunas métricas, como el NMB (sesgo normalizado) con valores entre -100 y 80% y el MB (sesgo promedio) con valores entre -10 y 10 a excepción de Tunal que obtuvo un valor inferior a -10. Mientras que los valores para el FAR inferiores a 40% para la mayoría de las estaciones a excepción de Usme, Usaquén y Puente Aranda, mostrando una proporción considerablemente alta de falsas alarmas generadas por el modelo en comparación con los casos positivos reales.

En general, estos valores sugieren que el modelo de predicción de O_3 tiene un buen rendimiento, con correlaciones positivas, un sesgo promedio cercano a cero y una alta precisión en términos de RMSE, FAC2, IOA y POC. Sin embargo, la tasa de falsas alarmas (FAR) es relativamente alta, lo que indica que el modelo puede generar algunas predicciones incorrectas en términos de alarmas de calidad del aire en comparación con las observaciones.

Es cierto que el modelo CNN podría beneficiarse de ser combinado con otras ML, como se hizo en el caso del modelo utilizado en la ciudad de Brasil, donde se combinaron ANN con SVM, logrando una correlación alta de 0,95 (Luna et al., 2014). Esta combinación de técnicas puede aprovechar las fortalezas individuales de cada enfoque y mejorar la precisión y robustez del modelo resultante.

Sin embargo, es importante tener en cuenta que cada ciudad presenta particularidades de geolocalización, características geográficas, patrones de contaminación específicos y condiciones meteorológicas únicas. Estas diferencias pueden influir en los resultados obtenidos incluso cuando se utilizan las mismas técnicas de ML. Por lo tanto, al aplicar un modelo desarrollado en una ciudad a otra, se deben considerar cuidadosamente las diferencias locales y adaptar el modelo según sea necesario para obtener resultados precisos y confiables.

3.4 Riesgo poblacional

Los resultados de las concentraciones promedio de $PM_{2.5}$ y O_3 por estación resaltaron que Carvajal, Kennedy, Tunal y Ciudad Bolívar presentaron los valores más elevados de $PM_{2.5}$, mientras que Centro de Alto Rendimiento, Carvajal, Guaymaral y Jazmín registraron las concentraciones más altas de ozono (ver Tabla 3-3).

Los límites establecidos por la OMS y de la norma colombiana fueron aplicados para el factor del riesgo dado que son más estrictos y con un control mayor para garantizar la salud de la población. Los niveles máximos de contaminantes según la OMS (2021) para los contaminantes estudiados son de 15 µg/m³ para el $PM_{2.5}$ y de 50 µg/m³ para el O_3 (ver Tabla 3-4).

Tabla 3-3: Concentraciones promedio anual $PM_{2.5}$ y promedio anual calculado del promedio diario máximo de 8 horas de O_3 para las estaciones de Bogotá entre los años 2013 y 2022.

Estación de monitoreo	Concentración promedio de $PM_{2.5}$ (µg/m³)	Concentración promedio de Ozono (µg/m³)
Bolivia	15.45	17.30
Carvajal	31.38	17.79
Centro de alto rendimiento	16.11	18.19
Ciudad Bolívar	18.91	15.54
Fontibón	18.53	16.63
Guaymaral	14.17	17.76
Jazmín	15.91	17.76
Kennedy	25.59	17.61
Las Ferias	15.65	16.70
Min Ambiente	14.60	16.97
Puente Aranda	18.45	15.88
San Cristóbal	11.44	16.09
Suba	17.33	17.12
Tunal	19.17	16.64
Usaquén	13.52	17.66
Usme	12.69	17.66

Fuente: Autor

Tabla 3-4: Indicador de exposición para $PM_{2.5}$ y O_3

Indicador		Valor máximo permisible (µg/m³)
OMS, 2021	Contaminante $PM_{2.5}$	15
Resolución 2254 de 2017	Contaminante O_3	50

Fuente: Autor

Con el fin de calcular la matriz de cálculo del índice de exposición se asigna un resultado a partir del factor de riesgo obtenido en cada estación dado por la clasificación (ver Tabla 3-5).

Tabla 3-5: Clasificación del riesgo para exposición.

Nivel de exposición	Resultado factor	Calificación
Alto	> 1.00	3
Medio	0,10- 1.00	2
Bajo	< 0,10	1

Fuente: Autor

A continuación, se llevó a cabo el cálculo del riesgo poblacional clasificado por grupos de edad y la concentración del contaminante expuesto, como se detalla en la Tabla 3-6. Los resultados revelaron consistentemente que los adultos en todas las estaciones de monitoreo presentan los porcentajes más altos de afectación en comparación con otros grupos. Dentro de estas afectaciones, las mujeres son las más perjudicadas, seguidas por los hombres jóvenes y los niños. Aunque estos últimos muestran valores elevados, el riesgo poblacional no es tan alarmante como en el caso de los adultos.

Tabla 3-6: Tabla riesgo poblacional por rango de edad para cada estación

Estaciones de monitoreo	Concentración $PM_{2.5}$ ($\mu g/m^3$)	Riesgo (%) bebés		Riesgo (%) Niños		Riesgo (%) Jóvenes		Riesgo (%) Adultos		Riesgo (%) Adultos de la tercera edad	
		Femenino	Masculino	Femenino	Masculino	Femenino	Masculino	Femenino	Masculino	Femenino	Masculino
Carvajal	31.38	1.10	0.03	3.56	3.72	7.91	8.11	10.37	9.20	2.00	1.41
Centro de alto rendimiento	16.11	0.40	0.43	1.10	1.19	2.98	3.13	4.48	4.19	1.19	0.82
Fontibón	19	0.42	0.50	1.50	1.56	3.72	3.59	5.28	4.45	1.06	0.70
Kennedy	26	0.72	0.74	2.33	2.43	5.17	5.30	6.78	6.01	1.31	0.92
Puente Aranda	18	0.52	0.48	1.38	1.48	3.39	3.73	4.98	4.67	1.27	0.85
Suba	17	0.49	0.52	1.49	1.56	3.44	3.38	4.84	4.04	1.01	0.68
Tunal	19	0.69	0.73	1.86	1.92	3.82	4.06	4.79	4.50	1.00	0.73
Usaquén	14	0.35	0.37	1.04	1.05	2.52	2.37	3.99	3.22	1.07	0.72

Fuente: Autor

Los resultados a partir del factor de riesgo obtenido en cada estación destacan la estación de Carvajal, registrando un índice de exposición alto para el $PM_{2.5}$ y medio para el O_3, posicionándola como una zona de mayor contaminación en Bogotá (ver Tabla 3-7). Esto indica que la población está expuesta a un riesgo significativo debido a la presencia de estos contaminantes.

Tabla 3-7: Resultados del cálculo de factor de exposición y clasificación de riesgo por exposición para cada estación de los contaminantes $PM_{2.5}$ y O_3.

Estación	Factor de exposición		Clasificación de riesgo por exposición	
	$PM_{2.5}$	O_3	$PM_{2.5}$	O_3
Bolivia	0.62	0.35	2	1
Carvajal	1.26	0.36	3	2
CDAR	0.64	0.36	2	2
Ciudad Bolívar	0.76	0.31	2	1
Fontibón	0.74	0.33	2	1
Guaymaral	0.57	0.36	2	2
Jazmin	0.64	0.36	2	2
Kennedy	1.02	0.35	3	2
Las Ferias	0.63	0.33	2	1
Min Ambiente	0.58	0.34	2	1
Puente Aranda	0.74	0.32	2	1
San Cristóbal	0.46	0.32	1	1
Suba	0.69	0.34	2	1
Tunal	0.77	0.33	2	1
Usaquén	0.54	0.35	2	2
Usme	0.51	0.35	2	2

Fuente: Autor

Adicionalmente, se observa que las estaciones de Carvajal, Kennedy y Centro de Alto Rendimiento presentan índices de exposición altos tanto para $PM_{2.5}$ como para ozono,

mientras que Usaquén, Usme, Jazmín y Guaymaral exhiben un riesgo de exposición medio en ambos contaminantes. Por lo que es crucial implementar medidas inmediatas en las estaciones que muestran índices medios y altos para estos contaminantes con respecto a la salud pública de la ciudad.

3.4.1 Impactos a la salud - Mortalidad

Siguiendo la metodología de Ballesteros et al, (2020), se procedió a la selección de los parámetros CRF *β*, tomando como referencia los trabajos de Liu et al. (2019) para estimar la asociación por cada aumento de 10 µg/m³ en la exposición a corto plazo de $PM_{2.5}$. Asimismo, los parámetros utilizados para evaluar la asociación entre la exposición a corto plazo al O_3 y los casos de mortalidad por todas las causas, cardiovascular y respiratoria, se utilizaron a partir del incremento diario de 10 ppb en la concentración de ozono, utilizando los valores proporcionados por el estudio de Bell, Dominici y Samet (2005).

Tabla 3-8: Casos confirmados de mortalidad en Bogotá.

Casos de muerte	Casos confirmados
Todas las causas	60.276
Cardiovascular	10.467
Respiratoria	3.496

Fuente: Dane, 2021

Como línea base de referencia, se emplearon datos detallados de población y mortalidad correspondientes al DANE (ver Tabla 3-8), asegurando una base sólida y actualizada para llevar a cabo las estimaciones asociadas a la contaminación atmosférica en el contexto colombiano, como se observa en la Tabla 3-9.

A partir de los datos, se estimaron los valores de mortalidad para las localidades de Bogotá, revelando un impacto significativo, como se muestra en la Tabla 3-10. La estación Carvajal destacó con los casos más elevados, registrando una mortalidad por todas las causas atribuible a la exposición media anual de $PM_{2.5}$ (β=0.68±0,11 %/10 µg/m3) de 55.58 casos, 14.20 casos adicionales para causas cardiovasculares (β=0.55 ± 0.10 %/10 µg/m³), y 7 casos adicionales para causas respiratorias (β=0.74 ± 0.2

%/10 µg/m3) en la población residencial cercana. Por otro lado, la estación de Kennedy también mostró un impacto significativo, con un total de 42.83 casos adicionales por todas las causas (β=0,68±0,11 %/10 µg/m^3). Ambas estaciones están ubicadas en la localidad de Kennedy, caracterizada por un alto flujo vehicular mixto, un uso de suelo con actividades industriales y una población significativa de más de un millón de residentes, lo que la convierte en la localidad con el índice de contaminación más alto en Bogotá.

En esta localidad, Herrera (2017) realizó una investigación para determinar el valor económico de la calidad del aire y su impacto en la salud de los habitantes, utilizando la función dosis-respuesta basada en los gastos per cápita para atención médica. Los resultados indicaron un costo estimado de aproximadamente $969.48 dólares estadounidenses.

Tabla 3-9: Datos de Incidencia inicial (por cien mil personas) de mortalidad por todas las causas, cardiovascular y respiratoria, y Coeficientes *β* para estimar la asociación por cada aumento en la concentración de contaminantes.

Casos de muerte	Y0 (por cada cien mil personas)	β $PM_{2.5}$	β O_3
Todas las causas	417.7	0.68 ± 0.11 %/10 µg/m³	0.87 ± 0.11 %/10 µg/m³
Cardiovascular	130.2	0.55 ± 0.10 %/10 µg/m³	0.110 ± 0.10 %/10 µg/m³
Respiratoria	50	0.74 ± 0.2 %/ 10 µg/m³	0.47 ± 0.2 %/ 10 µg/m³

Fuente: Autor

En cuanto a las estaciones de Fontibón y Puente Aranda, se observaron valores significativos asociados a la alta concentración media de $PM_{2.5}$ y O_3, generando una estimación de mortalidad por todas las causas de 23 y 25 casos, respectivamente. En específico, se registraron 6 casos cardiovasculares y 3 casos de mortalidad en exceso por casos de enfermedad respiratoria. Este fenómeno puede atribuirse a la ubicación de estas estaciones en la zona Sur Occidente de la ciudad, donde se concentran las mayores cantidades de contaminantes, como ha sido señalado previamente (Acevedo, 2021). A pesar de que ninguna de las estaciones supera la media anual máxima permisible de ozono, se identificaron algunos valores atípicos altos en la estimación del

modelo, indicando la posibilidad de días en los que este contaminante podría afectar a grupos sensibles a cambios en la calidad del aire.

Tabla 3-10: Total de casos de mortalidad excedentes al año 2021 por cada estación.

Estación por Localidad	Población	$\Delta PM_{2.5}$	Total casos de mortalidad Excedentes 2021		
			Todas las causas	Cardiovascular	Respiratoria
		µg/m³	(Casos)	(Casos)	(Casos)
Carvajal	1 092 110	21	55.58	14.20	7.20
Centro de alto rendimiento	891 530	6.1	16.97	4.30	2.21
Fontibón	395 122	8.91	24.56	6.23	3.19
Kennedy	1 092 110	15.91	42.83	10.91	5.55
Puente Aranda	257 242	8.45	23.32	5.91	3.03
Suba	1 124 692	7.33	20.31	5.14	2.64
Tunal	217 139	9.17	25.25	6.40	3.28
Usaquén	503 767	3.52	9.88	2.50	1.29

Fuente: Autor

En el caso de la estación de Suba, a pesar de ser la localidad más poblada y de históricamente superar los niveles de $PM_{2.5}$ (Molina, et al., 2020), se observaron valores relativamente bajos de ΔX asociados al exceso de contaminantes con respecto a la norma (ΔPM2.5). Sin embargo, según la Secretaría Distrital de Salud, durante el periodo comprendido entre 2015 y 2019, las localidades con mayor registro del número medio anual de muertes absolutas debido a enfermedades respiratorias inferiores crónicas (ERIC) en personas menores de 70 años, es la localidad de Kennedy con un total de 28 casos, Suba con 19 casos, y Ciudad Bolívar con 18 casos. Este hallazgo sugiere una discrepancia entre las mediciones de contaminación atmosférica y los impactos reales en la salud respiratoria de la población, respaldando investigaciones previas como la realizada por Morales (2023). En dicho estudio, se evaluó la calidad del aire en Bogotá y su relación con la incidencia en el número de hospitalizaciones por enfermedades respiratorias. Se encontró una incidencia inversamente proporcional entre las concentraciones de material particulado y la aparición de enfermedades respiratorias.

Esto se atribuye a la complejidad de dichas enfermedades y al tiempo que tardan en manifestarse los síntomas en relación con la exposición a PM_{10} y $PM_{2.5}$. Los resultados destacan que, si bien el material particulado no causa directamente estas enfermedades, sí potencia su incidencia.

En cuanto a la estación Tunal, su ubicación en la ciudad cerca de importantes avenidas, como la Caracas, la carrera 51 Venecia, la Carrera 25 Tunal y la Avenida Boyacá, generan un alto flujo de transporte público, camiones y automóviles, además de la actividad industrial de la zona, que se centra en la transformación de materias primas en la industria textil y de cueros. Esto la cataloga como una fuente significativa de emisiones, justificando así el resultado alto de 25 casos estimados de mortalidad por todas las causas en la localidad.

Adicionalmente, es crucial considerar otros factores que podrían contribuir a los altos casos de contaminación en las localidades mencionadas. Uno de estos factores, que se consideró en la realización del modelo de predicción de O_3, es la dirección y velocidad del viento, que ejerce una influencia directa en la calidad del aire. Por ejemplo, en las estaciones de Carvajal, Kennedy, Fontibón y Tunal, prevalecen vientos provenientes del noreste al sureste. Esto significa que el desplazamiento del material particulado se dirige hacia esas zonas, afectando a las localidades circundantes, trayendo los contaminantes de las áreas donde se desarrollan actividades económicas e industriales de los municipios cercanos.

Este primer acercamiento de la estimación de riesgo poblacional e impactos en la salud subraya la necesidad de contar con una mayor resolución temporal en los datos, preferiblemente a nivel semanal o diario. Esta recomendación se fundamenta en investigaciones previas, como la llevada a cabo por Camacho (2021), que evidenció la importancia de una temporalidad más detallada para comparar de manera más precisa el ingreso hospitalario y la presencia de contaminantes atmosféricos en la salud pública de Bogotá.

4. Conclusiones y recomendaciones

Este estudio tuvo como objetivo diseñar un modelo de predicción de O_3 troposférico a corto plazo en la ciudad de Bogotá. Esto se logra mediante la identificación de variables meteorológicas y de contaminación del aire relacionadas con el O_3 que no presentan multicolinealidad. Posteriormente, se estructuró el diseño del modelo predictivo *machine learning* y finalmente se validó el modelo por medio de parámetros estadísticos y categóricos.

En la ciudad de Bogotá, se ha observado una preocupante problemática en cuanto a la calidad del aire, especialmente con la tendencia positiva de los niveles de O_3, lo que tiene impactos significativos en la salud humana y en los ecosistemas. Este aumento se ha evidenciado de manera más notable en los años 2020 y 2021, principalmente debido a las restricciones de movilidad impuestas como medida de control durante la pandemia.

Sin embargo, en los años posteriores, los niveles de ozono siguen siendo superiores a los registrados en los años anteriores a la pandemia. Esto se atribuye a circunstancias relacionadas con la quema de biomasa y el aumento del uso de combustibles de baja calidad, factores que favorecen la producción de O_3. Estos incrementos se han observado en gran parte de la ciudad, especialmente en las estaciones de Bolivia, CDAR, Ciudad Bolívar, Fontibón, Guaymaral, Kennedy, Min Ambiente, Puente Aranda, Tunal, Usme y Móvil Fontibón.

Aunque los niveles de O_3 aún no superan los límites permisibles establecidos por las normativas, es preocupante la tendencia ascendente que se ha mostrado en los últimos años. Por lo tanto, es crucial continuar con el estudio y monitoreo de este contaminante criterio para implementar medidas efectivas de mitigación y protección de la salud pública y del medio ambiente.

Por esta razón, este estudio se enfocó en la generación de un modelo de predicción utilizando ML, donde se combinó el modelo convolucional y LSTM para mejorar su precisión y exactitud. Las variables de entrada del modelo utilizadas fueron NO, NO_2, CO, PM_{10}, $PM_{2.5}$, radiación, rapidez del viento, humedad relativa, temperatura y los valores calculados de los componentes del viento.

La validación del modelo generado se realizó con los datos promedios de 8 horas diarios de los 13.125 datos a partir de julio de 2021 a diciembre de 2022, calculando 9 parámetros estadísticos. Esta validación concluyó que el modelo es capaz de predecir valores precisos, replicando el comportamiento del O_3, obteniendo una correlación de 0,87, y adaptándose a los cambios abruptos que ocurren en el tiempo, como por ejemplo los valores altos presentados los primeros meses del año, teniendo en cuenta las anomalías meteorológicas y los cambios de concentración de los demás precursores, por condiciones externas presentadas en otras regiones, como las quemas de biomasa.

Sin embargo, los resultados obtenidos en la predicción del O_3 y en la validación del modelo mostraron que se sobreestiman los valores versus los valores observados, especialmente en los primeros 60 datos de entrenamiento, que representan 2 días, aproximadamente, y en los picos altos que ocurren a medio día en los momentos de mayor radiación. De este modo, la sobreestimación disminuye con el entrenamiento de varios valores, ajustándose a los valores observados. La dispersión de los resultados y el coeficiente de correlación, mostraron que la sobreestimación es baja, confirmando que el modelo es capaz de reproducir con exactitud los valores de O_3.

Se ha evidenciado en este estudio, así como en otros similares, que los precursores clave para la producción de O_3 incluyen NO, SO_2, la humedad relativa y el NO_2. Sin embargo, en Bogotá aún no se realizan mediciones del precursor NMHC dentro de la RMCAB, a pesar de que se considera clave para mejorar las herramientas de predicción y minimizar los errores del modelo.

Además, se observó que el modelo podría mejorarse mediante la adición de tiempo de entrenamiento. Se constató que, a mayor tiempo de entrenamiento, menores son los errores del modelo. Sin embargo, esto implica que los resultados requerirían más tiempo

de procesamiento, lo cual dependerá de la capacidad de operación del equipo de cómputo disponible.

En definitiva, para futuros estudios sería beneficioso considerar la combinación de la CNN con otras herramientas como el SVM para mejorar los resultados obtenidos. Esta estrategia puede ser especialmente útil al tener en cuenta las particularidades de la ciudad, como las características geográficas, las tendencias de los contaminantes y las condiciones meteorológicas, entre otros factores.

El SVM puede complementar la CNN al ofrecer capacidades de clasificación robustas y una mayor capacidad para manejar conjuntos de datos más pequeños o con características no lineales. Al combinar estas dos técnicas, se pueden aprovechar las fortalezas individuales de cada una para obtener un modelo más completo y preciso.

Además, es importante tener en cuenta que los futuros estudios deberían considerar la adaptación del modelo a las condiciones específicas de la ciudad en la que se aplicará, lo que podría implicar la personalización de ciertos parámetros o la inclusión de características específicas de la localidad en el proceso de entrenamiento del modelo. Esto garantizaría una mejor generalización y aplicabilidad del modelo en entornos del mundo real.

Respecto a la evaluación de riesgo poblacional, para las estaciones de monitoreo en Bogotá se destaca que Carvajal, Kennedy, Tunal y Ciudad Bolívar enfrentan concentraciones elevadas de $PM_{2.5}$ y O_3, siendo los adultos, especialmente mujeres, los más afectados. Por su parte, Carvajal se identifica como un punto crítico, con un fuerte impacto asociado al exceso de la mortalidad por todas las causas, cardiovascular y respiratorias por contaminantes atmosféricos.

A partir de los resultados, se resalta la necesidad urgente de implementar medidas en las estaciones con niveles medios y altos de contaminantes para proteger la salud y el medio ambiente. Además, se destaca la complejidad de los riesgos, afectando no solo a la salud respiratoria sino también a la esperanza de vida y años saludables, incluyendo enfermedades como diabetes. La sugerencia de datos con mayor resolución refuerza la

importancia de mejorar el monitoreo para comprender mejor la relación entre contaminación y salud pública en Bogotá.

Para estudios futuros, es importante enfocarse en mejorar las herramientas que se tienen para los sistemas de alertas de la ciudad, principalmente para los contaminantes criterio de más preocupación, como lo son el material particulado y el O_3. Así mismo, teniendo en cuenta los precursores que se pueden incrementar por hechos externos en regiones aledañas, como lo son los incendios forestales, los cuales han ido en incremento y que no son predecibles, ya que en su mayoría son causados por fuentes antrópicas.

Bibliografía

Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G.S.; Jeffrey Dean, D.; Devin, M.; Ghemawat, S.; et al. (2015) TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. arXiv. Doi: https://doi.org/10.48550/arXiv.1603.04467

Abdul-Wahab S, Al-Alawi S. (2002) Assessment and prediction of tropospheric ozone concentration levels using artificial neural networks. Environ Model Softw 17:219–228. https://doi.org/10.1016/S1364-8152(01)00077-9

Acevedo F. (2020) Influencia de la calidad del aire en la mortalidad y la morbilidad por enfermedades respiratorias en Colombia. http://hdl.handle.net/11634/33256

Adame, J., Gutiérrez-Álvarez, I., Cristofanelli, P., Notario, A., Bogeat, J., López, A., . . . Yela, M. (2022). Surface ozone trends over a 21-year period at El Arenosillo observatory (Southwestern Europe). Atmospheric Research, Vol. 269. doi: https://doi.org/10.1016/j.atmosres.2022.106048

Agatonovic-Kustrin, S. & Beresford, R. (2000). Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. Journal of Pharmaceutical and Biomedical Analysis. doi: https://doi.org/10.1016/S0731-7085(99)00272-1

Agencia Europea de Medio Ambiente. (2020). Contaminación atmosférica. Obtenido de https://www.eea.europa.eu/es/themes/air/intro

Agency, U. E. (s.f.). Systems Quality Assurance Handbook for Air Pollution Measurement (Vol. Vol II). Ambient Air Quality Monitoring Program.

Albawi, S., A. Mohammed, T. & Al-Zawi, S. (2017). Understanding of a convolutional neural network. 2017 International Conference on Engineering and Technology (ICET). doi: 10.1109/ICEngTechnol.2017.8308186

Alcaldía Mayor de Bogotá D. C. (2021). Plan de Ordenamiento Territorial Bogotá Reverdece 2022-2035. Proyecto de Acuerdo. https://bogota.gov.co/mi-ciudad/pot-bogota-reverdece-2022-2035/articulado-del-pot-bogota-reverdece-2022-2035

Ali, J., Khan, R., Ahmad, N. & Maqsood, I. (2012). Random Forests and Decision Trees. IJCSI International Journal of Computer Science Issues. ISSN (Online): 1694-0814

Aljarah, I., Faris, H., & Mirjalili, S. (2019). Evolutionary Machine Learning Techniques: Algorithms and Applications. Springer Nature.

AlOmar, M., Hameed, M. & AlSaadi, M. (2020). Multi hours ahead prediction of surface ozone gas concentration: Robust artificial intelligence approach. Atmospheric Pollution Research. doi: https://doi.org/10.1016/j.apr.2020.06.024.

Alpaydin, E. (2016). Machine Learning: The New AI. Cambridge, MA, USA. Obtenido de https://search-ebscohost-com.ezproxy.unal.edu.co/login.aspx?direct=true&db=nlebk&AN=1369420&lang=es&site=eds-live

Baena-Salazar, D., Jimenez, J., Zapata, C., & Ramirez-Cardona, A. (2019). Red neuronal artificial aplicado para el pronóstico de eventos críticos de PM2.5 en el Valle de Aburrá. Dyna rev.fac.nac.minas, ISSN 0012-7353. doi: https://doi.org/10.15446/dyna.v86n209.63228

Ballesteros-González, K., Sullivan, A., & Morales-Betancourt, R. (2020). Estimating the air quality and health impacts of biomass burning in northern South America using a chemical transport model. Science of the Total Environment. doi: https://doi.org/10.1016/j.scitotenv.2020.139755

Banja, M., Papanastasiou, D., Poupkou, A., & Melas, D. (2012). Development of a short-term ozone prediction tool in Tirana area based on meteorological variables. doi:10.5094/APR.2012.002

Baquero, J. (2023). Análisis de la variabilidad estacional, interanual y tendencia del ozono (O_3) y sus precursores, así como la dinámica oxidativa de la atmósfera en Bogotá entre 2010 y 2020. Universidad Nacional de Colombia. Repositorio: https://repositorio.unal.edu.co/handle/unal/84018

Barten, j., Ganzeveld, L., Visser, A., Jiménez, R., & Krol, C. (2020). Evaluation of nitrogen oxides (NOEvaluation of nitrogen oxides (NOx) sources and sinks and ozone production in Colombia and surrounding areas. Atmos. Chem. Phys. doi: https://doi.org/10.5194/acp-20-9441-2020

Bhattacharya, S., & Shahnawaz, S. (2021). Using Machine Learning to Predict Air Quality Index in New Delhi. doi:arXiv:2112.05753

Bigi, A. & Ghermandi, G. (2016). Trends and variability of atmospheric PM2.5 and PM_{10} – 2.5 concentration in the Po Valley, Italy. Atmos. Chem. Phys. doi: 10.5194/acp-16-15777-2016

Bolaño-Diaz, S., Camargo-Caicedo, Y., Tovar-Bernal, F. & Bolaño-Ortiz, T.R. (2022). The Effect of Forest Fire Events on Air Quality: A Case Study of Northern Colombia. Fire. doi:. https://doi.org/10.3390/fire5060191

Brockwell, P.J. & Davis, R.A. (2002). State-Space Models. In: (eds) Introduction to Time Series and Forecasting. Springer Texts in Statistics. Springer. doi: https://doi.org/10.1007/0-387-21657-X_8

Butković, V., Cvitaš, T., & Klasinc, L. (1990). Photochemical ozone in the mediterranean. Science of The Total Environment, Vol. 99. doi.org/10.1016/0048-9697(90)90219-K

Butt, E. W., Turnock, S. T., Rigby, R., Reddington, C. L., Yoshioka, M., Johnson, J. S., et al. (2017). Global and regional trends in particulate air pollution and attributable health burden over the past 50 years. Environmental Research Letters, 12(10), 104017. https://doi.org/10.1088/1748-9326/aa87be

Burnett, R., Chen, H., Szyszkowicz, M., Fann, N., Hubbell, B., Pope, C. A., et al. (2018). Global estimates of mortality associated with long-term exposure to outdoor fine particulate matter. Proceedings of the National Academy of Sciences of the United States of America, 115(38), 9592–9597. https://doi.org/10.1073/pnas.1803222115

Burnett, R. T., Pope, C. A., III, Ezzati, M., Olives, C., Lim, S. S., Mehta, S., et al. (2014). An integrated risk function for estimating the global burden of disease attributable to ambient fine particulate matter exposure. Environmental Health Perspectives, 122(4), 397–403. https://doi.org/10.1289/ehp.1307049

Camacho N (2021) Efectos de la temperatura y los contaminantes atmosféricos sobre la salud pública en Bogotá, Colombia. https://www.researchgate.net/profile/Natalia-Camacho-Castaneda/publication/354611482_Efectos_de_la_temperatura_y_los_contaminantes_atmosfericos_sobre_la_salud_publica_en_Bogota_Colombia/links/61424ce328667828a8981607/Efectos-de-la-temperatura-y-los-contaminantes-atmosfericos-sobre-la-salud-publica-en-Bogota-Colombia.pdf

Cañada Torrecilla, M. R. (2021). El riesgo de contaminación por ozono en dos ciudades españolas (Madrid y Sevilla): Un estudio realizado con técnicas de modelado espacial y SIG. Geographicalia, 73, 195-212.

Casallas, A., Cabrera, A., Guevara-Luna, M-A., Tompkins, A., González, Y., Aranda, J., Belalcazar, L., Mogollon-Sotelo, C., Celis, N., Lopez-Barrera, E., Peña-Rincon, C., Ferro, C. (2024). Air pollution analysis in Northwestern South America: A new Lagrangian framework. Science of The Total Environment. doi: https://doi.org/10.1016/j.scitotenv.2023.167350

Casallas, A., Castillo-Camacho, M., Sanchez, E., González, Y., Celis, N., Mendez-Espinosa, J., Belalcazar, L. & Ferro, C. (2023). Surface, Satellite Ozone Changes in Northern South America During Low Anthropogenic Emission Conditions: A Machine Learning Approach. SSRN. doi: http://dx.doi.org/10.2139/ssrn.4016140

Casallas, A., Córdoba, T., Sánchez-Cárdenas, L., Guevara-Luna, M., & Belalcázar, L. (2022a). Understanding the atmospheric characteristics of high polluted. EIA. Esc. Ing. Antioq. doi: https://doi.org/10.13140/RG.2.2.22694.65606

Casallas, A., Ferro, C., Celis, N., Guevara-Luna, M., Mogollón-Sotelo, C., Guevara-Luna, F., & Merchán, M. (2021). Long short-term memory artificial neural network approach to forecast meteorology and PM2.5 local variables in Bogotá, Colombia. Model. Earth Syst. Environ. doi: https://doi.org/10.1007/s40808-021-01274-6

Casallas, A., Ferro, C., Celis, N., Mogollon-Sotelo, C., & Belalcazar, L. (2022b). Design of a 3D convolutional neural network model to forecast PM2.5 and O_3 concentrations in Colombia. IEEE Xplore. doi:10.1109/CASAP54985.2021.9703414

Casallas, S., Celis, N., Ferro, C., López, E., Peña, C., Corredor, J., & Ballen, M. (2020). Validation of PM_{10} & PM2.5 early alert in Bogotá, Colombia, through the modeling software WRF-CHEM. Environmental Science and Pollution Research. doi:10.1007/s11356-019-06997-9

Casas, N. (2017). Deep Deterministic Policy Gradient for Urban Traffic Light Control. Cornell University. arXiv:1703.09035. https://doi.org/10.48550/arXiv.1703.09035

Castillo-Camacho, M.P., Tunarrosa-Grisales, I.C., Chacon-Rivera, L.M., Guevara-Luna, M. A. & Belalcazar-Cerón, L.C. (2020). Personal Exposure to PM2.5 in the Massive Transport System of Bogotá and Medellín, Colombia. Asian J. Atmos. Environ. https://doi.org/10.5572/ajae.2020.14.3.210

Castro, M., & Pires, J. (2019). Decision support tool to improve the spatial distribution of air quality monitoring sites. Atmospheric Pollution Research, Vol. 10. doi: https://doi.org/10.1016/j.apr.2018.12.011

Celis, N., Casallas, A., López, E., Martínez, H., Peña, C., Arenas, R., & Ferro, C. (2022). Design of an early alert system for PM2.5 through a stochastic method and machine learning models. Environmental Science & Policy, vol. 127. doi: https://doi.org/10.1016/j.envsci.2021.10.030

Chaparro, L., Bustos, J., Barrera, E., & Casallas, A. (2022). Evaluation of Recurring Neuronal Network For Prediction of Risk For Sulfur Oxides (SOx) in The City Of Bogotá. IEEE Xplore. doi:10.1109/CASAP54985.2021.9703418

Chai, T., Kim, H., Lee, P., Tong, D., Pan, L., Tang, Y., Huang, J., McQueen, J., Tsidulko, M. & Stajner, I. (2013). Evaluation of the United States National Air Quality Forecast Capability experimental real-time predictions in 2010 using Air Quality System ozone and NO_2 measurements. Geoscientific Model Development. doi:10.5194/gmd-6-1831-2013

Chen, Y., Chen, X., Xu, A., Sun, Q. & Peng, X. (2022). A hybrid CNN-Transformer model for ozone concentration prediction. Air Qual Atmos Health. https://doi-org.ezproxy.unal.edu.co/10.1007/s11869-022-01197-w

Cheng, M., Fang, F., Navon, I., Zheng, J., Zhu, J. & Pain, C.(2023). Assessing uncertainty and heterogeneity in machine learning-based spatiotemporal ozone prediction in Beijing-Tianjin- Hebei region in China. Science of The Total Environment. https://doi.org/10.1016/j.scitotenv.2023.163146.

Chiarvetto, L., Rey, F., & Brignole, N. (2008). Aplicación de Redes neuronales artificiales para la predicción de la calidad de aire. Mecánica Computaciona, vol. XXVII, pp. 3607-3625. Obtenido de https://cimec.org.ar/ojs/index.php/mc/article/viewFile/1654/1618

Cleveland, R. B., Cleveland, W. S., McRae, J., and Terpenning, I. (1990). STL: a seasonal-trend decomposition procedure based on Loess. J. Off. Stat. https://www.nniiem.ru/file/news/2016/stl-statistical-model.pdf

Chowdhury, P. H., He, Q., Carmieli, R., Li, C., Rudich, Y., & Pardo, M. (2019). Connecting the oxidative potential of secondary organic aerosols with reactive oxygen species in exposed lung cells. Environmental Science & Technology, 53(23), 13949–13958. https://doi.org/10.1021/acs.est.9b04449

Chowdhury, S., Pozzer, A., Dey, S., Klingmueller, K., & Lelieveld, J. (2020). Changing risk factors that contribute to premature mortality from ambient air pollution between 2000 and 2015. Environmental Research Letters, 15(7), 074010. https://doi.org/10.1088/1748-9326/ab8334

Clifton, J. & Laber, E. (2020). Q-Learning: Theory and Applications. Annual Review of Statistics and Its Application. https://doi.org/10.1146/annurev-statistics-031219-041220

Cooper, O., Parrish, D., Stohl, A., Trainer, M., Nédélec, P., Thouret, V., . . . Avery, M. (2009). Increasing springtime ozone mixing ratios in the free troposphere over western North America. Nature. doi:doi.org/10.1038/nature08708

Chollet, F. (2015). Keras. Disponible online: https://keras.io

Curto, J. D., & Pinto, J. C. (2011). The corrected VIF (CVIF). Journal of Applied Statistics. http://explore.bl.uk/primo_library/libweb/action/display.do?

Deroubaix, A., Brasseur, G., Gaubert, B., Labuhn, I., Menut, L., Siour, G. & Tuccella, P. (2021). Response of surface ozone concentration to emission reduction and meteorology during the COVID-19 lockdown in Europe. Royal Meteorological Society. https://doi.org/10.1002/met.1990

Dunner, A., Iglesias Álamos, V., & Ruiz Rudolph, P. (2018). El ozono troposférico como modificador del efecto de material particulado 2,5 sobre muertes diarias cardiovasculares en adultos mayores. Santiago, 2009-2013. https://www.researchgate.net/publication/327060086_El_ozono_troposferico_como_modif

icador_del_efecto_de_material_particulado_25_sobre_muertes_diarias_cardiovasculares_en_adultos_mayores_Santiago_2009-2013

EPA. (2007). Un resumen de la Ley de Aire Limpio. Publicación N.° EPA-456/K-07-001.

EPA. (2021). EPA Partners with Tribes to Deploy Air Sensors in Communities. Obtenido de https://www.epa.gov/sciencematters/epa-partners-tribes-deploy-air-sensors-communities

European Environment Agency. (2008). Medio Ambiente Europeo - La Evaluación Dobris.

European Environment Agency. (2016). Air quality in Europe 2016 report. Obtenido de http://www.eea.europa.eu/publications/air-quality-in-europe-2016

Fernández, S., Sánchez, J. & Córdoba, A. (2002). Estadística descriptiva. ESIC Editorial. ISBN: 84-7356-306-9

Fiore, A., Naik, V. & Leibensperger, E. (2015). Air Quality and Climate Connections. Journal of the Air & Waste Management Association. doi: 10.1080/10962247.2015.104052

Franceschi, F., Cobo, M., & Figueredo, M. (2018). Discovering relationships and forecasting PM_{10} and $PM_{2.5}$ concentrations in Bogotá, Colombia, using Artificial Neural Networks, Principal Component Analysis, and k-means clustering. Atmospheric Pollution Research. doi: https://doi.org/10.1016/j.apr.2018.02.006

Fu, TM., Zheng, Y., Paulot, F., Mau, J. & Yantosca, R. (2015). Positive but variable sensitivity of August surface ozone to large-scale warming in the southeast United States. Nature Clim Change. doi: https://doi-org.ezproxy.unal.edu.co/10.1038/nclimate2567

Ghahremanloo, M., Lops, Y., Choi, Y., & Yeganeh, B. (2021). Deep Learning Estimation of Daily Ground-Level NO_2 Concentrations From Remote Sensing Data. J. Geophys. Res Atmos. doi: https://doi-org.ezproxy.unal.edu.co/10.1029/2021JD034925

Fukushima, K. (1975). Cognitron: A self-organizing multilayered neural network. Biol. Cybern. doi: https://doi.org/10.1007/BF00342633

Ghahremanloo M, Lops Y, Choi Y, Jung J, Mousevinezhad S & Hammond D (2022) A comprehensive study of the COVID-19 impact on PM2.5 levels over the contiguous United States: a deep learning approach. Atmos Environ 118944. https://doi.org/10.1016/j.atmosenv.2022.118944

Gonzalez, C.M., Ynoue, R.Y., Vara-Vela, A., Rojas, N.Y. & Aristizabal, B.H. (2018). High-resolution air quality modeling in a medium-sized city in the tropical Andes: Assessment of local and global emissions in understanding ozone and PM_{10} dynamics. Atmospheric Pollution Research. doi: https://doi.org/10.1016/j.apr.2018.03.003

Goodfellow, I., Bengio, Y. & Courville, A. (2016). Deep Learning. https://www.deeplearningbook.org/

Guzmán, D., Ruíz, J.F. & Cadena, M. (2014). Regionalización de Colombia según la estacionalidad de la precipitación media mensual, a través análisis de componentes principales (ACP). IDEAM. Bogotá, Colombia. http://www.ideam.gov.co/documents/21021/21141/Regionalizacion+de+la+Precipitacion+Media+Mensual/1239c8b3-299d-4099-bf52-55a414557119

Hahad, O., Frenis, K., Kuntic, M., Daiber, A., & Münzel, T. (2021). Accelerated aging and age-related diseases (CVD and neurological) due to air pollution and traffic noise exposure. International Journal of Molecular Sciences, 22(5), 2419. https://doi.org/10.3390/ijms22052419

Henao, J., Rendón, A., Hernández, K., Giraldo-Ramirez, P., Robledo, V., Posada, J., . . . Mejia, J. (2021). Differential Effects of the COVID-19 Lockdown and Regional Fire on the Air Quality of Medellín, Colombia. Atmosphere. doi: https://doi.org/10.3390/atmos12091137

Hernandez-Deckers D (2021) Features of atmospheric deep convection in northwestern South America obtained from infrared satellite data. Quart J Royal Meteor Soc. https://doi.org/10.1002/qj.4208

Herrera, S. (2017). Valor económico de la calidad del aire en la localidad de Kennedy medido desde la salud de los habitantes. Universidad de los Andes. https://repositorio.uniandes.edu.co/entities/publication/de137266-925e-4b81-a76c-af190ae9369b

Huang, C.-J., & Kou, P.-H. (2018). A Deep CNN-LSTM Model for Particulate Matter (PM2.5) Forecasting in Smart Cities. Sensors . doi: https://doi.org/10.3390/s18072220

Inter Agency for Research on Cance. Outdoor air pollution. IARC. (2016). Monographs on the Evaluation of Carcinogenic Risks to Humans (Vol. Vol. 109). Lyon.

Ismail, M., Suroto, A., Ibrahim, T., & Abdullah, A. (2011). Tropospheric ozone trend in the Muda Irrigation Area, Kedah. Journal of Physical. Science. doi:web.usm.my/jps/22-2-11/22.2.7.pdf

Jiménez, P. (2019). Estudio de la contaminación. Elearning, S.L., Jin, Z., Yan, D., Zhang, Z., Li, M., Wang, T., Huang, X., Xie, M., Li, S., Zhuang, B. (2023). Effects of Elevated Ozone Exposure on Regional Meteorology and Air Quality in China Through Ozone-Vegetation Coupling. JGR Atmosfpheres. doi: https://doi-org.ezproxy.unal.edu.co/10.1029/2022JD038119

Kapadia, D. & Jariwala, N. (2022). Prediction of tropospheric ozone using artificial neural network (ANN) and feature selection techniques. Model. Earth Syst. Environ. doi: https://doi-org.ezproxy.unal.edu.co/10.1007/s40808-021-01220-6

Ke, Li., Jacob, D., Liao, H., Shen, L., Zhang, Q. & Bates, K. (2018). Anthropogenic drivers of 2013–2017 trends in summer surface ozone in China. PNAS. https://doi.org/10.1073/pnas.1812168116

Khalil, A., Barhoom, A., Abu-Nasser, B., Musleh, M. & Abu-Naser, S. (2019). Energy Efficiency Prediction using Artificial Neural Network. International Journal of Academic Pedagogical Research. https://philpapers.org/rec/KHAEEP

Kingman, D.P. & Ba, J. (2014) Adam: A Method for Stochastic Optimization. arXiv arXiv:1412.6980. https://doi.org/10.48550/arXiv.1412.6980

Kiranyaz, S., Avci, O., Abdeljaber, O., Ince, T., Gabbouj, M. & Inmanf, D.(2021). 1D convolutional neural networks and applications: A survey. Mechanical Systems and Signal Processing. https://doi.org/10.1016/j.ymssp.2020.107398.

Kodinariya, T. & Makwana, P. (2013). Review on determining number of Cluster in K-Means Clustering. International Journal of Advance Research in Computer Science and Management Studies. ISSN: 2321-7782

Kumar, J., Goomer, R. & Singh, A. (2018). Long Short Term Memory Recurrent Neural Network (LSTM-RNN) Based Workload Forecasting Model For Cloud Datacenters. Procedia Computer Science. https://doi.org/10.1016/j.procs.2017.12.087.

Kumar, A., Jiménez, R., Belalcázar, L. C., & Rojas, N. Y. (2016). Application of WRF-Chem Model to Simulate PM_{10} Concentration over Bogota. Aerosol and Air Quality Research. doi:10.4209/aaqr.2015.05.0318

Lee, H. & Park, R. (2022). Factors determining the seasonal variation of ozone air quality in South Korea: Regional background versus domestic emission contributions. Environmental Pollution. doi: https://doi.org/10.1016/j.envpol.2022.119645

Lelieveld, J., Evans, J. S., Fnais, M., Giannadaki, D., & Pozzer, A. (2015). The contribution of outdoor air pollution sources to premature mortality on a global scale. Nature, 525(7569), 367–371. https://doi.org/10.1038/nature15371

Lelieveld, J., Pozzer, A., Pöschl, U., Fnais, M., Haines, A., & Münzel, T. (2020). Loss of life expectancy from air pollution compared to other risk factors: A worldwide perspective. Cardiovascular Research, 116(11), 1910–1917. https://doi.org/10.1093/cvr/cvaa025

Liu, C., Chen, R., Sera, F., Vicedo-Cabrera, A. M., Guo, Y., Tong, S., ... García, S. O. (2019). Contaminación del aire ambiente por partículas y mortalidad diaria en 652 ciudades. DOI: 10.1056/NEJMoa1817364

Liu, F., Beirle, S., Zhang, Q., Dörner, S., He, K., & Wagner, T. (2016). NOX lifetimes and emissions of cities and power plants in polluted background estimated by satellite observations. Atmospheric Chemistry and Physics, 16(8), 5283–5298. https://doi.org/10.5194/acp-16-5283-2016

Liu, Y., Wang, T., Stavrakou, T., Elguindi, N., Doumbia, T., Granier, C., Bouarar, I., Gaubert, B. & Brasseur, G. (2021). Diverse response of surface ozone to COVID-19 lockdown in China. Science of The Total Environment. https://doi.org/10.1016/j.scitotenv.2021.147739.

MADS. (2010). Protocolo para el monitoreo y seguimiento de la calidad del aire.

Mendez-Espinosa, J. F., Belalcazar, L. C., & Betancourt, R. M. (2019). Regional Air Quality Impact of Northern South America Biomass Burning Emissions. Atmospheric Environment. doi:10.1016/j.atmosenv.2019.01.042

Mendez-Espinosa, J. F., Rojas, N., Vargas, J., Patchón, J., Belalcazar, L. & Ramírez, O. (2020). Air quality variations in Northern South America during the COVID-19 lockdown. Science of The Total Environment. doi: https://doi.org/10.1016/j.scitotenv.2020.141621

Mogollón-Sotelo, C., Casallas, A., Vidal, S., Celis, N., Ferro, C., & Belalcazar, L. (2021). A support vector machine model to forecast ground-level PM2.5 in a highly populated city with a complex terrain. Air Qual Atmos Health, vol. 14, pp. 399-409. doi:https://doi.org/10.1007/s11869-020-00945-0

Moghani, M. & Archer, C. (2020). The impact of emissions and climate change on future ozone concentrations in the USA. Air Quality, Atmosphere and Health. doi: https://doi.org/10.1007/s11869-020-00900-z

Monks, P., Archibald, A., Colette, A., Cooper, O., Coyle, M., Derwent, R., . . . Williams, M. (2015). Tropospheric ozone and its precursors from the urban to the global scale from air quality to short-lived climate forcer. Atmos. Chem. Phys.(15, 8889–8973). doi:10.5194/acp-15-8889-2015

Morales O (2023) Evaluación de la calidad de aire en Bogotá y su relación con la incidencia en el número de hospitalizaciones por enfermedades respiratorias.
chrome-
extension://efaidnbmnnnibpcajpcglclefindmkaj/https://repositorio.unbosque.edu.co/bitstream/handle/20.500.12495/11447/Trabajo%20de%20grado.pdf?sequence=1&isAllowed=y

Murray, C. J. L., Aravkin, A. Y., Zheng, P., Abbafati, C., Abbas, K. M., Abbasi-Kangevari, M., et al. (2020). Global burden of 87 risk factors in 204 countries and territories, 1990–2019: A systematic analysis for the Global Burden of Disease Study 2019. Lancet, 396(10258), 1223–1249. https://doi.org/10.1016/S0140-6736(20)30752-2

Ndiaye, E., Le, T., Fercoq, O., Salmon, J., & Takeuchi, I. (2019). Safe Grid Search with Optimal Complexity. Proceedings of the 36th International Conference on Machine Learning, in Proceedings of Machine Learning Proceedings of the 36th International Conference on Machine Learning. https://proceedings.mlr.press/v97/ndiaye19a.html

Noble, W. (2006). What is a support vector machine?. Nat Biotechnol. doi: https://doi.org/10.1038/nbt1206-1565

O'Shea, K. & Nash, R. (2015). An Introduction to Convolutional Neural Networks. Cornell University. https://doi.org/10.48550/arXiv.1511.08458 arXiv:1511.08458

Parrish, D., Millet, D., & Goldstein, A. (2009). Increasing ozone in marine boundary layer inflow at the west coasts. Atmospheric Chemistry and Physics. doi: doi.org/10.5194/acp-9-1303-2009

Peters, R., Ee, N., Peters, J., Booth, A., Mudway, I., & Anstey, K. J. (2019). Air pollution and dementia: A systematic review. Journal of Alzheimer's Disease, 70(s1), 145–163. https://doi.org/10.3233/JAD-180631

Pires, J., & Martins, F. (2011). Correction methods for statistical models in tropospheric ozone forecasting. Atmospheric Pollution Research. doi:10.1016/j.atmosenv.2011.02.011

Prechelt , L. (1998). Early Stopping - But When? Lecture Notes in Computer Science. doi: https://doi.org/10.1007/3-540-49430-8_3

Pope, C. A., Ezzati, M., & Dockery, D. W. (2009). Fine-particulate air pollution and life expectancy in the United States. New England Journal of Medicine, 360(4), 376–386. https://doi.org/10.1056/NEJMsa0805646

Pozzer, A., Anenberg, S. C., Dey, S., Haines, A., Lelieveld, J., & Chowdhury, S. (2022). Mortalidad atribuible a la contaminación del aire ambiente: una revisión de las estimaciones mundiales. https://doi.org/10.1029/2022GH000711

Prabhakaran, D., Mandal, S., Krishna, B., Magsumbol, M., Singh, K., Tandon, N., et al. (2020). Exposure to particulate matter is associated with elevated blood pressure and incident hypertension in urban India. Hypertension. https://www.ahajournals.org/doi/full/10.1161

Prechelt, L. (2002). Early Stopping - But When? LNCS, V 1524. doi: 10.1007/3-540-49430-8_3

Ramírez, O., Sánchez de la Campa, A., Amato, F., Moreno, T., Silva, L. & de la Rosa, J. (2019). Physicochemical characterization and sources of the thoracic fraction of road dust in a Latin American megacity. Science of The Total Environment. doi: https://doi.org/10.1016/j.scitotenv.2018.10.214

Río, A (2020). Cambio climático, la penalidad del ozono y la mortalidad asociada en la Ciudad de México. http://hdl.handle.net/11651/4342

Roderick, M., MacGlashan, J. & Tellex, S. (2017). Implementing the Deep Q-Network. Cornell University. arXiv:1711.07478. doi: https://doi.org/10.48550/arXiv.1711.07478

Rojas, N. (2007). Aire y problemas ambientales de Bogotá. Recuperado de: Microsoft Word - Aire y problemas ambientales de Bogotá.doc (bogota.gov.co)

Rojas, N., 2004. Revisión de las emisiones de material particulado por la combustion de Diesel y Biodiesel. Revista de Ingeniería. Disponible: http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S0121-49932004000200007&lng=en&tlng=es

Russo, F., Biancardo , S., Formisano , A., & Dell'Acqua, G. (2018). Predicting percent air voids content in compacted bituminous hot mixture specimens by varying the energy laboratory compaction and the bulk density assessment method. Construction and Building Materials. doi: https://doi.org/10.1016/j.conbuildmat.2017.12.174

Sayeed, A., Choi, Y., Eslami, E., Jung, J., Lops, Y., Khan, A., . . . Choi, M.-H. (2021). A novel CMAQ-CNN hybrid model to forecast hourly surface-ozone concentrations 14 days in advance. Scientific Reports, vol. 806. doi: https://doi.org/10.1016/j.scitotenv.2021.150149

Sayeed, A., Choi, Y., Eslami, E., Lops, Y., Roy, A., & Jung, J. (2020). Using a deep convolutional neural network to predict 2017 ozone concentrations, 24 hours in advance. Neural Networks. doi: doi.org/10.1016/j.neunet.2019.09.033

Sayeed A, Lops Y, Choi Y, Jung J, Salman AK (2021a) Bias correcting and extending the PM forecast by CMAQ up to 7 days using deep convolutional neural networks. Atmos Environ 253:118376. https://doi.org/10.1016/j.atmosenv.2021.118376

Secretaría Distrital de Ambiente. (2023). Informe anual de calidad del aire de Bogotá año 2022. Red de Monitoreo de Calidad del Aire de Bogotá D.C RMCAB. http://rmcab.ambientebogota.gov.co/Pagesfiles/Informe%20anual%202022.pdf

Secretaría Distrital de Ambiente, Dirección de Control Ambiental Subdirección de Calidad del Aire, Auditiva y Visual-SCAAV Red de Monitoreo de Calidad del Aire de Bogotá - RMCAB, (2021). Informe anual de calidad del aire en Bogotá 2020. Recuperado de: 120721Informe-Anual-de-Calidad-del-Aire-Ano-2020-v2 (1).pdf

Seinfeld, J., & Pandis, S. (2016). Atmospheric Chemistry and Physics: From Air Pollution to Climate Change (3rd Edition ed., Vol. Chapter 6). Wiley. Obtenido de http://www.ideam.gov.co/web/tiempo-y-clima/ozono-troposferico

Shewalkar, A., Nyavanandi, D & Ludwig, S. (2019). Performance Evaluation of Deep Neural Networks Applied to Speech Recognition: RNN, LSTM and GRU. Journal of Artificial Intelligence and Soft Computing Research. https://doi.org/10.2478/jaiscr-2019-0006

Singh, V., Singh, S. & Biswal, A. (2021). Exceedances and trends of particulate matter (PM2.5) in five Indian megacities. Sci. Total Environ. doi: https://doi.org/10.1016/j.scitotenv.2020.141461

Sokhi, R., Singh, V., Querol, X., Finardi, S., Targino, A., Andrade, M., . . . Zavala, M. (2021). A global observational analysis to understand changes in air quality during exceptionally low anthropogenic emission conditions,. Environment International. doi: doi.org/10.1016/j.envint.2021.106818

Sundermeyer, M., Schlüter, R. & Ney, H. (2012). LSTM Neural Networks for Language Modeling. In Thirteenth annual conference of the international speech communication association

Suthaharan, S. (2015). Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning. Springer

Telikani, A., Tahmassebi, A., Banzhaf, W., & Gandomi, A. (2021). Evolutionary Machine Learning: A Survey. ACM Computing Surveys, vol. 54(no. 8). doi: https://doi.org/10.1145/3467477

Turner, M. C., Andersen, Z. J., Baccarelli, A., Diver, W. R., Gapstur, S. M., Pope, C. A., et al. (2020). Outdoor air pollution and cancer: An overview of the current evidence and public health recommendations. CA: A Cancer Journal for Clinicians, 70(6), 460–479. https://doi.org/10.3322/caac.21632

Ultsch, A. (1993). Self-Organizing Neural Networks for Visualisation and Classification. In: Opitz, O., Lausen, B., Klar, R. (eds) Information and Classification. Studies in Classification, Data Analysis and Knowledge Organization. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-50974-2_31

University Colles London, Universidad de los Andes. (2013). Caracterización de la contaminación atmosférica en Colombia. Reino Unido - Colombia. Obtenido de https://prosperityfund.uniandes.edu.co/site/wp-content/uploads/Caracterizaci%C3%B3n-de-la-contaminaci%C3%B3n-atmosf%C3%A9rica-en-Colombia.pdf

Valencia, R., Sanchez, G., & Díaz I. (2016). A general regression neural network for modeling the behavior of PM_{10} concentration level in Santa Marta, Colombia. ARPN J Eng Appl Sci, ISSN 1819-6608

Von Schneidemesser, E., Monks, P., Allan, J., Bruhwiler, L., Forster, P., Fowler, D., . . . Sutton, M. (2015). Chemistry and the linkages between air quality and climate change. Chem. Rev. doi: https://doi.org/10.1021/acs.chemrev.5b00089

Wang, J., Wang, D., Ge, B., Lin, W., Ji, D., Pan, X., . . . Wang, Z. (2022). Increase in daytime ozone exposure due to nighttime accumulation in a typical city in eastern China during 2014–2020. Atmospheric Pollution Research, Vol. 13. doi: doi.org/10.1016/j.apr.2022.101387

Wang, SC. (2003). Artificial Neural Network. In: Interdisciplinary Computing in Java Programming. The Springer International Series in Engineering and Computer Science. https://doi.org/10.1007/978-1-4615-0377-4_5

World Health Organization. (2021). WHO global air quality guidelines: particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide. World Health Organization. https://iris.who.int/handle/10665/345329

World Health Organization. (2018). Burden of disease from ambient air pollution for 2016. World Health Organization. Obtenido de https://www.who.int/airpollution/data/AAP_BoD_results_May2018_final.pdf?ua=1

Wong, C. M., Tsang, H., Lai, H. K., Thomas, G. N., Lam, K. B., Chan, K. P., et al. (2016). Cancer mortality risks from long-term exposure to ambient fine particle. Cancer Epidemiology, Biomarkers & Prevention, 25(5), 839–845. https://doi.org/10.1158/1055-9965.EPI-15-0626

Wu, Y., & Feng, J. (2017). Development and Application of Artificial Neural. Network. Wireless Personal Communications. doi:10.1007/s11277-017-5224-x

Wu, K., Yang, X., Chen, D., Gu, S., Lu, Y., Jiang, Q., Wang, F., Ou, Y., Qian, Y., Shao, P. & Lu, S. (2020). Estimation of biogenic VOC emissions and their corresponding impact on ozone and secondary organic aerosol formation in China. Atmospheric Research. doi:10.1016/j.atmosres.2019.104656

Yafouz, A., Ahmed, A., Zaini, N., & El-Shafie, A. (2021). Ozone Concentration Forecasting Based on Artificial Intelligence Techniques: A Systematic Review. Water, Air, & Soil Pollution. doi:10.1007/s11270-021-04989-5

Yafouz, A., AlDahoul, N., Birima, A., Ahmed, A., Sherif, M., Sefelnasr, A., . . . Elshafie, A. (2022). Comprehensive comparison of various machine learning algorithms for short-term ozone concentration prediction. Alexandria Engineering Journal, Vol. 61. doi: https://doi.org/10.1016/j.aej.2021.10.021

Yilmaz, A. (2021). Ozone level prediction with Machine Learning Algorithms. Journal of Aeronautics and Space Technologies, vol. 14(no. 2), pp. 177 – 183. Obtenido de https://orcid.org/0000-0003-0038-7519

Zambrano-Monserrate, M., Ruano, M., & Sanchez-Alcalde, L. (2020). Indirect effects of COVID-19 on the environment. Science of the Total Environment. doi: https://doi.org/10.1016/j.scitotenv.2020.138813

Zheng, S., Schlink, U., Ho, K. F., Singh, R. P., & Pozzer, A. (2021). Distribución espacial de la mortalidad prematura relacionada con PM 2,5 en China. https://doi.org/10.1029/2021GH000532

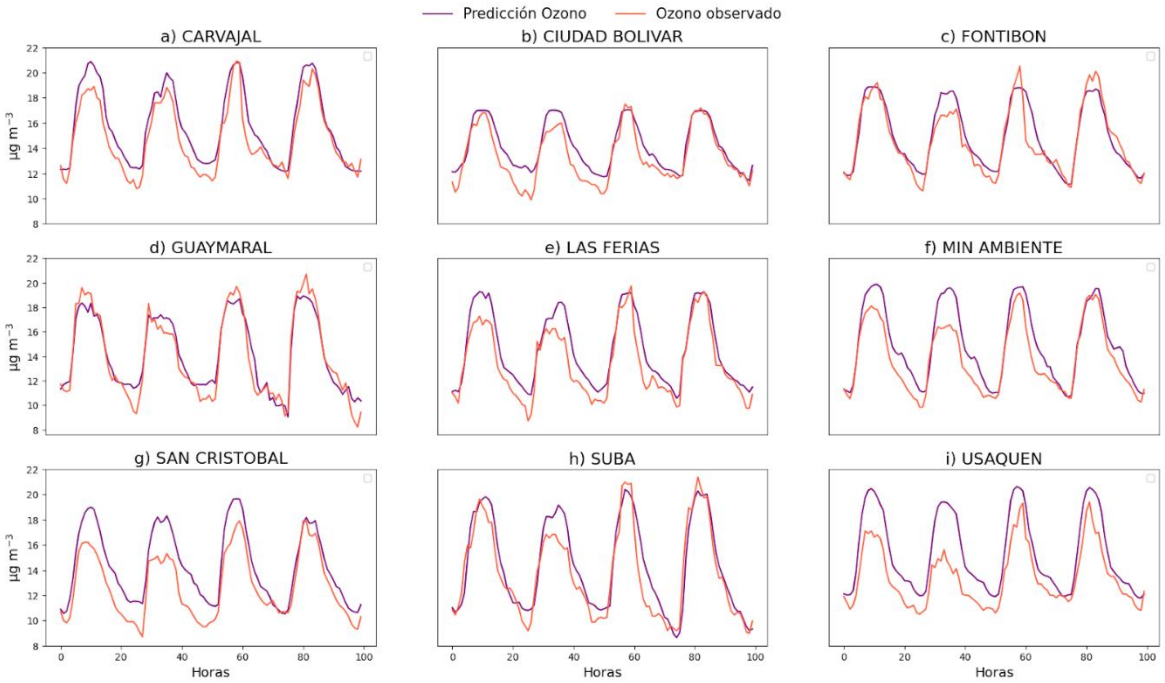
Zhang, J., Sun, L., Rainham, D., Dummer, T., Wheeler, A., Anastasopolos, A., . . . Johnson, M. (2021). Predicting intraurban airborne PM1.0-trace elements in a port city: Land use regression by ordinary least squares and a machine learning algorithm. Science of the Total Environment, vol. 806. doi: https://doi.org/10.1016/j.scitotenv.2021.150149

Zhang, Z., He, H.-D., Yang, J.-M., Wang, H.-W., Xue, Y., & Peng, Z.-R. (2022). Spatiotemporal evolution of NO_2 diffusion in Beijing in response to COVID-19 lockdown using complex network. Chemosphere. doi: https://doi.org/10.1016/j.chemosphere.2022.133631

Zihao, Z., Zhan, L., & Wang, Y. (2019). Outlier removal based on Chauvenet's criterion and dense disparity refinement using least square support vector machine. Journal of Electronic Imaging. doi:10.1117/1.JEI.28.2.023028

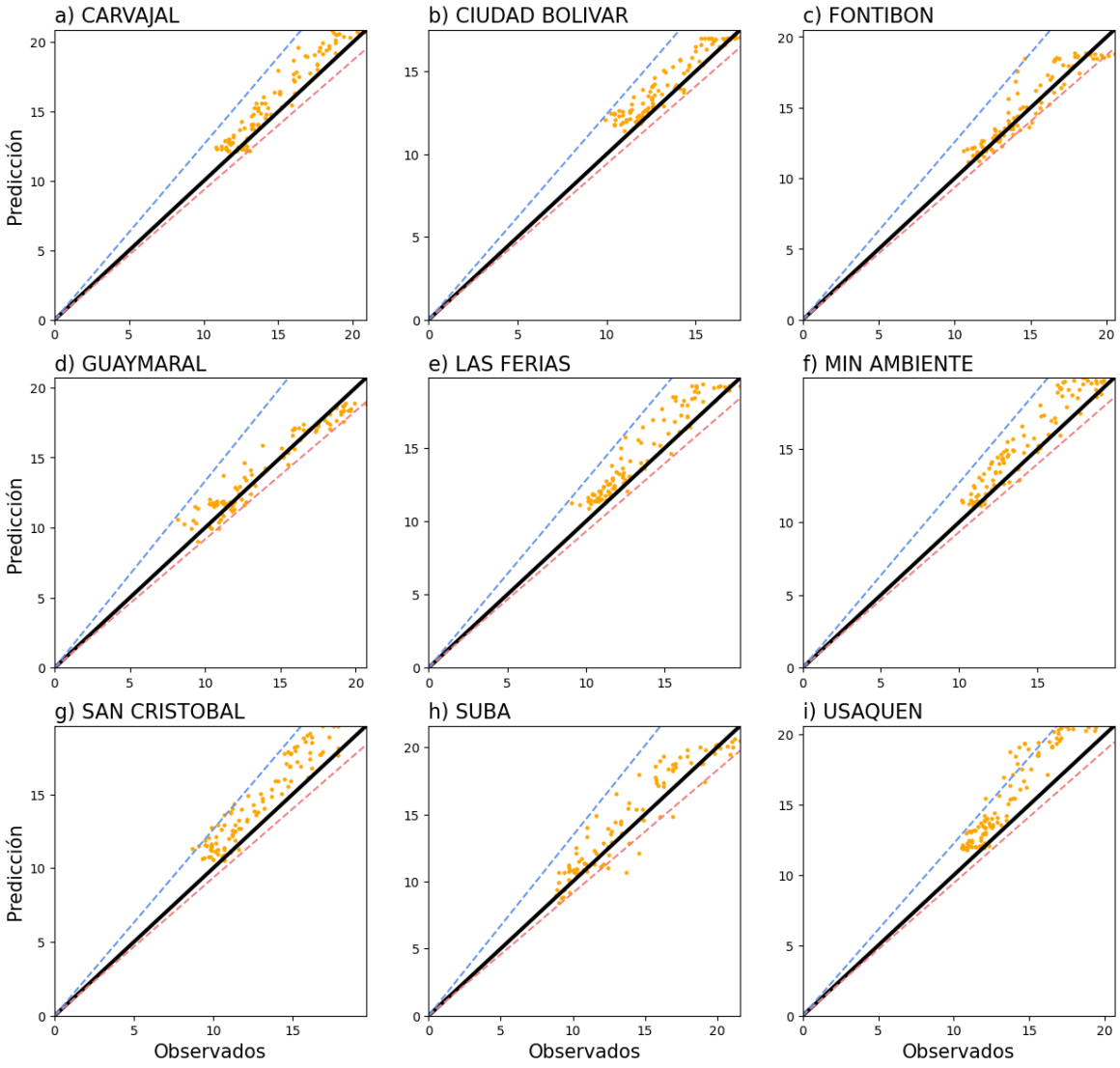
ANEXOS

Anexo 1: Valores de observación vs predicción de ozono para las estaciones que no realizan mediciones de todas las variables necesarias como entradas en el modelo.



Fuente: Autor

Anexo 2: Diagrama de dispersión de observación vs predicción de ozono de estaciones que no realizan mediciones de todas las variables necesarias como entradas en el modelo. La línea negra corresponde a la regresión lineal 1:1, las líneas punteadas azul y rojas corresponden a 1:2 y 1:0.5 respectivamente.



Fuente: Autor