Predicción de Contaminantes Atmosféricos en Bogotá utilizando Redes LSTM

Christian Alejandro Sarmiento Sánchez
Escuela Colombiana de Carreras Industriales (ECCI), Colombia
cs407083@gmail.com

Ingenio Tecnológico
vol. 6, e051, 2024
Universidad Tecnológica Nacional, Argentina
ISSN-E: 2618-4931
ingenio@frlp.utn.edu.ar

Recepción: 09 noviembre 2024
Aprobación: 12 diciembre 2024

Resumen: Esta investigación presenta el desarrollo e implementación de modelos basados en redes LSTM para predecir los niveles de contaminantes atmosféricos en Bogotá, utilizando datos de la estación meteorológica de Las Ferias, parte de la Red de Monitoreo de Calidad del Aire de Bogotá (RMCAB). Se recopilaron y analizaron datos entre 2021 y 2023, sumando más de 29,200 registros, que fueron empleados para entrenar, validar y probar los modelos. Se desarrollaron dos enfoques: el primero, un modelo univariable de predicción single-step para contaminantes específicos como PM10, PM2.5, CO y NO2, y el segundo, un modelo multivariable que integra variables meteorológicas como la dirección y velocidad del viento, temperatura, humedad, presión y precipitación. El rendimiento de ambos enfoques se evaluó utilizando el error cuadrático medio (RMSE), comparando las predicciones con mediciones reales. Los resultados muestran que el modelo multivariable ofreció mejores predicciones debido a la inclusión de factores atmosféricos adicionales, destacando su capacidad para mejorar la precisión en las estimaciones de la calidad del aire. Esto resalta la relevancia de considerar múltiples variables en la predicción de contaminantes, especialmente en contextos urbanos donde las interacciones entre factores ambientales son complejas.

Palabras clave: Predicción, contaminantes atmosféricos, LSTM, calidad del aire, series temporales, aprendizaje profundo.

Abstract: This research presents the development and implementation of LSTM-based models to predict atmospheric pollutant levels in Bogotá, using data from the Las Ferias meteorological station, part of Bogotá's Air Quality Monitoring Network (RMCAB). Data from 2021 to 2023, totaling over 29,200 records, were collected and analyzed, and used to train, validate, and test the models. Two approaches were developed: first, a single-step univariate prediction model for specific pollutants such as PM10, PM2.5, CO, and NO2; second, a multivariate model integrating meteorological variables such as wind direction and speed, temperature, humidity, pressure, and precipitation. The performance of both approaches was evaluated using root mean square error (RMSE), comparing predictions with real measurements. Results indicate that the multivariate model offered superior predictions due to the inclusion of additional atmospheric

AmeliCA
PDF generado a partir de XML-JATS4R
Modelo de publicación sin fines de lucro para conservar la naturaleza académica y abierta de la comunicación científica

factors, highlighting its ability to enhance accuracy in air quality estimations. This underscores the importance of considering multiple variables in pollutant prediction, particularly in urban contexts where interactions among environmental factors are complex.

Keywords: Prediction, air pollutants, LSTM, air quality, time series, Deep learning.

AmeliCA
PDF generado a partir de XML-JATS4R
Modelo de publicación sin fines de lucro para conservar la naturaleza académica y abierta de la comunicación científica

1. Introducción

La contaminación atmosférica representa un problema de salud pública cada vez más preocupante en las grandes ciudades (ONU, 2022), impulsado por el crecimiento de la población, el aumento del tráfico vehicular y las emisiones de fuentes industriales. Bogotá, la capital de Colombia, enfrenta esta problemática con niveles de contaminación que superan los umbrales recomendados para proteger la salud humana, particularmente en contaminantes como PM10 y PM2.5 (Guerrero-Rojas, 2020). Esto ha motivado la necesidad de desarrollar herramientas predictivas que permitan anticipar picos de contaminación y tomar medidas preventivas.

En este contexto, los modelos de aprendizaje profundo, como las redes de memoria a corto y largo plazo (LSTM, por sus siglas en inglés), han demostrado ser eficientes para el análisis de datos secuenciales complejos (Fang et al., 2021). Las LSTM son capaces de capturar patrones no lineales y temporales en conjuntos de datos extensos, lo que las convierte en una herramienta adecuada para la predicción de la calidad del aire (Zou et al., 2021).

Este estudio propuso el diseño e implementación de modelos LSTM para predecir los niveles de contaminantes atmosféricos en Bogotá, utilizando los datos de la estación meteorológica de Las Ferias, parte de la Red de Monitoreo de Calidad del Aire de Bogotá (RMCAB). Se comparan dos enfoques: un modelo univariable y single-step para la predicción de contaminantes específicos como PM10, PM2.5, CO y NO2, y un modelo multivariable y single-step que incorpora variables meteorológicas adicionales como la dirección y velocidad del viento, temperatura, humedad, presión y precipitación.

El objetivo del estudio fue evaluar la precisión y eficacia de ambos enfoques en la predicción de contaminantes atmosféricos, proporcionando una herramienta valiosa para la gestión de la calidad del aire en Bogotá. Los resultados podrían servir como base para la toma de decisiones por parte de las autoridades locales y la formulación de políticas públicas orientadas a mitigar la contaminación.

1.1. Conceptos relacionados

Para llevar a cabo el estudio de predicción de contaminantes atmosféricos utilizando modelos LSTM, es esencial definir ciertos conceptos clave que proporcionan un marco teórico sólido. En esta sección, se presenta el contexto conceptual y las características distintivas que hacen de las redes LSTM una herramienta adecuada para este tipo de predicciones.

1.2. Contaminantes Atmosféricos

Los contaminantes atmosféricos son sustancias del aire que pueden afectar a la salud humana y al medio ambiente. Estas sustancias pueden ser químicas o partículas suspendidas (Cantú, 2023). Una de las formas más comunes de medir la calidad del aire es mediante el Índice de Calidad del Aire (ICA), que clasifica la concentración de contaminantes en un rango de 0 a 500, con cinco niveles de riesgo, como se muestra en la figura 1. Cabe señalar que el ICA se calcula individualmente para cada contaminante, por lo que los valores pueden variar en función del tipo y la concentración de cada sustancia (Ameer et al., 2019).

Figura 1.
Rangos del índice de calidad del aire
Nota: Adaptado de Resolución 2254 de 2017 del Ministerio de ambiente.

En Colombia, el ICA se calcula siguiendo las directrices establecidas por el (Ministerio de Ambiente y Desarrollo Sostenible, 2017), que define los parámetros y los tiempos máximos de exposición permitidos para cada contaminante, como se ilustra en la figura 2. Este marco normativo proporciona una guía para evaluar la calidad del aire y gestionar los riesgos asociados a la contaminación atmosférica. Los principales contaminantes considerados en este estudio son las partículas en suspensión (PM10 y PM2.5), el monóxido de carbono (CO) y el dióxido de nitrógeno (NO2).

AmeliCA
PDF generado a partir de XML-JATS4R
Modelo de publicación sin fines de lucro para conservar la naturaleza académica y abierta de la comunicación científica

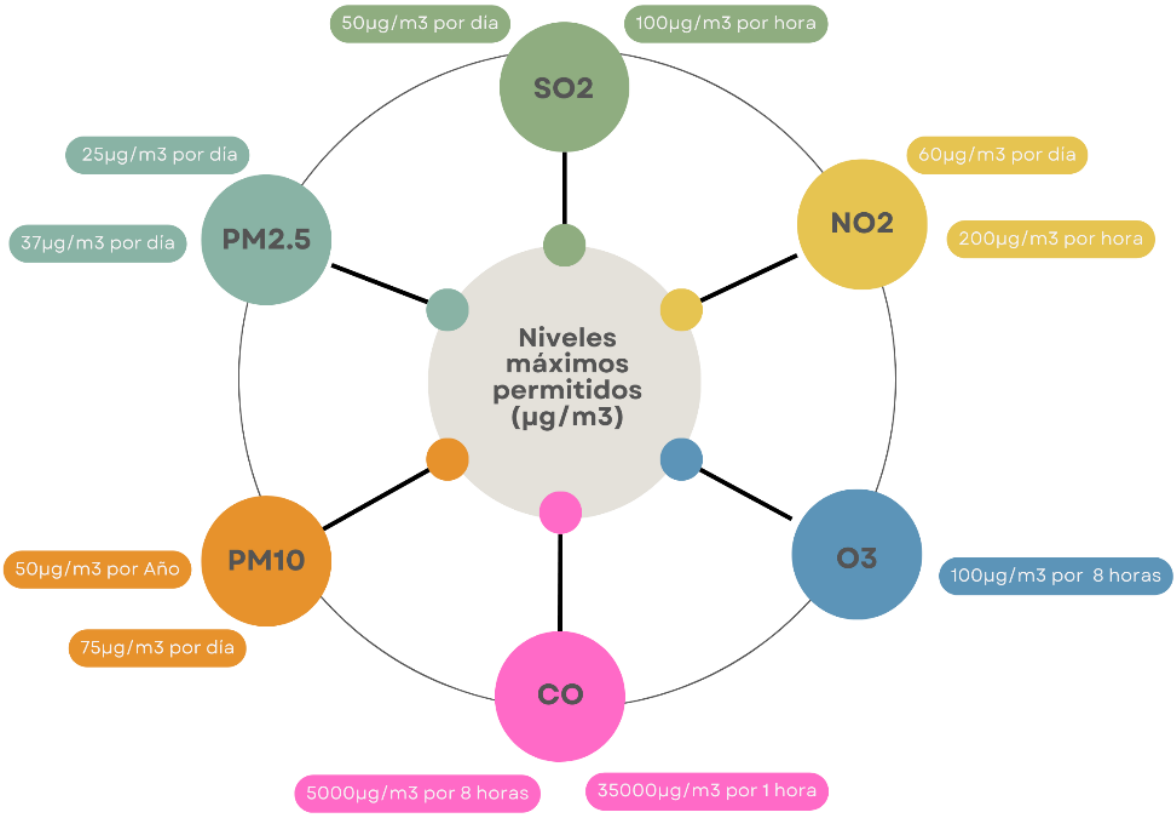

Figura 2.
Niveles máximos permitidos de contaminantes atmosféricos
Nota: Adaptado de Resolución 2254 de 2017 del Ministerio de ambiente.

Entre los principales contaminantes que afectan la calidad del aire se encuentran las partículas en suspensión (PM10 y PM2.5), el monóxido de carbono (CO) y el dióxido de nitrógeno (NO2), cada uno con características particulares y efectos sobre la salud humana y el medio ambiente.

- PM2.5 y PM10: son términos que se refieren a partículas en suspensión en el aire, pero se diferencian por su tamaño. PM10 incluye partículas con un diámetro de 10 micrómetros o menos, mientras que PM2.5 abarca partículas de 2.5 micrómetros o menos. Las partículas PM10 y PM2.5 difieren en tamaño y, por lo tanto, en su capacidad para penetrar el sistema respiratorio. PM2.5, debido a su menor tamaño, representa un mayor riesgo para la salud, ya que puede llegar a los pulmones y al torrente sanguíneo, incrementando las probabilidades de enfermedades respiratorias y cardiovasculares. PM10, aunque más grande, también es perjudicial, ya que puede irritar las vías respiratorias y contribuir al desarrollo de enfermedades como el asma (Zheng et al., 2018). Ambas formas de partículas se originan de la quema de combustibles fósiles, procesos industriales y otras actividades humanas, y su presencia en el aire está vinculada con problemas de salud pública y un mayor riesgo de mortalidad prematura (Sharma et al., 2020).

- CO: el monóxido de carbono (CO) es un gas incoloro, inodoro e insípido que se produce principalmente por la combustión incompleta de combustibles fósiles como gasolina y carbón. Este contaminante es altamente tóxico, ya que interfiere con la capacidad de la sangre para transportar oxígeno, causando síntomas como dolores de cabeza, mareos y, en casos extremos, la muerte (Bao & Zhang, 2020). Las fuentes comunes de CO incluyen vehículos de motor, calefacción a base de combustibles y maquinaria industrial. Para proteger la salud pública, es fundamental monitorear los niveles de CO en el aire y establecer regulaciones estrictas para reducir las emisiones de este gas.
- NO2: El dióxido de nitrógeno (NO2) es un gas de color marrón rojizo que se forma principalmente por la combustión de combustibles fósiles en vehículos y plantas industriales. Este gas no solo irrita el sistema respiratorio y agrava enfermedades como el asma, sino que también contribuye a problemas ambientales como el smog y la lluvia ácida, lo que resalta la necesidad de controlar sus emisiones (Bao & Zhang, 2020). El NO2 también contribuye a la formación de smog y lluvia ácida, ambos con implicaciones ambientales negativas. Dada su capacidad para afectar la calidad del aire y la salud, las regulaciones y monitoreo de sus niveles son esenciales para proteger a la población y al medio ambiente. Reducir las emisiones de NO2 implica cambios en las fuentes de energía, mejoras en la eficiencia de los motores de combustión y la adopción de tecnologías limpias (Ruggieri et al., 2021).

El monitoreo de estos contaminantes es esencial para mejorar la calidad del aire y reducir los riesgos asociados tanto para la salud pública como para el medio ambiente. Este estudio se centrará en predecir sus niveles mediante el uso de modelos LSTM, lo que podría proporcionar información valiosa para la toma de decisiones.

1.3. Datos Meteorológicos

Los datos meteorológicos juegan un papel crucial en la dispersión y acumulación de contaminantes atmosféricos (Han et al., 2020). La tabla 1 sintetiza el papel que juegan las condiciones climáticas con relación de los contaminantes mencionados.

Tabla 1.
Impacto de condiciones meteorológicas en la calidad del aire.

Condición meteorológica	Factor positivo	Factor negativo
Velocidad del viento	Mayor dispersión de contaminantes.	Puede llevar contaminantes a otras áreas.
Dirección del viento	Puede alejar contaminantes de áreas urbanas.	Puede arrastrar contaminantes a zonas sensibles.
Temperatura	Mayor circulación de aire, reduciendo acumulación.	Favorece la formación de ozono troposférico y el smog.
Humedad	Puede reducir partículas en suspensión.	Aumenta la acumulación de contaminantes.
Presión atmosférica	Baja presión favorece dispersión de contaminantes.	Alta presión puede atrapar contaminantes.
Precipitaciones	Limpia el aire de partículas y gases.	Puede trasladar contaminantes al suelo y agua.

Nota: construido mediante información obtenida de Sofowote et. al (2021).

Dado que estas condiciones ambientales afectan de manera directa o indirecta la calidad del aire, estos datos son elementos clave para tener en cuenta al diseñar redes LSTM para correlacionar estas variables y prever resultados concretos.

1.4. Características de un LSTM

Las redes Long Short-Term Memory (LSTM) son una variante avanzada de las redes neuronales recurrentes (RNN) diseñadas principalmente para abordar el problema del desvanecimiento y explosión del gradiente que suele presentarse en las RNN tradicionales (Brownlee, 2017). Gracias a esta capacidad, las LSTM han sido ampliamente utilizadas para procesar y analizar secuencias temporales complejas, lo que las convierte en una herramienta ideal para la predicción de contaminantes atmosféricos.

Una unidad LSTM consta de una celda de memoria, la cual es responsable de almacenar información a lo largo del tiempo, y tres puertas clave: la puerta de entrada (input gate), la puerta de olvido (forget gate) y la puerta de salida (output gate). Estas puertas controlan el flujo de información dentro y fuera de la célula de memoria, lo que permite a la LSTM mantener, modificar o descartar información según sea necesario. A continuación, se describen las funciones de cada puerta.

- Puerta de Entrada (Input Gate): regula la cantidad de nueva información que se permite en la célula de memoria. Utiliza una función de activación sigmoidea para decidir qué información es relevante, combinándola luego con el estado de memoria actual.
- Puerta de Olvido (Forget Gate): determina qué parte de la información almacenada debe olvidarse. A través de una función sigmoidea, esta puerta decide qué información se retiene o descarta, lo que es fundamental para evitar la acumulación de datos irrelevantes.
- Puerta de Salida (Output Gate): controla qué información almacenada en la célula de memoria se utiliza para la salida de la LSTM. Al igual que las otras puertas, emplea una función sigmoidea para decidir qué información debe ser enviada como resultado final.

Cabe mencionar que estas puertas poseen sus respectivos pesos ("input weight" y "output weight"), los cuales permiten ponderar la entrada y salida en su estado actual (time step), utilizando el estado interno (Internal State) para realizar la ponderación (Sak et al., 2014). Además, es importante considerar factores externos a la estructura básica de la red, ya que las LSTM dependen de diversos hiperparámetros que permiten ajustar y optimizar su rendimiento. En este estudio, se configuraron cuidadosamente estos parámetros para mejorar la precisión en la predicción de contaminantes atmosféricos. La tabla 2 presenta un resumen de los hiperparámetros utilizados en los modelos desarrollados.

Tabla 2.
Hiperparámetros de un modelo LSTM

Hiperparámetro	Descripción	Impacto en el Modelo
Número de Unidades LSTM	Número de celdas LSTM por capa. Puede variar según el diseño del modelo.	Aumentar el número de unidades puede mejorar la capacidad del modelo para capturar patrones, pero también aumenta el riesgo de sobreajuste y el costo computacional.

Número de Capas LSTM	Cantidad de capas LSTM en la red.	Más capas pueden aumentar la capacidad para aprender estructuras complejas, pero también incrementan la complejidad del modelo y el tiempo de entrenamiento.
Tasa de Aprendizaje (Learning Rate)	Velocidad con la que el optimizador ajusta los pesos durante el entrenamiento.	Tasas altas pueden conducir a una convergencia rápida, pero también a inestabilidad; tasas bajas son más estables, pero más lentas.
Tamaño del Lote (Batch Size)	Número de ejemplos utilizados en cada actualización durante el entrenamiento.	Un tamaño de lote grande puede mejorar la estabilidad y la precisión, mientras que uno pequeño puede permitir adaptaciones más rápidas, pero a veces con mayor variabilidad.
Función de Activación	Tipo de función utilizada para transformar la salida de una capa.	LSTM generalmente utiliza activaciones como ReLU o tanh, que pueden afectar el aprendizaje y la estabilidad del modelo.
Regularización	Métodos para prevenir el sobreajuste, como Dropout y L2 regularization.	Dropout desactiva aleatoriamente un porcentaje de unidades durante el entrenamiento para mejorar la generalización. L2 regularization ayuda a controlar el crecimiento de los pesos.
Optimización	Método utilizado para ajustar los pesos del modelo durante el entrenamiento.	Optimizadores como Adam y RMSprop son comunes en LSTM por su eficacia para manejar el entrenamiento en series temporales.
Funciones de Pérdida	Métrica utilizada para evaluar la diferencia entre predicciones y resultados reales durante el entrenamiento.	El error cuadrático medio (MSE) es común para modelos LSTM, proporcionando una medida clara de la precisión de las predicciones.

Nota: información recolectada de TensorFlow (2024).

2. Metodología

2.1. Set de datos

Los datos para este estudio se obtuvieron de la Red de Monitoreo de Calidad del Aire de Bogotá (RMCAB), específicamente de la estación de Las Ferias. Tras revisar las veinte estaciones operativas, se seleccionó esta estación debido a que disponía de la mayor cantidad de datos relacionados con contaminantes y condiciones atmosféricas. Además, se observó que la disponibilidad de datos en esta estación era superior en comparación con otras estaciones.

Particularmente, se recopilaron datos desde 2021 hasta 2023, con cerca de 29,200 registros, que se obtuvieron con una periodicidad de una hora. Dado que las redes LSTM requieren secuencias temporales continuas, era imperativo que el set de datos no contuviera valores nulos, ya que la secuencia se vería interrumpida. Por este motivo, el primer paso fue realizar un análisis exploratorio para validar la secuencialidad de los registros.

Esto incluyó:

1. Verificar que los registros estuvieran organizados cronológicamente, del más antiguo al más reciente.
2. Comprobar que no hubiera registros duplicados en la secuencia y, de ser necesario, corregirlos.
3. Identificar y cuantificar el número de registros faltantes por variable.

La tabla 3 presenta los resultados de esta validación.

Tabla 3.
Validación de estado de registros

Variable	registros vacíos	registros vacíos en secuencia
PM10	1978	76
PM2.5	91	85
CO	2723	730
OZONO	808	109
NO	1381	134
NO2	1383	134
NOX	1385	134
Velocidad del viento	570	212
Dirección del viento	460	102
Temperatura	312	76
Humedad	312	76
Precipitación	463	142
Presión barométrica	315	76

Nota: Elaboración propia

Al analizar la tabla 3, se concluye que la disponibilidad de datos de la estación para el intervalo determinado es de aproximadamente el 89.63%, con al menos treinta días consecutivos sin registros (para el caso del CO siendo este el más crítico de todos). Este hallazgo sugiere una oportunidad de mejora en la cobertura y disponibilidad del sistema, aspecto que podría abordarse en futuras investigaciones. Para completar los datos faltantes, se aplicó una interpolación lineal utilizando la ecuación 1, lo que permitió estimar los registros ausentes mediante una recta que conecta los puntos conocidos. Aunque el comportamiento real de los contaminantes no sigue una línea recta, se optó por este método debido a que otras técnicas de interpolación introducían valores atípicos y picos pronunciados, lo que podría afectar negativamente el proceso de escalamiento de los datos.

$$y = \frac{x - x_1}{x_2 - x_1}(y_2 - y_1) + y_1 \quad (1)$$

Una vez completados los datos, se realizó una validación gráfica para identificar posibles valores atípicos. La figura 3 muestra la representación gráfica de todas las variables, lo que permitió visualizar si existían anomalías o comportamientos inusuales en los registros interpolados.

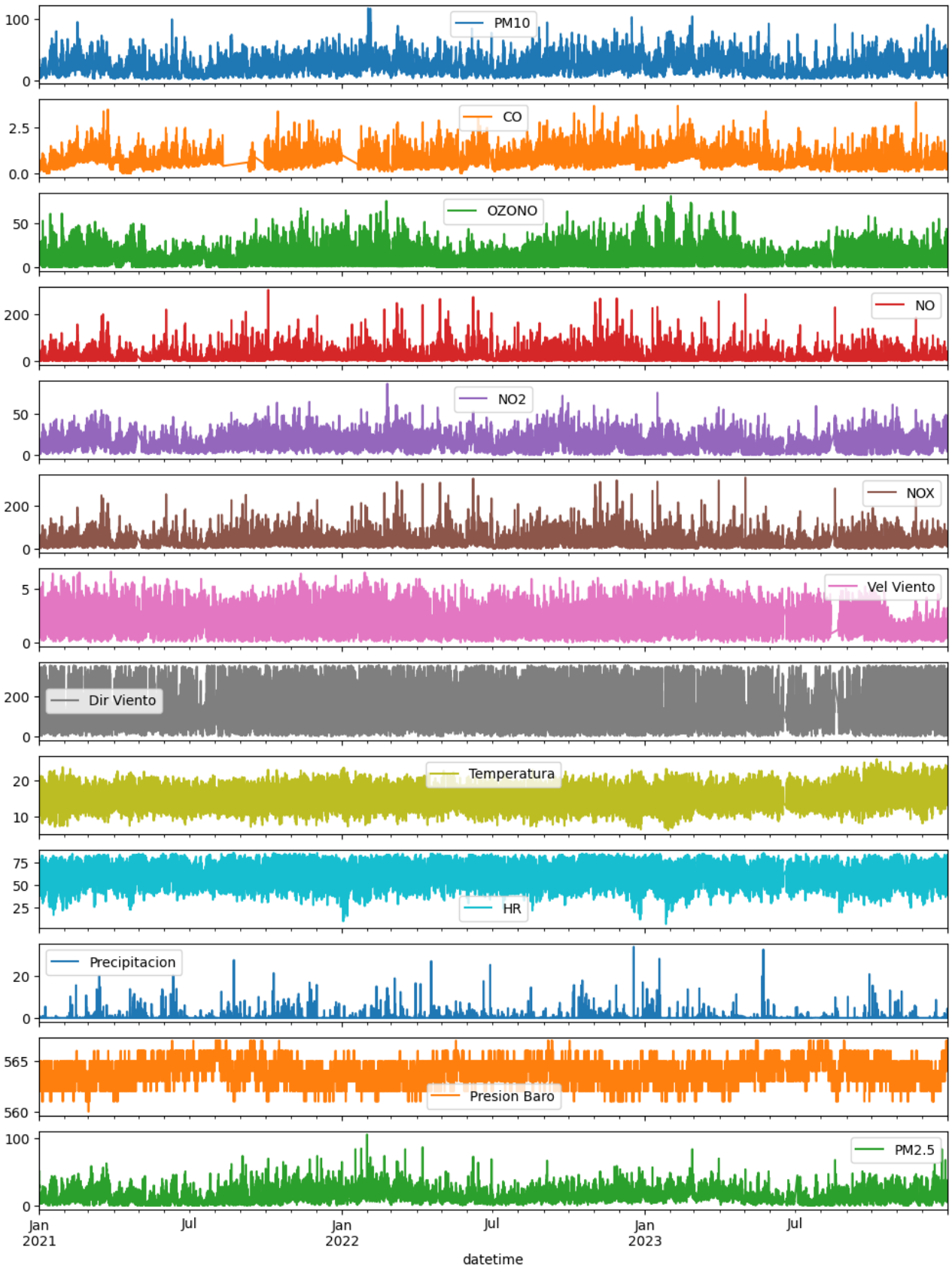
PM10
CO
OZONO
NO
NO2
NOX
Vel Viento
Dir Viento
Temperatura
HR
Precipitacion
Presion Baro
PM2.5
Jan 2021
Jul
Jan 2022
Jul
Jan 2023
Jul
datetime

Figura 3.
Validación grafica de las variables
Nota: Elaboración propia.

Al analizar la Figura 3, se observa que no se identificaron valores atípicos en ninguna de las variables, salvo en aquellos registros generados por la interpolación, siendo más evidente en la variable de CO debido a la mayor cantidad de datos faltantes. Esta situación probablemente incremente el error durante la validación del modelo y pone de manifiesto la necesidad de mejorar la cobertura y disponibilidad de datos del RMCAB.

2.2. Preprocesamiento de datos

Con el set de datos ajustado se procedió a formatear los datos de acuerdo con la estructura de un LSMT, en primera medida se distribuyó el conjunto de datos en tres subconjuntos, en la tabla 4 se visualiza esta distribución.

Tabla 4.
Distribución de los datos

Conjunto de datos	Número de registros
Set de datos de entrenamiento	21024
Set de datos de validación	2628
Set de datos de prueba	2628

Nota: Elaboración propia

Dado que la red LSTM pertenece al grupo de modelos de aprendizaje supervisado, el primer subconjunto de datos se destinó al entrenamiento del modelo, el segundo se utilizó para validar dicho entrenamiento, y el tercero, compuesto por datos no conocidos por el modelo, sirvió para evaluar su comportamiento final. Esta división permite comprobar la ausencia de sobreajuste (overfitting). La Figura 4 muestra la distribución de los datos conforme a esta metodología.

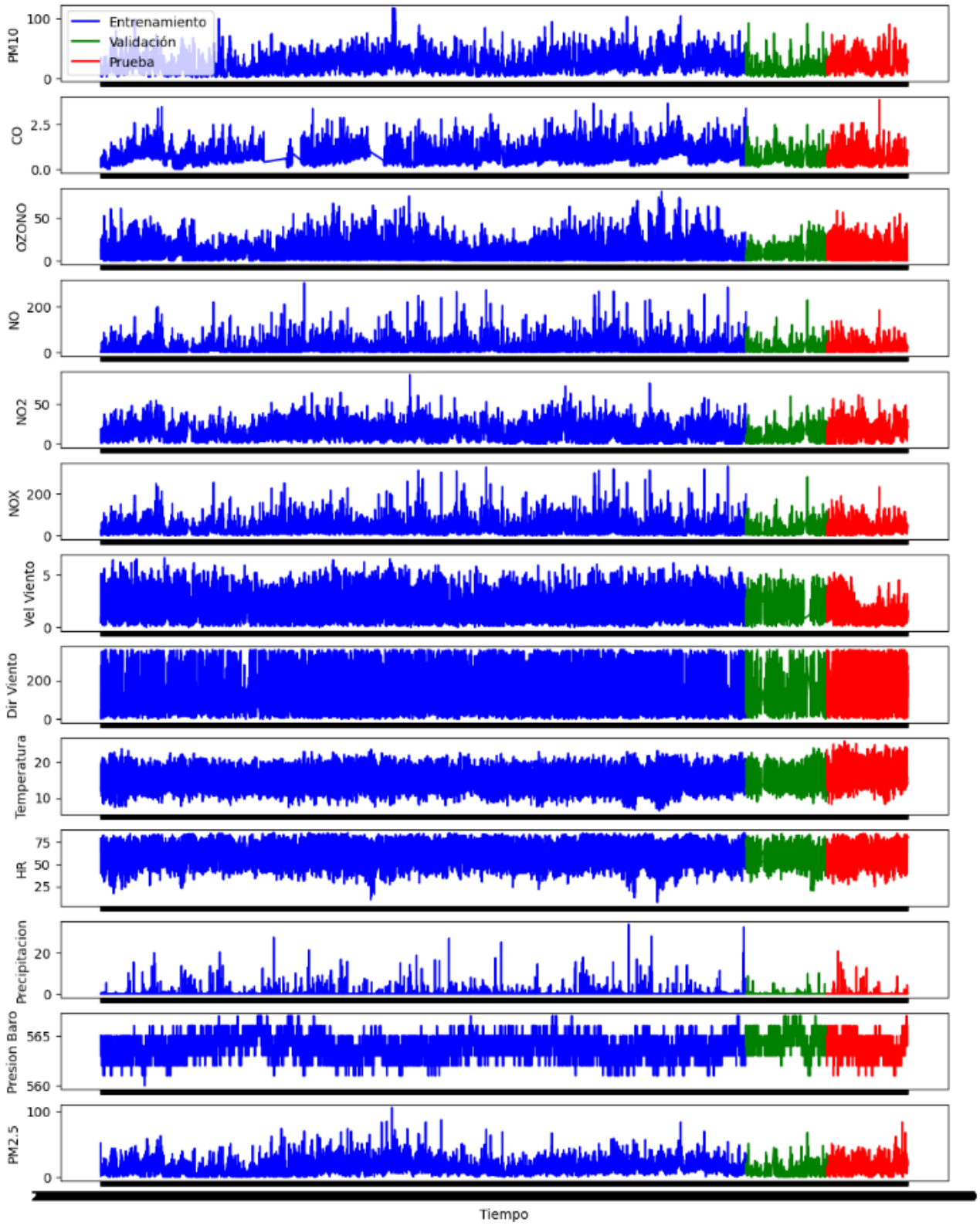

Figura 4.
Gráfica de variables con los sets de datos distribuidos.

Nota: elaboración propia.

La librería TensorFlow utilizada para la construcción de los modelos LSTM espera el valor X y Y como arreglos numpy con la siguiente estructura:

$$X = (Batches\ , Input\ length, Feature), Y = (Batches\ , Output\ length, Feature)$$

Donde:

- *Batches*: numero de bloques que ingresan al modelo
- *Input length*: número de registros secuenciales que ingresan al modelo
- *Features*: número de variables que se ingresan al modelo
- *Output length*: número de registros predichos

De esta manera la tabla 5 presenta la configuración para el modelo univariable-unistep.

Tabla 5

Definición del conjunto de datos para el modelo univariable-unistep.

Conjunto de datos	X			Y		
	Batches	Input length	Features	Batches	Output length	Features
Set de datos de entrenamiento	21011	12	1	21011	1	1
Set de datos de validación	2615	12	1	2615	1	1
Set de datos de prueba	2615	12	1	2615	1	1

Nota: Elaboración propia.

Como se muestra en la Tabla 5, el modelo fue diseñado para utilizar 12 registros como entrada (input length) de una única variable (Features) y, a partir de estos, predecir un registro de salida (output length) para la misma variable. En otras palabras, el modelo predice el valor correspondiente a una hora posterior a 12 horas consecutivas. No obstante, la hora 13 se empleará para la siguiente predicción, ya que los datos se desplazarán progresivamente, tal como se ilustra en la Figura 5.

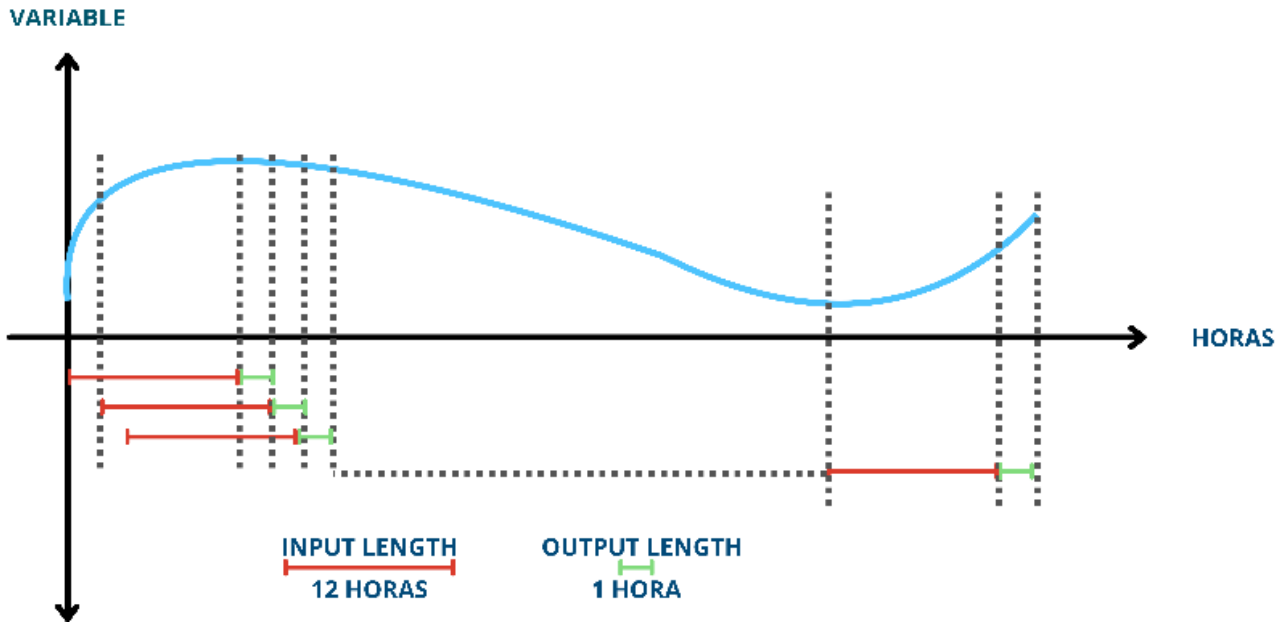

Figura 5
Predicciones en modelo univariado.
Nota: Elaboración propia.

Para el caso del modelo multivariado-unistep el comportamiento es similar, con la diferencia de las características de entrada (Features), ya que en este caso se introduce todo el conjunto de datos para predecir una única variable. La tabla 6 presenta la estructura definida y la figura 6 plasma el desplazamiento de predicción en la secuencia de bloques (batches).

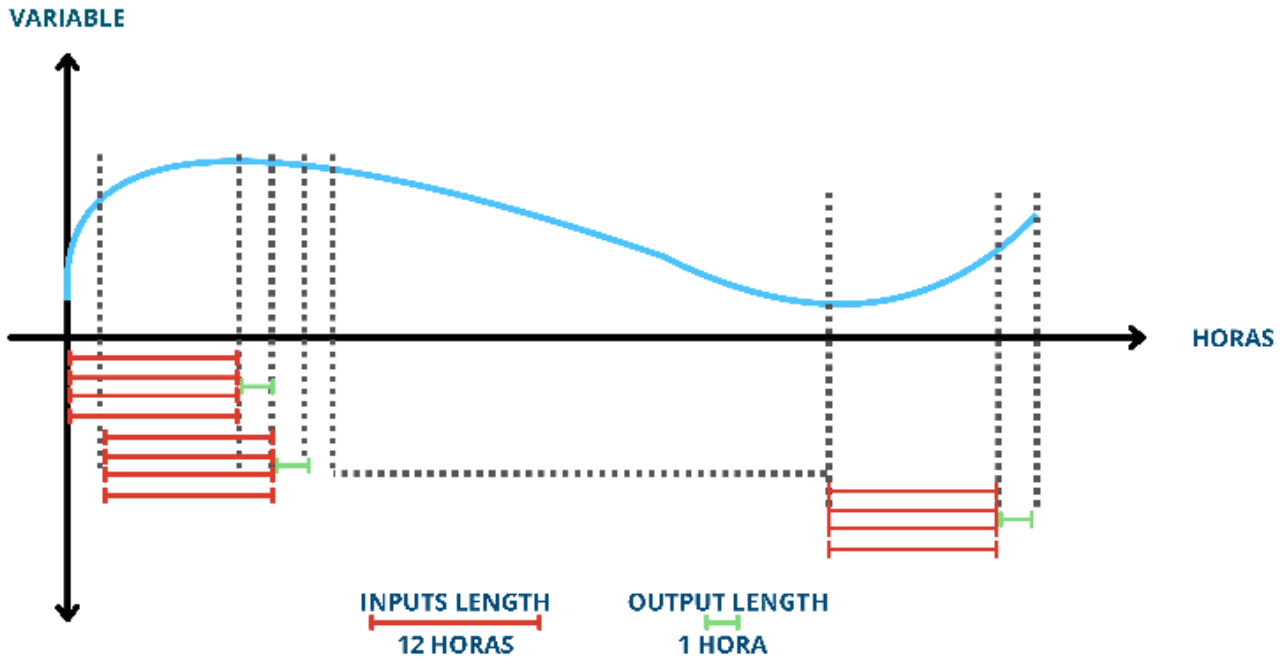

Figura 6
Predicciones en modelo multivariado
Nota: Elaboración propia.

Tabla 6
Definición del conjunto de datos para el modelo multivariable-unistep.

Conjunto de datos	X			Y		
	Batches	Input length	Features	Batches	Output length	Features
Set de datos de entrenamiento	21011	12	13	21011	1	1
Set de datos de validación	2615	12	13	2615	1	1
Set de datos de prueba	2615	12	13	2615	1	1

Nota: Elaboración propia.

Para este conjunto de modelos, se modificaron las variables de dirección y velocidad del viento, ya que ambas están correlacionadas en una única variable vectorial. En este caso, la dirección determina el ángulo del vector, mientras que la velocidad define su magnitud. Como resultado, se obtuvieron los componentes (x, y) para mejorar la interpretación de estas variables. Este proceso se llevó a cabo utilizando las ecuaciones 2 y 3, cuya representación gráfica se muestra en la Figura 7.

$$velX = \frac{\cos(dirViento)}{velociodad\ viento} \quad (2)$$

$$velY = \frac{seno(dirViento)}{velociodad\ viento} \quad (3)$$

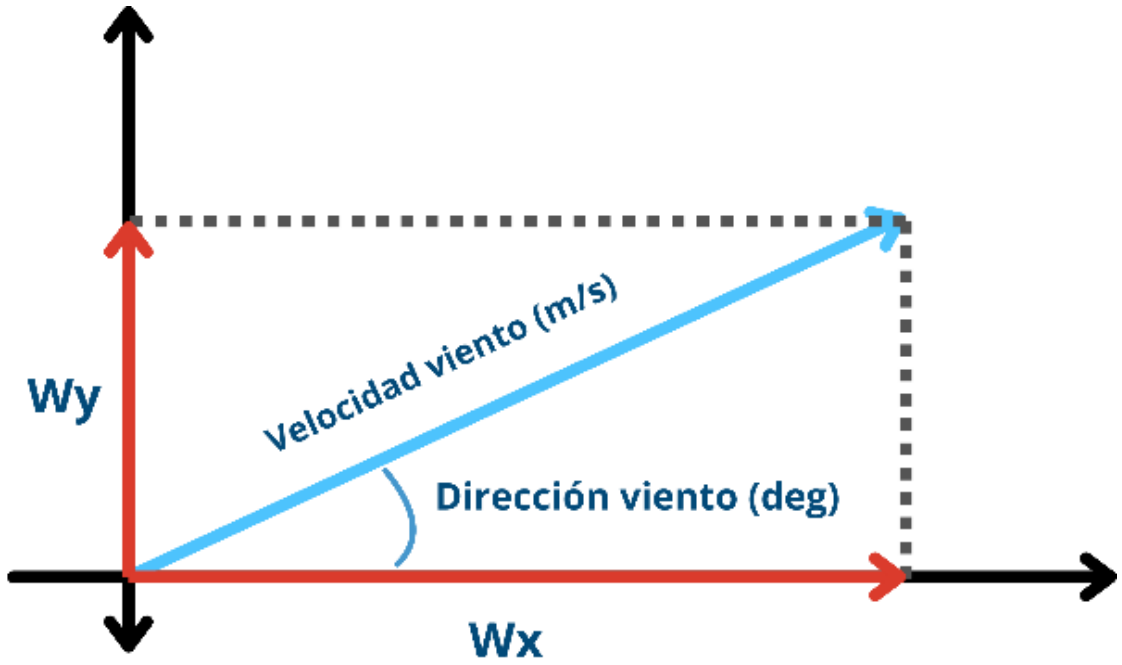

Figura 7
Relación trigonométrica velocidad y dirección del viento.
Nota: elaboración propia.

Finalmente, se procedió a escalar los datos dentro de un rango de cero a uno (0 a 1), con el fin de facilitar la obtención de los pesos. En este caso, se construyó un único escalador para el conjunto de datos de entrenamiento, y tanto el conjunto de validación como el de prueba utilizaron dicho escalador. Este proceso se aplicó en ambos tipos de modelos, tanto el multivariado como el univariado. Es importante mencionar que el escalado se realizó de forma individual para cada variable, por lo que se utilizaron escaladores independientes. La Figura 8 presenta una gráfica tipo violín que muestra el escalamiento de cada variable, considerando los registros de los conjuntos de entrenamiento, validación y prueba dentro del rango 0 a 1 mencionado anteriormente.

AmeliCA
PDF generado a partir de XML-JATS4R
Modelo de publicación sin fines de lucro para conservar la naturaleza académica y abierta de la comunicación científica

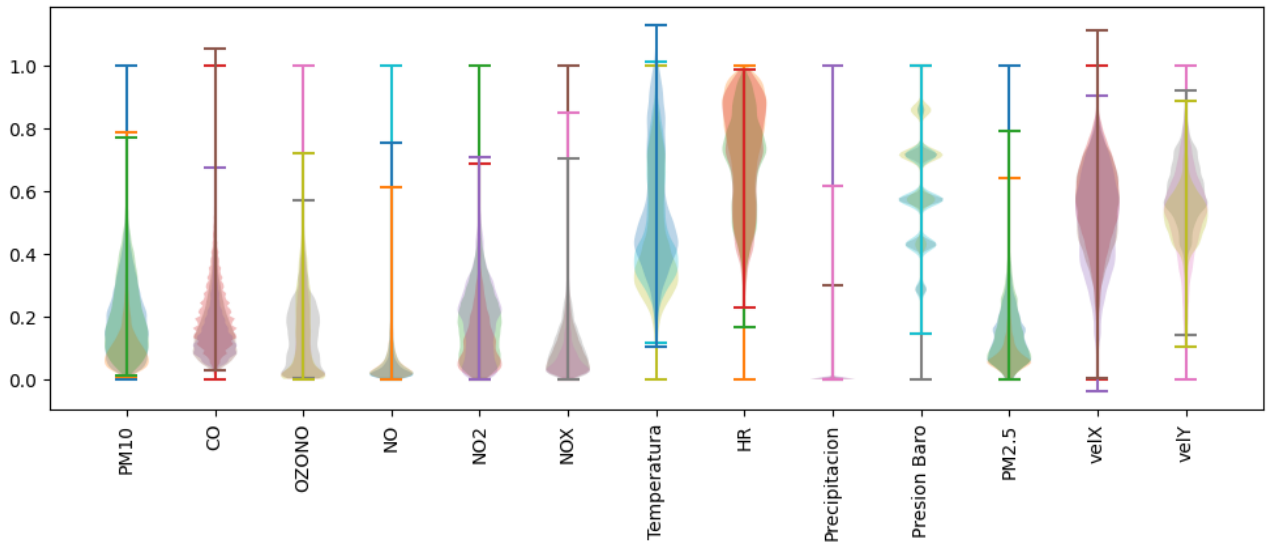

Figura 8
Distribución de datos escalados en gráfica tipo violín.
Nota: elaboración propia.

2.3. Entrenamiento de los modelos

El conjunto de modelos se construyó utilizando dos capas: la primera, correspondiente a la capa LSTM, y la segunda, a una capa densa. Se empleó una función de activación lineal para predecir los registros de salida. Las pérdidas se calcularon utilizando la raíz del error cuadrático medio (RMSE), y como optimizador se utilizó RMSprop. La Tabla 7 presenta la configuración de los hiperparámetros utilizados en ambos conjuntos de modelos.

Tabla 7
Hiperparámetros de los modelos construidos.

Hiperparámetro	Modelo Univariable	Modelo Multivariable
Input leangth	12	12
Output leangth	1	1
Unidad de memoria	256	256
Épocas de entrenamiento	80	140
Tasa de entrenamiento	$50x10^{-6}$	$5x10^{-6}$
Tamaño de lote (batch)	64	64
Función de activación	Lineal	Lineal
Optimizador	RMSprop	RMSprop

Nota: elaboración propia.

Se utilizó la librería TensorFlow para desarrollar la arquitectura de cada modelo, como se muestra en la figura 9, lo que confirma la coherencia de la configuración de hiperparámetros presentada en la tabla 7.

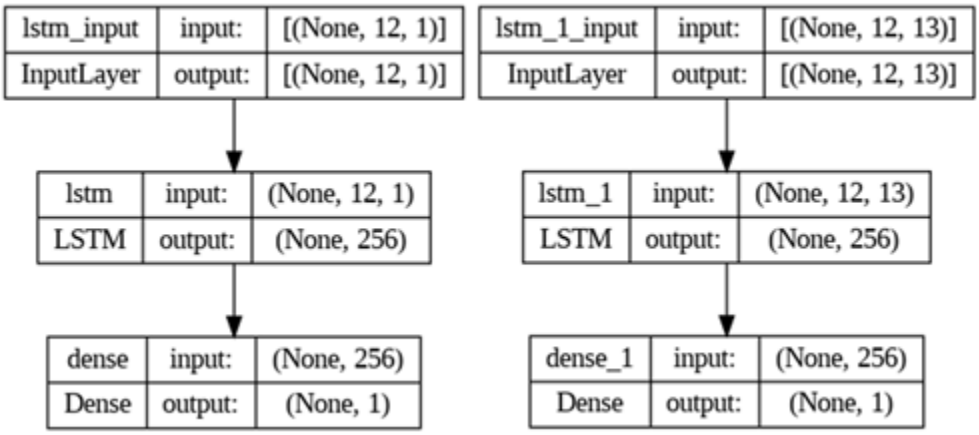

Figura 9
Arquitectura de los modelos creados
Nota: Elaboración propia, a la izquierda modelo univariable a la derecha modelo multivariable.

Con la arquitectura definida, se graficó el entrenamiento de cada modelo para visualizar el comportamiento de las pérdidas, calculadas mediante la raíz del error cuadrático medio (RMSE), de acuerdo con la ecuación 4.

$$RSME = \sqrt{\frac{1}{Total\ registros}\sum_{i=1}^{Total\ registros}\left(real_i - \overline{prediccion_i}\right)^2} \quad (4)$$

Se observa un comportamiento estable tanto en el entrenamiento como en la validación, con una pérdida menor en esta última sin llegar a sobre ajustar el modelo (ver figura 10 y figura 11).

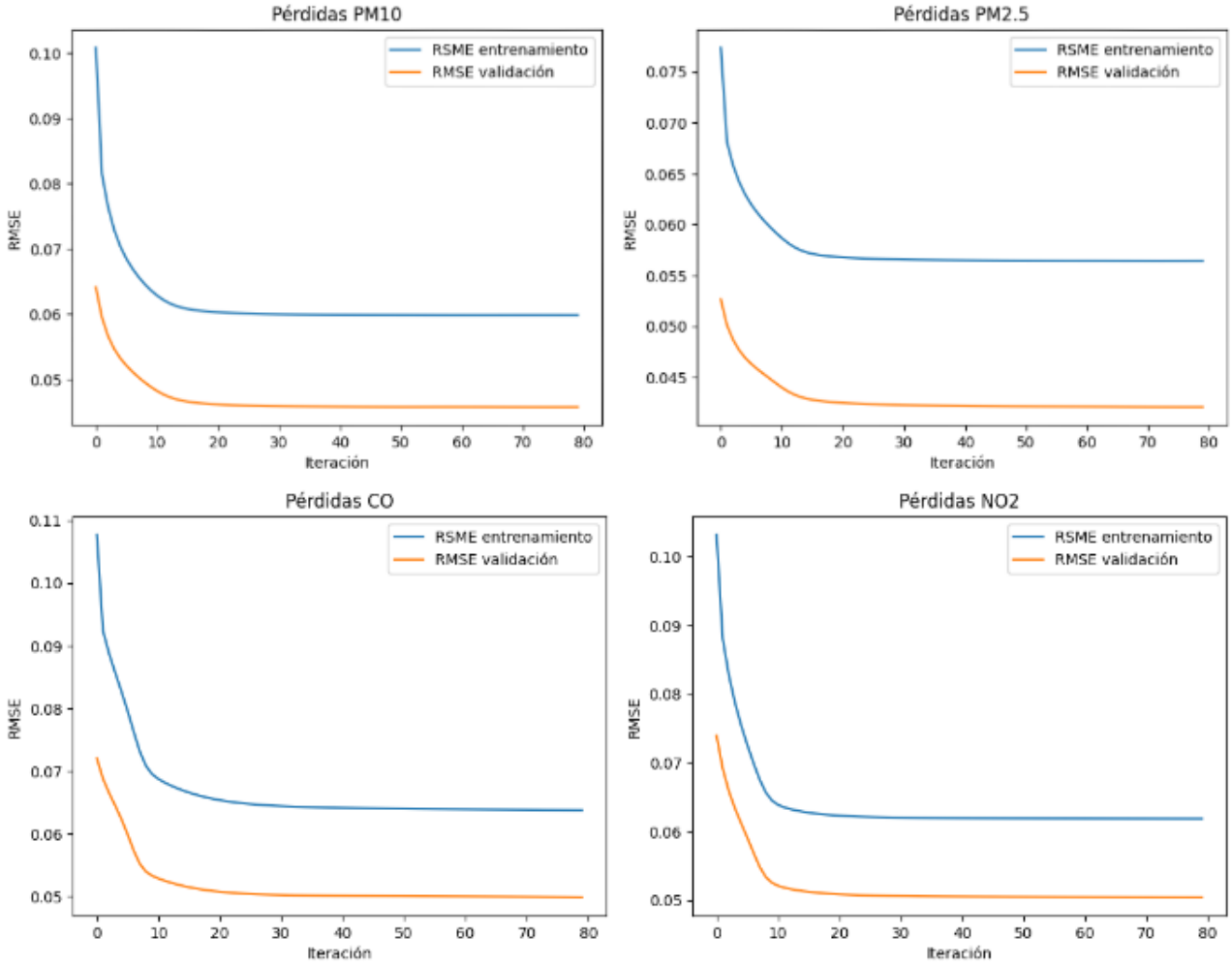

Figura 10

Entrenamiento de los modelos univariados, validación de perdidas.

Nota: elaboración propia.

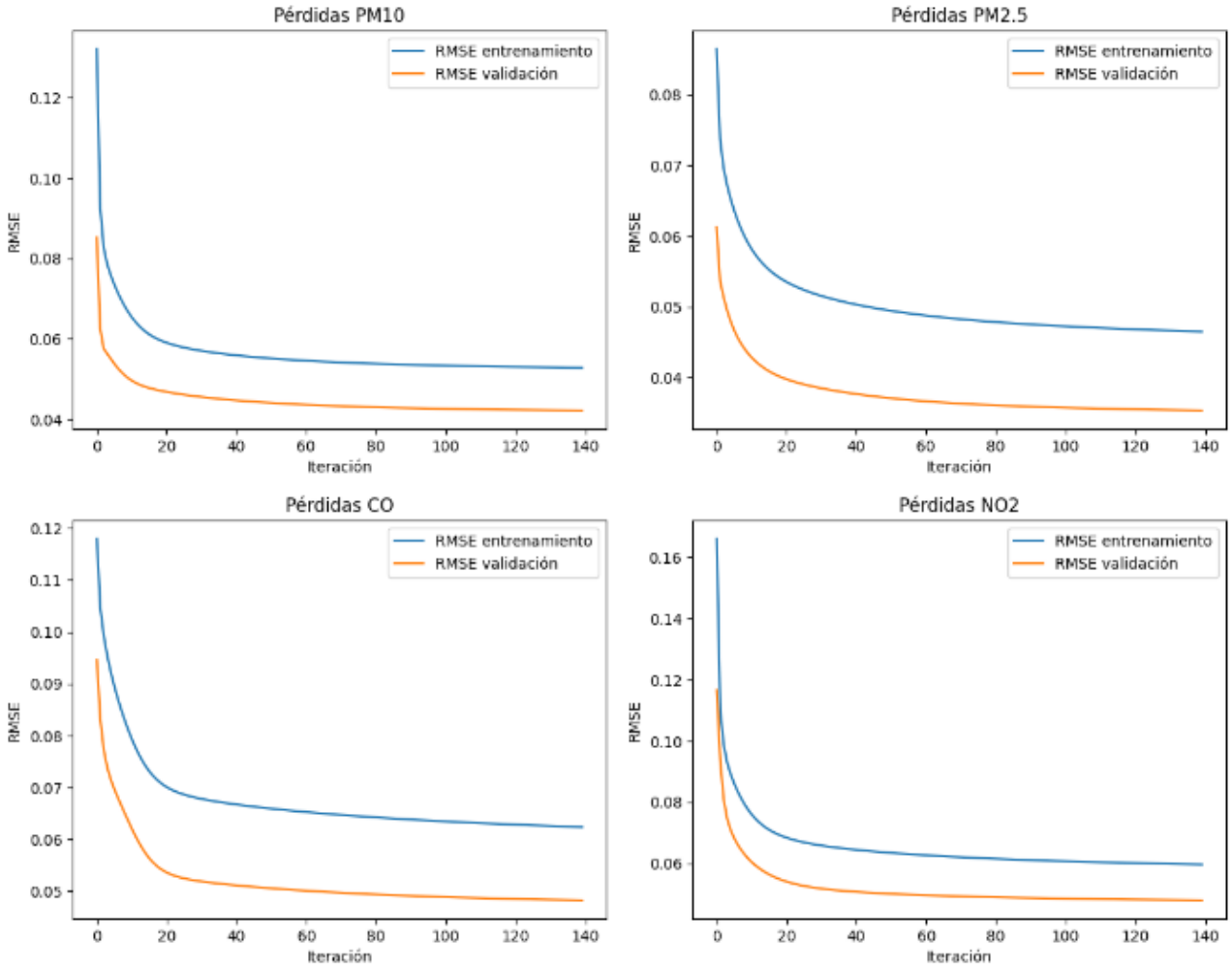

Figura 11
Entrenamiento de los modelos multivariados, validación de perdidas.
Nota: elaboración propia

2.4. Validación

Una vez que los modelos fueron entrenados, se utilizó el conjunto de datos de prueba para validar que no hubiera sobreajuste, recordando que este conjunto no formó parte del entrenamiento, por lo que representa datos desconocidos para los modelos. La tabla 8 presenta los valore RSME obtenidos para cada modelo, evidenciando un rendimiento superior en los modelos multivariados.

Tabla 8.
Comparación del rendimiento entre modelo univariado y multivariado en escala original

Variable	Univariado	Multivariado	Mejora evidenciada
PM10	7.95	6.94	12.7%
PM2.5	7.08	5.79	18.22%

CO	0.29	0.24	17.24%
NO2	6.2	5.73	7.58%

Nota: Elaboración propia.

Por último, se graficaron los primeros 250 datos de cada modelo para evidenciar el comportamiento de las predicciones. La figura 12 muestra las predicciones de los modelos univariados, mientras que la figura 13 presenta las predicciones de los modelos multivariados.

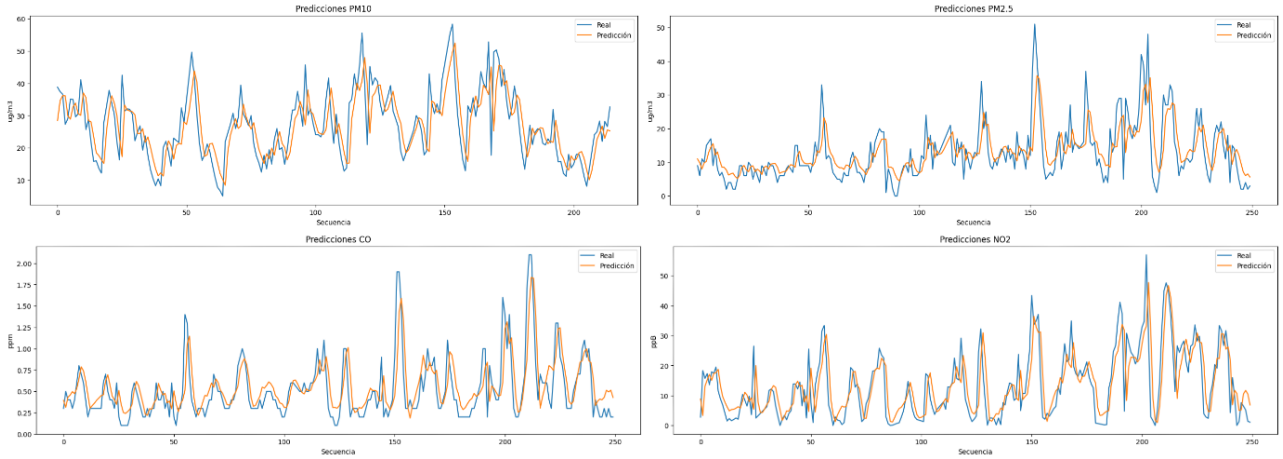

Figura 12
Predicciones modelo LSTM univariado
Nota: elaboración propia.

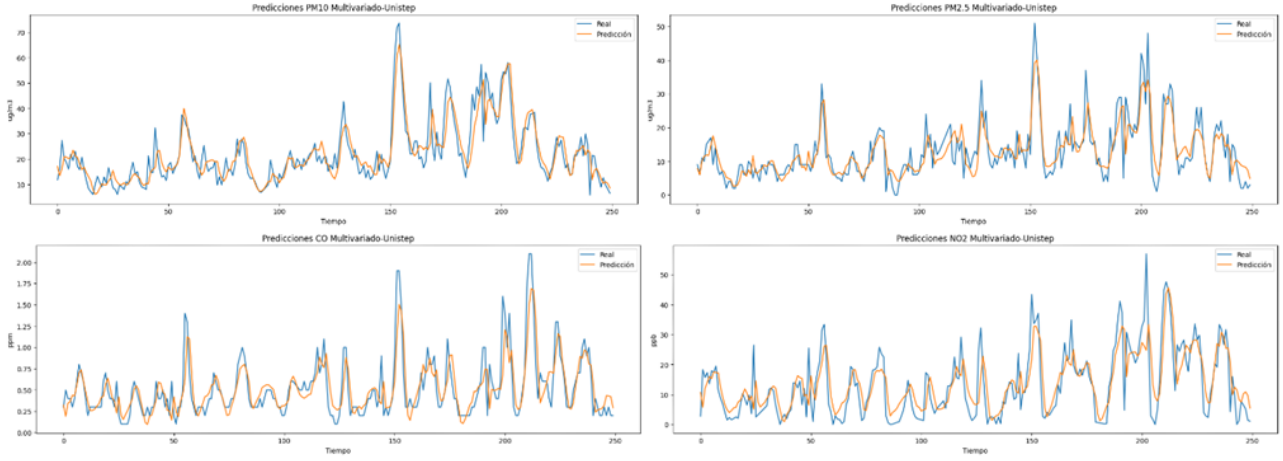

Figura 13
Predicciones modelo LSTM Multivariado.
Nota: elaboración propia.

3. Resultados y discusión

Los hallazgos de esta investigación destacan el impacto significativo de incluir covariables en los modelos LSTM para mejorar la precisión en las predicciones de contaminantes atmosféricos. Se observa una mejora notable en el rendimiento del modelo multivariable-unistep en comparación con el modelo univariable-unistep, con aumentos en la precisión que varían entre un 7% y un 18% en todos los casos analizados. Estos resultados subrayan la importancia de considerar múltiples variables en la modelización de la calidad del aire, sugiriendo que la interacción entre diferentes contaminantes influye en su comportamiento y, por tanto, en su capacidad predictiva. Esto pone en evidencia la complejidad inherente a los sistemas atmosféricos y la necesidad de emplear enfoques más sofisticados para capturar adecuadamente estas interacciones.

Además de los resultados generales, se presentan los siguientes resultados específicos:

- Análisis exploratorio de datos: Se realizó un análisis exhaustivo del conjunto de datos de la estación de Las Ferias, compuesto por más de 29,000 registros preparados para ser utilizados en modelos de predicción de series temporales. Este conjunto de datos se encuentra disponible para futuras investigaciones y actividades académicas.
- Modelos LSTM univariable-unistep: Se desarrollaron varios modelos LSTM univariable-unistep con un error cuadrático medio (RMSE) promedio de 5.38. Estos modelos son adecuados para hacer predicciones diarias de variables individuales, ya sea utilizando el sistema RMACAB u otros sistemas de monitoreo ambiental en Bogotá.
- Modelos LSTM multivariable-unistep: Se construyeron modelos LSTM multivariable-unistep con un RMSE promedio de 4.67. Estos modelos también son apropiados para realizar predicciones diarias de múltiples variables, utilizando el sistema RMACAB o cualquier otro sistema de monitoreo en la ciudad de Bogotá.
- Repositorio público: Se ha establecido un repositorio público que contiene los modelos desarrollados, junto con el código fuente correspondiente, a disposición de la comunidad. Esto facilita que otros investigadores y profesionales puedan utilizar, adaptar o mejorar estos modelos para sus propias investigaciones.

4. Conclusiones

El análisis comparativo entre los modelos univariable-unistep y multivariable-unistep mostró que los modelos multivariables ofrecen un rendimiento superior en la predicción de contaminantes del aire. En promedio, el error cuadrático medio (RMSE) del modelo multivariable-unistep fue un 18% menor que el del modelo univariable, lo que evidencia que la inclusión de múltiples variables mejora significativamente la precisión de las predicciones de series temporales LSTM para contaminantes como PM10, PM2.5, CO y NO2.

De cara al futuro, se sugiere implementar un sistema CNN-LSTM que permita incluir los datos de todas las estaciones de monitoreo y realizar predicciones espaciotemporales. Este enfoque combinaría la capacidad de las redes neuronales convolucionales (CNN) para extraer características espaciales con la habilidad de las LSTM para modelar relaciones temporales, incrementando tanto la precisión como el alcance de las predicciones sobre la calidad del aire en Bogotá. Sin embargo, para llevar a cabo este sistema, es crucial desarrollar primero un modelo de regresión robusto que permita completar los datos faltantes de todas las estaciones con mayor efectividad que una interpolación tradicional. Este paso garantizaría la consistencia y calidad de los datos de entrada, lo que es fundamental para el rendimiento del sistema CNN-LSTM.

Adicionalmente, se propone implementar los modelos LSTM directamente en los dispositivos de borde (edge) en cada estación de monitoreo, eliminando así la dependencia de procesar las predicciones en un servidor central. Este enfoque no solo reduciría la carga computacional del servidor, sino también la necesidad de conectividad constante, aumentando la resiliencia del sistema ante interrupciones de red. Sin embargo, uno de los principales desafíos radica en identificar y seleccionar sistemas embebidos con la capacidad computacional suficiente para ejecutar estos modelos en tiempo real, manteniendo un equilibrio entre rendimiento, costo y eficiencia energética, se sugiere arquitectura ARM-R o FPGA Artyx-7.

Además, se recomienda explorar métodos alternativos para abordar fluctuaciones localizadas en el tiempo. Por ejemplo, la descomposición empírica de modos (EMD), que descompone la señal en modos intrínsecos (IMFs) de forma adaptativa, podría proporcionar un análisis más detallado de las características temporales. Una posibilidad interesante sería aplicar la Transformada de Hilbert (HHT) sobre los IMFs generados, lo que permitiría obtener un espectro tiempo-frecuencia y analizar propiedades instantáneas como frecuencia y amplitud. Este enfoque, aunque diferente al de las LSTMs, podría complementar el análisis al proporcionar información adicional sobre las propiedades oscilatorias de las señales.

La incorporación de estos modelos en el sistema RMCAB mejoraría la generación de informes y la emisión de alertas tempranas, proporcionando información crucial para la formulación de políticas públicas y la planificación urbana. Una comprensión más precisa de la distribución espaciotemporal de los contaminantes permitiría a las autoridades implementar medidas más efectivas para mitigar los efectos adversos en la salud pública y en el medio ambiente.

Finalmente, se sugiere evaluar la posibilidad de combinar los modelos predictivos de calidad del aire con análisis socioeconómicos y de salud pública, para obtener una comprensión más integral de los impactos a largo plazo. De esta forma, se podría contribuir a la creación de políticas más holísticas que promuevan no solo la mitigación de la contaminación, sino también la equidad en salud y bienestar en zonas urbanas.

5. Referencias

Ameer, S., Shah, M. A., Khan, A., Song, H., Maple, C., Islam, S. U., & Asghar, M. N. (2019). Comparative analysis of machine learning techniques for predicting air quality in smart cities. *IEEE Access, 7*, 128325–128338. https://doi.org/10.1109/ACCESS.2019.2925082

Bao, R., & Zhang, A. (2020). Does lockdown reduce air pollution? Evidence from 44 cities in northern China. *Scienceof The Total Environment, 731*, 139052. https://doi.org/10.1016/j.scitotenv.2020.139052

Brownlee, J. (2017). *Long short-term memory networks with Python: Develop sequence prediction models with deep learning* (1.0).

Cantú, P. (2023). La contaminación del aire y los riesgos de la salud. *Revista UANL, 122*. https://ojs.biblio.uanl.mx/index.php/ojs/article/view/296

Fang, Z., Wang, Y., Peng, L., & Hong, H. (2021). Predicting flood susceptibility using LSTM neural networks. *Journal of Hydrology, 594*, 125734. https://doi.org/10.1016/j.jhydrol.2020.125734

Guerrero-Rojas, N. K. (2020). *Alternativas para la reducción de contaminantes atmosféricos emitidos por el sistema vehicular en Bogotá D.C.*https://hdl.handle.net/10983/24784

Han, H., Liu, J., Shu, L., Wang, T., & Yuan, H. (2020). Local and synoptic meteorological influences on daily variability in summertime surface ozone in eastern China. *AtmosphericChemistry and Physics, 20*(1), 203–222. https://doi.org/10.5194/acp-20-203-2020

Ministerio de Ambiente y Desarrollo Sostenible. (2017). *Resolución 2254 de 2017*. https://www.minambiente.gov.co/documento-entidad/resolucion-2254-de-2017/

ONU (2022). El 99% de la población mundial respira aire contaminado. *https://news.un.org/es/story/2022/04/1506592.*

Ruggieri, R., Ruggeri, M., Vinci, G., & Poponi, S. (2021). Electric mobility in a smart city: European overview. *Energies, 14*(2), 315. https://doi.org/10.3390/en14020315

Sak, H., Senior, A., & Beaufays, F. (2014). Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In H. Li & P. Ching (Eds.), *15th Annual Conference of the International Speech Communication Association (INTERSPEECH 2014)* (pp. 338–342). International Speech Communication Association (ISCA). https://www.isca-archive.org/interspeech_2014/sak14_interspeech.pdf

Sharma, S., Zhang, M., Anshika, Gao, J., Zhang, H., & Kota, S. H. (2020). Effect of restricted emissions during COVID-19 on air quality in India. *Science of The Total Environment, 728*, 138878. https://doi.org/10.1016/j.scitotenv.2020.138878

Sofowote, U. M., Healy, R. M., Su, Y., Debosz, J., Noble, M., Munoz, A., Jeong, C.-H., Wang, J. M., Hilker, N., Evans, G. J., Brook, J. R., Lu, G., & Hopke, P. K. (2021). Sources, variability and parameterizations of intra-city factors obtained from dispersion-normalized multi-time resolution factor analyses of PM2.5 in an urban environment. *Science of The Total Environment, 761*, 143225. https://doi.org/10.1016/j.scitotenv.2020.143225

TensorFlow. (2024). tf.keras.layers.LSTM. https://www.tensorflow.org/api_docs/python/tf/keras/layers/LSTM

Zheng, B., Tong, D., Li, M., Liu, F., Hong, C., Geng, G., Li, H., Li, X., Peng, L., Qi, J., Yan, L., Zhang, Y., Zhao, H., Zheng, Y., He, K., & Zhang, Q. (2018). Trends in China's anthropogenic emissions since 2010 as the consequence

of clean air actions. *AtmosphericChemistry and Physics, 18*(19), 14095–14111. https://doi.org/10.5194/acp-18-14095-2018

Zou, X., Zhao, J., Zhao, D., Sun, B., He, Y., & Fuentes, S. (2021). Air quality prediction based on a spatiotemporal attention mechanism. *Mobile Information Systems, 2021*, 1–12. https://doi.org/10.1155/2021/6630944

Repositorio público: https://github.com/christiancass/modelos-LSTM-LasFerias.git

AmeliCA

Disponible en:
https://portal.amelica.org/ameli/journal/266/2664941006/2664941006.pdf

Cómo citar el artículo

Número completo

Más información del artículo

Página de la revista en portal.amelica.org

AmeliCA
Ciencia Abierta para el Bien Común

Christian Alejandro Sarmiento Sánchez
Predicción de Contaminantes Atmosféricos en Bogotá utilizando Redes LSTM

Ingenio Tecnológico
vol. 6, e051, 2024
Universidad Tecnológica Nacional, Argentina
ingenio@frlp.utn.edu.ar

ISSN-E: 2618-4931

CC BY-NC-SA 4.0 LEGAL CODE
Licencia Creative Commons Atribución-NoComercial-CompartirIgual 4.0 Internacional.