Predicción de los niveles de NO_2 en el aire de Seúl mediante el modelo Temporal Fusion Transformer (TFT)

Carlos López Pérez[a] (carlosdl@uninorte.edu.co),
Miguel Herrera Rocha[a] (mherreras@uninorte.edu.co)

[a] Msc. en Estadística Aplicada, Departamento de Matemáticas y Estadística, Universidad del Norte, Barranquilla, Colombia

Predicción de los niveles de NO_2 en el aire de Seúl mediante el modelo Temporal Fusion Transformer (TFT)

Carlos López Pérez[a], Miguel Herrera Rocha[a]

[a] *Msc. en Estadística Aplicada, Departamento de Matemáticas y Estadística, Universidad del Norte, Barranquilla, Colombia.*

Resumen

El monitoreo de la calidad del aire es esencial para comprender los efectos de la contaminación atmosférica en la salud pública y el medio ambiente. En este estudio, se aborda la predicción de los niveles de dióxido de nitrógeno (NO_2) en el aire de Seúl, Corea del Sur, utilizando datos de 1988 a 2021. A pesar de la aplicación de modelos de predicción en series temporales, como los transformers en la predicción de contaminantes como el ozono (O_3), PM y CO_2, se han encontrado pocos estudios que utilicen específicamente el modelo Temporal Fusion Transformer (TFT) para este propósito. En este trabajo, se emplea el modelo TFT para predecir los niveles de NO_2, evaluando su capacidad para capturar dependencias temporales y realizar predicciones precisas. Los resultados muestran que el modelo TFT puede identificar patrones complejos y no lineales en los datos históricos, superando a otros modelos tradicionales en términos de precisión. Este estudio destaca la efectividad de los transformers, y en particular el TFT, para predecir contaminantes atmosféricos, sugiriendo su potencial para ser implementado en la predicción y gestión de la calidad del aire a nivel global.

Palabras Clave: NO_2, Temporal Fusion Transformer, calidad del aire, serie temporal, transformers, modelos de predicción.

Email addresses: carlosdl@uninorte.edu.co (Carlos López Pérez), mherreras@uninorte.edu.co (Miguel Herrera Rocha)

1. Datos

Los datos utilizados en este estudio provienen del conjunto de datos histórico sobre la calidad del aire en Seúl, Corea del Sur desde 1988 hasta 2021. Este conjunto de datos contiene más de 5 millones de observaciones en diferentes puntos de medición por cada hora relacionadas con varios contaminantes atmosféricos tales como Ozono (O_3), PM2.5, PM10 y SO_2 y los niveles de dióxido de nitrógeno (NO_2). La variable de interés en este análisis es la concentración de NO_2, que se mide en partículas por millón [ppm]. Este conjunto de datos está estructurado en formato de serie temporal lo que facilita su uso para modelos de predicción basados en este tipo de datos. Los autores proporcionan acceso completo al conjunto de datos, permitiendo su replicación para futuras investigaciones y puede ser consultado en la bibliografía (Hyun, 2021).

El dióxido de nitrógeno NO_2 es un contaminante atmosférico común que puede tener efectos negativos en la salud humana y el medio ambiente ya que la exposición a altos niveles ambientales puede aumentar el riesgo de infecciones del tracto respiratorio en la población debido a la interacción del contaminante con el sistema inmune (Stieb et al., 2021). El análisis de las concentraciones de NO_2 en el aire es crucial para entender los patrones de contaminación y tomar medidas adecuadas para su control.

1.1. Análisis exploratorio de los datos - EDA

El objetivo del presente análisis exploratorio de datos es comprender la estructura de los datos históricos de concentraciones de NO_2 en el aire de Seúl, identificar patrones temporales, detectar datos atípicos y determinar la existencia de tendencia o estacionalidad de la variable objetivo. Este proceso es fundamental para preparar los datos antes de aplicar modelos predictivos, ya que permite entender la distribución, tendencias y estacionalidad de la serie, así como manejar valores faltantes o atípicos.

El conjunto de datos inicial contenía un total de 5,984,782 registros, con mediciones de concentraciones. Una de las variables clave en la base de datos es

'*loc*', que hace referencia al punto de medición donde se tomaron las muestras de NO_2. Esta variable indica el punto de cada medición, por tanto, se decidió trabajar con una única ubicación. Para ello, se identificaron los valores únicos de loc y se seleccionó la ubicación con el mayor número de registros, que resultó ser la ubicación 124. Esta locación contaba con 288,736 mediciones, lo que la convierte en la más representativa y adecuada para realizar un análisis detallado de las concentraciones de NO_2 a lo largo del tiempo.

Durante el análisis, se identificó una observación atípica en la concentración de NO_2, cuyo valor era mucho más alto que el resto de los datos, registrando 0.687 ppm. Este valor (contrastaba notablemente con el percentil 99 que era de 0.08 ppm.) Dada esta discrepancia, se decidió tratar este valor como un outlier y reemplazarlo por un valor NaN para evitar distorsiones en el análisis. Posteriormente, este valor faltante fue imputado junto con otros valores faltantes en la serie. En total, se encontraron 6,898 valores faltantes en la variable de estudio, lo que representó aproximadamente el 2.39% del total de observaciones. Estos valores fueron imputados utilizando el método de interpolación cuadrática, lo que permitió mantener la integridad de la serie temporal sin perder información relevante. A continuación, se presenta un resumen estadístico de las concentraciones de NO_2 después de la imputación:

Métrica	**Valor**
Cantidad de datos	288736.00
Media	0.0318
Desviación estándar	0.0171
Valor Mínimo	0.0000
Cuartil 1	0.0190
Mediana	0.0290
Cuartil 3	0.0420
Valor Máximo	0.2040
Rango Intercuartílico (IRQ)	0.0230

Table 1: Resumen estadístico de las concentraciones de NO_2.

La distribución de las concentraciones de NO_2, visualizada en la gráfica de densidad, muestra un sesgo a la izquierda, con la mayoría de los valores concentrados en el rango inferior (entre 0.00 y 0.05 ppm), pero con una cola que se extiende hacia valores más altos. Esto indica que, aunque la mayoría de las mediciones son bajas, existen muy ocasionales picos de concentración que elevan la media.

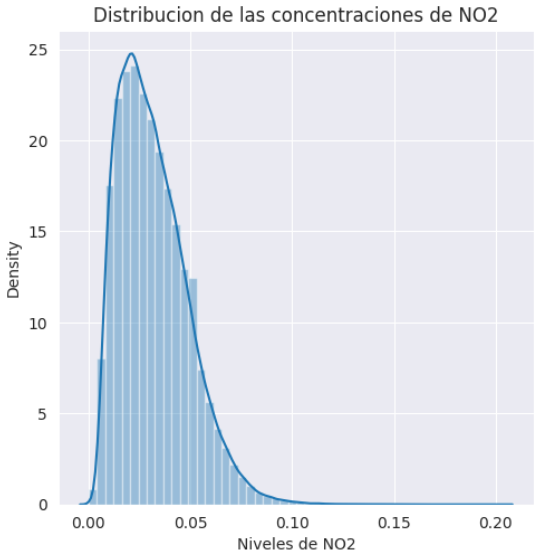

Figure 1: Distribución de la concentración de NO_2 en *ppm*

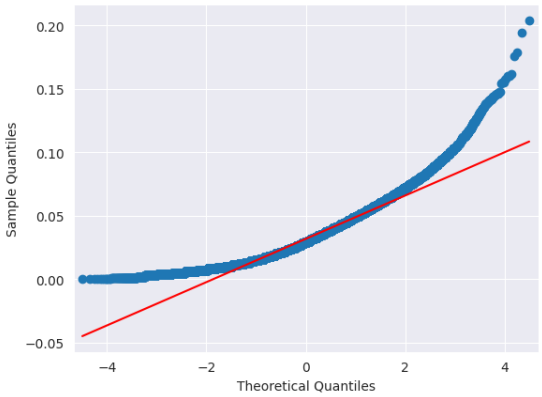

Figure 2: Grafico de cuantiles - NO_2

Por otro lado, el gráfico de cuantiles revela una desviación significativa de la normalidad, especialmente en los extremos, donde los valores observados no siguen la línea teórica de una distribución normal. Esto se confirma con la prueba de Shapiro-Wilk, que arrojó un p-valor extremadamente bajo (p ¡ 0.0001) y un estadístico de 0.949, lo que lleva a rechazar la hipótesis nula de normalidad. En conclusión, las concentraciones de NO_2 no siguen una distribución normal.

A continuación se presenta en las gráficas 3 y 4 una visualización de la serie temporal con el objetivo de observar la tendencia general y posibles anomalías. En la siguiente gráfica se observan tres series superpuestas que representan diferentes niveles de agregación temporal de las concentraciones de NO_2.

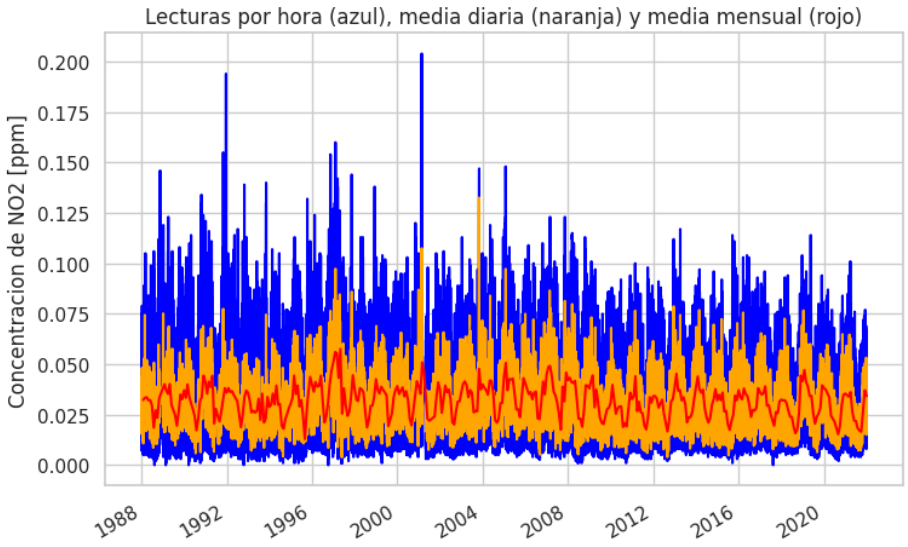

Figure 3: Visualización concentraciones de NO_2 en el tiempo

La serie graficada en azul muestra las mediciones originales de NO_2 tomadas por hora. Se puede apreciar una alta variabilidad, con fluctuaciones significativas a lo largo del tiempo. Para suavizar la variabilidad horaria y resaltar patrones a más largo plazo, se calculó la media diaria (naranja) y mensual (rojo) de las concentraciones de NO_2, en esta gráfica podemos sospechar que no se encuentra una tendencia marcada (ya sea creciente o decreciente) pero que

con las medias mensuales (rojo) se sospecha de un patrón estacional, esto será verificado mas adelante.

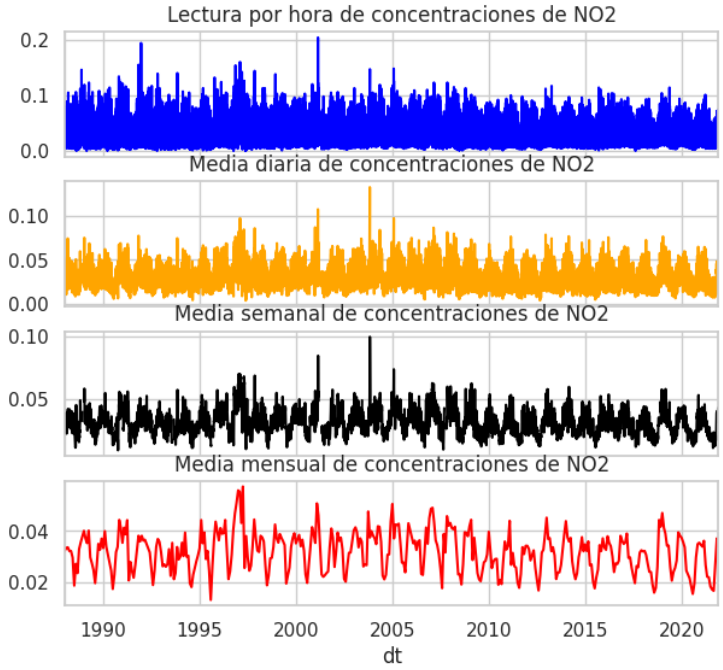

Figure 4: Concentraciones de NO_2 variando su agrupación temporal

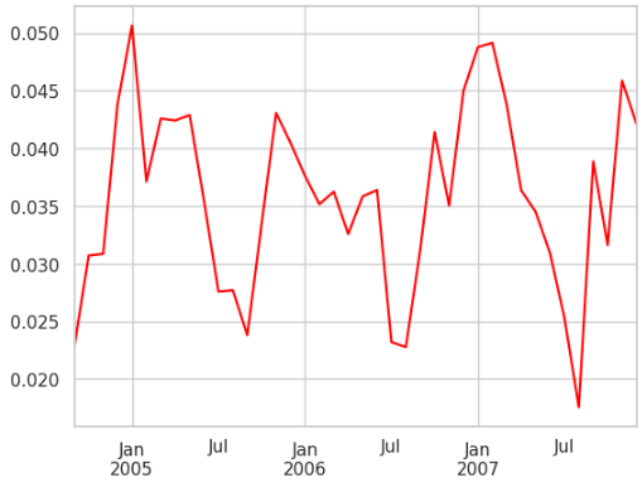

Figure 5: Ampliación de la serie de medias mensuales entre 2005-2008

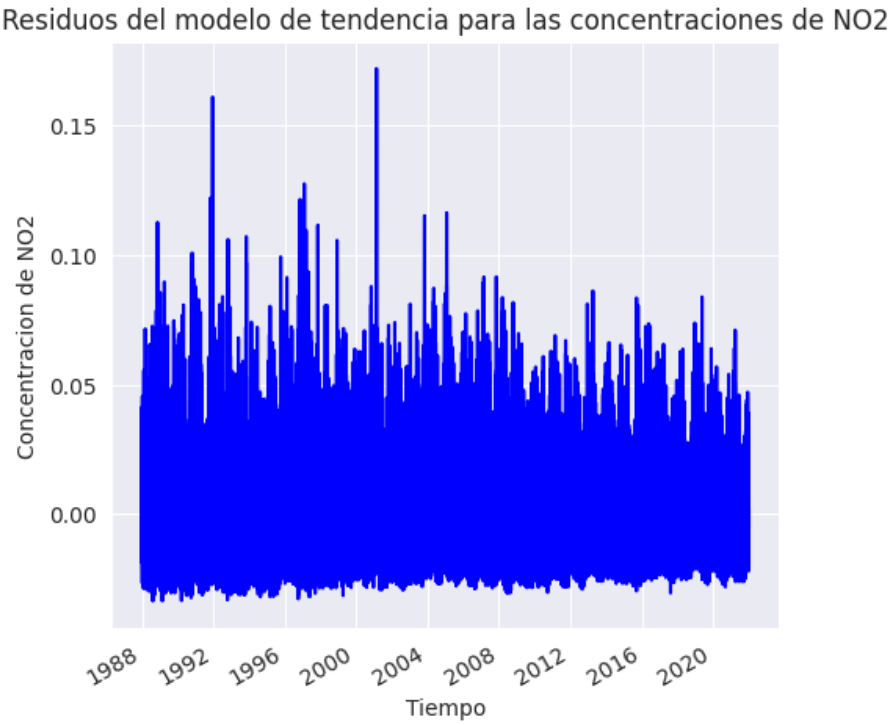

Figure 6: Residuos del modelo de tendencia para las concentraciones de NO_2

Se modelo la tendencia de la serie de tiempo con una regresion lineal, en esta, los residuos del modelo de tendencia aplicado a las concentraciones de NO_2 oscilan entre 0.15 y -0.05. Es importante verificar si los residuos muestran algún patrón o autocorrelación que indique la presencia de estructura adicional no capturada por el modelo. Se recomienda realizar pruebas de autocorrelación y ruido blanco en los residuos para validar la adecuación del modelo y, en caso de ser necesario, incorporar términos adicionales como componentes estacionales o de media móvil.

Si se amplia en una sección de la serie con las medias mensuales, es posible ver un patrón que se repite de forma anual en donde las concentraciones de NO_2 parecen llegar a su punto mas bajo en los meses de Julio y agosto (ver gráfica 5).

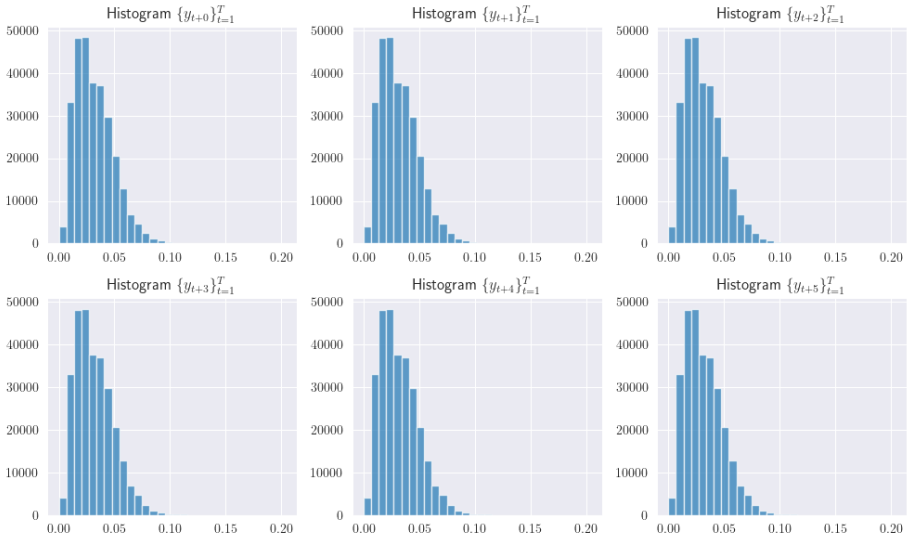

Figure 7: Distribución de NO_2 para Diferentes Horizontes Temporales

El análisis de los histogramas de la serie temporal y_{t+k} para $k = 0, 1, 2, 3, 4, 5$ sugiere que la distribución de la variable se mantiene estable a lo largo del tiempo. La forma de los histogramas no muestra cambios significativos, lo que indica una dinámica estacionaria o con baja variabilidad en su estructura. No se observa un desplazamiento evidente en la media o la dispersión de los datos, lo que sugiere que la serie no tiene tendencias marcadas ni cambios abruptos. Esto es relevante para el modelado, ya que implica que métodos estadísticos que asumen estacionalidad (La cual se confirmara con la prueba de Dickey Fuller) pueden ser adecuados para su análisis y predicción.

Se desea verificar la estacionalidad de las mediciones por hora de las concentraciones de NO_2 por lo que se realiza una prueba Dickey fuller para confirmar la estacionalidad. En los resultados de la prueba de Dickey-Fuller para la concentración de NO_2, obtuvimos un valor p menor a 0.0001 sin diferenciar los datos, lo que indica que podemos rechazar la hipótesis nula de que la serie tiene una raíz unitaria (es decir, que la serie no es estacionaria) al nivel de significancia del 5%. Esto nos confirma que la serie es estacionaria.

El análisis de la función de autocorrelación (ACF) y la autocorrelación par-

cial (PACF) sugiere que la serie de NO_2 por hora presenta una fuerte dependencia temporal y una estructura estacional clara. La ACF de la serie original muestra un decaimiento exponencial y/o sinusoides amortiguadas, mientras que la PACF presenta un pico significativo en el primer y segundo rezago, lo que segun lo visto graficamente, un modelo auto-regresivo podría ser útil para modelar esta serie de tiempo pero debe ser confirmado con pruebas estadisticas.

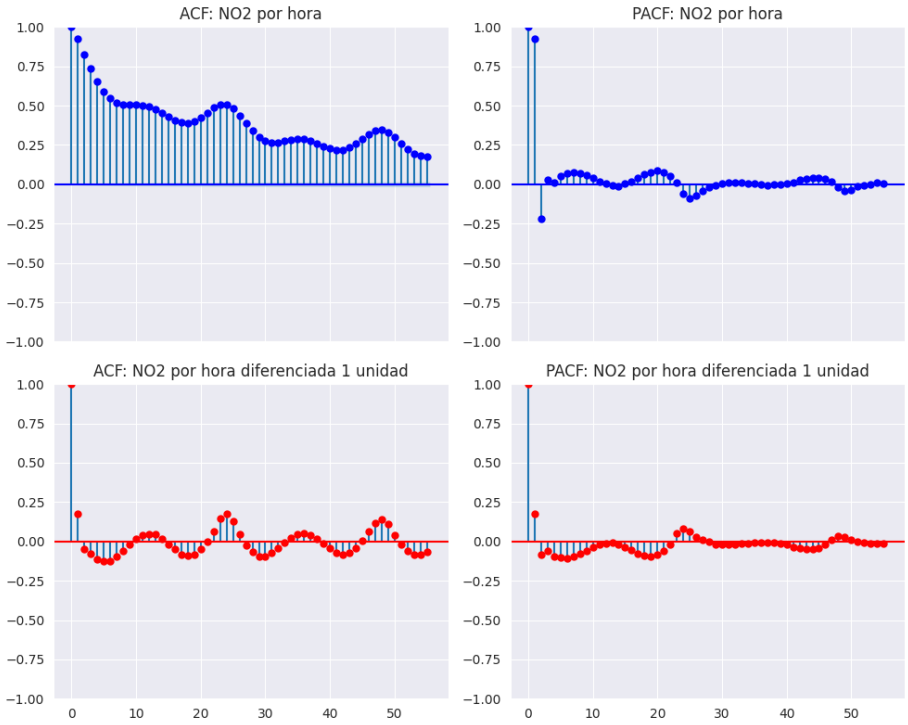

Figure 8: ACF y PACF de la serie de tiempo original y diferenciada 1 unidad

La prueba de Dickey-Fuller (DF) confirma que la serie es estacionaria, lo que indica que no presenta tendencia o raíz unitaria. Sin embargo, la función de autocorrelación (ACF) muestra autocorrelación significativa (En la prueba de Ljung-Box) en varios lags, lo que sugiere la presencia de una estructura de dependencia temporal en la serie. Por otro lado, la función de autocorrelación parcial (PACF) no muestra autocorrelación parcial significativa, lo que indica la ausencia de un componente autorregresivo (AR) claro. Estos resultados no

son contradictorios, sino que reflejan que la serie, aunque estacionaria, tiene una dependencia temporal que podría ser capturada por un componente de media móvil (MA).

En adición a lo visto con anterioridad, se presentan las concentraciones de NO_2 por trimestre en donde vemos que dependiendo del trimestre de medición se obtuvieron diferentes medias de concentración de NO_2, cabe resaltar que al tomar las medias por mes y por trimestre la serie temporal se comporta de forma diferente a si se toman los datos por hora (el cual es la serie original).

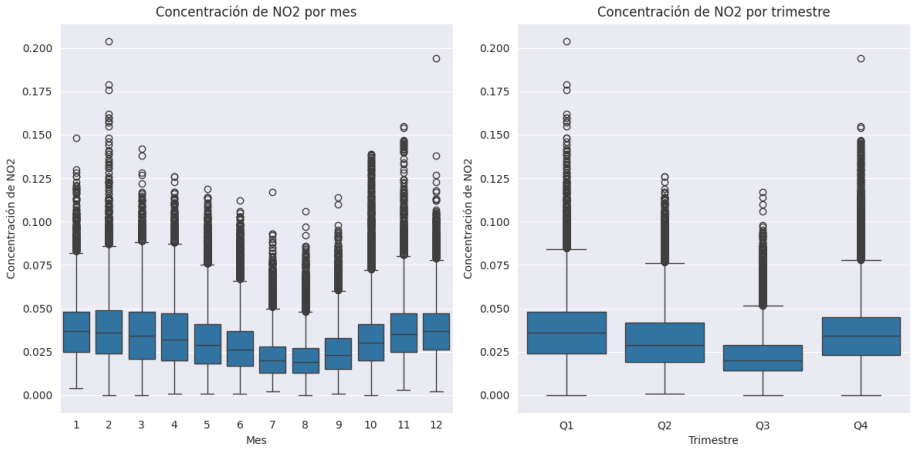

Figure 9: Concentración de NO_2 por trimestre

2. Métodos

(Catalano, 2024) (Mattioli, 2024)

3. Resultados

4. Conclusiones

5. Apéndices

References

Catalano, L. (2024). *A Transformer-based approach to air quality prediction in Milan through satellite imagery combined with meteorological and morpho-*

logical data. Ph.D. thesis Politecnico di Torino.

Hyun, W. (2021). Seoul air quality historical data. URL: `https://www.kaggle.com/datasets/williamhyun/seoulairqualityhistoricdata/data`.

Mattioli, F. (2024). *Enhancing Urban Air Quality Predictions with Temporal Fusion Transformers: A Methodological Evaluation*. Master's thesis Universidad Politecnica de Madrid Madrid, España.

Stieb, D. M., Berjawi, R., Emode, M., Zheng, C., Salama, D., Hocking, R., Lyrette, N., Matz, C., Lavigne, E., & Shin, H. H. (2021). Systematic review and meta-analysis of cohort studies of long term outdoor nitrogen dioxide exposure and mortality. *PloS one*, *16*, e0246451.