Visualizador interactivo para el análisis de imputación espacio temporal en datos de calidad del Aire

Albert Daniel Llica Alvarez

Docente: Mag. Ana Maria Cuadros Valdivia

UNSA - Universidad Nacional de San Agustín de Arequipa
Junio de 2025

Índice

1. Motivación y Contexto

La calidad del aire es un factor crítico para la salud pública, el medio ambiente y la formulación de políticas urbanas. En regiones urbanas densas, como las ciudades de China, los contaminantes como PM2.5, PM10, NO2, CO, O3 y SO2 tienen un impacto significativo en la calidad de vida [World Health Organization, 2016]. El dataset analizado, que abarca mediciones horarias de calidad del aire en 43 ciudades chinas entre mayo de 2014 y abril de 2015, presenta desafíos como valores nulos (8.16 % en PM10 y 14.74 % en presión atmosférica) y datos sucios, incluyendo valores extremos (PM25 hasta 1463 $\mu g/m^3$) y valores no físicos [Zheng et al., 2015].

El data wragling del dataset reveló patrones importantes: una ciclicidad estacional en PM2.5, con picos en invierno debido a factores como la quema de carbón y condiciones de baja dispersión [He et al., 2017], y correlaciones significativas entre contaminantes, como PM2.5 y PM10 (0.873) o SO2 con otros contaminantes (0.529–0.570). Además, variables geográficas (latitud y longitud) y meteorológicas (temperatura, velocidad del viento) influyen en las concentraciones de contaminantes [Tai et al., 2010]. Estos patrones sugieren que los modelos de imputación deben capturar dependencias temporales, espaciales y correlaciones entre variables.

Dado que los modelos de imputación, como interpolación lineal, k-vecinos más cercanos o redes neuronales, varían en precisión dependiendo de las características del dataset [Junninen et al., 2004, Gong and Ordieres-Meré, 2018], comparar sus resultados visualmente puede mejorar la calidad de los datos imputados. Este proyecto propone desarrollar un visualizador interactivo que facilite la comparación de modelos de imputación preentrenados, permitiendo a los usuarios evaluar cuál modelo preserva mejor los patrones temporales, espaciales y correlaciones del dataset, apoyando la toma de decisiones informadas.

2. Definición del Problema

El dataset presenta valores nulos significativos (8.16 % en PM10, 1.59 % en PM2.5, 14.74 % en presión) y datos sucios, como valores extremos (PM25 = 1463 $\mu g/m^3$) y valores no físicos (PM25 = 0 $\mu g/m^3$) [Zheng et al., 2015]. Estos problemas dificultan el cálculo del Índice de Calidad del Aire (AQI) y la construcción de modelos predictivos confiables [Bai et al., 2018]. Los modelos de imputación, como los basados en estadísticas, aprendizaje automático o redes neuronales profundas, varían en su capacidad para manejar dependencias temporales (ciclicidad estacional), espaciales (latitud/longitud) y correlaciones entre variables (por ejemplo, PM2.5 y PM10) [Junninen et al., 2004, Yi et al., 2018].

Sin una herramienta que facilite la comparación visual de los resultados de imputación, los usuarios no pueden evaluar fácilmente cuál modelo preserva mejor las propiedades estadísticas y los patrones del dataset. El problema a resolver es: *¿Cómo diseñar una herramienta interactiva que permita visualizar y comparar los resultados de múlti-*

ples modelos preentrenados de imputación para datos de calidad del aire, considerando patrones temporales, espaciales y correlaciones, para facilitar la selección del modelo más adecuado?

3. Justificación

La necesidad de un visualizador de resultados de imputación se fundamenta en los siguientes puntos:

1. **Alta prevalencia de valores nulos**: Variables clave como PM10 (8.16 % nulos) y presión (14.74 % nulos) presentan datos faltantes que afectan los análisis [Zheng et al., 2015]. La imputación es esencial, pero los modelos varían en precisión según las características del dataset [Junninen et al., 2004].
2. **Patrones complejos**: El dataset muestra ciclicidad estacional (picos de PM2.5 en invierno), correlaciones entre contaminantes (PM2.5 y PM10: 0.873) y variaciones geográficas (medianas de PM2.5 de 200 µg/m^3 a 800 µg/m^3 entre ciudades) [Zheng et al., 2015]. Los modelos de imputación deben preservar estas propiedades.
3. **Variabilidad en modelos de imputación**: Métodos como k-vecinos capturan relaciones espaciales, mientras que las redes neuronales modelan dependencias temporales [Gong and Ordieres-Meré, 2018, Yi et al., 2018]. Comparar visualmente sus resultados es crucial para evaluar su idoneidad.

4. Objetivos

4.1. Objetivo General

Diseñar y desarrollar un visualizador interactivo que compare los resultados de imputación de múltiples modelos preentrenados para datos de calidad del aire, destacando su capacidad para preservar patrones temporales, espaciales y correlaciones, facilitando la selección del modelo más adecuado.

4.2. Objetivos Específicos

1. Implementar un visualizador de imputación con al menos 2–3 modelos preentrenados que consideren ciclicidad estacional, coordenadas geográficas y correlaciones entre variables.
2. Desarrollar una interfaz interactiva con gráficos comparativos (series temporales, boxplots, mapas de calor, scatterplots geográficos) para evaluar los resultados de imputación.

3. Comparar estadísticas descriptivas y patrones (estacionalidad, correlaciones) de los datos imputados con los originales para validar la calidad de la imputación.

4. Permitir a los usuarios seleccionar el modelo de imputación más adecuado según criterios visuales y estadísticos, considerando condiciones extremas y ubicaciones específicas.

Referencias

[Bai et al., 2018] Bai, L., Wang, J., Ma, X., and Lu, H. (2018). Air pollution forecasts: An overview. *International Journal of Environmental Research and Public Health*, 15(4):780.

[Gong and Ordieres-Meré, 2018] Gong, B. and Ordieres-Meré, J. (2018). Prediction of daily maximum ozone threshold exceedances by preprocessing and ensemble artificial neural networks in the beijing-tianjin-hebei region. *Atmospheric Environment*, 181:158–167.

[He et al., 2017] He, J., Gong, S., Yu, Y., Yu, L., Wu, L., Mao, H., Song, C., Zhao, S., Liu, H., Li, X., and Li, R. (2017). Air pollution characteristics and their relation to meteorological conditions during 2014–2015 in major chinese cities. *Environmental Pollution*, 223:484–496.

[Junninen et al., 2004] Junninen, H., Niska, H., Tuppurainen, K., Ruuskanen, J., and Kolehmainen, M. (2004). Methods for imputation of missing values in air quality data sets. *Atmospheric Environment*, 38(18):2895–2907.

[Tai et al., 2010] Tai, A. P. K., Mickley, L. J., and Jacob, D. J. (2010). Correlations between fine particulate matter (pm2.5) and meteorological variables in the united states: Implications for the sensitivity of pm2.5 to climate change. *Atmospheric Environment*, 44(32):3976–3984.

[World Health Organization, 2016] World Health Organization (2016). Ambient air pollution: A global assessment of exposure and burden of disease. *World Health Organization Report*. Accessed: June 2025.

[Yi et al., 2018] Yi, X., Zhang, J., Wang, Z., Li, T., and Zheng, Y. (2018). Deep distributed fusion network for air quality prediction. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2493–2502.

[Zheng et al., 2015] Zheng, Y., Yi, X., Li, M., Li, R., Shan, Z., Chang, E., and Li, T. (2015). Forecasting fine-grained air quality based on big data. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 2267–2276. ACM.