Análisis comparativo de concentraciones de PM2.5 entre estaciones de referencia y sensores de bajo costo

Pedro Alejandro Osma Porras-22414431-Física
Yasser Jasid Holguin Soto-2240688-Física

Septiembre 2025

1 Introducción

La medición de la calidad del aire constituye un aspecto esencial para comprender el impacto de los contaminantes en la salud pública y en el medio ambiente. Entre dichos contaminantes, el material particulado fino (PM2.5) destaca por su relevancia, debido a su capacidad de penetrar en el sistema respiratorio humano y generar efectos adversos. Para cuantificar este contaminante se emplean tanto estaciones de monitoreo oficiales, administradas por entidades ambientales, como sensores de bajo costo, cuya creciente implementación ha permitido ampliar la cobertura espacial de las mediciones. No obstante, surge la necesidad de establecer comparaciones que permitan determinar la confiabilidad y la precisión de los sensores frente a las estaciones de referencia.

El presente trabajo, desarrollado en el marco de la asignatura *Métodos Matemáticos para Físicos*, tiene como propósito aplicar herramientas matemáticas al estudio del error existente entre las mediciones de PM2.5 provenientes de estaciones oficiales y aquellas registradas por sensores independientes. La metodología se fundamenta en el cálculo de distancias entre funciones, lo que posibilita cuantificar de manera objetiva el grado de discrepancia entre ambos conjuntos de datos a lo largo del tiempo.

Se deben tener en cuenta conceptos como El RMSE (Root Mean Square Error) hace referencia al error cuadrático medio; es una medida que indica qué tan grandes son, en promedio, las desviaciones entre los valores observados y los estimados, dándole más peso a los errores grandes. El MAE (Mean Absolute Error) corresponde al error absoluto medio; refleja el promedio de las diferencias en valor absoluto entre lo observado y lo estimado, sin importar el signo.

De esta manera, el proyecto ilustra la pertinencia del uso del formalismo físico y matemático en la validación de instrumentos de medición. Asimismo, ofrece una medida clara del error asociado, aspecto que resulta indispensable tanto para la interpretación científica de los datos como para apoyar la toma de decisiones en materia de calidad del aire y salud pública.

2 Metodología

El siguiente proyecto se realizó en dos fases metodologicas, enfocadas en el correcto analisis y revisión de los datos suministrados:

2.1 Fase I: Análisis y tratamiento de datos

Inicialmente se realizó una revisión detallada del enunciado del taller y de los documentos base proporcionados. Entonces, se procede a organizar las bases de datos obtenidas de las estaciones de referencia y de los sensores. Para ello, se emplean herramientas computacionales en `Python`, haciendo uso de librerías como `pandas` y `matplotlib` para:

Reporte de investigación del subgrupo 3, grupo f1, presentado al profesor Rogelio Ospina Ospina en la asignatura de Laboratorio de Física 2. Fecha: 14/09/2025.

- **Limpieza y normalización de datos.**
  - El estudio realiza una homogeneización los nombres de columnas (tiempo y $PM_{2.5}$), convierte las marcas de tiempo mediante un parseo robusto (seriales de Excel y cadenas con `dayfirst`/`yearfirst`) y normaliza a *naive UTC*.
  - El conjunto de datos elimina filas con valores nulos, "NoData" u otras entradas no numéricas; además ordena por tiempo y usa el tiempo como índice. Cuando existen mediciones múltiples en el mismo instante, promedia dichas observaciones.
  - La comparación entre referencia y sensor se realiza tras una **alineación por cercanía temporal**, implementada con `merge_asof` y una tolerancia adaptativa basada en el paso de muestreo típico de ambas series (acotada a una fracción de la ventana de suavizado). El número de emparejamientos válidos se reporta como *pares*.
- **Promedios móviles (12, 24, 36, 48 h).**
  - La metodología ofrece el siguiente suavizado:
    Modo loose: `rolling` por ventana de tiempo (`"12h"`, `"24h"`, ...) con `min_periods=1`; aprovecha muestreos irregulares aunque introduce sesgos de borde.
  - Las gráficas de series se presentan como *nubes de puntos* en los instantes observados (sin líneas interpoladas) para no inducir tendencias artificiales.
- **Distancia entre funciones (sensor vs. referencia).**
  - Sea $e_i = y_i^{(\text{sensor})} - x_i^{(\text{ref})}$ para $i = 1, \ldots, n$ (pares alineados).
  - La **norma euclídea cruda** se define como

$$D \;=\; \sqrt{\sum_{i=1}^{n} e_i^2}.$$

  - Para comparar ventanas con distinto número de pares, el análisis emplea **RMSE** y **MAE**:

$$\text{RMSE} = \sqrt{\tfrac{1}{n}\sum e_i^2} = \tfrac{D}{\sqrt{n}}, \qquad \text{MAE} = \tfrac{1}{n}\sum |e_i|.$$

  - En presencia de muestreo irregular, el estudio considera la **norma L^2 ponderada por tiempo**:

$$D_{L^2(t)} = \sqrt{\sum_{i=1}^{n-1} \frac{e_i^2 + e_{i+1}^2}{2}\,\Delta t_i}, \quad \Delta t_i = t_{i+1} - t_i.$$

- Se grafica la evolución de la discrepancia (RMSE/MAE/D) en función del tamaño de ventana, destacando la ventana que minimiza el error con un número de *pares* suficiente.

- **Calibración por mínimos cuadrados y tolerancias.**
  - El procedimiento ajusta dos modelos: (i) OLS con intercepto, $y = a + b\,x$, y (ii) ajuste forzado por el origen, $y = \lambda x$; se reportan R^2, RMSE y MAE de los residuos.
  - Cuando se dispone de incertidumbres instrumentales, el análisis contempla *WLS* (pesos $w_i = 1/\sigma_{y,i}^2$) y, si existen errores en ambas variables, *Deming/ODR* (cociente $\lambda = \sigma_y^2/\sigma_x^2$).
  - La **banda de tolerancia** alrededor de la recta ajustada $\hat{y}$ se define punto a punto como

$$\mathrm{tol}_i = \max\Big(\underbrace{\mathrm{ABS}}_{\mu\mathrm{g/m}^3},\ \underbrace{r\,|\hat{y}_i|}_{\text{relativa}},\ \underbrace{k\,\sqrt{\sigma_y^2 + b^2\sigma_x^2}}_{\text{propagación}} \Big),$$

    y clasifica como *fuera de tolerancia* a los puntos con $|y_i - \hat{y}_i| > \mathrm{tol}_i$. Dichos puntos se destacan en rojo y se exportan a un CSV con sus tiempos, valores y residuos.

- **Visualizaciones y entregables.**
  - Gráficas de series originales y suavizadas (sólo puntos) para referencia y sensor.
  - Curvas de discrepancia (RMSE/MAE/D) frente al tamaño de ventana.
  - Diagrama de dispersión sensor–referencia con línea $y = x$, recta ajustada, banda de tolerancia y marcación explícita de outliers.
  - Figuras y tablas se guardan automáticamente en una carpeta de salida, junto con un resumen de métricas por ventana y un CSV de fuera de tolerancia.

- **Aspectos de robustez.**
  - El flujo controla zonas horarias (conversión a UTC y deslocalización), elimina duplicados y entradas "NoData".
  - La alineación por cercanía usa tolerancia *creciente* hasta conseguir emparejamientos; en ausencia de solape suficiente, se incluye un gráfico de respaldo en el dominio temporal.
  - Los dos modos de suavizado (*loose*/*strict*) permiten equilibrar cobertura y sesgo de borde.

2.2 Fase II: Realización del informe

Esta fase organiza y condensa los productos de la Fase I en una narrativa coherente. Primero se fija la notación y unidades (con `siunitx`) y se describe brevemente el flujo de datos limpio–alineado. Luego, el informe se estructura en cuatro pasos: (i) selección de escalas mediante la curva RMSE–ventana (barrido 2–120 h), que orienta qué promedios destacar; (ii) láminas comparativas E–S para ventanas representativas (12, 24, 72, 96 y 120 h), siempre como nubes de puntos; (iii) tablas sintéticas por ventana con pares, alcances y distancias (RMSE/MAE/D), facilitando la lectura cruzada con las figuras; y (iv) calibración por mínimos cuadrados con su banda de tolerancia, marcando puntos fuera de especificación. Finalmente, se incluyen anexos de reproducibilidad (código y datos agregados) y referencias internas claras a tablas y figuras; las conclusiones se incorporarán en la sección correspondiente.

3 Tratamiento de datos

En el contexto del trabajo realiza, definimos alcance como la proporción de mediciones en las que el sensor logra reproducir los valores de la referencia dentro de un margen de tolerancia previamente definido. En otras palabras, mide hasta qué punto el sistema de medición es capaz de mantenerse "dentro de lo aceptable" frente a la referencia.

Table 1: Comparación de alcances (ref/sen) y distancia D para distintas ventanas (horas).

Horas	Pares	Alcance ref	Alcance sen	Distancia
4	3819	24.48	41.08	478.69
8	3819	22.20	37.21	439.34
12	3819	20.02	36.43	420.82
16	3819	19.96	34.24	410.95
20	3819	19.22	33.26	403.46
24	3819	18.63	33.11	398.12
28	3819	17.97	32.71	394.77
32	3819	17.57	32.16	392.22
36	3819	17.28	32.36	389.71
40	3819	17.06	31.74	387.46
44	3819	16.72	31.08	385.12
48	3819	16.04	30.75	382.70

Table 2: Distancia D y alcances de los promedios móviles por ventana.

Ventana	D	Alcance ref	Alcance sen
12 h	420.81	20.02	36.43
24 h	398.11	18.63	32.36
36 h	389.71	17.28	29.14
48 h	369.23	16.04	382.70

4 Análisis de resultados

En esta sección se presentan los resultados obtenidos al aplicar diferentes tamaños de ventana para el cálculo de promedios móviles sobre las series de concentración de PM2.5 provenientes de la estación de referencia y del sensor. El análisis se centra en la distancia entre funciones definida como

$$D = \sqrt{\sum_j \big(\text{ref}(t_j) - \text{sen}(t_j)\big)^2},$$

donde $\text{ref}(t_j)$ corresponde al promedio móvil de la estación y $\text{sen}(t_j)$ al promedio móvil del sensor, emparejados en función del tiempo.

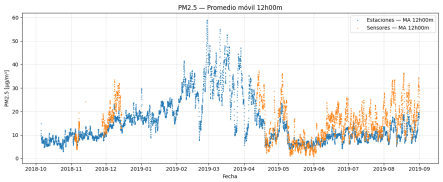
4.1 Resultados cuantitativos

La Tabla 2 resume los valores de D para distintos tamaños de ventana, los cuales permiten observar de forma clara el decrecimiento de la distancia conforme aumenta el promedio. Se observa que conforme la ventana aumenta, la distancia tiende a disminuir.

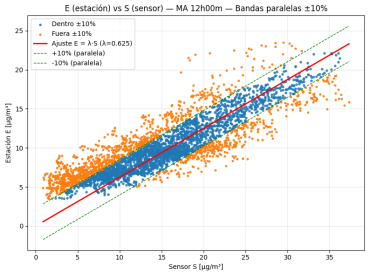
4.2 Interpretación de las gráficas

Las gráficas de promedios móviles permiten visualizar cómo se suavizan las variaciones de corto plazo junto con el alcance de cada promedio mediante minimos cuadrados conforme se amplía el tamaño de la ventana:

- **Ventanas de 12 h:** Se nota claramente unos cuantos picos locales de concentración, además de una gran variabilidad entre los datos. Se pueden evidenciar las diferencias entre sensor y estación, es decir, un mayor error(o según la definición usada en este trabajo, una mayor distancia). Tengase en cuenta que la cantidad de datos dentro de la tolerancia para este promedio es de 2391/3819 además de que los valores λ =0.625216, $RMSE$ =2.521, MAE =2.032



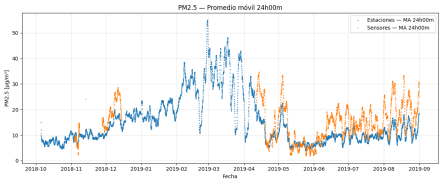
(a) Gráfica tomando ventana para el promedio de 12h.



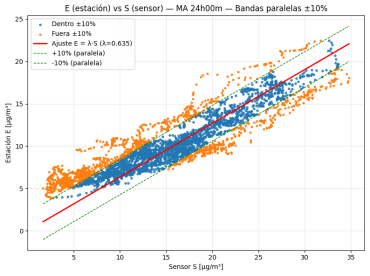
(b) Gráfica del alcance para una ventana movil de 12h.

Figure 1: Comparación de resultados obtenidos en el análisis.

- **Ventanas de 24h:** Al realizar una comparación con su contraparte de un promedio mucho menor, se puede ver claramente el como los datos varian de menor manera, además, las curvas tienden a aproximarse, por ende, la distancia disminuye. En cambio, notese que la cantidad de datos dentro de la tolerancia para este promedio es de 2467/3819 además de que los valores λ =0.63482, $RMSE$ =2.111, MAE =1.712



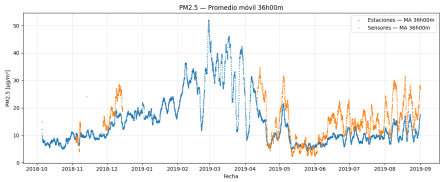
(a) Gráfica tomando ventana para el promedio de 24h.



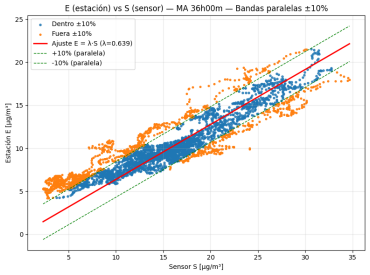
(b) Gráfica del alcance para una ventana movil de 24h.

Figure 2: Comparación de resultados obtenidos en el análisis.

- **Ventana de 36h:** Se obtiene una variación bastante lineal con respecto a los anteriores promedios, el suavizado aumenta y por ende la ditancia aumenta.notese que la cantidad de datos dentro de la tolerancia para este promedio es de 2553/3819 además de que los valores λ =0.638602 , $RMSE$ =1.982 , MAE =1.596



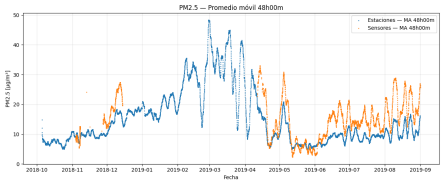
(a) Gráfica tomando ventana para el promedio de 36h.



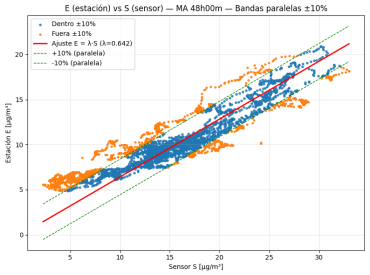
(b) Gráfica del alcance para una ventana movil de 36h.

Figure 3: Comparación de resultados obtenidos en el análisis.

- **Ventana de 48h:** producen un suavizado fuerte, de modo que ambas series convergen a la tendencia global de fondo. La distancia entre funciones alcanza aquí sus valores mínimos dentro del rango de promedio máximo que se manejó en el presente trabajo.notese que la cantidad de datos dentro de la tolerancia para este promedio es de 2549/3819 además de que los valores λ =0.64172 , $RMSE$ =1.873, MAE =1.511



(a) Gráfica tomando ventana para el promedio de 48h.



(b) Gráfica del alcance para una ventana movil de 48h

Figure 4: Comparación de resultados obtenidos en el análisis.

- **Comparación:** Cada punto corresponde a una ventana distinta. La RMSE cae rápido para ventanas pequeñas (por alta frecuencia y desfases locales) y se estabiliza al aumentar w —el promedio móvil filtra ruido—, mostrando un “codo” alrededor de 12–24 h que equilibra reducción de ruido y resolución temporal.

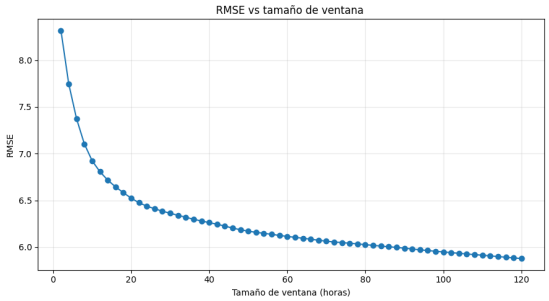


Figure 5: Distancia RMSE vs. tamaño de ventana.

4.3 Discusión sobre la variación de la distancia

Alcance. El análisis se circunscribe a evaluar la coherencia y calibración entre la estación de referencia (E) y el sensor (S) mediante promedios móviles, alineación temporal por cercanía y métricas de discrepancia. Se estudian ventanas de 12, 24, 36 y 48 horas; se comparan series originales y suavizadas; y se incluye una calibración por mínimos cuadrados con banda de tolerancia alrededor de la recta ajustada. No se persigue explicar causalmente los episodios de contaminación ni corregir el instrumento, sino cuantificar su concordancia en el intervalo observado.

Tolerancia y criterios de concordancia. La recta de mínimos cuadrados se expresa como $E = \lambda\, S$. Un par (S_i, E_i) se clasifica *fuera de tolerancia* si $|E_i - \hat{y}_i| > \mathrm{tol}_i$. En la figura mostrada, la banda paralela a la recta refleja una tolerancia constante (o proporcional a $\hat{y}$) que permite distinguir visualmente los puntos concordantes (azul) de los discrepantes (naranja). El porcentaje de puntos fuera de tolerancia y su distribución a lo largo del rango de S constituyen indicadores prácticos de consistencia y sesgos locales.

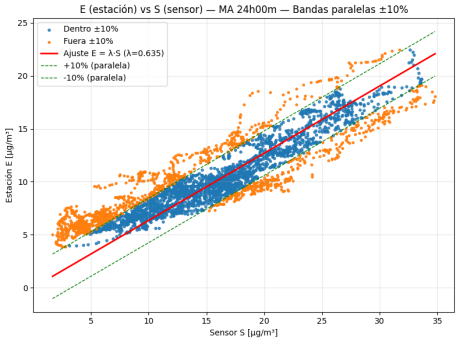


Figure 6: Minimos cuadrados.

Desde la perspectiva de filtrado, el promedio móvil actúa como *pasa–bajos*:

al crecer la ventana se atenúan las componentes de alta frecuencia de $E - S$, disminuye el peso de los residuos y, por tanto, D. En paralelo, las métricas normalizadas (RMSE y MAE) permiten comparar ventanas con distinto número de pares, evitando que D (sensibles a n) sesgue la selección.

Lecturas a partir de la calibración. La pendiente (b o λ) resume un *factor de escala* entre sensor y estación; el intercepto (a), un *offset* sistemático. Una banda de tolerancia bien definida alrededor de $\hat{y}$ ayuda a separar discrepancias esperables (ruido y variabilidad natural) de *outliers* operativos. Típicamente, las ventanas intermedias (12-36 h) ofrecen un equilibrio entre reducción de ruido, estabilidad de los parámetros (a, b o λ) y fracción razonable de puntos dentro de tolerancia, mientras que ventanas excesivamente cortas maximizan D y la tasa de fuera de tolerancia, y ventanas muy largas suavizan en exceso los episodios de interés.

5 Conclusiones

Al limitar el barrido de ventanas móviles a **0.5–48 h** (pasos de 0.5 h) y evaluar la concordancia entre la estación (E) y el sensor (S) mediante la recta de mínimos cuadrados $E = \lambda S$, se obtienen las siguientes conclusiones:

La distancia disminuye al aumentar las ventanas. Pues,sea $e_i = E_i - S_i$ la diferencia en pares alineados y $D = \sqrt{\sum e_i^2}$ la distancia euclídea. La mayor magnitud de D para ventanas cortas obedece a:

1. **Alta frecuencia conservada:** promedios cortos filtran poco; picos y oscilaciones rápidas quedan presentes y amplifican $|e_i|$.
2. **Desfases locales:** pequeños corrimientos temporales entre series (subminutales o de pocos minutos) impactan más cuando la ventana es reducida.
3. **Eventos puntuales:** irrupciones breves (p. ej., incrementos súbitos) pesan relativamente más; al aumentar la ventana, su contribución se diluye.
4. **Efectos de borde y muestreo irregular:** con ventanas pequeñas y datos no uniformes, los extremos contienen menos información efectiva, lo que vuelve más volátiles los residuales.

Además, en terminos más generales podemos llegar a:

1. **Tendencia del error con la ventana.** Al aumentar el tamaño de la ventana hasta 48 h, las métricas de discrepancia (*RMSE* y *MAE*) *descienden de forma monótona* y muestran un *claro aplanamiento* a partir de **24–36 h**. Por encima de ese intervalo, las mejoras adicionales son marginales (rendimientos decrecientes), pero continúan disminuyendo levemente hasta 48 h.

2. **Compromiso error–alcance.** El suavizado más fuerte reduce la distancia entre funciones, pero *comprime el alcance* de ambas series. En el rango analizado, el compromiso más equilibrado entre *baja discrepancia* y *preservación de variabilidad* se observa en **24–36 h**. Las ventanas **36–48 h** maximizan la concordancia, aunque con una pérdida de detalle mayor en episodios breves.

3. **Calibración y tolerancia operativa.** La pendiente λ permanece estable al pasar de 24 h a 48 h, lo que indica coherencia de escala entre S y E tras el filtrado. Con la banda *paralela* de $\pm 10\%$ alrededor de la recta, la fracción de puntos *dentro de tolerancia* crece claramente con la ventana y se estabiliza cerca del extremo superior (36–48 h), coherente con la disminución del residuo.

4. **Recomendación práctica.**

   - Para *seguimiento operativo* y reportes diarios donde se priorice robustez frente a ruido: usar **36–48 h**.
   - Para *investigación* que requiera balancear detalle temporal y concordancia: preferir **24–36 h** como ventana de trabajo por defecto.
   - Para *análisis de eventos rápidos* (picos de corta duración): considerar ventanas **12–24 h**, asumiendo un RMSE mayor y mayor proporción de excedencias al 10%.

5. **Definición de paso y el como afecta a los resultados.** La elección del paso de muestreo influye directamente en la calidad de la comparación entre referencia y sensor. Con pasos muy pequeños, se conserva gran detalle temporal, pero aumenta la presencia de huecos y disminuye la cobertura de pares válidos. En cambio, con pasos demasiado amplios, se gana estabilidad y mayor número de coincidencias, pero se pierde información fina y se suavizan en exceso las variaciones reales. Por ello, definir un paso intermedio resulta fundamental para equilibrar precisión, correlación y cobertura, permitiendo una representación más fiel de los datos sin sacrificar su comparabilidad.

Al no explorar ventanas > 48 h, no se persigue la tendencia de fondo de muy baja frecuencia; las conclusiones son válidas para escalas diarias a sub–semanales. Si se requiriera caracterizar fondo de larga duración, habría que extender el barrido y aceptar una reducción adicional del alcance.

References

[1] W. McKinney, "Data Structures for Statistical Computing in Python," *Proceedings of the 9th Python in Science Conference (SciPy 2010)*, pp. 51–56, 2010.

[2] C. R. Harris, K. J. Millman, S. J. van der Walt, *et al.*, "Array programming with NumPy," *Nature* **585**, 357–362, 2020.

[3] J. D. Hunter, "Matplotlib: A 2D graphics environment," *Computing in Science & Engineering* **9**(3), 90–95, 2007.

[4] P. Virtanen, R. Gommers, T. E. Oliphant, *et al.*, "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python," *Nature Methods* **17**, 261–272, 2020.

[5] C. J. Willmott y K. Matsuura, "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance," *Climate Research* **30**, 79–82, 2005.

[6] J. H. Seinfeld y S. N. Pandis, *Atmospheric Chemistry and Physics: From Air Pollution to Climate Change*, 3. ed., Wiley, 2016.

[7] J. Wright, "`siunitx`: A comprehensive (SI) units package," *TUGboat* **39**(1), 2018.