Taller de Distancias y Calibración Lineal de Sensores de $PM_{2.5}$

Métodos Matemáticos para la Física I

Jefferson Meza (2231482)

22 de septiembre de 2025

Resumen

Este informe documenta un protocolo cuantitativo, reproducible y trazable para comparar y calibrar lecturas de un sensor de bajo costo de material particulado fino ($PM_{2.5}$) respecto de una estación de referencia (patrón). El procedimiento sigue los lineamientos del *Taller de Distancias*: (i) armonización temporal y construcción de promedios móviles sobre ventanas comunes, (ii) selección del ancho de ventana mediante la minimización de la distancia euclidiana entre series suavizadas, (iii) calibración lineal sin intercepto $f \approx \alpha\,\hat{f}$, (iv) validación temporal fuera de muestra, y (v) delimitación del alcance de validez bajo una tolerancia especificada. Aplicado al conjunto de datos suministrado, se obtiene una ventana óptima de 180.00 min, distancia mínima $d(W)^{=497.44}$ y un factor de calibración global $\alpha = 0.59$. El desempeño agregado es MAE = 3.05 y RMSE = 3.83; en validación mitad/mitad, $\alpha_{\text{train}} = 0.64$ con $\text{MAE}_{\text{test}} = 2.68$ y $\text{RMSE}_{\text{test}} = 3.45$. Con tolerancia absoluta $\tau = 5.00\,\mu\text{g}\,\text{m}^{-2}$, el 81.80 % de los puntos queda dentro de tolerancia y el rango operativo del sensor (suavizado) donde la calibración es confiable es $[0.03, 40.39]\ \mu\text{g}\,\text{m}^{-2}$.

1. Introducción y motivación

La proliferación de sensores de bajo costo ha democratizado el monitoreo de calidad del aire; sin embargo, sus lecturas suelen exhibir ruido, sesgos sistemáticos, deriva y sensibilidad a variables ambientales. Una *calibración* frente a una estación de referencia es indispensable para convertir lecturas brutas en estimaciones útiles para análisis científico. Este trabajo implementa un flujo *end-to-end* de limpieza, alineación, suavizado, comparación, calibración, validación y reporte, con énfasis en reproducibilidad y trazabilidad, en el espíritu de la asignatura Métodos Matemáticos para la Física I.

2. Datos, supuestos y preprocesamiento

2.1. Descripción de los datos

- **Patrón (referencia):** columna `PM2.5` del archivo `Datos Estaciones AMB.xlsx`. La columna temporal presenta variantes (`Date&Time`, `Fecha y Hora`, etc.).
- **Sensor (IoT):** colección de archivos `mediciones_*.csv`, potencialmente con separadores y codificaciones heterogéneas.

2.2. Normalización temporal y de tipo

Para prevenir errores de fusión por zona horaria (`datetime64[ns]` vs `datetime64[ns,UTC]`), toda marca temporal t se normaliza a **UTC naive** mediante el mapeo

$$t \mapsto \mathrm{NaiveUTC}(t) := \text{tz-drop}\big(\mathrm{toUTC}(t)\big). \tag{1}$$

Se convierten todas las magnitudes a tipo numérico, coercionando cadenas y descartando filas sin fecha o valor válidos.

2.3. Grilla temporal común y suavizado

Sea Δt la mediana del paso del patrón; se remuestrean patrón y sensor por mediana (robusto a atípicos) sobre la grilla $t_k = t_0 + k\,\Delta t$. Para cada W (minutos) se define el *promedio móvil centrado*:

$$\overline{f}_k(W) = \frac{1}{m(W)} \sum_{j \in \mathcal{W}_k(W)} f_j, \qquad \overline{\hat{f}}_k(W) = \frac{1}{m(W)} \sum_{j \in \mathcal{W}_k(W)} \hat{f}_j, \tag{2}$$

donde $\mathcal{W}_k(W)$ es la ventana centrada en t_k e $m(W)$ el número de puntos en la ventana. El centrado reduce desfases; W controla el compromiso *sesgo–varianza*.

2.4. Emparejamiento temporal

Se aplica una unión *nearest* con tolerancia δ (típicamente 10–20 min) para obtener pares coetáneos $(\overline{\hat{f}}_i(W), \overline{f}_i(W))$. Esto preserva estructura temporal con pequeños desajustes.

3. Selección de ventana por distancia euclidiana

Para cada W se evalúa

$$d(W) = \left\| \overline{\mathbf{f}}(W) - \overline{\hat{\mathbf{f}}}(W) \right\|_2 = \sqrt{\sum_{i=1}^{n(W)} \big(\overline{f}_i(W) - \overline{\hat{f}}_i(W) \big)^2}. \tag{3}$$

Se define $W^{=\arg\min_W d(W)}$. En los datos analizados: $W^{=180.00\,\mathrm{min}}$ con $d(W^{)=497.44}$. La Figura 1 ilustra la superposición suavizada en W. (Si se dispone de la curva $d(W)$, incluirla como Figura 2 y analizar el posible codo/elbow).

4. Calibración lineal sin intercepto

Se plantea $f \approx \alpha\, \hat{f}$ sin término independiente. Derivando la solución de mínimos cuadrados por el origen:

$$\alpha = \arg\min_{\alpha} \sum_i \left(f_i - \alpha\, \hat{f}_i\right)^2 = \arg\min_{\alpha} \left(\sum_i f_i^2 - 2\alpha \sum_i f_i \hat{f}_i + \alpha^2 \sum_i \hat{f}_i^2 \right), \tag{4}$$

$$\frac{d}{d\alpha}(\cdot) = 0 \;\Rightarrow\; -2\sum_i f_i \hat{f}_i + 2\alpha \sum_i \hat{f}_i^2 = 0 \;\Rightarrow\; \boxed{\alpha = \frac{\sum_i f_i \hat{f}_i}{\sum_i \hat{f}_i^2}}. \tag{5}$$

En nuestro caso, $\alpha_{\mathrm{global}} = 0.59$. La Figura 3 muestra la nube $(\overline{\hat{f}}, \overline{f})$ y la recta $y = \alpha x$.

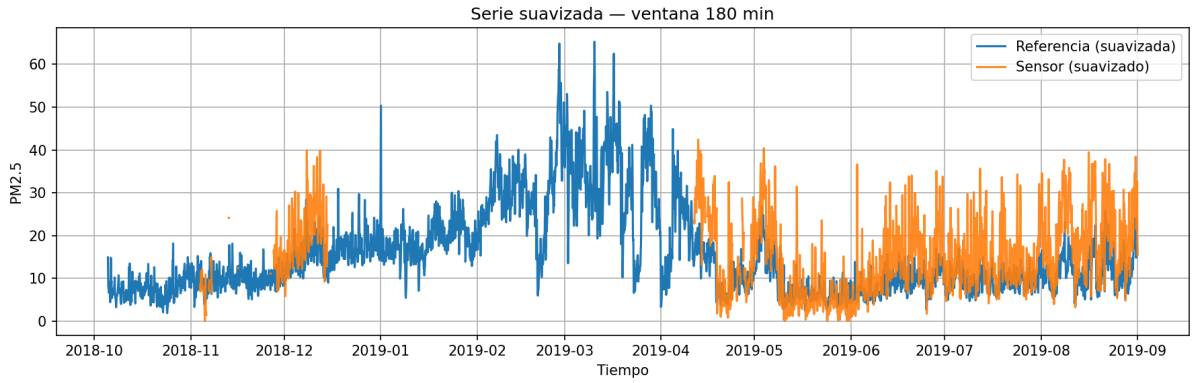

Figura 1: Series suavizadas del patrón y del sensor con $W^{=180.00\,\mathrm{min}}$.

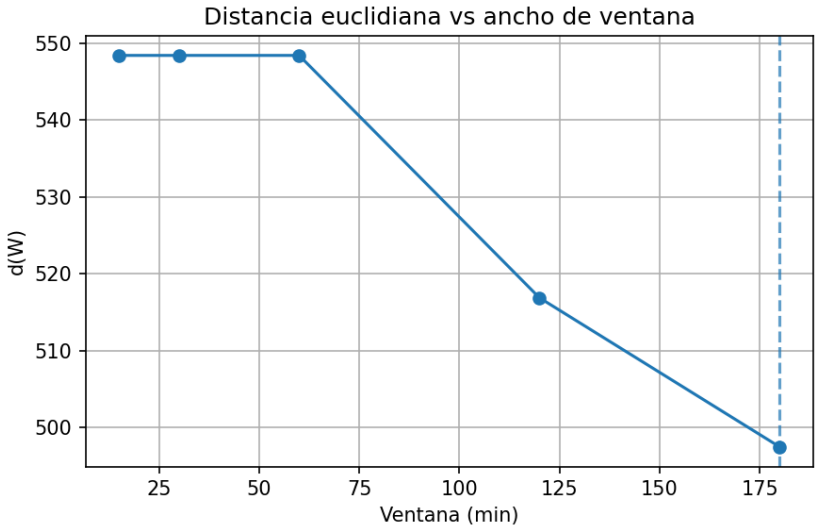

Figura 2: (Opcional) Curva $d(W)$ vs. W.

4.1. Métricas de error

Sea $e_i = f_i - \alpha\,\hat{f}_i$. Reportamos

$$\mathrm{MAE} = \frac{1}{n}\sum_i |e_i|, \qquad \mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_i e_i^2}. \tag{6}$$

Con α global: MAE = 3.05 y RMSE = 3.83.

5. Validación temporal (fuera de muestra)

Se divide cronológicamente el conjunto en dos mitades: entrenamos α_{train} en la primera y evaluamos en la segunda. Se obtienen:

	MAE	RMSE	Observaciones
Global ($\alpha = 0.59$)	3.05	3.83	todas las fechas
Train ($\alpha_{\mathrm{train}} = 0.64$)	3.49	4.31	estima α
Test (con α_{train})	2.68	3.45	generalización temporal

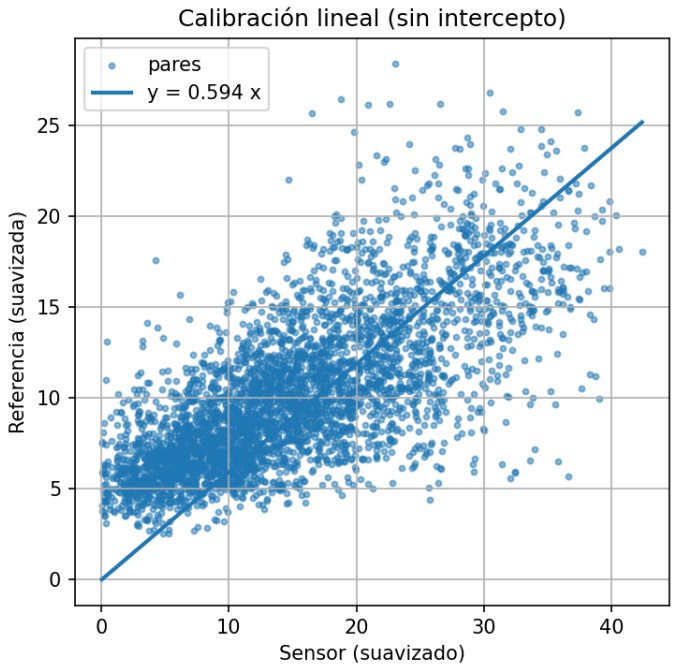

Figura 3: Dispersión y recta de calibración ($y = \alpha x$) en W.

Diagnóstico de residuos. (Opcional) Incluir la serie temporal $e(t)$ y el histograma de residuos (Figuras 4 y 5), comentando simetría, colas y autocorrelación.

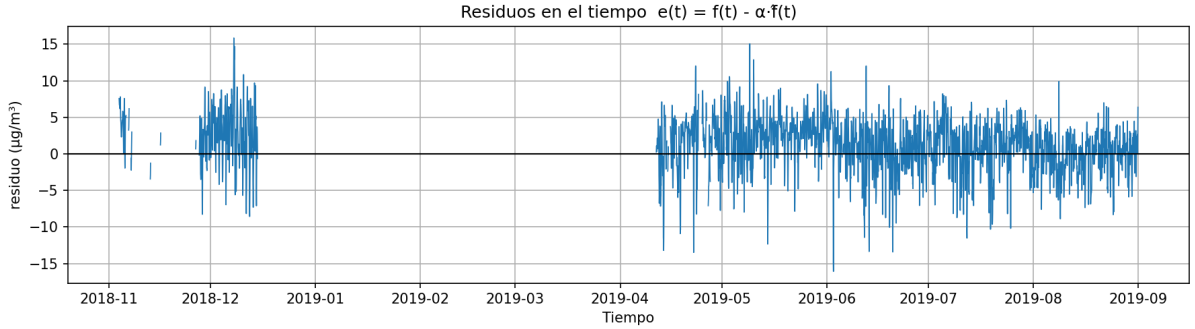

Figura 4: (Opcional) Residuos en el tiempo $e(t) = f(t) - \alpha \hat{f}(t)$.

6. Alcance de validez y tolerancias

Con tolerancia absoluta $\tau = 5.00\,\mu\mathrm{g\,m^{-2}}$, el 81.80 % de los pares cumple $|e_i| \le \tau$. Definimos el *rango operativo* del sensor (suavizado) como

$$[\hat{f}^{(\tau)}_{\text{mín}},\ \hat{f}^{(\tau)}_{\text{máx}}] = \text{mín./máx. de } \overline{\hat{f}}_i \text{ tales que } |e_i| \le \tau. \tag{7}$$

En nuestro caso: [0.03, 40.39] $\mu\mathrm{g\,m^{-2}}$. Si se desea una tolerancia relativa (porcentaje), reemplazar la condición por $|e_i| \le \rho\,|f_i|$.

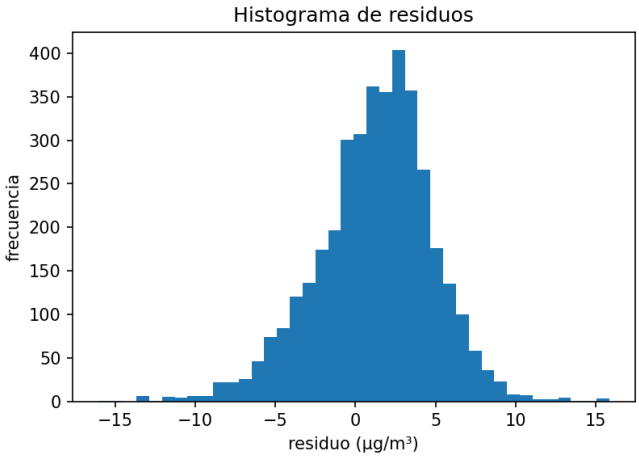

Figura 5: (Opcional) Histograma de residuos.

7. Análisis de sensibilidad y consideraciones prácticas

- **Elección de W:** W pequeño reduce sesgo y aumenta varianza; W grande reduce varianza y aumenta sesgo (desfase/atenuación de picos). El óptimo hallado sugiere estructura de variabilidad de escala horaria (3 h).
- **Estabilidad de α:** comparar α global vs. α_{train} informa sobre deriva estacional o dependencia de humedad/temperatura.
- **Grilla y tolerancia temporal:** una tolerancia de emparejamiento demasiado estricta elimina pares; demasiado laxa mezcla episodios disímiles.
- **Robustez:** la mediana en el remuestreo mitiga atípicos frente a la media; se sugiere evaluar también Huber/biweight si hay outliers persistentes.

8. Conclusiones

El protocolo implementado provee una **calibración lineal simple y operacional** del sensor respecto del patrón, con una ventana óptima $W^{=180.00\,\text{min}}$ que minimiza la discrepancia suavizada y un factor $\alpha = 0.59$ aplicable en el rango operativo estimado. La validación temporal muestra consistencia en MAE/RMSE, lo que respalda el uso de α en periodos similares. Este flujo es reproducible e integrable en pipelines de monitoreo.

Reproducibilidad

El notebook `TallerDistancias.ipynb` automatiza: carga (Excel/CSV), normalización temporal a UTC naive, remuestreo, ventana–distancia, ajuste de α, validación, tolerancia y exportes (`figuras/`, `resultados/`). Este informe se compila con LaTeX; basta colocar $\texttt{serie}_w in 180.pngy$

A. Pseudocódigo del pipeline

Paso 1. Cargar patrón y sensor (auto-detección de columnas de tiempo/valor).

Paso 2. Normalizar tiempo → UTC naive (ambas fuentes).

Paso 3. Remuestrear por mediana en grilla común $\Delta t = \text{mediana}(\Delta t_{\text{patrón}})$.

Paso 4. Para cada W en $\{15, 30, 60, 120, 180\}$ min:

a) Calcular promedios móviles centrados de patrón y sensor.

b) Emparejar por tiempo (nearest, tolerancia δ).

c) Calcular $d(W)$.

Paso 5. Seleccionar $W^{=\arg\min_W d(W)}$.

Paso 6. Con W: estimar α con (5); computar MAE/RMSE.

Paso 7. Validar mitad/mitad: α_{train} en la primera mitad; MAE/RMSE en la segunda.

Paso 8. Alcance de validez: porcentaje dentro de tolerancia y rango operativo.

B. Tabla de símbolos

Símbolo	Descripción
f	serie del patrón (referencia)
$\hat{f}$	serie del sensor (bajo costo)
W	ancho de ventana (minutos)
$d(W)$	distancia euclidiana entre series suavizadas
α	factor de calibración (sin intercepto)
e	residuo $f - \alpha\hat{f}$
τ	tolerancia (absoluta o relativa)