Análisis Comparativo y Calibración de Sensor IoT para PM2.5

Jorge Steven Plata Berdugo / Samuel Alexis Gonzalez Martinez

26 de septiembre de 2025

Índice

1. Introducción

Se analiza la correspondencia entre una estación de referencia (AMB) y un sensor de bajo costo (IoT) para PM2.5. El objetivo es cuantificar diferencias, proponer una calibración lineal y estimar el mínimo conjunto de datos necesarios para obtener una calibración con una tolerancia aceptable.

2. Datos

- Estación de referencia (AMB): período analizado 2018-10-01 a 2019-08-31.
- Sensor IoT: período disponible 2018-11-03 a 2019-09-01.
- Archivos organizados en el repositorio: `data/AMB/`, `data/IoT/`, código en `code/` y figuras en `figures/`.

3. Preprocesamiento y sincronización

1. Conversión de columnas de fecha a `datetime` y normalización de zona horaria: se removió información de timezone en el índice del IoT (`tz_localize(None)`).
2. Filtrado de valores inválidos (NaN) en ambas series.
3. Determinación del período común: **2018-11-03 – 2019-08-31** (superposición usada para el análisis).
4. Alineación por timestamps y eliminación de filas sin datos en ambos sensores (operación `dropna` sobre las dos columnas).

4. Estadísticas descriptivas

4.1. Estación de referencia (AMB)

- Período: 2018-10-01 a 2019-08-31.
- Total de puntos: 8040.
- Puntos válidos: 7921 (98.5 %).
- Media: $15{,}23 \pm 10{,}20\,\mu\mathrm{g\,m^{-3}}$.

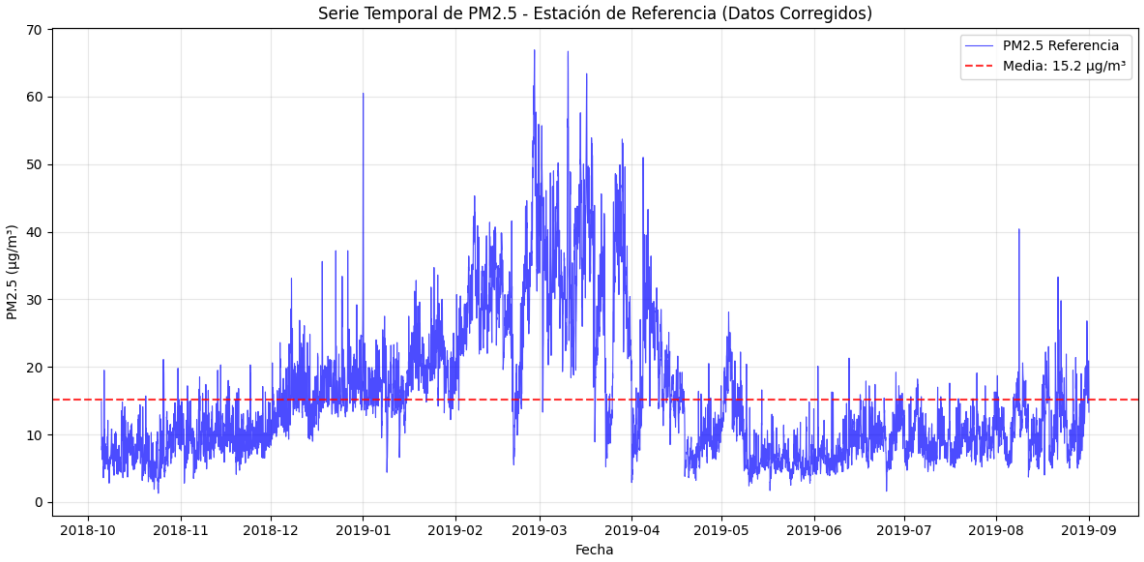

Figura 1: Serie temporal de PM2.5 registrada por la estación de referencia (AMB), tras corrección y filtrado de datos.

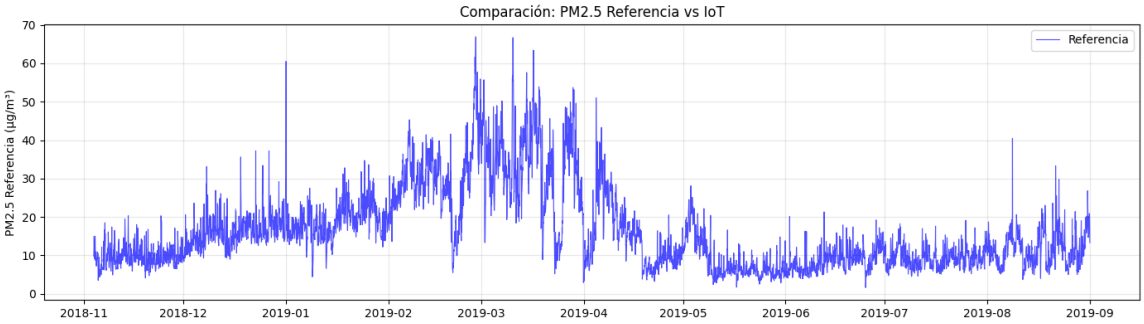

Figura 2: Serie temporal de PM2.5 registrada por el sensor IoT.

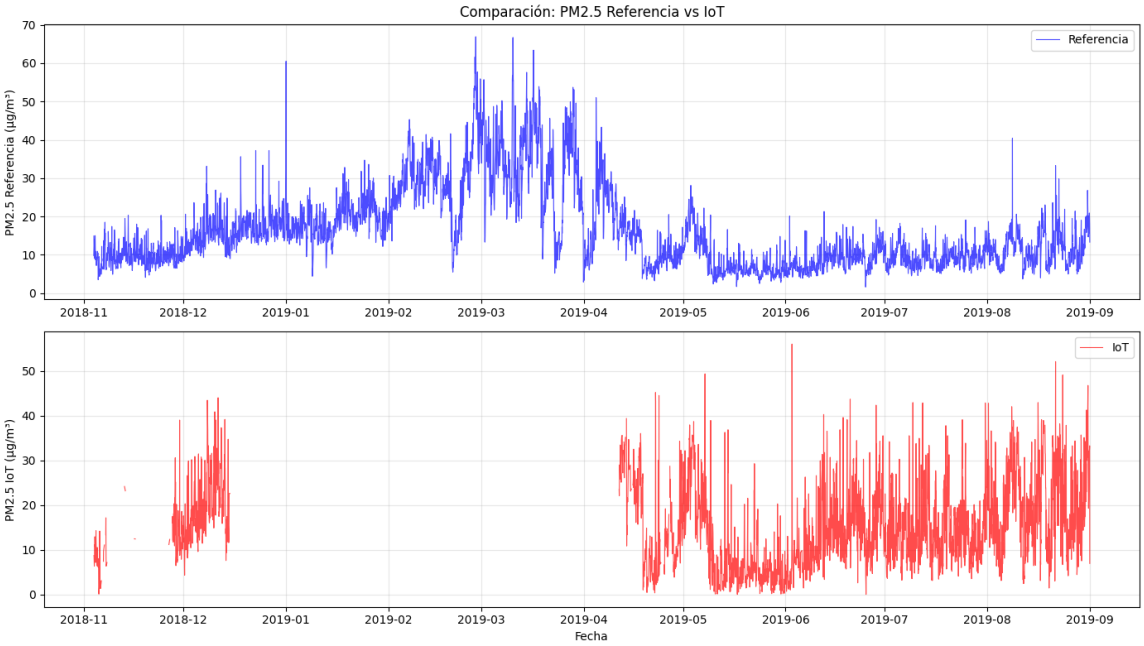

Figura 3: Comparación temporal de PM2.5 entre estación de referencia e IoT en el periodo común.

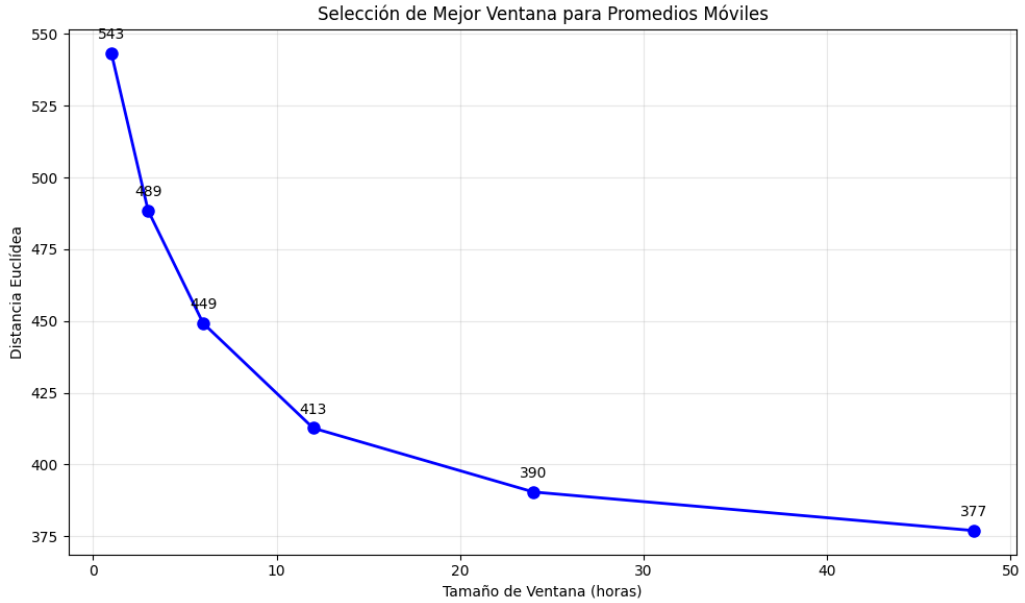

Figura 4: Distancia euclídea entre promedios móviles de referencia e IoT en función del tamaño de ventana.

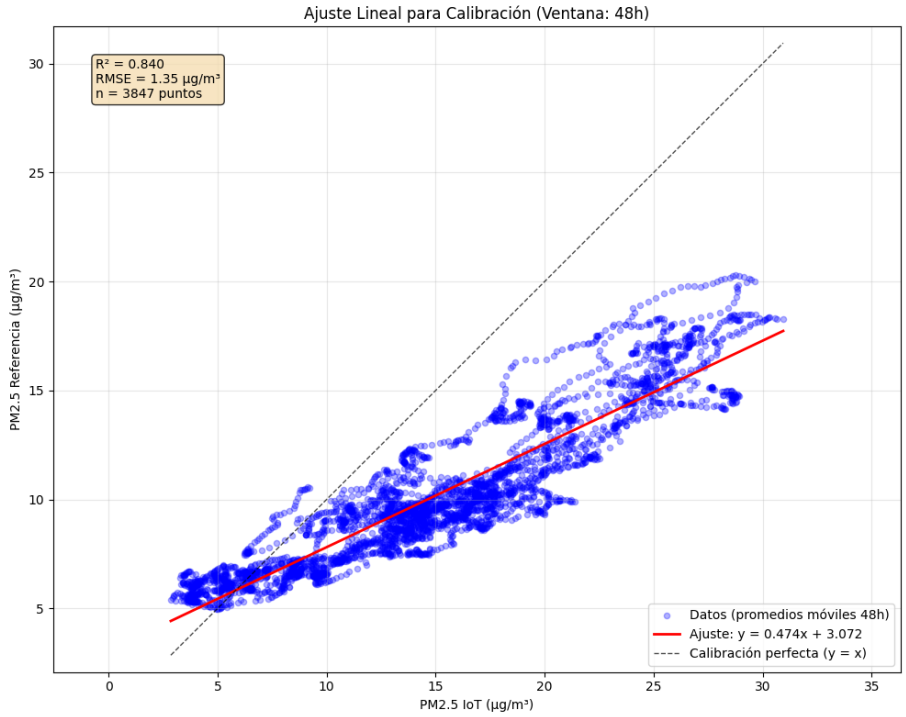

Figura 5: Diagrama de dispersión (IoT vs. referencia, suavizado a 48h). Línea roja: ajuste lineal; línea punteada: calibración perfecta ($y = x$).

5. Distancia euclídea entre promedios móviles

Se calculó la distancia euclídea entre las medias móviles (centradas) de referencia e IoT para ventanas: 1, 3, 6, 12, 24 y 48 horas. La métrica usada:

$$D = \sqrt{\sum_i (R_i - S_i)^2}$$

con R_i y S_i las medias móviles correspondientes.

Cuadro 1: Distancia euclídea según tamaño de ventana (promedios móviles).

Ventana (h)	Distancia	Puntos válidos
1	543.34	3847
3	488.53	3847
6	449.24	3847
12	412.70	3847
24	390.44	3847
48	376.99	3847

Resultado: Ventana óptima = **48 horas** (mínima distancia).

6. Calibración lineal (ajuste y validación)

Se trabajó con series suavizadas por la ventana óptima (48 h). El ajuste lineal usado:

$$PM_{25}^{ref} = \alpha\, PM_{25}^{iot} + \beta$$

6.1. Ajuste con todos los datos (comprobación)

Ajuste sobre todo el conjunto suavizado (verificación):

$$PM_{25}^{ref} = 0{,}4738 \times PM_{25}^{iot} + 3{,}0721$$

R^2 (entrenamiento completo): 0.8405. RMSE (completo): 1.3524 μg m^{-3}.

6.2. Ajuste con partición temporal (train/test)

División temporal: 70 % train / 30 % test (sin mezcla aleatoria).

- Datos en train: 2692 puntos.
- Datos en test: 1155 puntos.

Ajuste final (modelo entrenado solo en train):

$$PM_{25}^{ref} = 0{,}5092 \times PM_{25}^{iot} + 2{,}8874$$

Métricas:

- R^2 en training (comprobación): 0.8405
- R^2 en test: 0.6355
- RMSE en test: 1.4952 μg m^{-3}
- Error medio absoluto (test): 1.28 μg m^{-3}
- Error máximo (test): 3.75 μg m^{-3}
- Percentil 95 de error: 2.82 μg m^{-3}
- Puntos dentro de la tolerancia definida ($\pm$5.0 μg m^{-3}): **100 %** (1155/1155)

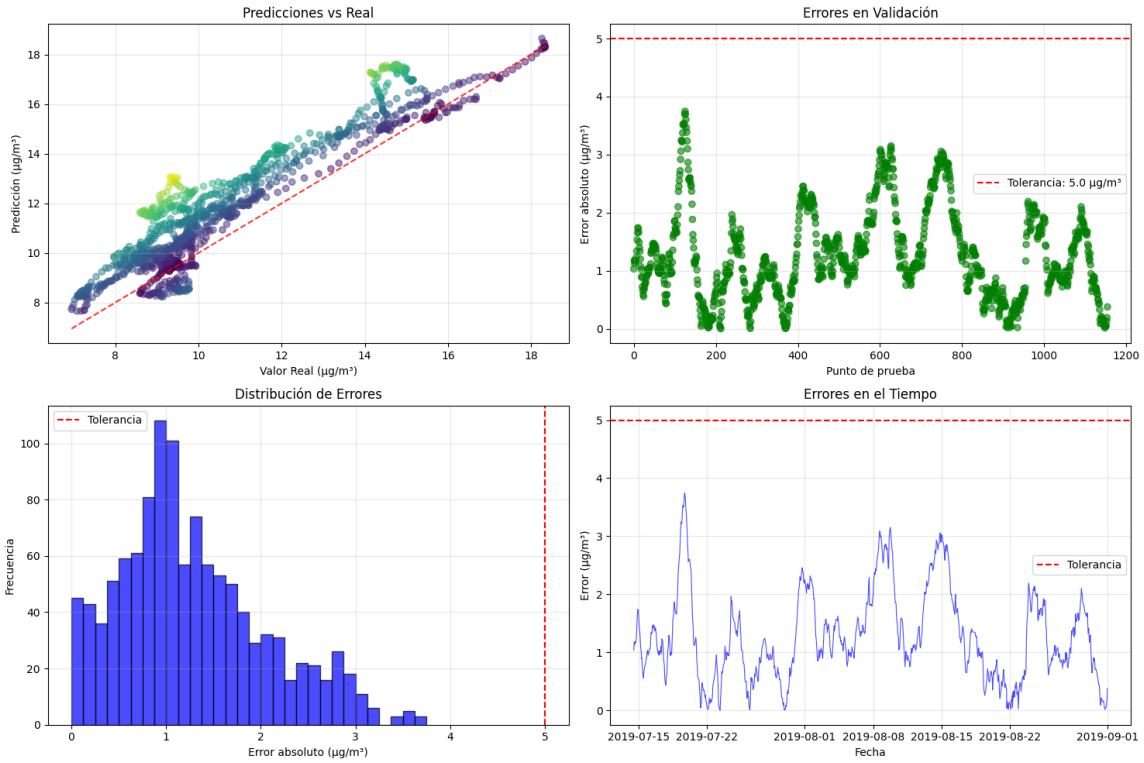

Figura 6: Panel de validación del modelo. (a) Predicciones vs observaciones en conjunto de test, (b) errores absolutos con banda de tolerancia de $\pm 5\,\mu g/m^3$, (c) histograma de errores, (d) evolución temporal de errores.

Cuadro 2: Resultados por tamaño de entrenamiento (fracciones seleccionadas).

Tamaño (fracción)	Puntos train	Puntos test	Precisión (%)	Error medio (g/m^3)
0.1	384	3463	97.98	2.176
0.2	769	3078	100.00	1.713
0.3	1154	2693	100.00	1.711
0.4	1538	2309	100.00	1.868
0.5	1923	1924	100.00	1.910
0.6	2308	1539	100.00	1.466
0.7	2692	1155	100.00	1.278

Resultado clave: criterio de precisión mínimo 95 % alcanzado con **10 %** de los datos (384 puntos). Equivalencia temporal (datos horarios): 384 puntos ≈ 16 días. Con 20 % ya se alcanza 100 % de precisión en el experimento.

7. Recomendación de calibración

Ecuación recomendada (factor de corrección inverso aplicado a las lecturas IoT para estimar referencia):

$$PM_{25}^{corregido} = 1{,}964 \times PM_{25}^{iot} - 5{,}670$$

(este factor surge aplicando la inversa de la pendiente del modelo final entrenado).

8. Discusión

8.1. Organización del código y reproductibilidad

El código se organizó en módulos/etapas: lectura y limpieza, sincronización temporal, cálculo de promedios móviles, cálculo de distancia (grilla de ventanas), ajuste lineal, validación (train/test), análisis de tamaño mínimo y generación de figuras. Las figuras se exportaron como archivos PNG en `figures/`; el código y los datos crudos deben mantenerse en el repositorio (`code/`, `data/`) para asegurar trazabilidad y reproducibilidad.

8.2. Limitaciones

- El estudio utiliza un solo periodo y un solo par estación–sensor; no se evaluó robustez frente a condiciones meteorológicas extremas ni distintos dispositivos IoT.
- El ajuste es lineal; no se exploraron modelos no lineales (por ejemplo, modelos polinómicos, regularizados o basados en árboles) que podrían mejorar performance en condiciones particulares.
- Huecos en la serie IoT reducen la cobertura efectiva; decisiones sobre interpolación/resampleo pueden cambiar resultados.

8.3. Aplicaciones prácticas

Los resultados son aplicables para:

- Calibración inicial de sensores de bajo costo instalados en red colocalizados con estaciones de referencia.
- Estimación rápida de rendimientos de sensores en campañas cortas ($\sim$2–3 semanas).
- Implementación de corrección en tiempo casi real en sistemas de monitoreo con despliegue de IoT.

8.4. Trabajo futuro

- Evaluar modelos alternativos y añadir características (meteorología) al ajuste.
- Analizar estabilidad temporal de la calibración (recalibrar periódicamente).
- Automatizar pipeline para regenerar figuras y tablas (Makefile o scripts de CI).
- Integrar tests unitarios mínimos y documentación en el repositorio.

9. Conclusiones

- La ventana de promedios móviles más adecuada para este dataset es 48 h.
- El ajuste lineal final (train/test) presenta pendiente $\alpha \approx 0{,}509$ e intercepto $\beta \approx 2{,}887$; el modelo aporta errores medios bajos y una proporción muy alta de predicciones dentro de $\pm 5\ \mu g/m^3$.
- El tamaño mínimo de datos horarios para una calibración operativa es de ≈384 puntos (∼16 días).
- Se recomienda aplicar la ecuación de corrección propuesta y mantener un procedimiento de recalibración periódico.

Referencias

- D. J. Hunter "Matplotlib: A 2D graphics environment", Computing in Science Engineering (2007).
- Scikit-learn documentation — regresión lineal y métricas (`https://scikit-learn.org`).
- Harris et al., .ªrray programming with NumPy", Nature (2020).