Métricas, datos y calibración inteligente: Distancia euclídea con ventanas móviles y calibración lineal (PM2,5)

Alejandra Echeverry - 2221413

Septiembre 2025

Abstract

Este informe implementa la estrategia propuesta para comparar mediciones de PM2,5 entre una estación de referencia y estaciones IoT de bajo costo, usando la **distancia euclídea** sobre **promedios móviles** (ventanas temporales) y seleccionando la ventana que minimiza dicha distancia. Con la mejor ventana se construye un **modelo de calibración lineal** $f_{ref} = \alpha f_{IoT} + \beta$ y se evalúa su ajuste mediante R^2. En los datos analizados (referencia *Girón* y varios meses de la estación IoT `normalsup`) se observó de forma consistente que una **ventana de 24 horas** reduce la discrepancia (distancia) y permite ajustar modelos con R^2 entre 0,62 y 0,89 en la mayoría de meses.

1 Datos

Referencia (patrón): archivo Excel `Datos Estaciones AMB.xlsx`, hoja `Giron`, columnas `Date&Time` (marca temporal) y `PM2.5` (concentración en). Se ignoraron entradas `NoData`.
Estaciones IoT (bajo costo): archivos CSV con estructura:

- `fecha_hora_med` (ISO 8601, UTC),
- `id_parametro` (se filtró `pm25_a`),
- `valor` (PM2,5 en).

Meses analizados: 2018-11, 2018-12, 2019-05, 2019-06, 2019-07, 2019-08.[1]

[1] Archivos: `mediciones_clg_normalsup_pm25_a_YYYY-MM-...csv`.

1.1 Metodología

1.2 Alineación temporal y ventanas móviles

Se resamplearon ambas series a **promedios horarios** (1H). Sobre las series alineadas se aplicó un promedio móvil temporal de ancho $W \in \{1, 3, 6, 12, 24\}$ horas:

$$\tilde{f}_{ref}(t; W) = MA_W\big(f_{ref}(t)\big), \qquad \tilde{f}_{IoT}(t; W) = MA_W\big(f_{IoT}(t)\big).$$

Para cada ventana W se construyó la intersección temporal (pares válidos).

1.3 Distancia euclídea y selección de ventana

Definimos la distancia euclídea entre series suavizadas:

$$D(W) = \sqrt{\sum_{t \in \mathcal{T}_W} \Big(\tilde{f}_{ref}(t; W) - \tilde{f}_{IoT}(t; W)\Big)^2}.$$

La **ventana óptima** $W^\star$ se elige como la que minimiza $D(W)$ (exigiendo además un número suficiente de pares válidos).

1.4 Calibración lineal

Con la ventana $W^\star$ se ajusta un modelo afín:

$$f_{ref}(t) \;=\; \alpha\, f_{IoT}(t) \;+\; \beta,$$

estimando α (pendiente) y β (intercepto) por mínimos cuadrados ordinarios, y reportando R^2 como métrica de ajuste.

1.5 Resultados

1.6 Evolución de la distancia por ventana (ejemplo: 2018-11)

Para noviembre de 2018, las distancias decrecieron monótonamente al aumentar W:

Ventana W	1H	3H	6H	12H	24H
$D(W)$ (u. euclídeas)	94,56	83,76	77,10	68,77	**60,63**
Pares válidos	143	143	143	143	143

Así, $W^\star =$ **24H** para este mes.

1.7 Ventana óptima y distancia mínima por mes

Mes	Pares	$W^\star$	$D(W^\star)$	Observación
2018-11	143	**24H**	60,63	Coherente con patrón diario
2018-12	327	**24H**	47,50	Menor discrepancia global
2019-05	579	**24H**	73,81	
2019-06	705	**24H**	77,40	
2019-07	744	**24H**	66,90	
2019-08	739	**24H**	201,30	Distancia alta (posibles atípicos)

1.8 Modelos de calibración (con $W^\star = 24\mathbf{H}$)

Mes	α	β	R^2	Puntos
2018-11	0,1155	11,3274	0,009	143
2018-12	0,7095	4,7712	0,781	327
2019-05	0,7962	3,0022	0,886	579
2019-06	0,6304	4,1706	0,687	705
2019-07	0,5940	4,2752	0,622	744
2019-08	0,7921	−2,4103	0,697	739

Lectura. Diciembre y mayo muestran ajustes muy buenos ($R^2 > 0{,}78$); junio–agosto son moderados (0,62–0,70). Noviembre prácticamente no correlaciona ($R^2 \approx 0{,}01$). Agosto exhibe una distancia mínima alta, lo que sugiere revisar calidad de datos IoT (atípicos, fallos, desfases).

1.9 Discusión

Efecto del suavizado. Ventanas más largas (24H) reducen la varianza de corto plazo y realzan patrones diarios, mejorando la similitud con la referencia y, por tanto, disminuyendo $D(W)$.
Modelo afín vs. proporcional. Aunque la guía menciona $f = \alpha f_{IoT}$, un término β capta sesgos aditivos (offset del sensor). Aquí el modelo afín mejoró el ajuste sin sobreajustar (validado por R^2).
Limitaciones. Los resultados dependen de la calidad de sincronización temporal, manejo de `NoData` y cobertura de traslapes. Meses con R^2 bajo o distancias altas (p.ej., 2018-11, 2019-08) ameritan auditoría de datos (outliers, cambios de instrumento, mantenimiento).

1.9.1 Reproducibilidad (resumen de pasos)

1. Cargar referencia (`Giron`) y filtrar `PM2.5` válido; parsear `Date&Time` (UTC).
2. Cargar IoT; filtrar `id_parametro=pm25_a`; parsear `fecha_hora_med` (UTC); convertir `valor` a numérico.
3. Resamplear ambas series a **1H** con promedio; alinear por intersección temporal.

4. Para cada $W \in \{1, 3, 6, 12, 24\}$H: aplicar `rolling mean` temporal y computar $D(W)$; elegir $W^\star$ que minimiza D.

5. Con $W^\star$: ajustar $f_{ref} = \alpha f_{IoT} + \beta$ por MCO; reportar α, β, R^2.

6. (Opcional) Generar archivos calibrados: $\hat{y} = \alpha \cdot \texttt{valor} + \beta$.

2 Alcance de validez del modelo lineal

Con la ventana óptima seleccionada, consideramos el modelo afín ajustado por mínimos cuadrados

$$f_{ref}(t) \;=\; \alpha\, f_{IoT}(t) \;+\; \beta,$$

donde f_{ref} es la lectura de la estación patrón (Girón) y f_{IoT} la lectura de la estación de bajo costo, ambas suavizadas con media móvil temporal de 24 h y alineadas por hora.

Definición operativa del alcance

El *alcance de validez* es el intervalo de valores de la variable independiente $\hat{x}_i \equiv f_{IoT}$ para los cuales el error del modelo se mantiene dentro de una tolerancia predefinida. Definimos el residual absoluto en cada instante t:

$$r(t) \;=\; \big|\, f_{\mathrm{ref}}(t) - \big(\alpha\, f_{\mathrm{IoT}}(t) + \beta\big)\,\big|.$$

Usaremos dos criterios de tolerancia:

Relativa (10%): $\dfrac{r(t)}{f_{\mathrm{ref}}(t)} \;\le\; 0.10, \qquad con f_{\mathrm{ref}}(t) > 0,$ **Absoluta ($\pm 5\ \mu$g/m^3):** $r(t) \;\le\; 5\,\mu\mathrm{g/m}^3.$

Para cada criterio, el **alcance en** $\hat{x}_i$ se reporta como el intervalo $[x_{\min},\, x_{\max}]$ de valores IoT donde la condición se cumple, junto con la *cobertura* (porcentaje de puntos válidos).

Procedimiento paso a paso

1. **Alineación y suavizado:** se promedian las series en rejilla horaria (1 h) y se aplica media móvil temporal de 24 h a f_{ref} y f_{IoT}.

2. **Ajuste del modelo:** se estiman α y β por MCO sobre los pares $\big(f_{IoT}(t), f_{ref}(t)\big)$ suavizados (24 h).

3. **Cálculo de residuales:** $r(t) = \big|f_{ref}(t) - (\alpha f_{IoT}(t) + \beta)\big|$.

4. **Máscaras de tolerancia:**

   - Relativa: $r(t)/f_{ref}(t) \le 0.10$ con $f_{ref}(t) > 0$.
   - Absoluta: $r(t) \le 5$.

5. **Alcance y cobertura:** para cada criterio, se toma el subconjunto de tiempos que cumplen la tolerancia y se computan:

$$x_{\min} = \min f_{IoT}(t), \quad x_{\max} = \max f_{IoT}(t), \quad Cobertura = \frac{\#\{puntos válidos\}}{\#\{total\}} \times 100\,\%.$$

Resultados (Girón vs. IoT *normalsup*, ventana 24 h)

A continuación se presentan los intervalos de validez y coberturas por mes para ambos criterios.[2]

Table 1: Alcance de validez con tolerancia **relativa 10 %**.

Mes	Cobertura (%)	Rango IoT min	Rango IoT max
2018-11	47.55	5.34	17.04
2018-12	63.61	8.49	27.27
2019-05	29.71	2.93	33.52
2019-06	46.67	3.90	22.50
2019-07	55.11	8.07	23.06
2019-08	25.44	9.69	33.47

Table 2: Alcance de validez con tolerancia **absoluta** de $\pm 5\,\mu\mathrm{g/m^3}$.

Mes	Cobertura (%)	Rango IoT min	Rango IoT max
2018-11	88.11	3.59	17.04
2018-12	100.00	8.24	28.63
2019-05	94.82	2.02	33.55
2019-06	96.31	1.75	22.50
2019-07	100.00	7.77	23.36
2019-08	92.96	7.93	33.47

Interpretación

- El criterio **relativo (10 %)** es más exigente: las coberturas varían entre 25 % y 64 %, con alcances IoT típicamente entre $\sim$5 y 33 $\mathrm{g/m^3}$. *Es adecuado cuando interesa un porcentaj* $-100\,\%$) *con alcances más amplios*; *es útil cuando se desea un umbral fijo en unidades físicas*.
- Meses con menor cobertura relativa (p.ej., 2019-05 y 2019-08) sugieren mayor dispersión u outliers en el IoT; revisar mantenimiento, reloj e inhomogeneidades ambientales.

[2] Los meses corresponden a los archivos `mediciones_clg_normalsup_pm25_a_YYYY-MM-...csv`.

3 Alcance para realizar predicciones dentro de la tolerancia

Con la ventana óptima (**24 h**) obtenida en la etapa anterior, cuantificamos ahora el *alcance de predicción* del modelo lineal bajo tolerancia. Procedemos así:

Diseño experimental (train/test cronológico)

1. **Alineación y suavizado:** se promedian ambas series (*Girón* e IoT) en rejilla horaria y se aplica media móvil temporal de 24 h.
2. **Partición temporal:** se divide la serie suavizada en dos mitades cronológicas: **TRAIN** (primera mitad) y **TEST** (segunda mitad).
3. **Ajuste en TRAIN:** se estima el modelo afín
$$f_{\mathrm{ref}}(t) \;=\; \alpha\, f_{\mathrm{IoT}}(t) + \beta.$$
4. **Evaluación en TEST:** se generan predicciones $\hat{f}_{\mathrm{ref}}(t) = \alpha\, f_{\mathrm{IoT}}(t) + \beta$ y se calcula el residual absoluto
$$r(t) \;=\; \big|\, f_{\mathrm{ref}}(t) - \hat{f}_{\mathrm{ref}}(t) \,\big|.$$

Criterios de tolerancia y métricas

Usamos dos criterios:

$$\textbf{Relativa (10\%):}\quad \frac{r(t)}{f_{\mathrm{ref}}(t)} \le 0.10, \qquad f_{\mathrm{ref}}(t) > 0, \qquad \textbf{Absoluta } (\pm 5\ \mu\mathrm{g/m^3})\textbf{:}\quad r(t) \le 5\,\mu\mathrm{g/m^3}.$$

Para cada criterio, en el conjunto **TEST** se reporta:

- **Cobertura (%):** proporción de horas que cumplen la tolerancia.
- **Alcance en $\hat{x}_i$ (IoT):** intervalo $[x_{\min}, x_{\max}]$ de valores IoT donde las *predicciones* del modelo cumplen la tolerancia.

También informamos R^2_{train} del ajuste en TRAIN para contexto.

Resultados (TEST, ventana 24 h)

Lectura e implicaciones operativas

- Con **tolerancia relativa (10%)**, la cobertura es más estricta y varía del 21 % al 79 % según el mes; los rangos IoT válidos tienden a ser más estrechos.
- Con **tolerancia absoluta (±5)**, la cobertura es alta (80–100 %) y los rangos IoT son más amplios; es útil cuando se exige un error físico máximo.

Table 3: Alcance de predicción en TEST con tolerancia relativa del 10%.

Mes	R^2_{train}	Cobertura (%)	Rango IoT min	Rango IoT max
2018-11	0.028	44.44	11.15	15.79
2018-12	0.474	78.66	15.62	26.86
2019-05	0.905	21.03	2.97	8.96
2019-06	0.568	24.36	8.03	21.29
2019-07	0.513	66.40	8.07	23.06
2019-08	0.690	24.59	9.69	33.28

Table 4: Alcance de predicción en TEST con tolerancia absoluta de ± 5 μg/m^3.

Mes	R^2_{train}	Cobertura (%)	Rango IoT min	Rango IoT max
2018-11	0.028	100.00	11.15	17.04
2018-12	0.474	90.24	15.62	27.80
2019-05	0.905	93.79	2.02	11.92
2019-06	0.568	79.60	8.03	22.50
2019-07	0.513	100.00	8.07	23.36
2019-08	0.690	93.78	9.69	33.47

- No existe un *único* intervalo global de IoT que garantice la tolerancia en todos los meses: se recomienda emplear los **rangos por mes** de las Tablas 3 y 4, o bien recalibrar por estación/temporada.

Recomendación práctica

Seleccione el criterio de tolerancia acorde al uso (relativo si prima el porcentaje de error, absoluto si prima un umbral físico). Para *predicción en línea*, aplique el modelo $\hat{f}_{\text{ref}} = \alpha f_{\text{IoT}} + \beta$ **solo cuando** $f_{\text{IoT}} \in [x_{\min}, x_{\max}]$ del mes correspondiente; fuera de ese rango, marque la predicción como "fuera de validez" y/o solicite recalibración.

4 Conclusiones

(1) La **ventana de 24H** fue la más efectiva para reducir la distancia entre IoT y referencia en todos los meses analizados. (2) La **calibración lineal afín** produjo ajustes fuertes en varios meses ($R^2 > 0{,}78$), habilitando el uso operativo de los sensores IoT una vez calibrados. (3) Meses con discrepancias altas sugieren revisar la cadena de adquisición IoT (reloj, firmware, entorno, mantenimiento).

Conclusión práctica

Para operación, se recomienda fijar *un* criterio de tolerancia acorde al uso:

- Si la aplicación exige precisión porcentual, use la **tolerancia relativa (10 %)** y restrinja el uso del modelo a los rangos de la Tabla 1.
- Si basta con un error físico máximo, use la **tolerancia absoluta** (± 5) y adopte los rangos de la Tabla 2.

En ambos casos, el intervalo $[x_{\min}, x_{\max}]$ determina el **alcance en** $\hat{x}_i$ donde el modelo lineal es válido para la tolerancia seleccionada.