# Calibración de sensores IoT usando el método de mínimos cuadrados

**Andrés Felipe Rubio**[*]
**Andres Felipe Vargas**[**]
**Carlos Andrés Laguado**[***]
*Univrsidad Industrial de Santander*
*Carrera 27 calle 9*

Versión 1 12/09/2021

## Índice



**Resumen**

En la actualidad, las ciudades enfrentan limitaciones con sus sistemas de monitoreo de calidad del aire, debido principalmente a su costo. Una estrategia para solucionar parcialmente este problema es ampliar la red de estaciones de monitoreo utilizando sensores "Low Cost"(basados en IoT) e implementar un correcto algoritmo de calibración. Para el presente trabajo, se tomaron

[*]e-mail: andres2218426@correo.uis.edu.co
[**]e-mail: andres2218420@correo.uis.edu.co
[***]e-mail: carlos2047095@correo.uis.edu.co

como referencia los datos de concentración de PM2.5 de la estación de la escuela Normal del municipio de Bucaramanga, Colombia, el cual hace parte del sistema de Monitoreo del AMB. Se procesaron los datos suministrados por un sensor IoT ubicado en la misma estación por el método de promedio móvil y se realizó un análisis de sensibilidad modificando tanto el tamaño de la ventana (WS) como el tamaño del paso (SS) para encontrar el menor error. Para validar la fiabilidad del modelo, se tomó como base un total del 70 % de los datos como datos de entrenamiento, y el 30 % restante se dejó como rango de prueba para validar el modelo planreado. Se observó que el error de la curva ajustada por mímimos cuadrados con relación a los datos de referencia varía entre el 29.04 % y el 50.56 % dependiendo de los parámetros WS y SS. Una vez utilizado el modelo ajustado para contrastar los resultados con el conjunto de datos de prueba, se observan porcentajes de error entre el 60.07 % y el 67.04 %. Se concluye que, para el método de mínimos cuadrados, la escogencia del tamaño de la ventana (WS) genera mejores resultados que la disminución del paso de muestreo (SS). También se observa que el paso de muestreo (SS) disminuye el error en una proporción muy pequeña en comparación con su costo computacional. Se recomienda revisar otros métodos para realizar el ajuste (por ejemplo el planteado en [2], ya que el error es alto.

## 1. Introducción

Históricamente, las áreas urbanas se han visto limitadas a la hora de implementar estaciones de monitoreo de la calidad del aire, principalmente por el alto costo de la instrumentación. Debido a esto, dichas áreas urbanas dependen de pocos monitores para evaluar la exposición a determinados contaminantes que pueden afectar a la población. En los últimos años ha habido un aumento en el desarrollo y las tecnologías de monitoreo de la calidad del aire basadas en sensores de bajo costo [1], [2].

El uso de sensores de bajo costo permitiría tener redes más densas para el monitoreo de la calidad del aire con un costo cercano al de un equipo moderno. Con redes de monitoreo más densas se podría cuantificar y caracterizar de mejor manera la información en las áreas urbanas para respaldar los modelos epidemiológicos. Además, al contar con mayor densidad de redes de monitoreo se pueden identificar áreas potencialmente más peligrosas y tomar las medidas correspondientes. Generalmente, estos sensores de bajo costo tienen un tamaño compacto y consumen poca energía. Estas características permiten el transporte de estos sensores de manera sencilla a regiones donde existe un monitoreo limitado. Uno de los requisitos fundamentales en los sensores de bajo costo es la calibración.

Debido a esto se hace necesario encontrar una forma de calibración para los sensores de bajo costo, de manera que mejore las mediciones de dichos sensores y se aproxime más a las mediciones de los equipos modernos. Como una posible solución a esta problemática se presenta en este estudio una estrategia de calibración inteligente, usando el método de mínimos cuadrados. Para el desarrollo del modelo de calibración contamos con dos dataset, uno proveniente de una estación con sensores de bajo costo ubicado en la escuela Normal de Bucaramanga y el otro proveniente de una estación con un equipo moderno del AMB. Para garantizar la validez del modelo de calibración se dividió el dataset en dos partes, una de entrenamiento (70 %) y otra de prueba (30 %). En el estudio se

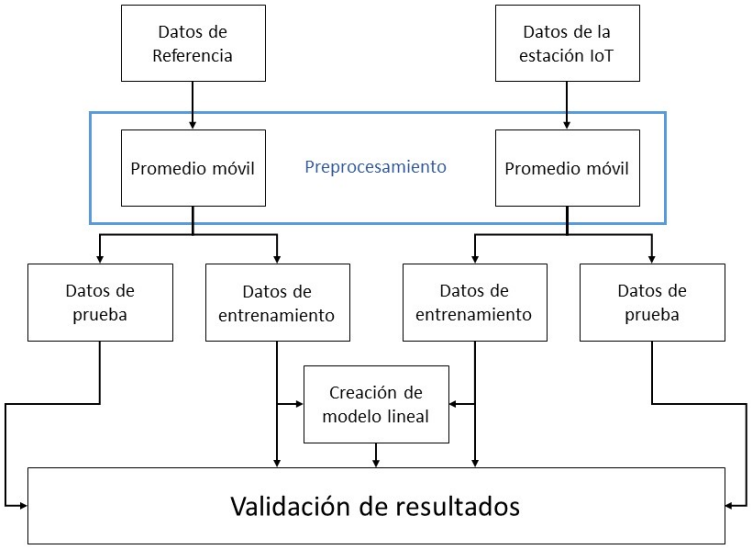


Figura 1: Diagrama de metodología

usaron datos tomados por las estaciones en un período de aproximadamente 4 meses (27 de abril de 2019 a 31 de agosto de 2019). Se muestra en detalle el ajuste de los algoritmos para determinar los mejores datos de calibración, además se presenta el rendimiento de dicho modelo.

## 2. Metodología

El desarrollo de este estudio se dividió en etapas. En la primera etapa se seleccionan los datos de la estación $IoT$ y de la estación AMB. En la segunda etapa se realiza un preprocesamiento a los datos, aplicando el promedio móvil con ventanas de tiempo, esto debido al desbalanceo de los datos. En la tercera etapa se realiza un modelo lineal usando la técnica de aproximación por mínimos cuadrados y por último se evalúa el rendimiento del modelo lineal dividiendo el conjunto de datos en conjunto de entrenamiento (*training set*) y conjunto de prueba (*test set*). En la figura 1 se puede observar de manera resumida la metodología a idesarrollar.

### 2.1. Selección de datos

El sistema de monitoreo de calidad de aire del Área Metropolitana de Bucaramanga (AMB) comprende cinco (5) estaciones (*Acualago*, *Caldas*, *Girón*, *Normal*, *Pilar*) distribuidas en puntos estratégicos del área, las cuales miden los siguientes parámetros: $PM_{10}$[3] ($\mu g/m^3$), $PM_{2,5}$[4] ($\mu g/m^3$), dióxido de Nitrógeno ($NO_2$) ($\mu g/m^3$), Ozono ($O_3$) ($\mu g/m^3$), Temperatura del Aire ($C$), Precipitaciones ($mm$), Humedad Relativa ( %), Dirección del Viento ($grados$), velocidad del viento ($m/s$) y Radiación

solar ($W/m^2$). Los datos suministrados para realizar el análisis[5] están tabulados cada hora, desde el 01 de octubre de 2018 a las 00:00h hasta el 31 de agosto de 2019 a las 23:00h.

En la estación *Normal* se instaló un sensor de prueba de bajo costo ($IoT$) el cual toma mediciones del parámetro $PM_{2,5}$ ($\mu g/m^3$). Las mediciones de concentración [6],[7],[8],[9],[10],[11],[12] están tabuladas cada hora, desde el 11 de abril de 2019 a las 17;00h hasta el 31 de agosto de 2019 a las 23:00h.

Al analizar los datos del sensor ($IoT$) se observa que no hay continuidad en la toma de las muestas. Para el periodo entre el 11 de abril y 27 de abril de 2019 hay rangos de hasta 27 horas consecutivas en los cuales no hay información. Para efectos del análisis de datos a realizar, se toma como muestra base los datos desde el 27 de abril de 2019 a las 01:00 hasta el 31 de agosto de 2021 a las 23:00.

## 2.2. Preprocesamiento de los datos

Las mediciones realizadas por los sensores de bajo costo de la estación $IoT$ ubicada en la Escuela Normal Superior (estación *Normal*), no están sincronizadas con las medidas realizadas por los sensores de la estación AMB (patrón de referencia). Por lo tanto, se usa la técnica de promedio móvil, para que los conjuntos de datos que representan las medidas de $PM_{2,5}$ de la estación $IoT$ y de la estación de la AMB sean del mismo tamaño.

El promedio móvil analiza un conjunto de datos en modo de puntos para crear series de promedios. De tal forma que se obtiene un conjunto de datos resultantes donde, cada dato corresponde al promedio de un subconjunto de los datos originales [13]. Los parámetros que se usaron para hacer el promedio móvil fueron: tamaño de la ventana (WS) y paso de la ventana (SS). Los parámetros WS y SS se establecen en horas. La figura 2 muestra el proceso de cálculo de promedio móvil de manera gráfica, teniendo en cuenta los parámetros WS y SS.

A manera de ejemplo la figura 3 y la figura 4 muestran las mediciones de $PM_{2,5}$ y el promedio móvil en función del tiempo de los datos de referencia y de la estación $IoT$, respectivamente, para el periodo desde el 06-junio-2019 hasta el 31-agosto-2019, WS = 24, SS = 3.

## 2.3. Método de calibración

El método de calibración utilizado fue el ajuste por mínimos cuadrados. Este ajuste consiste en intentar encontrar la función continua que mejor aproxime una cierta cantidad de datos, tomando en cuenta el criterio de mínimo error cuadrático [14]. El ajuste por mínimos cuadrados o regresión lineal parte de que los datos pueden representarse mediante una recta de la forma:

$$y = \alpha + \beta x \tag{1}$$

Donde $\alpha$ y $\beta$ representan los coeficientes del modelo de regresión. Esta sería la recta teórica del modelo. Como la distribución de valores no se ajusta a una recta perfecta de esa manera cuando se

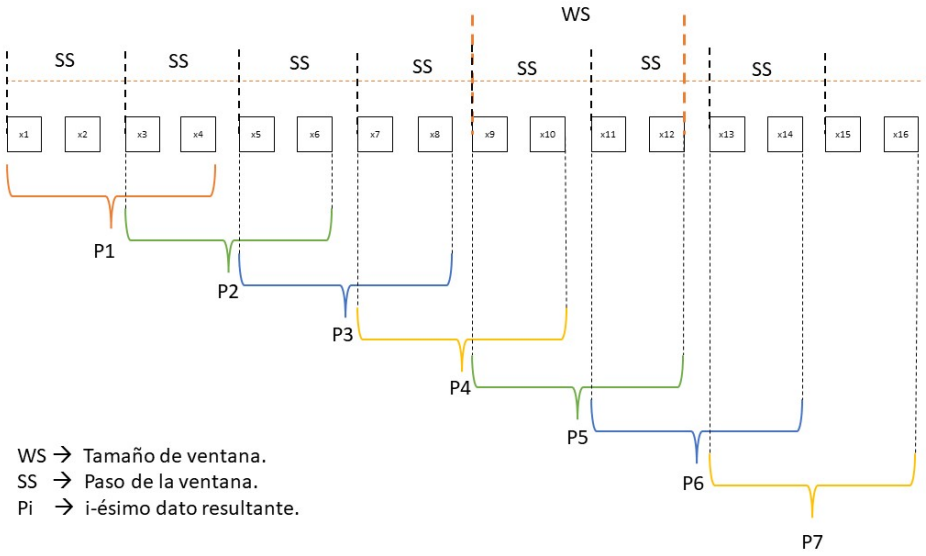


Figura 2: Ilustración gráfica del promedio móvil

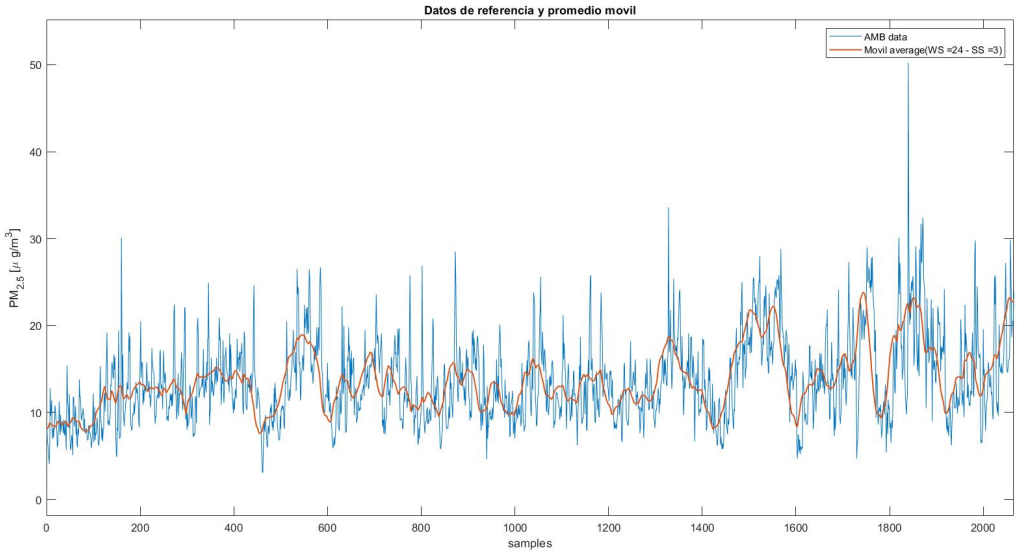


Figura 3: Ejemplo de promedio movil a los datos de referencia (Estación AMB).

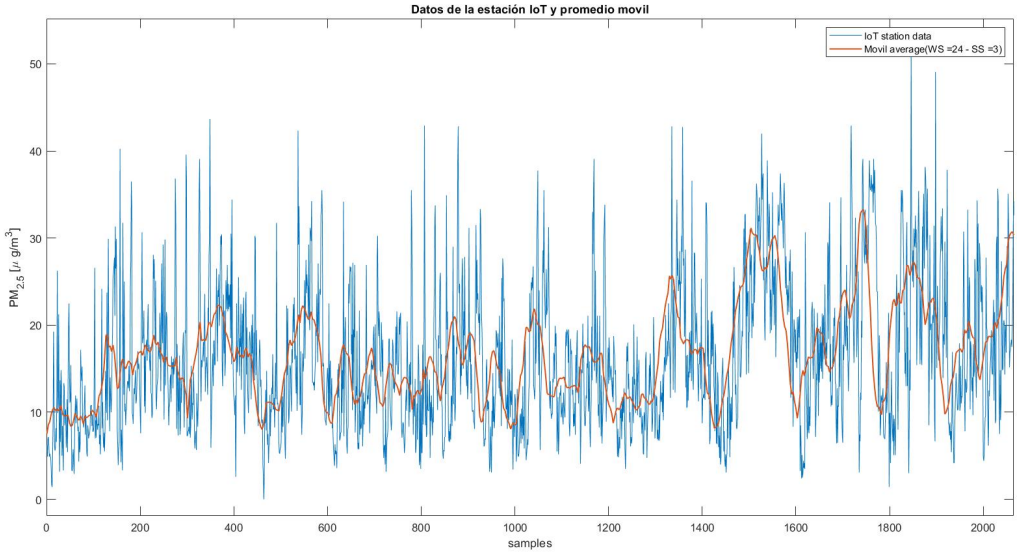


Figura 4: Ejemplo de promedio movil a los datos a calibrar (Estación IoT).

calcule un y(i) a partir de un x(i) habrá una diferencia entre el valor real y el valor obtenido con la fórmula de la recta. De esta forma aparece un nuevo componente en la ecuación relacionado con el error:

$$y = \alpha + \beta x + e \tag{2}$$

Este error se puede representar de la forma:

$$e = y_i - y_i* \tag{3}$$

Ahora, el objetivo es encontrar la recta que tenga el menor error cuadrático medio, lo cual se consigue obteniendo la recta con menor $\epsilon$.

$$\sum_{n=1}^{n} e_i^2 = \sum_{n=1}^{n} (y_i - y_i*)^2 \tag{4}$$

Si hacemos $y_i* = a + bx$, reemplazando en 4 nos queda:

$$\sum_{n=1}^{n} e_i^2 = \sum_{n=1}^{n} (y_i - a - bx_i)^2 \tag{5}$$

Ahora, hay que encontrar los valores de a y b que minimicen la función. Esto se hace de la forma:

$$b = S_{xy}/S_x^2 \tag{6}$$

Donde tenemos en el numerador la covarianza de las dos variables y en el denominador, la varianza de la variable independiente. Para hallar a se hace:

$$a = \bar{y} - \bar{b} \tag{7}$$

De esta forma podemos construir nuestra recta que pasa por los valores medios de x e y.

### 2.4. Validación de resultados

Para la validación de resultados se dividió el conjunto de datos en dos partes, el 70 % de los datos se usa para la implementación del modelo de aproximación lineal (*Training set*). El 30 % restante de los datos se usa para comparar la predicción del modelo (*test set*). Los resultados de esta validación, ingresando los datos de entrenamiento (*training set*) y los datos de prueba (*test set*) se discuten en la secciones 3.1 y 3.2, respectivamente.

Como métrica de error usamos la distancia entre dos conjuntos de datos. Para ello usamos la distancia euclídea definida como:

$$D(X_i, Y_{\hat{i}}) = \sqrt{\sum_{i,\hat{i}} (X_i - Y_{\hat{i}})^2} \tag{8}$$

Donde $X_i$ y $Y_{\hat{i}}$ corresponden a los i-ésimos valores promedio de cada conjunto de datos en el mismo intervalo de tiempo.

Para los resultados usamos un error porcentual. Este error porcentual se obtiene de la normalización de la distancia entre los conjuntos de datos de referencia y los datos aplicando el modelo lineal, con respecto a la distancia de entre los datos de referencia y los datos de la estación $IoT$.

$$E_{\%} = \frac{D}{D_0} 100 \tag{9}$$

Donde D es la distancia entre los conjuntos de datos de referencia y los datos aplicando el modelo lineal, $D_0$ es la distancia de entre los datos de referencia y los datos de la estación $IoT$.

## 3. Resultados y discusión

En esta sección presentamos los resultados en términos del error de la predicción del modelo lineal presentado en la sección 2.3, con respecto al conjunto de datos de referencia. Primero evaluamos la predicción del modelo con los datos con los que se creó el modelo lineal (*training set*) 3.1, luego se evalua la predicción del modelo con un conjunto independiente (*test set*) 3.2.

### 3.1. Comparación de las predicicones del modelo

En la primera prueba de validación, realizamos una comparación del modelo usando el conjunto de datos con los que se entrenó el modelo. Los resultados de error en distancia de los datos, se

Cuadro 1: Mediciones de error usando los datos de entrenamiento para evaluar la predicción del modelo

| **Tamaño de ventana [horas]** | **Paso de muestreo [horas]** | **Error [%]** |
|---|---|---|
| 48 | 24 | **29.0418** |
| 48 | 12 | 29.4366 |
| 48 | 6 | 29.2016 |
| 48 | 3 | 29.2554 |
| 48 | 1 | 29.2373 |
| 24 | 12 | **32.0225** |
| 24 | 6 | 31.9275 |
| 24 | 3 | 32.1177 |
| 24 | 1 | 32.0810 |
| 12 | 6 | 36.0157 |
| 12 | 3 | 37.2395 |
| 12 | 1 | 37.1716 |
| 6 | 3 | 46.4476 |
| 6 | 1 | 46.2920 |
| 4 | 2 | 50.4969 |
| 4 | 1 | 50.5646 |

pueden ver en el cuadro 1. Se observa que el menor error se obtiene en los casos donde se usa ventana de 48 horas, aunque no difiere mucho con respecto a los resultados usando una ventana de 24 horas. También se puede observar que el paso de muestreo no afecta de manera considerable el error, mientras que el tamaño de ventana si afecta en gran manear, si se modifica en gran magnitud el tamaño de la ventana.

En la figura 5 se muestran los datos (originales, referencia, y aproximados usando el modelo lineal), en función del tiempo, para los parámetros para los que se obtuvo el menor error. En la figura 6 se muestran los datos (originales, referencia, y aproximados usando el modelo lineal), en función del tiempo, para una ventana de 24 horas y un paso de muestreo de 6 horas, que obtuvo el menor error para un tamaño de ventana de 24 horas. Ya que no se observa una diferencia considerable con respecto al menor error en general (tamaño de ventana de 48 horas y paso de muestreo de 24 horas).

### 3.2. Evaluación del modelo usando datos de prueba

De igual manera que en la sección 3.1 en el cuadro 2 se presentan los resultados de error para diferentes tamaños de ventana y pasos de muestreo. Se puede observar que los errores aumentan considerablemente con respecto a los resultados de la sección 3.1. En las figuras 7 y 10 se muestran

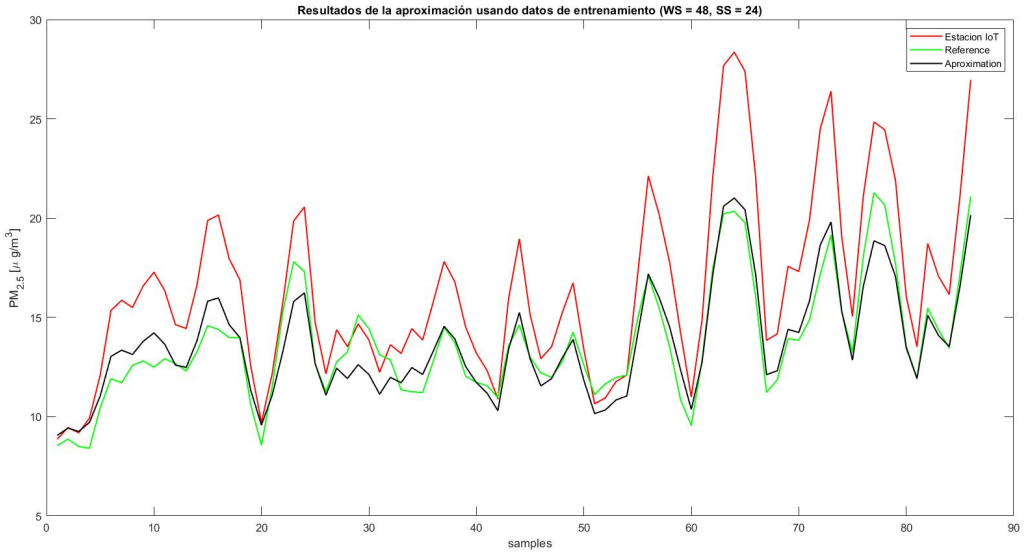


Figura 5: Gráfica de resultados de aproximación del modelo lineal usando datos de entrenamiento (WS = 48, SS = 24).

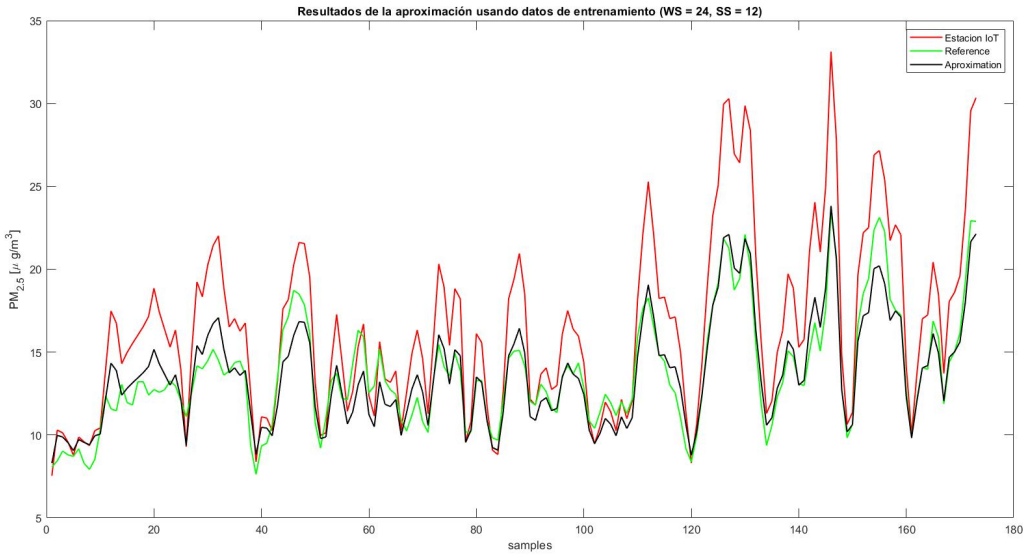


Figura 6: Gráfica de resultados de aproximación del modelo lineal usando datos de entrenamiento (WS = 24, SS = 12).

Cuadro 2: Medidas de error usando los datos de prueba para evaluar la predicción del modelo

| **Tamaño de ventana [horas]** | **Paso de muestreo [horas]** | **Error [%]** | **Tiempo [seg]** |
|---|---|---|---|
| 48 | 24 | 60.1590 | 1.5677 |
| 48 | 12 | 60.3773 | 1.6125 |
| 48 | 6 | **60.0709** | **1.8573** |
| 48 | 3 | 60.1883 | 2.0531 |
| 48 | 1 | 60.2259 | 2.9042 |
| 24 | 12 | 61.9054 | 1.6106 |
| 24 | 6 | 61.5356 | 1.7504 |
| 24 | 3 | 61.6111 | 1.9873 |
| 24 | 1 | 61.6014 | 3.0683 |
| 12 | 6 | 64.8771 | 1.7421 |
| 12 | 3 | 64.4072 | 2.1145 |
| 12 | 1 | **64.3252** | **3.0718** |
| 6 | 3 | 66.8303 | 1.9726 |
| 6 | 1 | 66.9010 | 3.0715 |
| 4 | 2 | 67.0420 | 2.3929 |
| 4 | 1 | 67.0054 | 2.9566 |

los datos (originales, referencia, y aproximados usando el modelo lineal), en función del tiempo, para los parámetros con los cuales se obtiene el menor error, usando una ventana de 48 horas y de 24 horas, respectivamente.

## 4. Conclusiones

Los resultados de este estudio muestran que la calibración por mínimos cuadrados reduce el error de la medición entre la estación de sensores de bajo costo y la estación con un equipo moderno, sin embargo, sigue teniendo un error considerable, por lo que se recomienda usar otros algoritmos de ajuste que minimicen aún más el error.

Se concluye que, para el método de mínimos cuadrados, la escogencia del tamaño de la ventana (WS) genera mejores resultados que la disminución del paso de muestreo (SS). También se observa que el paso de muestreo (SS) disminuye el error en una proporción muy pequeña en comparación con su costo computacional.

El desarrollo de estudios de este tipo serían muy beneficiosos, ya que permitirían aumentar la densidad de las redes de monitoreo con sensores de bajo costo obteniendo resultados de calidad.

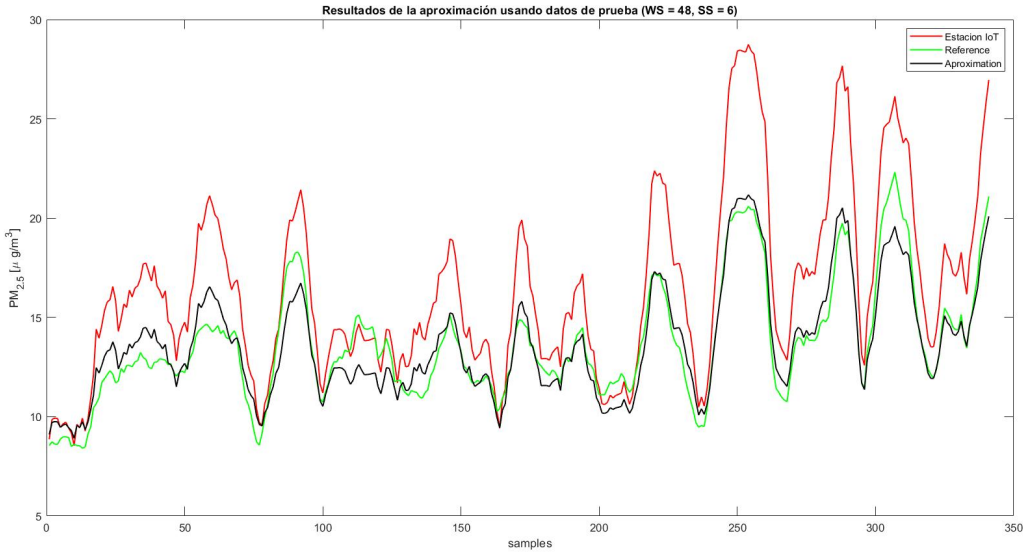


Figura 7: Gráfica de resultados de aproximación del modelo lineal usando datos de prueba (WS = 48, SS = 6).

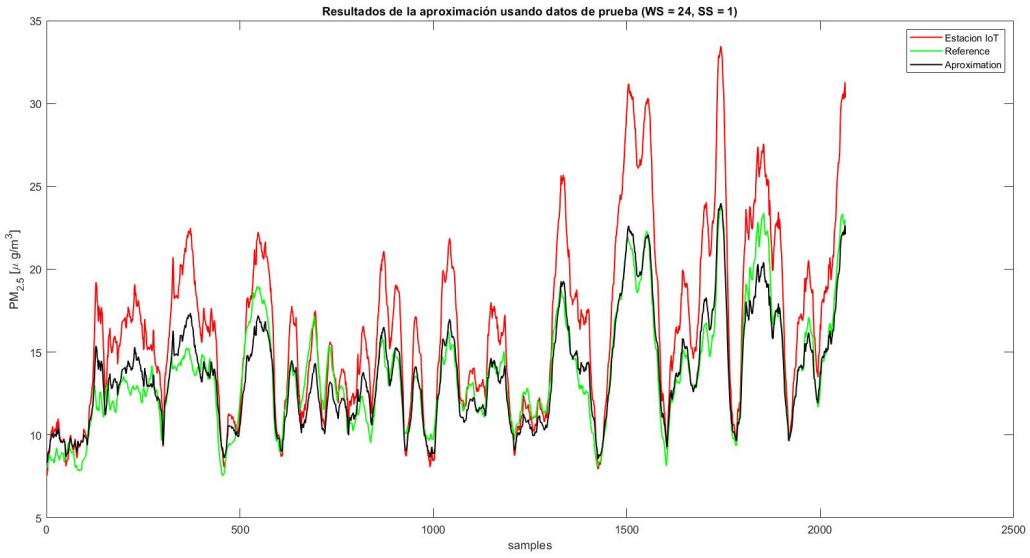


Figura 8: Gráfica de resultados de aproximación del modelo lineal usando datos de prueba (WS = 24, SS = 1).

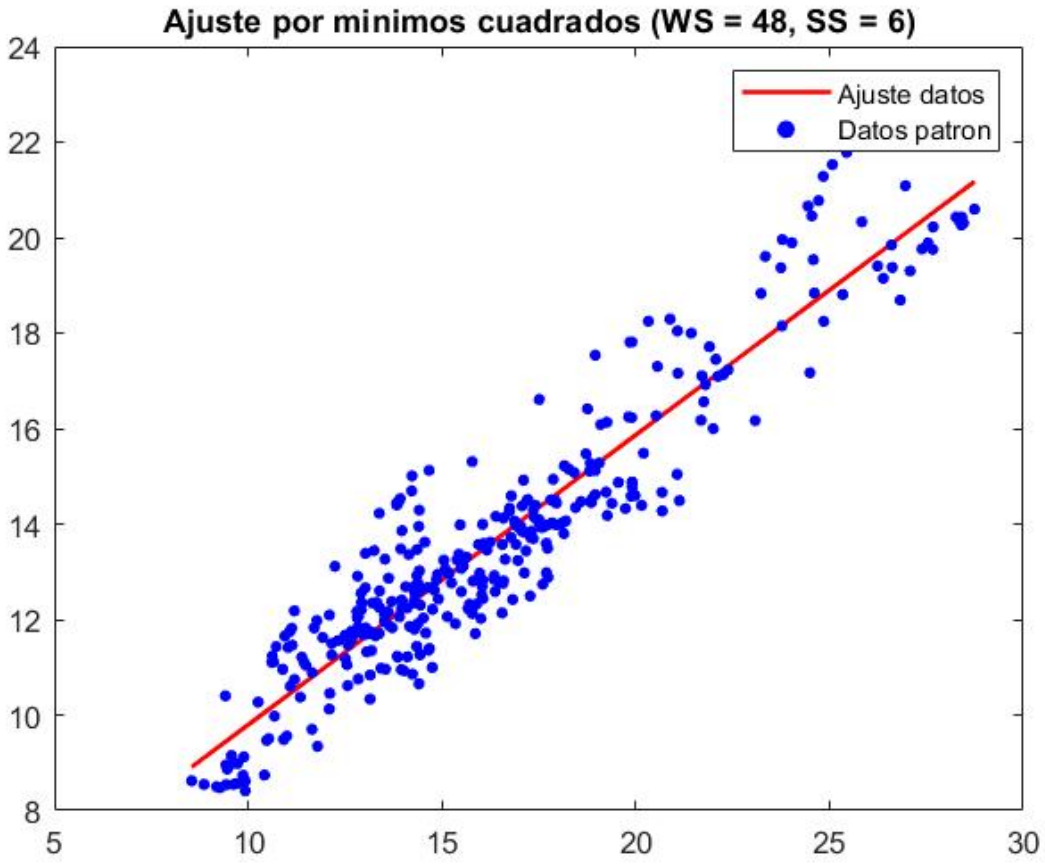


Figura 9: Recta de regresión para WS = 48, SS = 6.

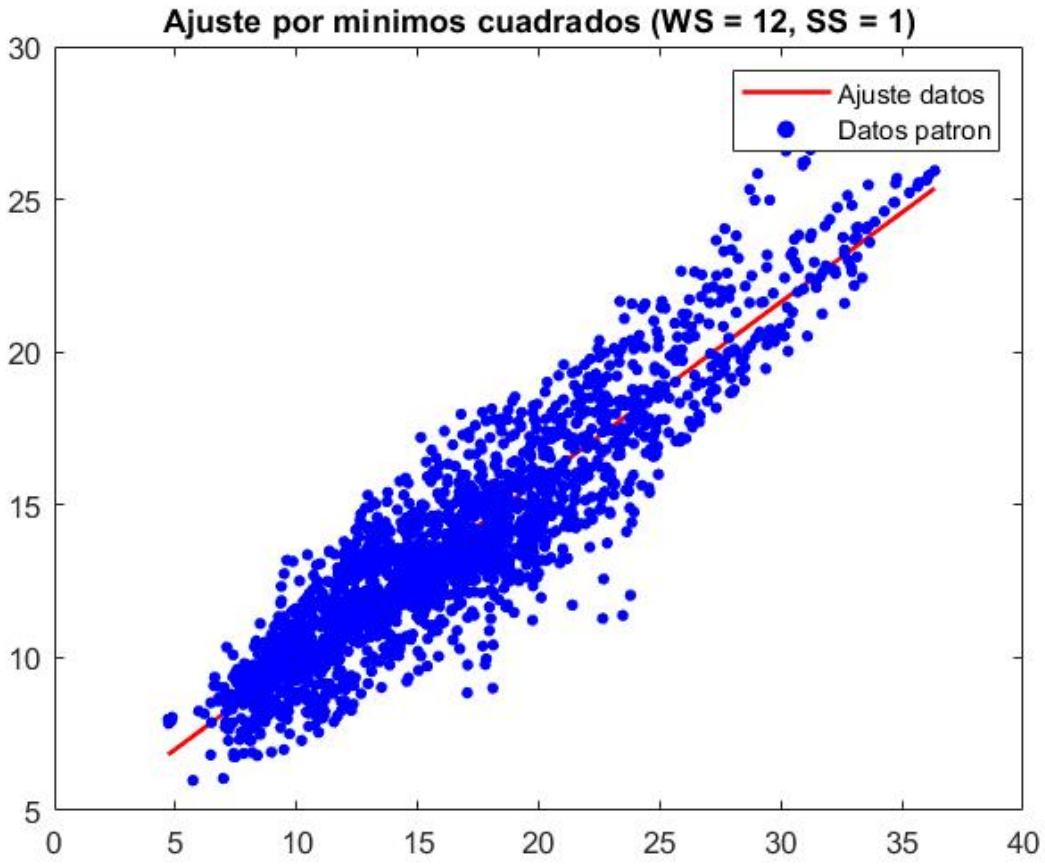


Figura 10: Recta de regresión para WS = 12, SS = 1.

# 5. Referencias

## Referencias

[1] https://www.nature.com/articles/535029a.pdf

[2] https://amt.copernicus.org/articles/11/291/2018/amt-11-291-2018.pdf

[3] https://es.wikipedia.org/wiki/PM10

[4] https://es.wikipedia.org/wiki/PM2.5

[5] https://github.com/nunezluis/MisCursos/blob/main/MisMateriales/Asignaciones/TallerDistancias/DatosDistancias/Datos%20Estaciones%20AMB.xlsx

[6] https://github.com/nunezluis/MisCursos/blob/main/MisMateriales/Asignaciones/TallerDistancias/DatosDistancias/mediciones_clg_normalsup_pm25_a_2018-11-01T00_00_00_2018-11-30T23_59_59.csv

[7] https://github.com/nunezluis/MisCursos/blob/main/MisMateriales/Asignaciones/TallerDistancias/DatosDistancias/mediciones_clg_normalsup_pm25_a_2018-12-01T00_00_00_2018-12-31T23_59_59.csv

[8] https://github.com/nunezluis/MisCursos/blob/main/MisMateriales/Asignaciones/TallerDistancias/DatosDistancias/mediciones_clg_normalsup_pm25_a_2019-04-01T00_00_00_2019-04-30T23_59_59.csv

[9] https://github.com/nunezluis/MisCursos/blob/main/MisMateriales/Asignaciones/TallerDistancias/DatosDistancias/mediciones_clg_normalsup_pm25_a_2019-05-01T00_00_00_2019-05-31T23_59_59.csv

[10] https://github.com/nunezluis/MisCursos/blob/main/MisMateriales/Asignaciones/TallerDistancias/DatosDistancias/mediciones_clg_normalsup_pm25_a_2019-06-01T00_00_00_2019-06-30T23_59_59.csv

[11] https://github.com/nunezluis/MisCursos/blob/main/MisMateriales/Asignaciones/TallerDistancias/DatosDistancias/mediciones_clg_normalsup_pm25_a_2019-07-01T00_00_00_2019-07-31T23_59_59.csv

[12] https://github.com/nunezluis/MisCursos/blob/main/MisMateriales/Asignaciones/TallerDistancias/DatosDistancias/mediciones_clg_normalsup_pm25_a_2019-08-01T00_00_00_2019-08-31T23_59_59.csv

[13] https://es.wikipedia.org/wiki/Media_m%C3%B3vil

[14] https://es.wikipedia.org/wiki/M%C3%ADnimos_cuadrados