Métrica, datos y calibración inteligente

Gabriela Sánchez Ariza
Nicolas Toledo Parra
*

Universidad Industrial de Santander
Cl. 9 Cra 27, Bucaramanga, Santander

10 de Diciembre de 2021

Índice

Resumen

En este trabajo se busca cuantificar el error de medición del sensor de bajo costo y calibrarlo para que se puedan establecer nuevas lecturas mucho más precisas, para lo que fue necesario realizar un tratamiento de datos en el que se eliminaron filas vacías y en el que se permitiera tener la misma cantidad de datos entre los datos de la estación AMB y los medidos por el sensor de bajo costo. También, se calculó la distancia euclídea entre los dos conjuntos de datos para los tiempos de horas, días y meses, para así por medio de una regresión lineal determinar la mejor calibración, siendo esta la correspondiente a la escala de horas, con un error del 2.62 % y una exactitud del 97.38 %. Adicionalmente, se realizó un análisis por medio del promedio móvil para ventanas de tamaños variables y saltos de 1 hora, obteniendo como resultado la regresión lineal más precisa con una ventana de ancho de 30, con un error del 0.29 % y una exactitud del 99.71 %. Tanto el resultado obtenido de la calibración en horas como el obtenido del promedio móvil son bastante tolerables, lo cual indica una buena predicción.

*e-mail: gabriela2200816@correo.uis.edu.co; nicolas2200017@correo.uis.edu.co

1. Introducción

Los sensores se han convertido en instrumentos capaces de convertir información del mundo real al mundo digital, permitiendo aumentar la eficiencia, calidad y velocidad de los procesos industriales, la investigación y el desarrollo científico. En algunas ciudades del mundo, con el fin de monitorear la calidad del aire, se implementan redes de monitoreo que permiten tener un panorama general del estado actual del recurso, por ejemplo los sensores de bajo costo, los cuales forman parte de los dispositivos de la llamada revolución de la *internet de las cosas*, *IoT*. No obstante, estos sensores son incapaces de caracterizar adecuadamente la exposición de los ciudadanos, debido a su baja resolución espaciotemporal y al alto número de datos faltantes. Teniendo en cuenta lo anterior, el problema de este trabajo está en cuantificar cuál es el error de medición del sensor de bajo costo y como calibrarlo para poder establecer nuevas lecturas más precisas.

A lo largo de este trabajo se podrá observar la metodología 2, en la que se plantean diferentes formas de dar una solución, la reorganización de los datos y el promedio móvil, y se explica como obtener la distancia euclídea entre los datos medidos y de referencia, y cómo obtener un ajuste determinado para calibrar el sensor. También, se presentan los resultados 3, en donde se muestra todo el proceso realizado y los resultados obtenidos. Además, se presentan las conclusiones 4, en donde se hace un balance de la comprensión lograda entre los distintos resultados obtenidos. Para finalizar, se hace uso de diferentes bibliografías. Stanley Gudder dijo alguna vez: "*La esencia de las matemáticas no es hacer las cosas simples complicadas, sino hacer las cosas complicadas simples*"; este trabajo es importante porque permite ver que un problema general, como la calibración de los datos obtenidos por un sensor, puede ser abordado de diferentes formas con el fin realizar lo más óptimo.

2. Metodología

Este trabajo tiene como objetivo principal cuantificar el error de medición del sensor de bajo costo y calibrarlo para poder establecer nuevas lecturas más precisas, para esto se hizo uso del archivo *Datos Estaciones AMB*, que contiene las medidas de referencia de concentración de material particulado PM 2.5, y los archivos etiquetados por *mediciones*, los cuales contienen los registros de las estaciones de bajo costo [1]. Teniendo en cuenta lo anterior, a lo largo de este informe se presentan dos diferentes formas de dar solución a lo planteado, la primera forma consiste en realizar la calibración de los datos a partir de una reorganización de los mismos, y la segunda está basada en el criterio de promedio móvil, abarcadas de forma simulada por medio de Python [**?**] haciendo uso de la librería *Pandas*, esta es una herramienta muy poderosa a la hora de trabajar con manipulación de datos de alto nivel, es por eso que se decidió ir por este camino.

Antes de llevar a cabo cada una de las formas mencionadas anteriormente se realizó un tratamiento al conjunto de datos de referencia y al de los datos a calibrar, esto con el fin de tener un mejor

[1] Aquí podrá encontrar los datos de referencia y los de las estaciones IoT.

desarrollo en la realización del código, se cambiaron los nombres de algunas columnas de los archivos originales y se limpiaron los conjuntos de datos eliminando filas vacías; los archivos finales que fueron usados en este trabajo se pueden encontrar en el siguiente link [2].

En ambas formas se realizó una estimación de la distancia entre las medidas de las estaciones de referencia *AMB* y de las de bajo costo, y esto se hizo por medio de la distancia euclídea (1) entre los dos conjuntos de datos; en donde se ha definido a $\mathbb{D}_i$ como el conjunto de datos a calibrar y como $\hat{\mathbb{D}}_{\hat{i}}$ al conjunto de datos de referencia.

$$\mathcal{D}(\mathbb{D}_i, \hat{\mathbb{D}}_{\hat{i}}) = \sqrt{\sum_{i,\hat{i}} \left(\mathbb{D}_i - \hat{\mathbb{D}}_{\hat{i}}\right)^2} \tag{1}$$

También, se realizó un ajuste de mínimos cuadrados para determinar un modelo de ajuste lineal, de tal forma que se tenga conocimiento de las operaciones y cambios que se deben realizar para calibrar el sensor de bajo costo. Adicionalmente, se determinó el alcance de validez del modelo lineal definiendo una tolerancia de forma que estableciera un criterio tal que la calibración obtenida fue correcta.

3. El experimento y los resultados

3.1. Distribución de datos medidos y de referencia

Para ver cómo estaban distribuidos los conjuntos de datos en cuanto a su frecuencia de aparición se emplearon los siguientes histogramas. Como es notable, la mayoría de los datos de ambos conjuntos se encuentran aproximadamente entre 8 y 13 medidas de concentración de material particulado $PM_{2,5}$.

3.2. Tratamiento de datos

Para comenzar con el desarrollo del problema, se realizó el tratamiento de los datos antes de ser importados a *Python*, organizando de una mejor manera los datos y eliminando las filas que estaban vacías, posterior a esto, se importaron los datos de referencia de PM 2.5 de las estaciones AMB y los datos medidos por el sensor de bajo costo, para así poder continuar con el tratamiento de datos esta vez en Python. Los archivos de mediciones se unieron en uno solo y se eliminaron datos sobrantes, quedando al final únicamente 2 dataframes, uno con los datos de PM 2.5 de referencia y otro con los datos medidos, sin embargo, el número de elementos de estos dataframes eran diferentes, por lo que realizar un análisis temporal emparejando cada dato entre ambos dataframes no era posible. En parte, esto se solucionó haciendo uso de la función resample de *Pandas* para horas, días y meses; un resample básicamente consiste en un remuestreo de datos en intervalos iguales de tiempo, entonces se utilizó la función merge para fusionar ambos dataframes y de esta manera se obtuvo un único dataframe con el mismo número de datos de PM 2.5 medidos por el sensor de bajo costo y de las

[2] Aquí podrá encontrar los archivos tratados.

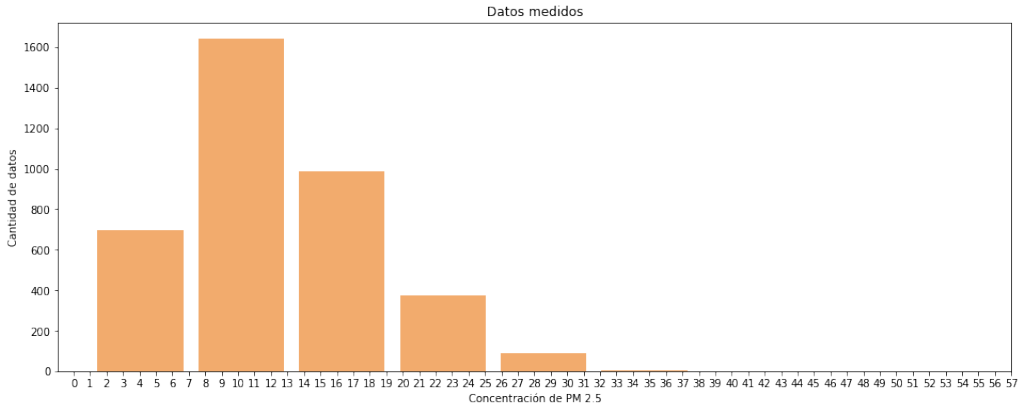

Figura 1: Histograma que muestra cómo están distribuidos los valores de concentración de $PM_{2,5}$ en el conjunto de datos medidos.

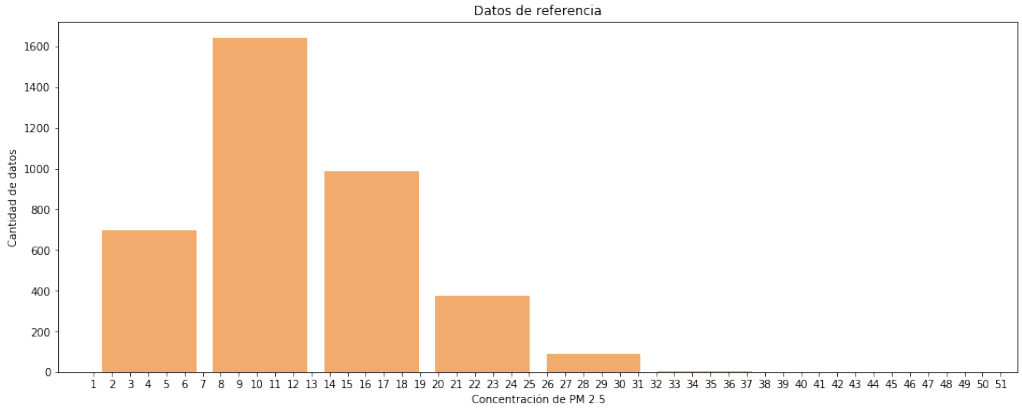

Figura 2: Histograma que muestra cómo están distribuidos los valores de concentración de $PM_{2,5}$ en el conjunto de datos de referencia.

estaciones AMB de referencia, correspondientes a los tiempos en horas (3), días (4) y meses (5).

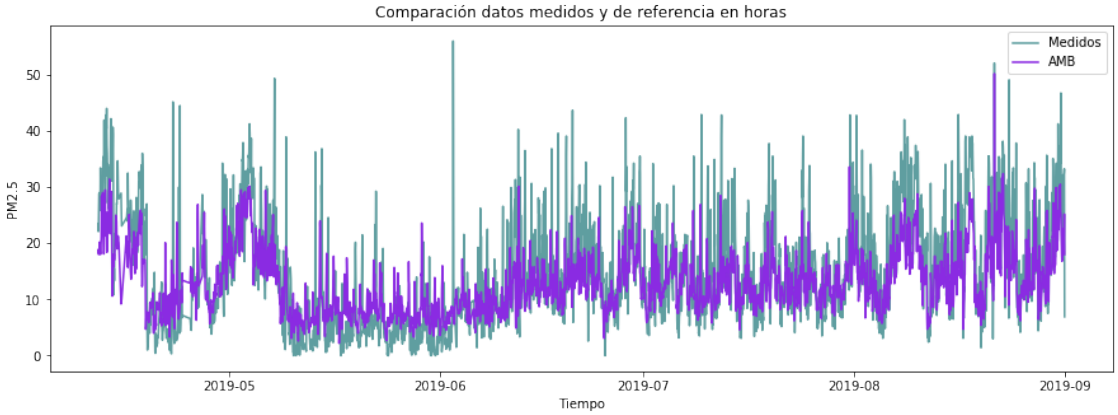

Figura 3: En esta gráfica se realiza una comparación entre los datos tratados de AMB (color morado) y de bajo costo (color azul) con respecto al tiempo en horas.

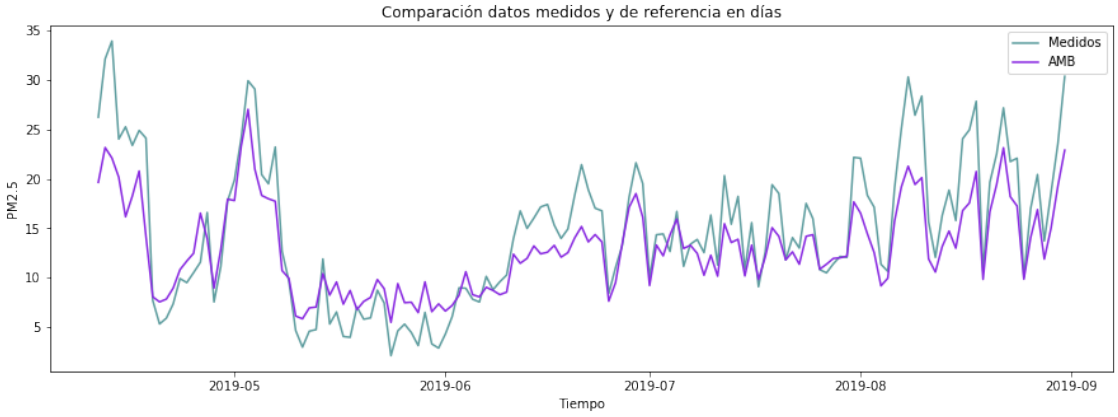

Figura 4: En esta gráfica se realiza una comparación entre los datos tratados de AMB (color morado) y de bajo costo (color azul) con respecto al tiempo en días.

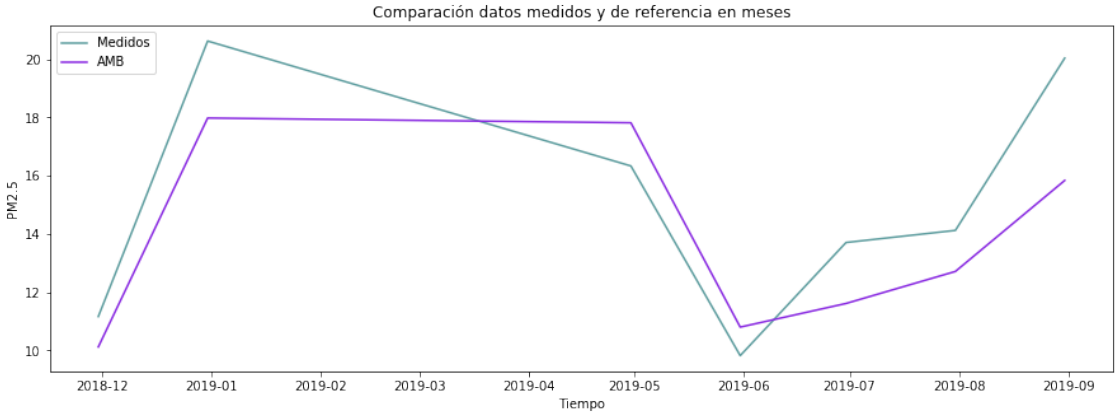

Figura 5: En esta gráfica se realiza una comparación entre los datos tratados de AMB (color morado) y de bajo costo (color azul) con respecto al tiempo en meses.

3.3. Distancia y calibración

Con el tratamiento de datos realizado anteriormente, se obtuvieron dataframes con el mismo numero de datos de PM 2.5 medidos por el sensor de bajo costo y de las estaciones AMB, para el tiempo en horas, días y meses. Teniendo esto en cuenta, fue posible calcular la distancia euclídea entre ambos conjuntos de datos para cada uno de los tiempos mencionados por medio de la ecuación 1, obteniendo los siguientes resultados:

Cuadro 1: En esta tabla se muestran los resultados obtenidos de distancia entre los datos de las estaciones AMB y de bajo costo para el tiempo de horas, días y meses.

Tiempo	Distancia euclídea
Horas	443.7818
Días	50.4503
Meses	5.9392

Para continuar con la calibración de los datos se realizó un scatterplot en Python, en donde los puntos conrresponden a la concentración de material particulado PM 2.5 de los datos medidos contra los datos de referencia de las estaciones de AMB. Posteriormente, se realizó un ajuste de mínimos cuadrados con el fin de determinar un modelo de ajuste lineal para los datos graficados en el scatterplot. Con el fin de predecir si el modelo lineal otorgaba una buena predicción para las mediciones, se dividieron los datos en dos mitades, para la primera mitad de los datos (puntos azules) se aplicó el modelo lineal y se obtuvo una pendiente, la cual se representa con una *línea azul* , esta recta predice aproximadamente en dónde deberían ubicarse los puntos de PM 2.5 venideros; ahora, se realiza lo mismo pero esta vez para la segunda mitad de los datos (*puntos verdes*), obteniendo con el modelo lineal la *línea roja*. Si las pendientes de las rectas de ambas regresiones lineales tienen un valor cercano, implica que los datos tendieron a tener el mismo comportamiento para ambas mitades y por lo tanto se puede concluir una buena predicción. A continuación, se presentarán las

gráficas mencionadas anteriormente en el orden de meses, días y horas.

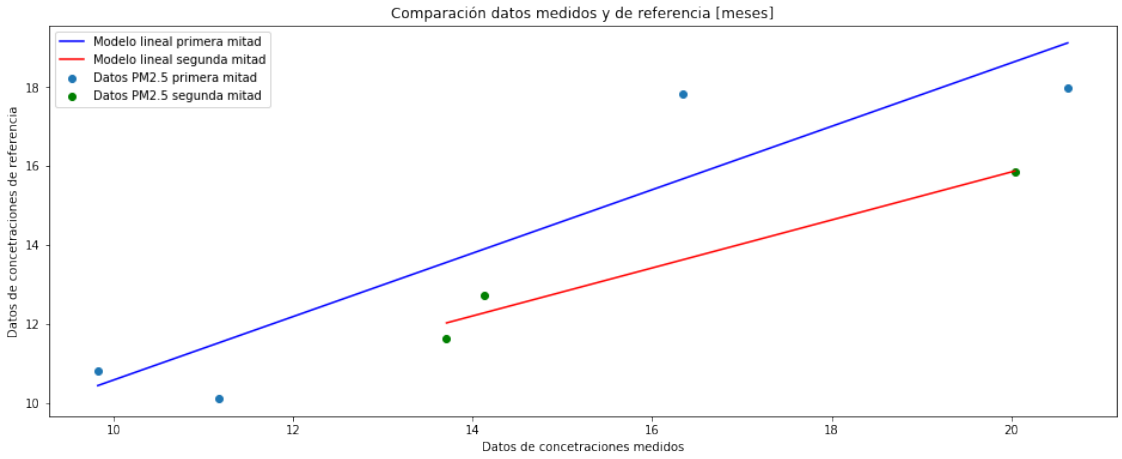

Figura 6: Gráfica en meses de PM2.5 medido por el sensor de bajo costo vs PM2.5 de las estaciones AMB de referencia. La pendiente azul es de 0.8025 y la roja de 0.6072, con un error respectivo de 24.33 %.

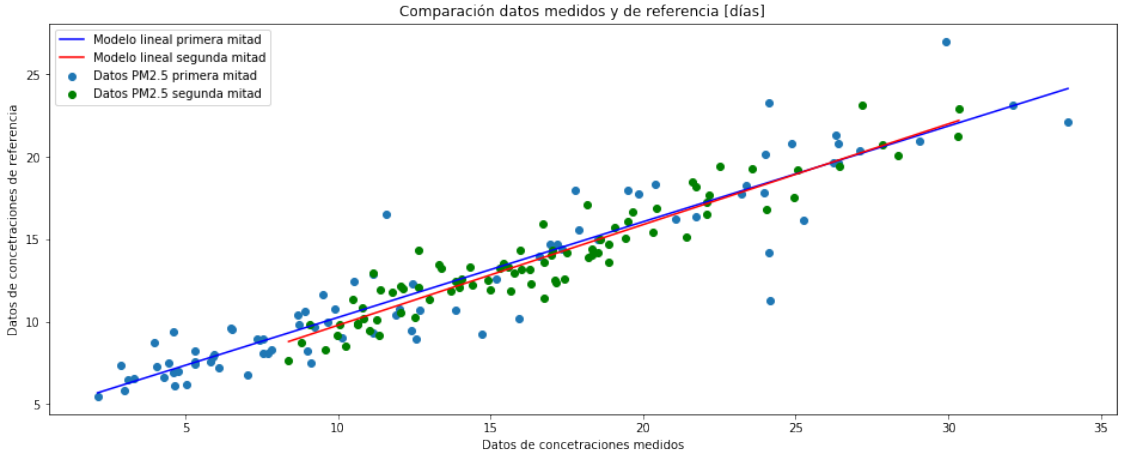

Figura 7: Gráfica en días de PM2.5 medido por el sensor de bajo costo vs PM2.5 de las estaciones AMB de referencia. La pendiente azul es de 0.5801 y la roja de 0.6102, con un error respectivo de 5.19 %.

Como se puede observar, entre más pequeña sea la escala de tiempo estudiada, se tendrá una mayor cantidad de datos, de manera que la escala de horas (con un error del 2.62 % y una exactitud del 97.38 %) es el conjunto de datos generador del modelo que nos brinda la mejor predicción.

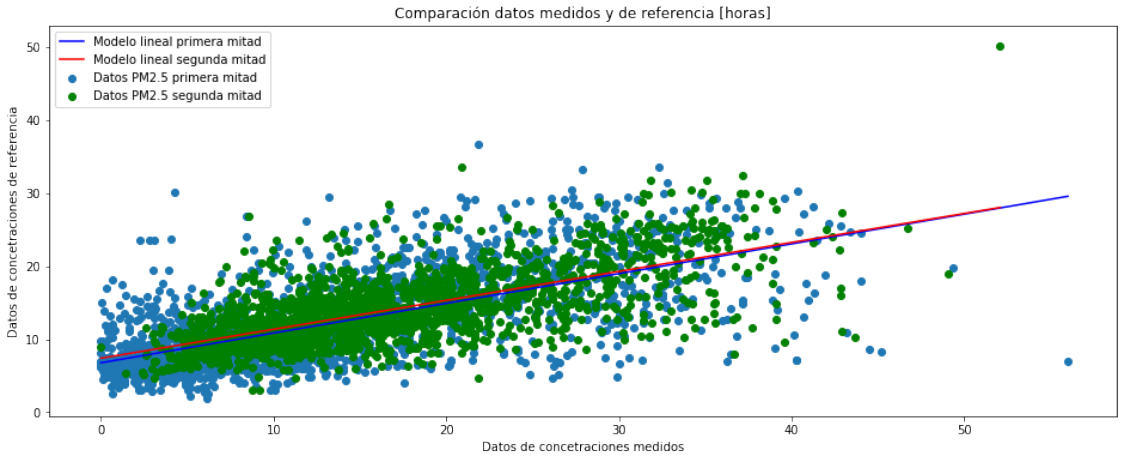

Figura 8: Gráfica en horas de PM2.5 medido por el sensor de bajo costo vs PM2.5 de las estaciones AMB de referencia. La pendiente azul es de 0.4069 y la roja de 0.3962, con un error respectivo de 2.62 %.

3.4. Promedio móvil

Tras haber realizado todo el proceso del remuestreo y de haber limpiado los datos, se procedió a mejorar aún más el comportamiento de los datos, controlándolos al hacerlos menos fluctuantes a través de una media móvil. Se decidió que el número de pasos a utilizar sería 1 hora y se realizaron múltiples pruebas para encontrar el ancho de la ventana que destacara más en precisión para los datos, como se puede observar en la tabla 2, este ancho fue de 30 datos.

Cuadro 2: En esta tabla se muestran los resultados obtenidos de distancia entre los datos de las estaciones AMB y de bajo costo para ventanas en las que se varían su tamaño y que tienen saltos de 1 hora.

Tamaño ventana	Distancia	%error entre regresiones
10	246.20	6.50 %
20	215.20	1.41 %
25	207.74	0.72 %
30	203.85	0.29 %
40	195.99	0.81 %
50	190.02	1.86 %
60	185.06	3.26 %

Entonces, utilizando una función de Python para el promedio móvil (*rolling*), se calcularon estos para cada uno de los tiempos (comenzando desde n=30 por como está definida la ventana), tanto de los valores medidos por el sensor de bajo costo como de los valores de referencia de las estaciones AMB, además, se calculó el error entre estos 2 promedios móviles.

Con el fin de visualizar mejor lo obtenido, se muestra una sección de los datos en horas, esto porque en la escala de horas el número de datos es mayor y por lo tanto podemos tener una mejor idea de cómo es el comportamiento de los datos a través del tiempo.

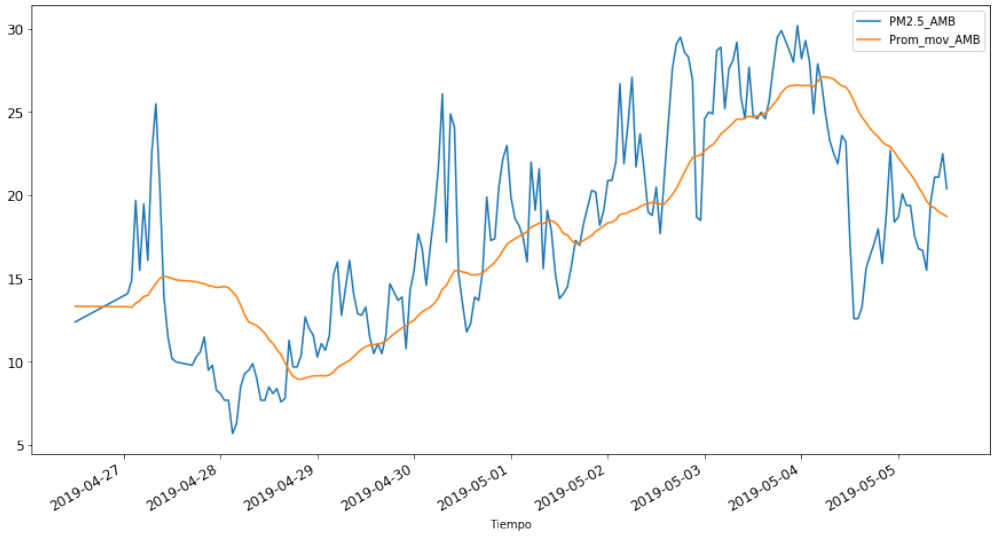

Figura 9: Comparación entre el comportamiento de los datos y de la media móvil a través del tiempo para las estaciones de referencia AMB.

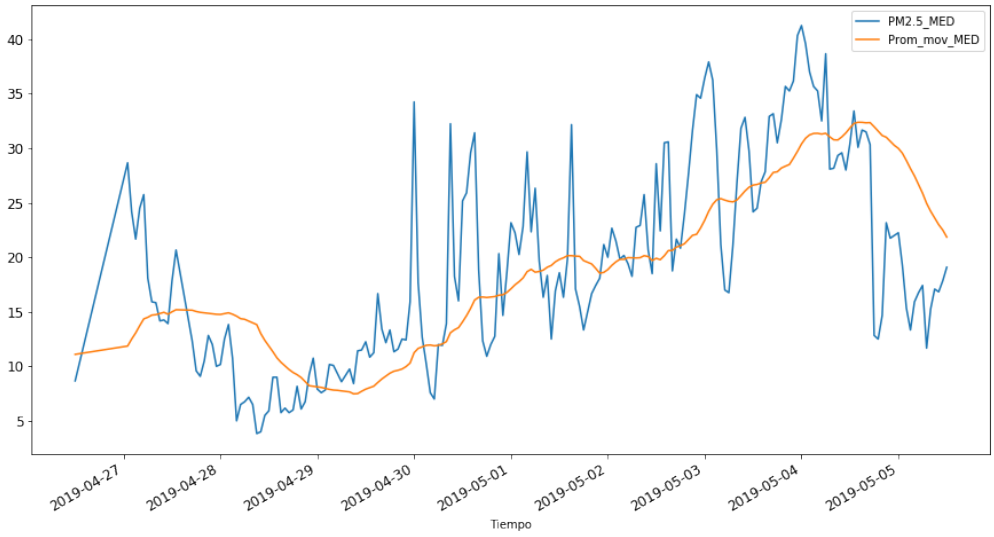

Figura 10: Comparación entre el comportamiento de los datos y de la media móvil a través del tiempo para los datos medidos por el sensor de bajo costo.

Como se puede observar, la variación del promedio móvil es menor que la de los datos como tal, haciendo así una gráfica más suave. Cabe destacar que si se toman los valores de los promedios móviles para hacer una regresión lineal la predicción es más precisa, arrojando un error del 0.29 % y una exactitud del 99,71 %, a continuación se mostrará lo dicho en escala de horas:

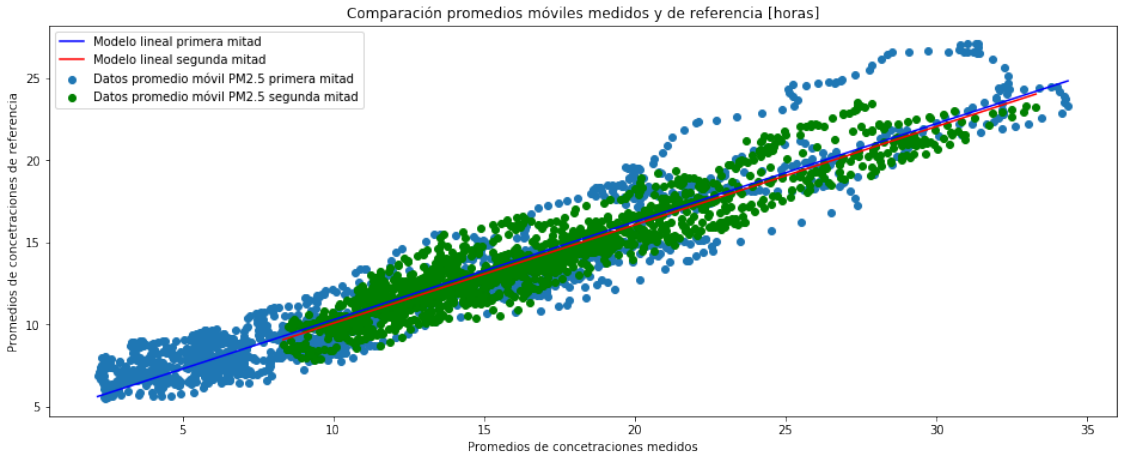

Figura 11: Gráfica en horas del promedio móvil de PM2.5 medido por el sensor de bajo costo vs promedio móvil PM2.5 de las estaciones AMB de referencia. La pendiente azul es de 0.5974 y la roja de 0.5991, con un error respectivo de 0.29 %.

4. Conclusiones

Teniendo en cuenta los resultados obtenidos a lo largo de este trabajo se llegaron a las siguientes conclusiones:

- En primera instancia, se realizó un ajuste lineal por medio del método de mínimos cuadrados para la relación entre los conjuntos de datos medidos y de referencia, obteniendo así que la mejor predicción correspondía a la escala de horas, ya que en esta se tienen una mayor cantidad de datos para analizar, con un error del 2.62 % y una exactitud del 97.38 %.
- En segunda instancia, a partir del tratamiento y reorganización de los datos se realizó también un ajuste lineal por medio del método de mínimos cuadrados para los valores promedios de las ventanas móviles, obteniendo como la mejor predicción la ventana con tamaño 30 y saltos de una hora, con un error del 0.29 %, una exactitud del 99.71 % y una distancia euclídea de 203.85.
- Finalmente, se llegó a que la mejor calibración correspondía al obtenido por medio del promedio de la ventana móvil, es decir, la ventana de ancho 30 y saltos de 1 hora, lo cual correspondía con la tolerancia establecida.

Referencias