Metricas, datos y calibración inteligente

Angel Blanco[*], **M.Tarazona-Alvarado**[†], **Yerimi Gamboa**[‡]
Universidad Industrial de Santander, Bucaramanga, Colombia.

Resumen

El desarrollo de sensores esta en su punto más alto, se encuentran sensores para casi todas las variables que encontramos en la cotidianidad y que sean medibles; esta proliferación de los sensores conlleva a que estos no sean lo suficientemente precisos y se hace necesario calibrarlos mediante un patrón de referencia. En este proyecto se busca calibrar un sensor de material partículado 2.5 [μm] medidos por un sensor *de bajo costo* teniendo como referencia los datos de las estaciones de *AMB* usando la noción de *distancia euclidiana* para estimar distancias entre los datos y posteriormente se usa el medo de *mínimos cuadros* para determinar el modelo lineal que permita calibrar las mediciones.

1. Introducción

Desde su surgimiento los sensores han sido protagonistas en la generación de datos, en la actualidad se ha llegado a un punto en que estos sensores se han hecho mas accesibles al publico, uno de los problemas de esto es que los sensores de bajo costo no son lo suficientemente precisos por lo tanto deben ser calibrados de aquí la importancia de la calibración inteligente. Los datos a calibrar fueron los obtenidos por estaciones de bajo costo y se compararon con datos de referencia obtenidos por el AMB (Acueducto Metropolitano de Bucaramanga), estos datos miden la concentración de material particulado $PM_{2,5}$ de $2,5\mu m$

2. Datos y métodos

Para realizar la calibración del sensor se proporcionaron datos desde el **2018-10** hasta el **2019-8** tanto para la estación de la **AMB** como para la del **sensor** y mediante el uso de un ajuste usando mínimos cuadrados se busca el modelo para que las mediciones del sensor de bajo casto sean lo más cercanas a las de la AMB.

2.1. Datos disponibles

Para poder comparar los datos se hizo necesario tenerlos en una misma estampa temporal, el formato utilizado entonces fue **Y-m-d H:M:00** de este modo si se hacen comparables los datos, también se observó que la frecuencia de muestreo del sensor de bajo costo era de varios datos por hora, este comportamiento se ve en el periodo de tiempo de 2018-11-08 en la sección A de la figura 1. Para el lapso de tiempo de 2018-11-15 y 2018-11-15 se observó que no se registraron datos, no es hasta 2018-12-01 que los registros del sensor de bajo costo pasó a ser un dato por hora, todos

[*]angelblanco43@gmail.com
[†]miguelta281@gmail.com
[‡]jeremi0112@gmail.com

estos comportamientos se muestran en la sección A de la figura 1.

En la sección B de la figura 1 se observa que los registros tienen mas consistencia y la frecuencia de muestreo es similar al de la AMB por lo tanto los datos escogidos para calibrar el sensor corresponden al lapso temporal de 2019-05 a 2019-9 y son los datos sobre los que se elaborara el modelo predictivo. Para hacer una depuración a los datos escogidos previo a la calibración del sensor se eliminaron entonces los datos duplicados que se tenían.

Luego de escoger el set de registros a trabajar se buscaron los datos correspondientes al mismo lapso de tiempo, pero esta vez obtenidos por el AMB esto con el fin de que los datos fuesen comparables, esto se puede observar en la sección C de la figura 1.

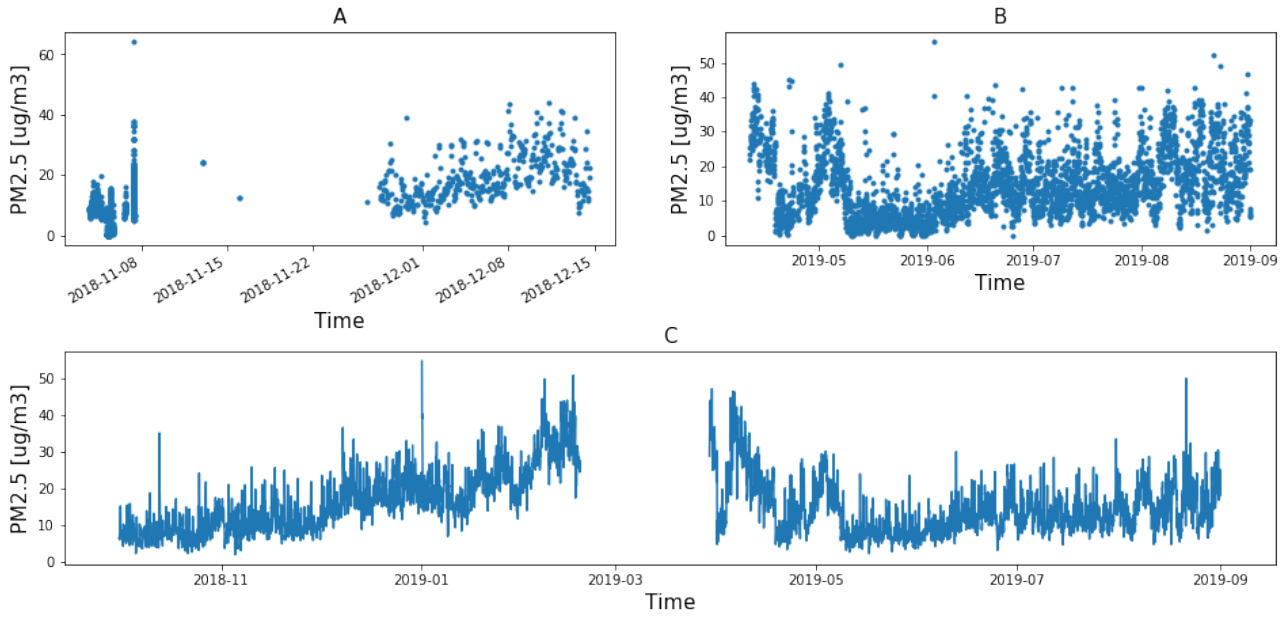

Figura 1: Visualización general de los datos

2.2. Metodología

Con los datos ya depurados, se usa un promedio móvil (1) para disminuir variaciones drásticas en las mediciones y tener una mejor idea del comportamiento de los datos, como se ve en la figura 2. Ya aplicando el promedio móvil se puede calcular la distancia euclidiana (1) entre los dos vectores (**AMB** y **sensor**), la cual brinda la noción de que tan cerca están los datos, una distancia de **0.0** representaría que los datos coinciden perfectamente lo cual en este caso sería una medición perfecta por parte del sensor de bajo costo.

$$D = \sqrt{\sum (Dj - Di)^2} \tag{1}$$

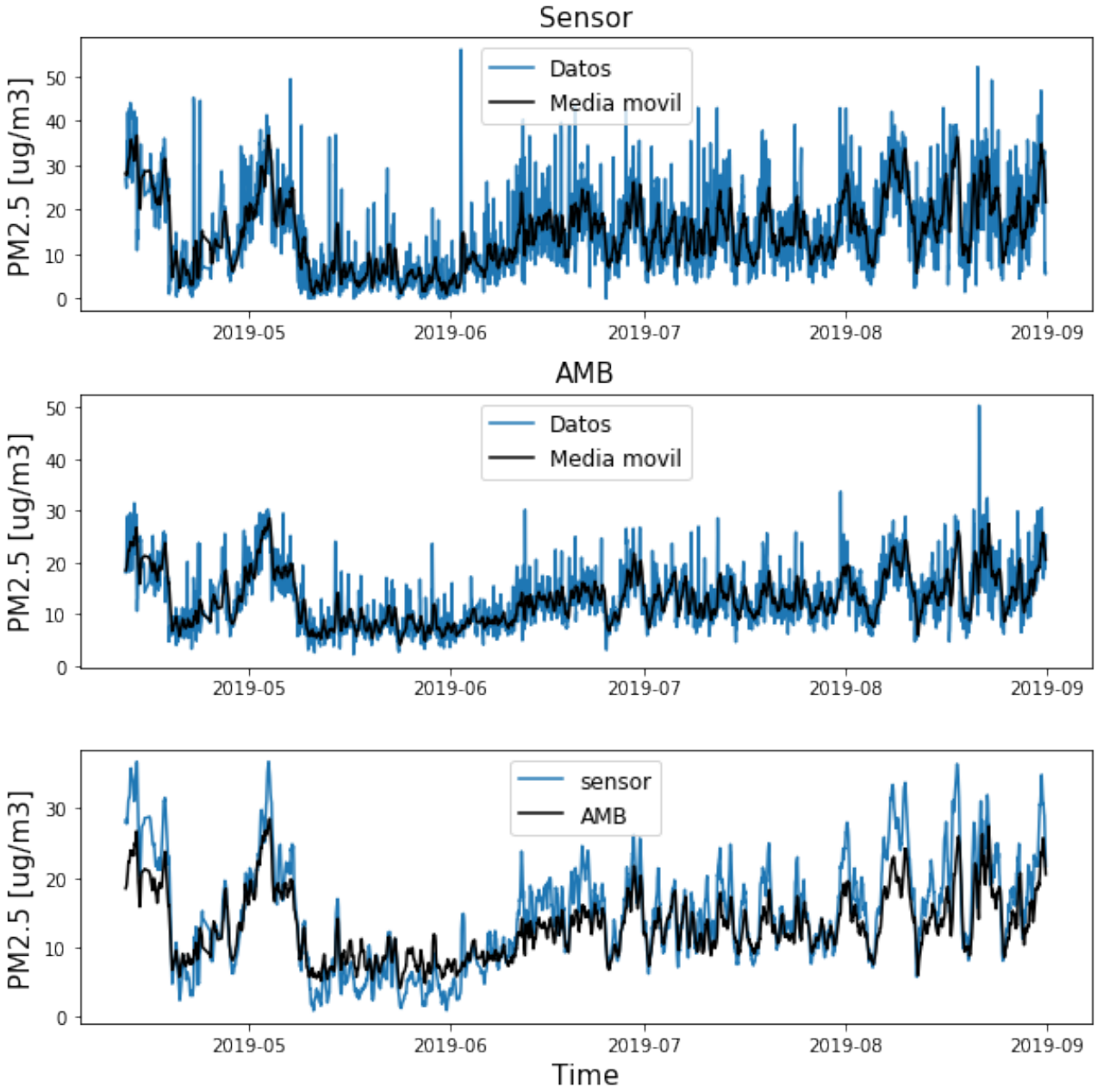

Figura 2: Visualización de los resultados de la aplicación de la media móvil a cada una de los sets de datos y superposición de las resultantes

Para encontrar el modelo que permita calibrar el sensor, se hace una correlación entre los datos de la **AMB** y el **sensor**, buscando un modelo lineal de la forma $f(\xi_j) = \alpha \hat{f}(\xi_j) + b$ usando el método de mininos cuadrados.

Para evaluar que tan bueno es el modelo encontrado se calcularon los promedios relativos y posteriormente teniendo este error como referencia, se calcularon varios modelos lineales de igual forma pero solo con un porcentaje de los datos y se probó el modelo prediciendo el faltante.

Para determinar el alcance de las predicciones dentro de una tolerancia con el fin de optimizar el menor porcentaje de datos que generan un modelo aceptable, se usó entonces el error cuando el modelo se calcula con la totalidad de los datos como una medida de la tolerancia.

3. Discusión

Cuando se calcula el promedio móvil, el valor de la ventana repercute directamente con el valor de la distancia euclidiana, como se ve en el cuadro 1. Es evidente que a medida que la ventana aumenta la distancia tiende a ser mínima, pero hay que tener cuidado con aumentar el tamaño deliberadamente pues esto representa un mayor tiempo de computo, y lo más importante una perdida de información ya que puede cambiar por completo el comportamiento del fenómeno (cambio de la cantidad de MP2.5 μm, por esta razón elegimos una ventana de **12/24** con la cual se obtiene una distancia de **208.3405 / 198.0105**.

Ventana	**Distancia entre los datos**
12	232.6
24	208.3405
36	198.0105
48	196.8
72	179.6

Cuadro 1: Tabla que contiene las distancias obtenidas entre los datos para diferentes ventanas en la medias móvil

Ya habiendo elegido la ventana buscamos la correlación entre los datos de **AMB** y el **sensor**, sobre estos hacemos el ajuste lineal y obtenemos el siguiente modelo, que tal como se muestra en la gráfica 3 representa un muy buen ajuste.

$$Modelo = 0{,}576 \times sensor + 4{,}493 \tag{2}$$

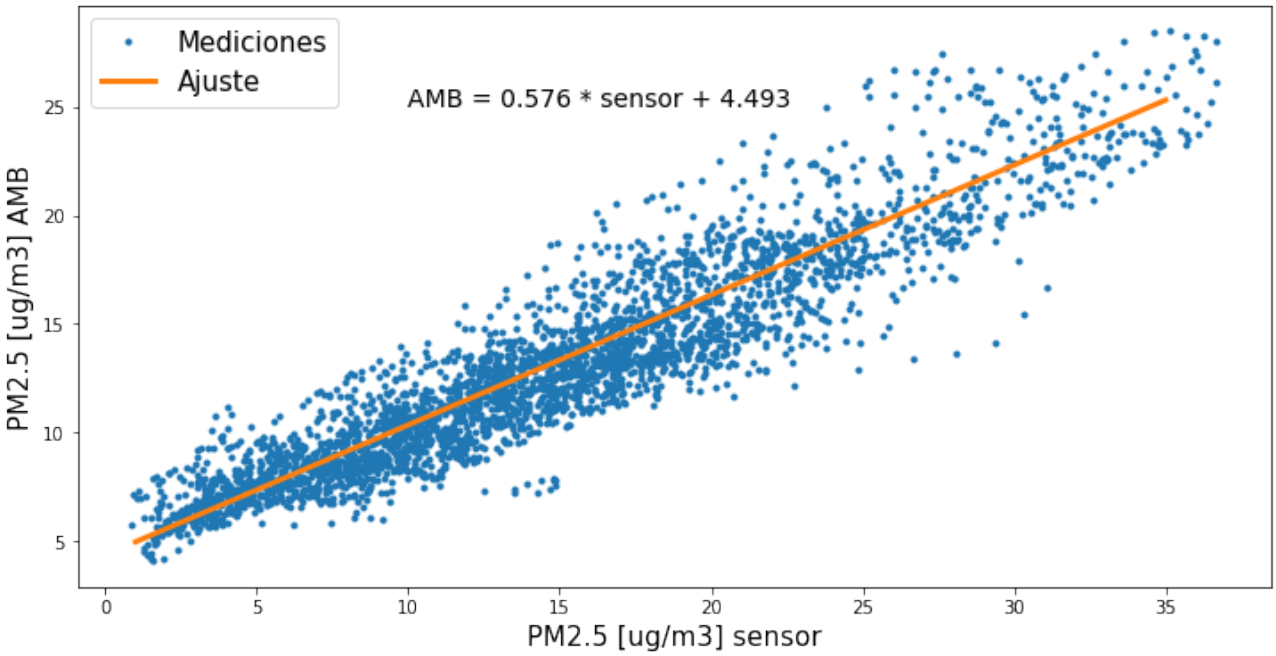

Figura 3: Correlación entre los datos del AMB y el sensor, y el ajuste lineal entre estos.

Este modelo es el que permite hacer la calibración del sensor, luego de su posterior aplicación a los datos se obtiene que efectivamente el modelo de la ecuación (2) se ajusta mejor a los datos de referencia, las mediciones del sensor de bajo costo son muy similares a las de la AMB, esto se muestra en la figura (4)

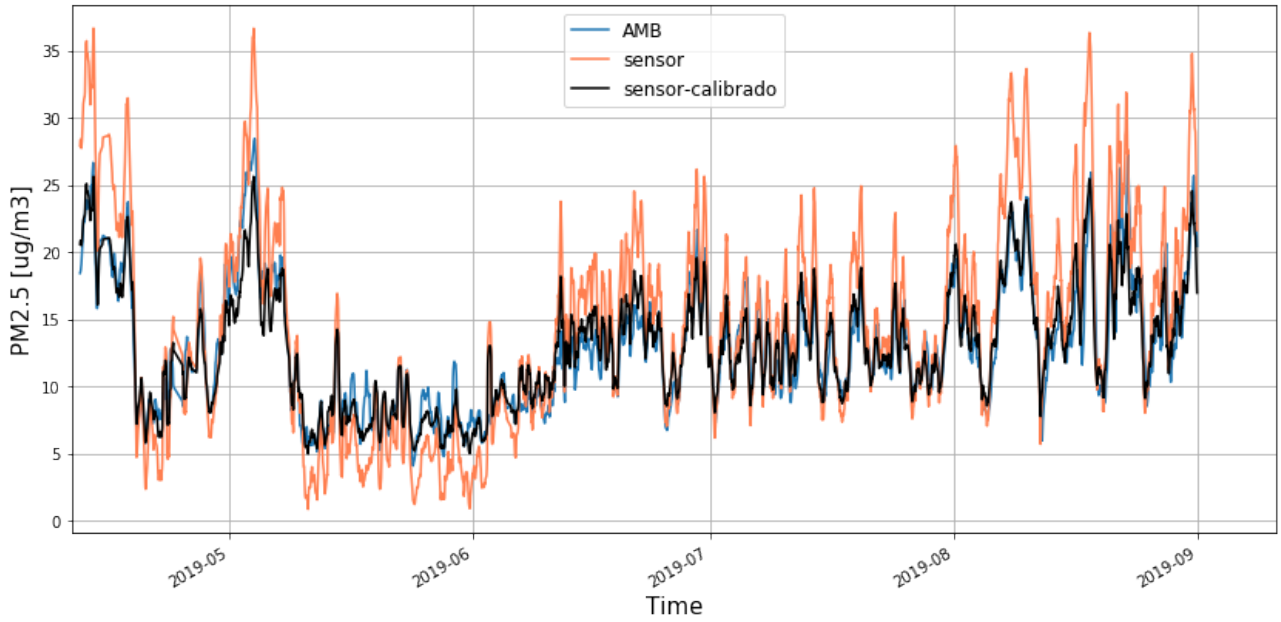

Figura 4: Muestra los datos del AMB, sensor y los datos del sensor ya calibrado

Haciendo el calculo del error relativo entre los datos del **AMB** y el **sensor calibrado** se obtiene que el error máximo es de **72% / 39%** y que en efecto el error esta cerca de cero en la mayor parte de los datos tal y como se ve en el histograma de la figura (5)

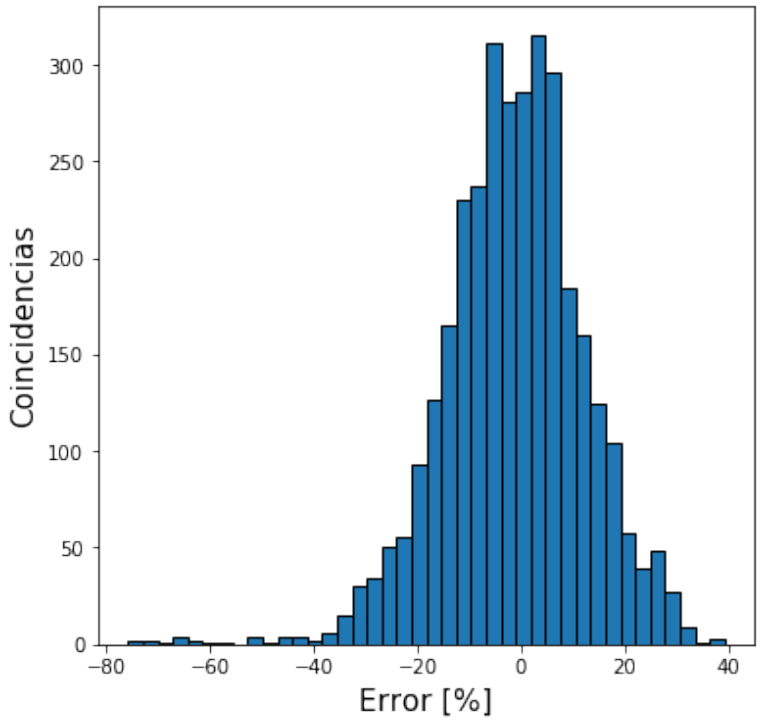

Figura 5: Error relativo entre los datos del AMB y los del sensor calibrado

Ahora, para tener un criterio de tolerancia se utilizo el modelo como medida de predicción, es decir, a cierto porcentaje de los datos se le aplicará el modelo con el fin de obtener la parte faltante de los datos, se calculó el error relativo tanto a la parte que se sometió al modelo así como a la

parte predicha con el fin de comparar estos errores y observar si el modelo es una buena forma de predecir datos.

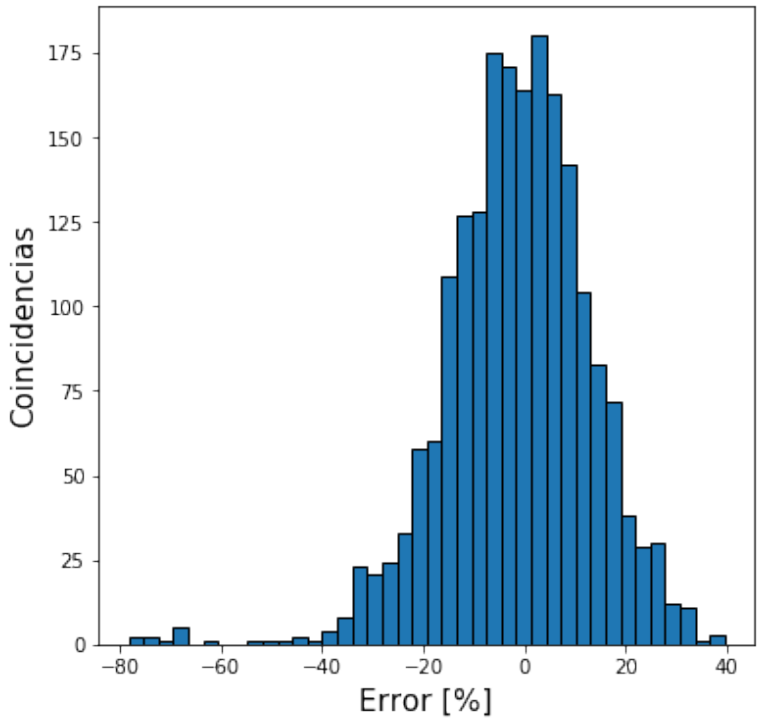

Figura 6: Error relativo entre un porcentaje de los datos del sensor y los datos correspondientes al AMB

Como se observa en la figura (6) se observa el error relativo entre el porcentaje de datos (60 %) y su parte correspondiente al AMB

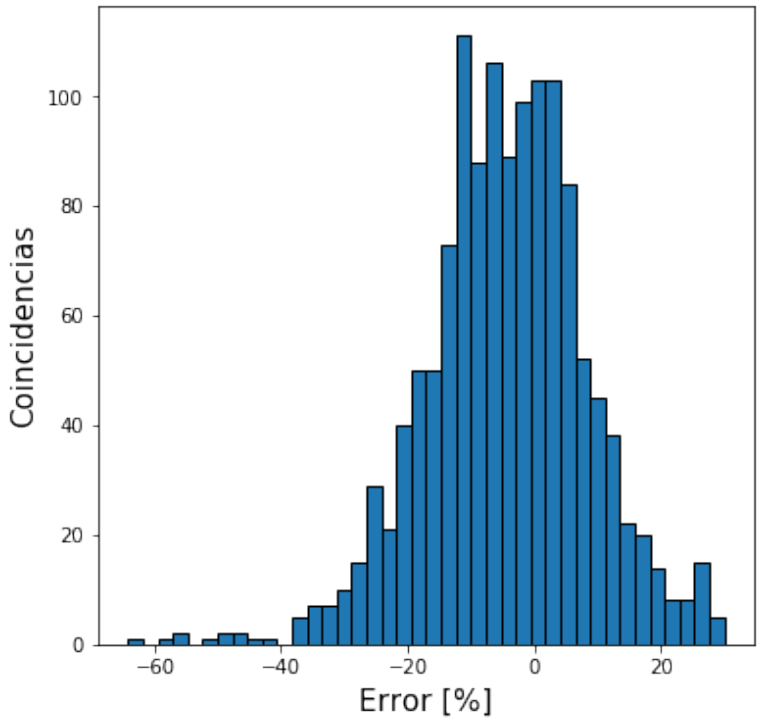

Figura 7: Error relativo entre los datos predichos y los datos del AMB

como se puede observar en la figura (7) se puede observar el error relativo que hay entre los datos faltantes que fueron predichos por el modelo y se comparan con los del AMB ahora al calcular el modelo con un porcentaje de los datos, se tomo como tolerancia cuando el error de los datos predichos sea igual o menor al error de los datos con los cuales se genero el modelo. En nuestro caso un buen modelo se obtuvo al usar (60 %) de los datos para generar el modelo y se hizo la predicción con el (40 %) restante, como se puede ver en la figura (7) el mayor error es de (60 %) el cual es inferior al (72 %) que se obtuvo para los datos que generan el modelo, como se ve en la figura (5), y esto obedece al modelo mostrado a continuación:

$$AMB = 0,592 \times sensor + 4,266$$

4. Conclusiones

El modelo lineal utilizado a la totalidad de los datos del sensor de bajo costo genera un buen ajuste con los datos de referencia **AMB** y se muestra a continuación: $AMB = 0,576 \times sensor + 4,493$ gracias a esto se pueden utilizar los datos del **AMB** para calibrar un sensor de bajo costo que tomará medidas del material particulado, que como se observó entrega medidas comparables a las de una estación (**AMB**).

Otra posibilidad es generar un modelo a partir de un porcentaje, la validez de esto se verá verificado al comparar el error generado por este porcentaje y el faltante, en este caso el mejor modelo obtenido de esta forma se orginó partiendo de un 60 % del total de datos, se podrían considerar porcentajes mas pequeños de hasta un 40 % y funcionaria bien pero consideramos que el porcentaje idóneo es del 60 % y el error máximo producido por esta configuración del modelo es del 60 %.

Es importante mencionar que con un modelo lineal como se mostró, se creó una buena calibración y consideramos que el siguiente paso seria aplicar un modelo que permita la inclusión de mas variables correspondientes a la fenomenología, se esperaría entonces la utilización de modelos autorregresivos o incluso usar machine learning.

Referencias

[1] L. A. Nuñes. Metricas, datos y calibración inteligente (2020, junio 6), de $https://drive.google.com/drive/folders/1Bc_EFntLU1VrIY4Zk4SliEfH8zv9di2M$

[2] Wikipedia contributors. (2020, June 6). Moving average. In Wikipedia, The Free Encyclopedia. Retrieved 17:23, June 9, 2020, from $https://en.wikipedia.org/w/index.php?title=Moving_averageoldid=961100045$

[3] Base de datos. AMB y sensor de bajo costo (2020, junio 6) de $https://www.dropbox.com/sh/97lqlzsac7qpykz/AAAeA0t1PC_5eRlBCvC5f1eSa?dl=0$