Métricas, datos y calibración

Barientos Bermeo D. [*]
Rodriguez Diaz J.[**]
Universidad Industrial de Santander
Calle 9 # carrera 27, Bucaramanga, Santander

6 de diciembre de 2022

Índice

Resumen

En el presente articulo se abordo el tema de la medición del error y la calibración de sensores de bajo costo respecto a unos datos de referencia "Datos Estaciones AMB" para la concentración de material particulado de dimensiones $\leq$ 2,5µm. El material particulado forma parte de la contaminación del aire, su composición es muy variada sus principales componentes son sulfatos, nitratos, el carbón, el polvo de minerales, cenizas metálicas y agua. Para la medición del error se uso la definición de distancia euclídea para ello fue necesario usar el criterio del promedio móvil (con solapamiento) donde se encontró un conjunto de datos con el que se pudo medir la distancia, también se evaluó la propuesta de utilizar un promedio sin solapamiento. Para la calibración se uso el método de mínimos cuadrados con la cual se encontró un modelo lineal que con un mínimo conjunto de datos describió la tendencia del conjunto total. Como resultado se obtuvo la estimación del error y una calibración para los sensores de bajo costo respecto a los datos de referencia.

[*]e-mail: diego2210713@correo.uis.edu.co
[**]e-mail: juan2211704@correo.uis.edu.co

1. Introducción

Los sensores permiten detectar, medir, analizar y procesar una gran cantidad de información como la alteración de la posición, la longitud, la altura, la concentración, entre muchas otras.

El desarrollo tecnológico ha traído consigo un desarrollo explosivo en sensores de bajo costo, estos sensores ya forman parte de la vida cotidiana de todos (como ejemplo de esto están los giroscopios implementados en los teléfonos celulares) sin embargo, estos sensores no poseen la precisión de uno de alto costo, y muchas veces esa imprecisión en la recolección de datos de este tipo de sensores se encuentra fuera de las magnitudes de error tolerables. Usualmente la definición del error en las mediciones de los sensores se encuentra íntimamente relacionado con la definición de métrica para espacios vectoriales de datos, en particular para el desarrollo del presente articulo el error se define como, la diferencia entre el valor verdadero y el obtenido de las mediciones (o distancia euclídea para el espacio vectorial de datos) y su cálculo es importante para determinar si un sensor de bajo costo necesita de algún tipo de calibración bajo un patrón de referencia.

El objetivo del presente articulo es cuantificar la distancia euclídea (el error) de un conjunto de mediciones de un sensor de bajo costo y como calibrarlo para establecer lecturas más precisas. La calibración que se desarrolla es sobre la concentración de material particulado de dimensiones $\leq$ 2,5µm obtenidos a través de sensores de bajo costo respecto a unos datos de referencia "Datos Estaciones AMB".

Los contenidos del presente se encuentran dispuestos de la siguiente forma: Una primera sección 2 en la que se describen los métodos para la solución del problema, en segundo lugar se tiene dispuesta la sección 3 que contiene los resultados obtenidos con gráficos que justifican y describen la información tratada, y por ultimo se encuentra dispuesta una sección 4 que concluye con el contenido del articulo.

2. Metodología

Para el desarrollo del proyecto se utilizo Python como herramienta para la lectura y el tratamiento de los datos, se dividió en dos fases que fueron la estimación del error entre los datos de los sensores de bajo costo y las estaciones AMB, y la calibración de dichos sensores.

2.1. El error

Para la estimación del error se definió el intervalo de tiempo para los cuales el conjunto de datos de la mediciones y la referencia eran continuos, es decir, para que intervalo de tiempo existían mediciones y datos de referencia. Posterior a esto se utilizo la definición de distancia euclídea para el calculo de la distancia entre las medidas de las estaciones de referencia y de las de bajo costo.

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

Figura 1: Donde esta definido como $\mathbb{D}_i = \{(x_1, y_1), (x_2, y_2) \cdots (x_n, y_n)\}$ el conjunto de datos de referencia y como $\hat{\mathbb{D}}_{\hat{i}} = \{(\hat{x}_1, \hat{y}_1), (\hat{x}_2, \hat{y}_2) \cdots (\hat{x}_m, \hat{y}_m)\}$ el conjunto de datos a calibrar, obsérvese que $f(x_i) = y_i$ y $\hat{f}(\hat{x}_i) = \hat{y}_i$.

Sin embargo los conjuntos $\mathbb{D}_i$ y $\hat{\mathbb{D}}_{\hat{i}}$ poseen dimensiones distintas. Para poder comparar los datos se realizaron dos propuestas, una fue utilizar un promedio sin solapamiento (promedio por horas), la otra fue utilizar el criterio del promedio móvil (con solapamiento) donde se encontró un conjunto o ventana ξ_j donde es posible comparar los promedios locales de ambos conjuntos $f(\xi_j)$ y $\hat{f}(\xi_j)$ y así estimar el error, para este caso $f(\xi_j) = f(x_i)$ es decir los datos de referencia se mantuvieron en las condiciones que se cargaron y se realizo el promedio móvil para los datos de las mediciones, de esta forma se identificaron los datos "más cercanos" a la referencia. Para la estrategia del promedio móvil fue necesario encontrar el solapamiento cuyo error fuera el menor, para poder comparar los resultados de ambas propuestas.

2.2. La calibración

Para la calibración se graficaron los puntos $(\hat{f}(\xi_j), f(\xi_j))$ y se usó un ajuste de mínimos cuadrados para determinar un modelo de ajuste lineal tal que $f(\xi_j) = \alpha \hat{f}_1(\xi_j)$ donde $\hat{f}_1(\xi_j)$ es el ajuste lineal y α se aproxima a 1. Este modelo se determinó a partir de un mínimo conjunto de $\hat{x_j}$ que describen la tendencia de el conjunto total de datos y cuyo error esta dentro de lo tolerable.

3. Los resultados

Los datos de las mediciones y de las estaciones AMB no son continuos en todo el intervalo del tiempo, sin embargo, hay grandes discontinuidades especialmente para los datos de las mediciones, por ello se opto por trabajar con los datos que hay a partir del 2019-04-11 hasta el 2019-08-31 ya que en este intervalo se encuentra dispuesta una relativa continuidad para los dos conjuntos de datos. En los siguientes gráficos se observa el contraste del total de datos recolectados (Véase fig 2).

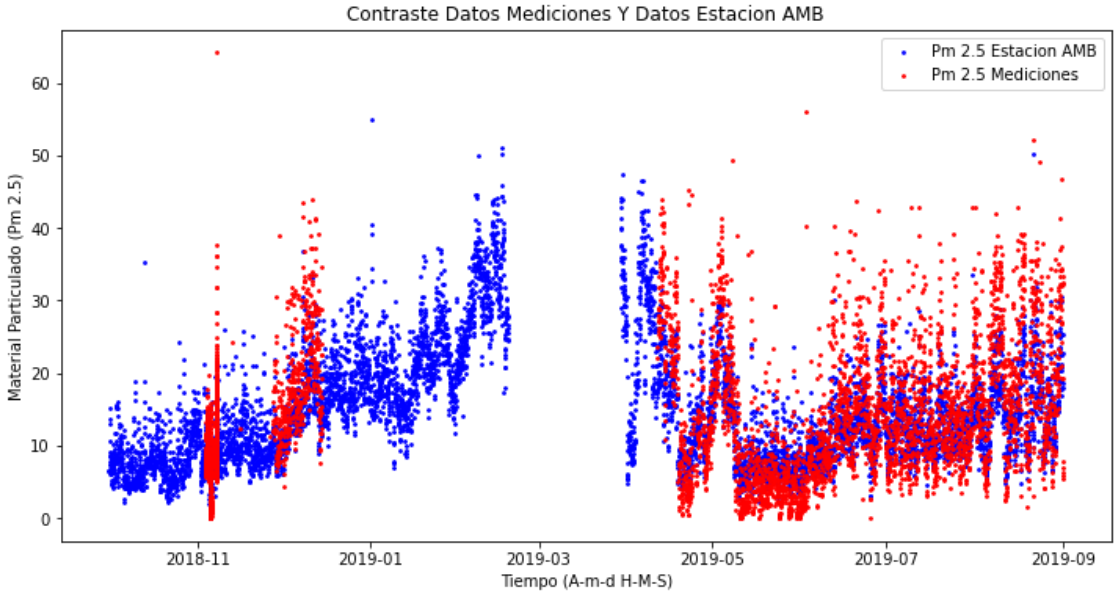

Figura 2: Contraste entre los datos de las mediciones y las estaciones AMB, donde el eje x es el tiempo y el eje y la concentración de material particulado PM2.5, los puntos azules corresponde a los datos de las estaciones y los rojos a los de las mediciones.

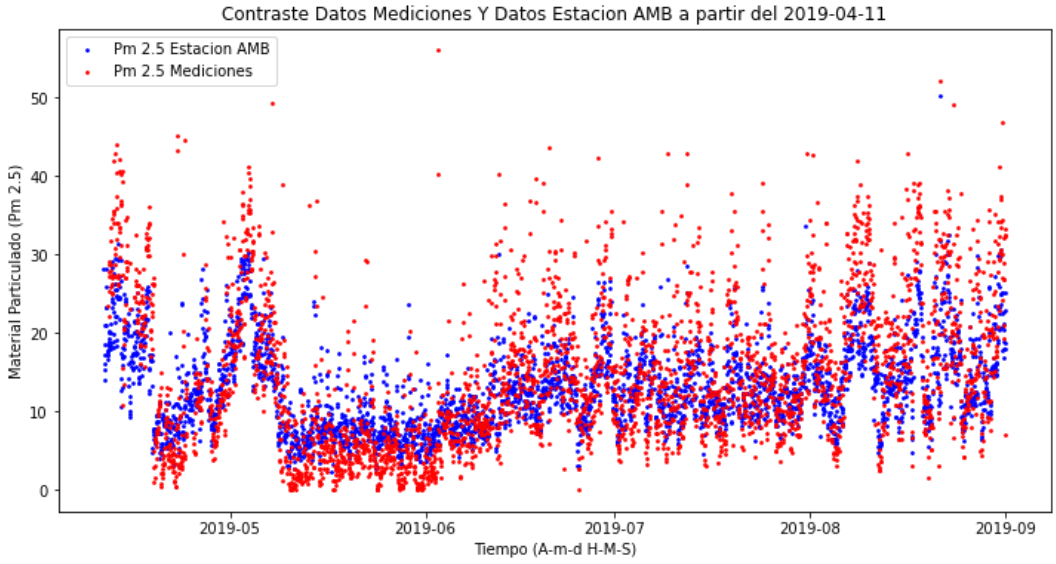

Figura 3: Contraste entre los datos de las mediciones y las estaciones AMB a partir del 2019-04-11, donde el eje x es el tiempo y el eje y la concentración de material particulado PM2.5, los puntos azules corresponde a los datos de las estaciones y los rojos a los de las mediciones.

3.1. Medición de distancias

Teniendo en cuenta la naturaleza del problema analizado "Concentración de material particulado" y la distribución horaria de los datos de referencia, el promedio móvil desarrollado utilizo ventanas de $\xi_j = 1h$. La siguiente gráfica muestra la distancia para la estrategia del promedio móvil con distintos solapamientos.

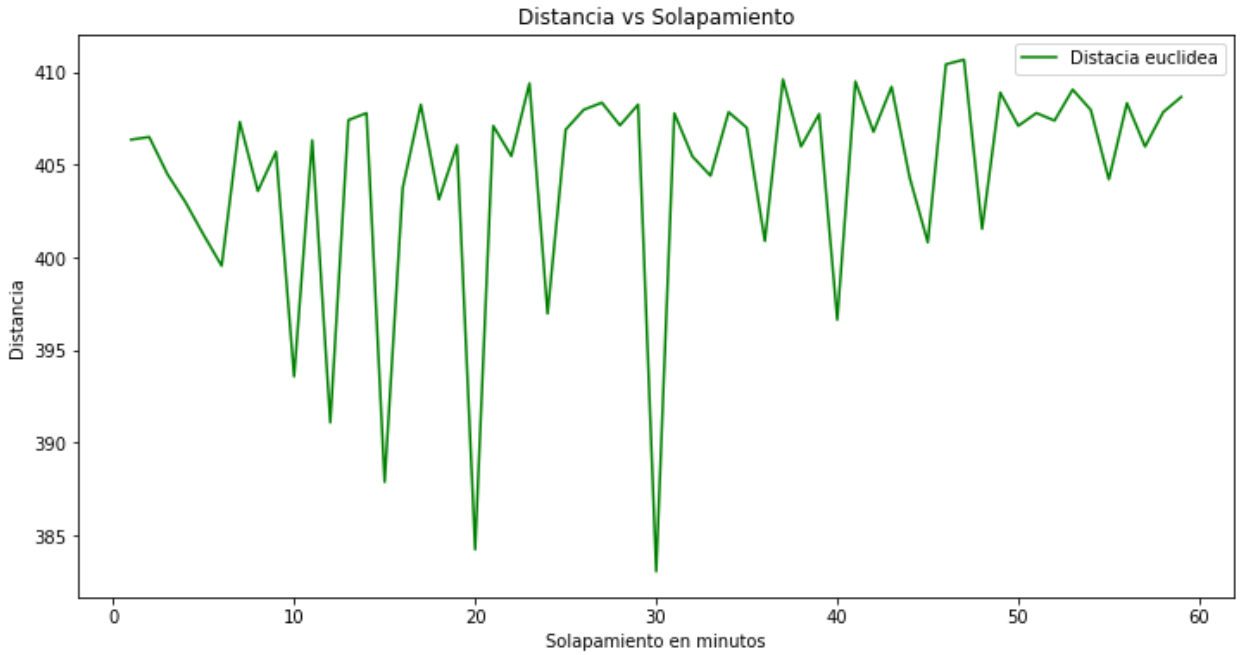

Figura 4: Calculo de la distancia euclídea para solapamientos de un minuto hasta 59 minutos, todos con una ventana de una hora.

De la gráfica anterior se observa que la menor distancia se encuentra para un solapamiento de 30 minutos y una ventana de una hora. En la siguiente gráfica se observa el contraste para las mediciones sin solapamiento, con solapamiento, y las estaciones AMB.

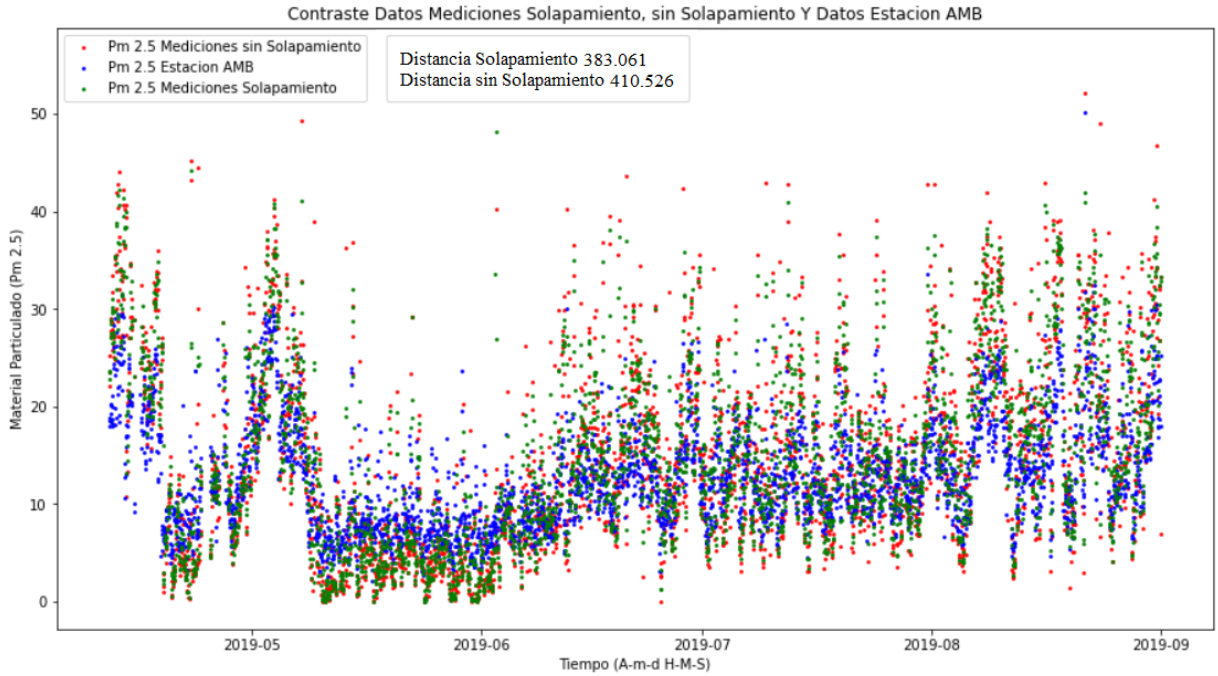

Figura 5: Contraste entre los datos de las mediciones con solapamiento, sin solapamiento y las estaciones AMB, donde el eje x es el tiempo y el eje y la concentración de material particulado PM2.5, los puntos azules corresponde a los datos de las estaciones, los rojos a los datos de las mediciones sin solapamiento y los verdes con solapamiento.

La distancia para el promedio con solapamiento es menor con un valor de 383.061 mientras que la distancia sin solapamiento es mayor con un valor de 410.526, la distancia entre estos dos métodos es de 172.956, dado que el error es menor para la estrategia del promedio móvil se escogió trabajar con estos datos para hacer la calibración.

3.2. Calibración de mediciones

Como se enuncio anteriormente en la metodología, se quiso desarrollar un ajuste por mínimos cuadrados para un modelo lineal. Sabiendo que dos magnitudes x y y se pueden relacionar a través de una ecuación lineal de la forma (1). [1]

$$y = ax + b \tag{1}$$

En la siguiente gráfica se muestran los puntos $(\hat{f}(\xi_j), f(\xi_j))$ donde $\hat{f}(\xi_j)$ son los datos para el eje X y corresponde a el promedio móvil para una ventana de una hora con un solapamiento de 30 minutos, y $f(\xi_j)$ son los datos para el eje Y y corresponden a los valores de la estación AMB.

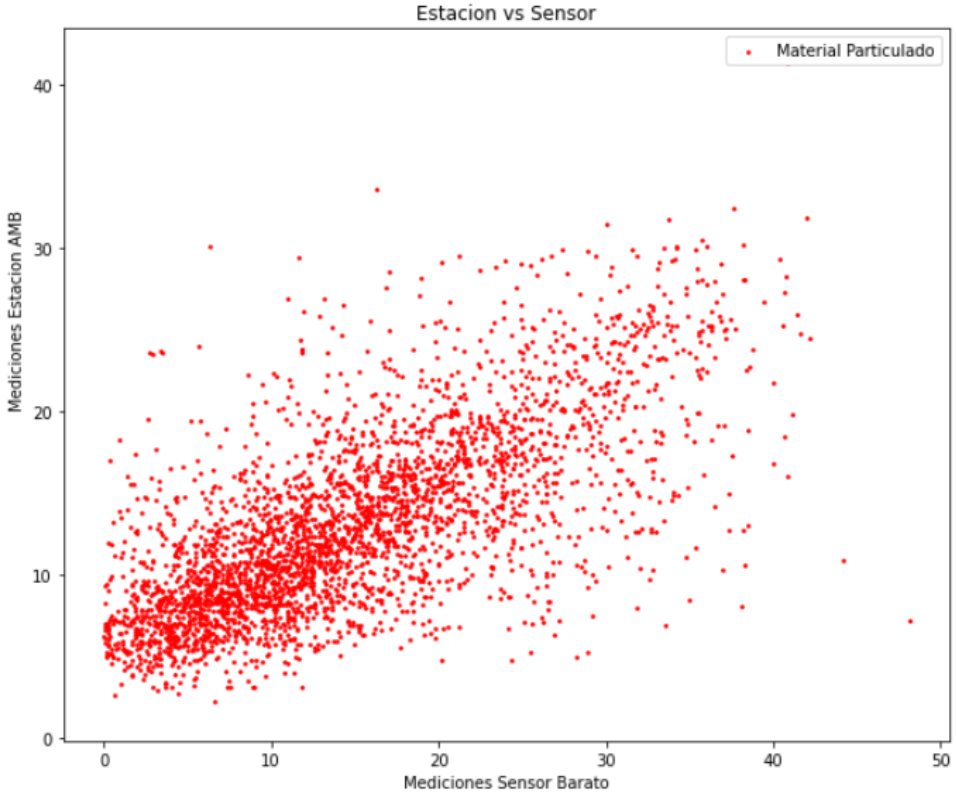

Figura 6: Donde el eje x corresponde a los valores del promedio móvil para las mediciones y el eje y corresponde a los datos de la estación AMB.

Para encontrar el valor de a y b que aparecen en la ecuación 1 se uso la siguiente definición 2.

$$\begin{aligned} a &= \frac{n(\sum x_i y_i) - (\sum x_i)(\sum y_i)}{n(\sum x_i^2) - (\sum x_i)^2} \\ b &= \frac{(\sum y_i) - a(\sum x_i)}{n} \end{aligned} \tag{2}$$

Con el fin de encontrar el intervalo de x_j mas pequeño que mejor se ajuste a los datos de la estación AMB se hizo la siguiente gráfica 7 que muestra la distancia entre los datos de referencia y las regresiones lineales según el numero de datos con el que se hizo cada regresión.

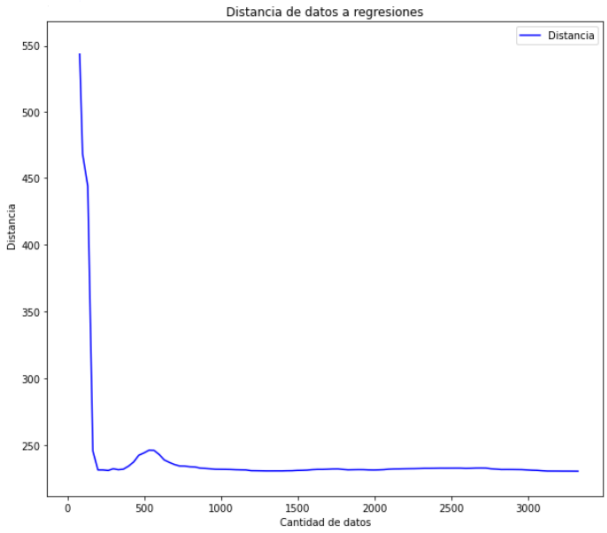

Figura 7: El eje X es el numero de datos con los que se hizo la regresión, y el eje Y es la distancia.

El siguiente gráfico muestra la regresión para los primeros 1000 datos.

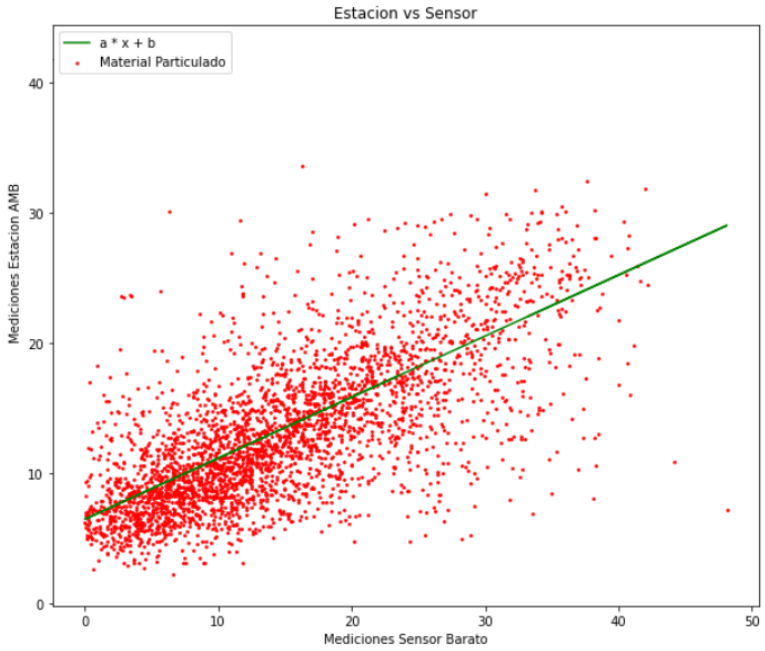

Figura 8: El eje X son los datos del sensor barato con el promedio móvil, el eje Y son los datos de la estación AMB, los puntos rojos son la concentración del material particulado, la linea verde es la regresión lineal con los primeros mil datos.

La calibración es igual al modelo lineal entonces $\hat{f}_1(x) = ax + b$ donde $\hat{f}_1(x)$ es la calibración en función de los datos de las mediciones con el promedio móvil, a tiene un valor aproximado de 0.468 y b uno aproximado de 6.444, a continuación se muestra un gráfico de los puntos $\hat{f}_1(\xi_j), f(\xi_j))$ con una regresión lineal usando todos los datos.

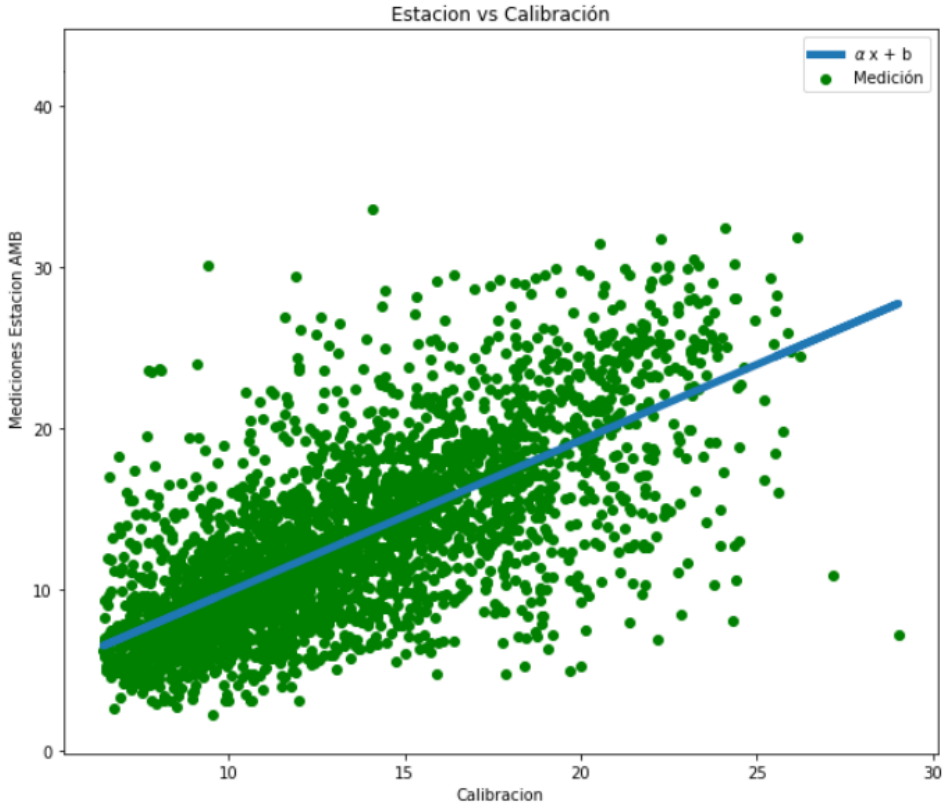

Figura 9: El eje X son los datos de la calibración el eje Y son los datos de la estación AMB los puntos verdes son la concentración de material particulado y la linea azul es la regresión lineal para todos los datos.

De la regresión lineal se obtuvo que $f(\xi_j) = \alpha \hat{f}_1(\xi_j)$ donde $\alpha = 0{,}94$ esto significa que los dato calibrados tienen la tendencia de los datos de la estación AMB.

4. Conclusiones y Recomendaciones

En conclusión el desarrollo del anterior material permite afirmar la relación entre la calibración de un sensor y la idea de métrica de espacios vectoriales.

En particular para el caso tratado, gracias a la definición de métrica fue posible identificar los datos más cercanos a la referencia y estimar un error de aproximadamente igual a 383.0611, también

fue posible determinar un mínimo conjunto de datos de mediciones $\approx 1/3$ del total de los datos, cuya tendencia es capaz de describir el comportamiento del total de datos, de los que a partir de ellos se logro estimar un α y β que representan una calibración de las mediciones de los sensores de bajo costo que se ajusta a las mediciones reales.

Por ultimo se recomienda desarrollar este tipo de mediciones y calibraciones con un conjunto continuo y uniformemente distribuido de datos de sensores de bajo costo, con el fin de evitar perdidas de información en el proceso de estimación de distancias y promedio de datos.

Referencias

[1] Torre la vega. Ajuste por mínimos cuadrados. *Escuela Politécnica de Ingeniería de Minas y Energía*, pages 7–8, s.f.