Asignación 2 Métricas, datos y calibración inteligente

Juan Diego Figueroa Hernández
Juan Andrés Guarín Rojas
*

Universidad Industrial de Santander
CL. 9 Cra 27, Bucaramanga, Santader

17 de diciembre de 2021

Índice

Resumen

En este reporte se realizo la calibración de datos de un sensor de PM 2.5, que corresponde a la concentración de partículas menores o iguales a 2.5 micrómetros. Para esto, se compararon los datos tomados como referencia de AMB con los de un sensor de bajo costo ('Low cost'). Se tomaron los datos en bruto de ambos sensores, se limpiaron y se encontró, usando el promedio de ventanas móviles, el error o distancia entre ambos datos. La distancia entre ambos datos correspondió a la raíz de la suma en cuadratura de la resta entre el promedio de una base de

*e-mails Juan Figueroa: juan2200815@correo.uis.edu.co; Juan Guarin: juan2201870@correo.uis.edu.co

datos en un intervalo de tiempo con el de la otra en el mismo intervalo. Así mismo, la calibración de los datos se realizó realizando un ajuste lineal de los datos de Low-cost vs los de AMB. Luego, se fue más allá y se probaron distintos modos de ajustar los datos para obtener la menor distancia luego de la calibración, como por ejemplo, realizar el ajuste con solo una fracción de los datos y a la final, ver cual sería el menor conjunto de datos para tener una calibración dentro del rango de error que se esté dispuesto a aceptar. Como resultados, se obtuvo que los parámetros de las ventanas que mejor capturan la información que se requerían es cuando se tienen ventanas de 1 hora y pasos de 1 hora también. La distancia entre las datos originales fue entonces de 410.6 $\mu g/m^3$ y luego de calibrar los datos se redujo un 57 % hasta 235.1. Igualmente, basados en una tolerancia igual a 10 $\mu g/m^3$ se determinó que el modelo lineal funciona bastante bien en los datos completos, y al realizar el ajuste con un conjunto menor se determinó que los datos posteriores a este conjunto se ajustan bien cuando se toma al menos un conjunto de 21 días con datos de al menos uno cada hora. Del mismo modo, el alcance de la mejor predicción fue variado pues se ajustó bien al primer y tercer 1/4 de los datos posteriores al ajuste, pero se ajustó mal en el segundo y cuarto 1/4 de los datos.

1. Introducción

En esta era se vive un desarrollo inmesurable de sensores que se encuentran en lo más cotidiano o íntimo de el dia a dia. Estos sensores acompañados de las nuevas tecnologías hacen parte de lo que hoy se conoce como Internet of Things (IoT), los cuales en muchas ocasiones carecen de presición y necesitan ser calibrados con un sistema de referencia dado. Así que a continuación se presenta un problema el cual consiste en cuantificar el error de un sensor específico y buscar cómo calibrarlo para obtener datos más certeros. Este problema es de gran importancia pues ilustra el porqué la idea de calibrar y buscar errores está muy ligada intrínsecamente con la idea de métrica en un espacio vectorial dado [1] .

A continuación se presenta la metodología 2, sección en la que se explica como obtener la distancia entre los datos medidos y de referencia, cómo obtener un ajuste determinado para calibrar el sensor y cómo implementar la ídea de promedios móviles por medio de ventanas y pasos temporales específicos para calcular según la métrica. En la sección de resultados se muestra a lo que se llegó por medio de python 3, y finalmente en esta sección 7 se mostrarán las conclusiones del presente trabajo.

2. Metodología

Antes de poder utilizar los datos para hallar distancias y plantear las ventanas, se revizó la cantidad de datos medidos y la cantidad de datos de referencia, se cambiaron los nombres de las columnas para cada conjunto de datos con el fin de no tener problema al hacer un merge con los datos, se concatenaron el conjunto de mediciones para poder tener todo en un mismo archivo, y se limpiaron los conjuntos de datos eliminando filas en las que no había información por medio del comando dropna.

Posteriormente se procede a plantear distintas ventanas móviles con distintos pasos, a fin de ver el comportamiento de la distancia entre los datos medidos y los de referencia, y a su vez se halló el valor de distancia media por unidad de dato, puesto que evidentemente entre más datos para comparar mayor será la distancia, para lo anterior se hizo uso de la siguiente definición de distancia.

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

Después se realizó un ajuste de datos de manera lineal con el fin de saber que operaciones han de hacerse a los datos medidos para calibrar el instrumento, y se comparó de nuevo el valor de distancia y de distancia por unidad de dato obteniendo mayor precisión al comparar los datos con los valores de referencia. Para definir los valores óptimos de ventana y pasos temporales se comparó los valores de distancia y distancia por unidad de dato entre 6 combinaciones de ventanas y pasos específicos, con el fin de que en los siguientes incisos se pueda resolver aplicando esta ventana y pasos óptimos.

Para los análisis de los alcances se procedió a tomar el intervalo temporal de interés y se dividió en cuatro o seis partes. Después de eso, en cada sub-intervalo se aplico la función de distancia para conocer el parecido de los puntos de datos de AMB con los de calibración en ese mismo sub-intervalo. Como resultado de esto se obtuvieron los valores de distancia y distancia relativa que nos permitieron saber si la calibración fue buena en cada pedazo del conjunto de datos.

Para las predicciones, se usaron un conjunto de datos en un intervalo temporal al que se le aplicó la función de ajuste lineal.Con esto se obtuvieron los mejores parámetros del ajuste lineal. Luego de esto se tomaron los datos ubicados temporalmente en el futuro de los datos del ajuste lineal y con ellos se halló el conjunto de datos calibrados. Luego de esto, se aplicó la función distancia a este conjunto de datos y el valor obtenido nos permite concluir que tan bueno fue la predicción, en ese conjunto de datos en el futuro.

3. Calibración y los resultados

Primero se realiza un tratamiento de datos el cual consiste en concatenar los datos, realizar un merge, eliminar filas vacias y conseguir que los datos de referencia y de mediciones tengan la misma cantidad de elementos, logrando obtener la siguiente gráfica.

Posteriormente se realiza un ajuste lineal al comparar los datos entre ellos mismos según la escala temporal, y con este ajuste se obtienen los parámetros que nos permiten calibrar las mediciones, obteniendo mejoras en las gráficas como se muestra a continuación.

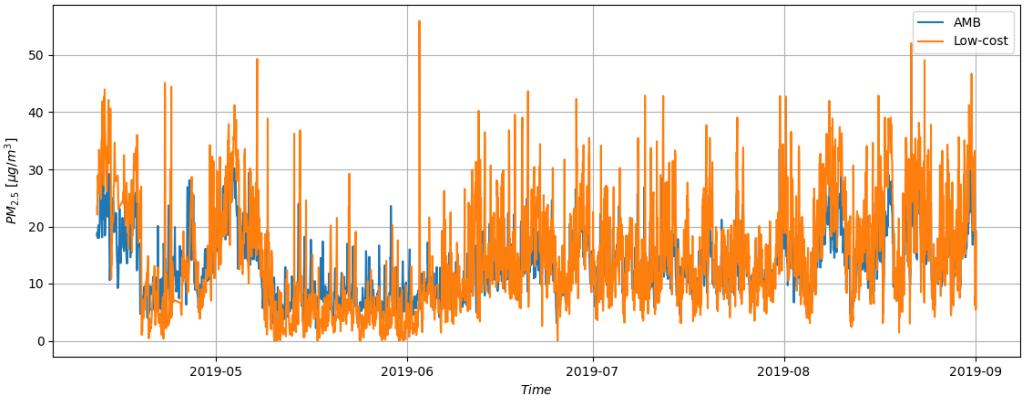

Figura 1: En esta gráfica se muestran los datos originales (sin calibrar) de low-cost y de AMB con respecto al tiempo.

También se realiza una optimización en el valor de ventana y pasos a emplear para medir distancia entre los conjuntos de datos, así que se emplearon 6 configuraciones distintas para estos valores temporales, y se busco cual de ellos realizó una mejor calibración, como se aprecia en la siguiente tabla.

Cuadro 1: En esta tabla se muestran los nuevos valores de distancia tras haber realizado la calibración con distintas configuraciones de ventanas y pasos temporales en horas.

Epsilon [$horas$]	L [$horas$]	Distancia [$\mu g/m^3$]	Distancia relativa [$\mu g/m^3$]
0.5	1	235.01	4.075
1.0	1	235.01	4.075
1.0	2	235.86	4.09
1.0	10	248.53	4.31
6.0	24	253.47	4.395
12.0	24	253.95	4.403

De manera adicional se presenta como se verían los datos organizados según la escala temporal para poder hacer el ajuste lineal, como se aprecia en la siguiente gráfica (obsérvese que en este caso se usaron los valores de ventana y pasos óptimos).

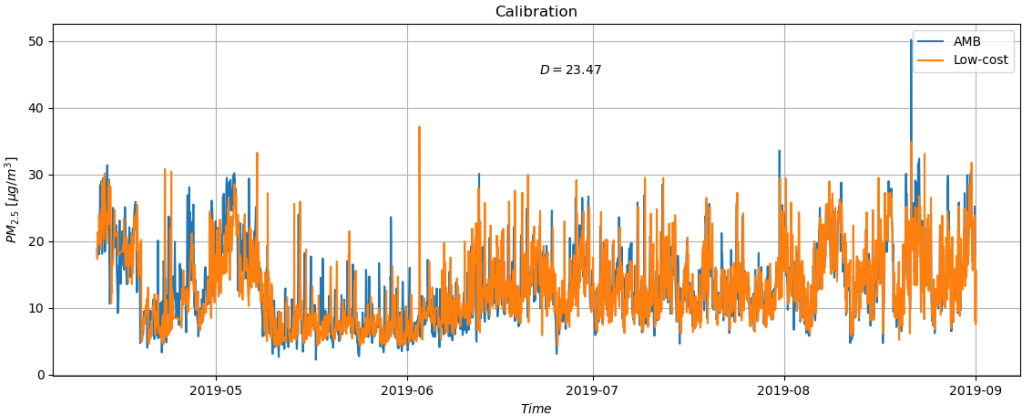

Figura 2: En esta gráfica se muestra un ejemplo al tener los datos de low-cost calibrados gracias a un ajuste lineal vs los datos de AMB con respecto al tiempo, de manera adicional aparece el valor de distancia pero el que corresponde al calculado con la ventana y valores reducidos.

También se busca visualizar ese ajuste óptimo por medio de una gráfica para poder proceder a analizar el alcance y la predictibilidad que tiene este modelo lineal.

En síntesis, el ajuste óptimo fue: , donde las desviaciones estándar entre los datos fueron $m = 0{,}01$, $b = 0{,}13$ con m y b siendo la pendiente y ordenada. Y a su vez, el error RMSE (Error cuadrático medio por sus siglas en inglés) del ajuste fue igual a 4,07 [$\mu g/m^3$].

4. Tolerancia dispuesto a aceptar y alcance del modelo lineal

Una vez que se tuvo el valor de la tolerancia establecido, que en este caso fue de 10 $\mu g/m^3$ pues coincidía con el orden de magnitudes de los datos de AMB. Se pudo analizar el alcance del ajuste lineal presentado hace un momento. Se procedió a tomar la mejor calibración de datos, que en este caso fue la que tuvo menor distancia relativa. Esta calibración fue de $data\_Calibracin_i = data\_Low\_cost_i \times 0{,}41 + 6{,}94$. Para analizar el alcance de este modelo se procedió a dividir el intervalo de tiempo de los datos en cuatro sub-intervalos. En cada una de las partes se halló la distancia relativa entre los datos de AMB con los calibrados para ese mismo sub-intervalo y este valor se comparo con la tolerancia puesto que la distancia relativa representa la variación media que tiene cada dato de una base de datos con respecto a la otra (al igual que la desviación estándar). De esta manera, se buscó que la distancia relativa fuera menor al valor de tolerancia. Ahora, Los resultados encontrados fueron:

Se observo que la distancia relativa seguía estando menor al valor de tolerancia en cada uno de los sub-intervalos de tiempo. Por lo cual, el alcance del modelo corresponde a todo el rango de tiempo de los datos originales. Adicionalmente, nótese que la distancia relativa en los sub-intervalos 3 y 4 es más pequeña que la distancia relativa con todos los datos. Indicando que si bien el ajuste

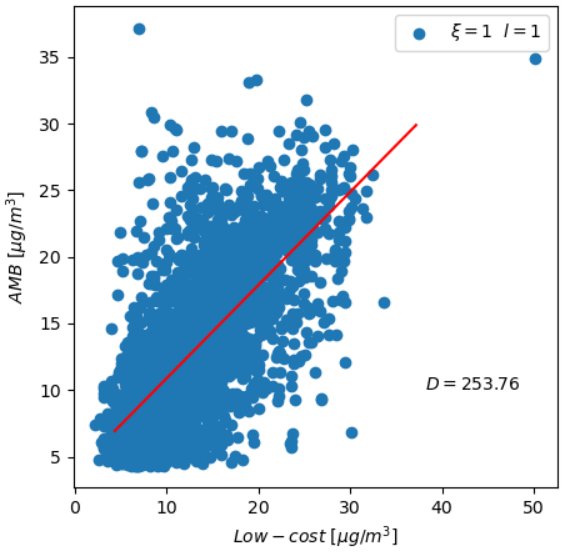

Figura 3: En esta gráfica se muestra Como se plantean los datos para hallar la recta de ajuste por medio de mínimos cuadrados y poder usar los parámetros de la recta para calibrar las mediciones.

lineal se realizo con todos los datos, este ajuste fue mejor para los datos ubicados en la segunda mitad del conjunto de datos.

5. Predicción

La primer predicción que fue realizada se hizo tomando en cuenta solo la primera mitad de los datos para hacer el ajuste lineal. Con este ajuste luego se encontraron los datos calibrados con la siguiente expresión:

$$data\_Calibracion_i = data\_Low\_cost_i \times m + b$$

Donde m y b son la pendiente y la ordenada del ajuste lineal. Luego de esto, se analizó la distancia entre los datos de calibrados con los de AMB obteniendo 273.74 $[\mu g/m^3]$ de distancia y 6.62 $[\mu g/m^3]$ de distancia relativa. Lo cual indica que el ajuste se mantuvo dentro de la tolerancia, pues la distancia relativa fue menor a 10 $[\mu g/m^3]$.

En este ajuste los parámetros encontrados fueron $m = 1{,}10 \pm 0{,}03$ y $b = -0{,}51 \pm 0{,}42$, y el error RMSE fue 7.50 $[\mu g/m^3]$. Cabe destacar que el resultado de distancia relativa fue peor que al hacer el ajuste con todos los datos.

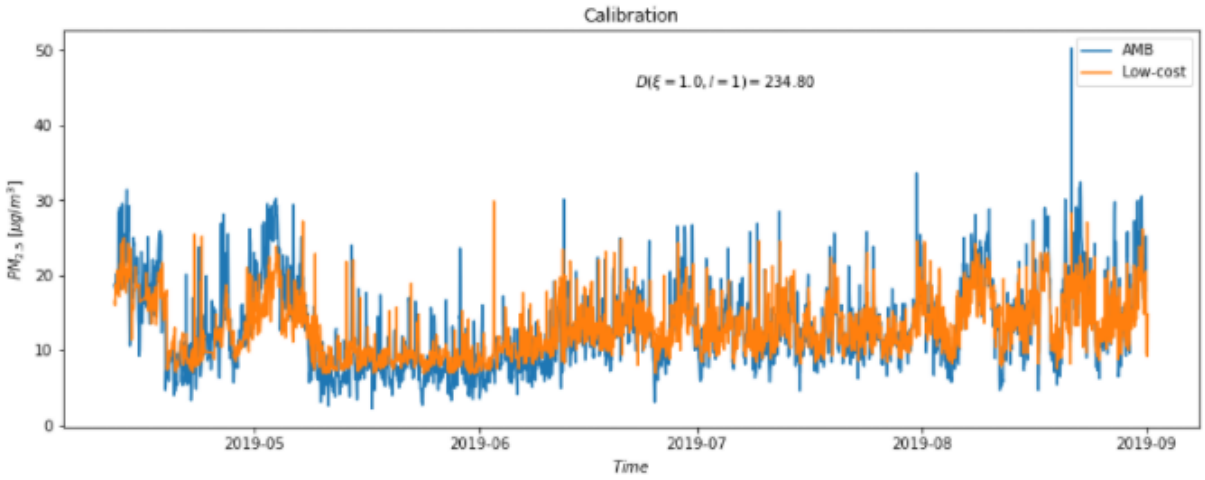

Figura 4: En esta gráfica se muestra la comparación entre los datos de AMB vs los datos calibrados con los parámetros del ajuste óptimo.

Cuadro 2: En esta tabla se ven las distancias por horas y distancias relativas entre los datos de calibración lineal y los de AMB para cuatro subconjuntos de datos. Observe que en todos ellos el valor de distancia relativa es menor que la tolerancia establecida.

Sub-intervalo	Distancia [$\mu g/m^3$]	Distancia relativa [$\mu g/m^3$]
1	139.97	5.01
2	119.72	4.14
3	110.33	3.78
4	95.22	3.26

5.1. Alcance de la predicción

Para conocer el alcance de la predicción se decidió tomar los datos posteriores a los usados en el ajuste, dividiéndolos en seis partes iguales (similar a como se hizo anteriormente). En cada uno de estos sub-intervalos se halló la distancia y la distancia relativa entre los datos calibrados y de AMB.

Por tanto, se observa que el ajuste se mantiene dentro de la tolerancia para cada sub-intervalo, pues en cada uno la distancia relativa es menor a 10 [mug/m^3]. Adicionalmente, el mejor rango donde los datos se ajustaron mejor fue en el tercero. Por tanto, el alcance de esta predicción correspondió a todos los datos analizados en la predicción.

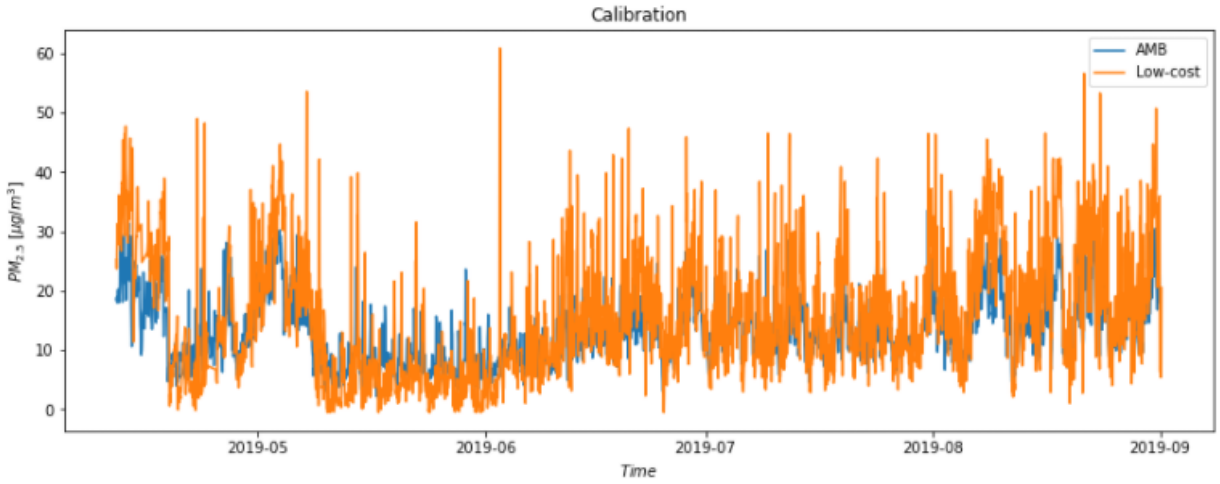

Figura 5: En esta imagen se ven los datos calibrados con solamente la primer mitad de los datos de Low cost. Observe que la segunda mitad de los datos también se calibró gracias a que la predicción fue buena.

6. Mínimo conjunto de datos para una buena predicción y su alcance

Para este caso, se realizaron pruebas con cuatro anchos o tamaños del conjunto de datos y tres tiempos de inicio de toma de datos. Por ejemplo. una prueba se realizó con unos datos de ancho igual a 21 días iniciando desde 0.3 veces el tamaño del conjunto completo de datos (ver Figura 6). Estos valores de ancho e inicio se testearon uno a uno usando una matriz de parámetros para testear. Los resultados pueden verse en la siguiente tabla (ver tabla 4)

Como pudo observarse, en general se obtuvo que los mejores ajustes de la predicción fueron cuando el punto de inicio era mayor. Esto puede explicarse debido a que según la gráfica, los primeros datos no se parecen tanto a los que siguen, y por tanto la predicción mejora al ignorar estos datos. Adicionalmente, el tamaño de la ventana generó resultados dispares, pues en algunos

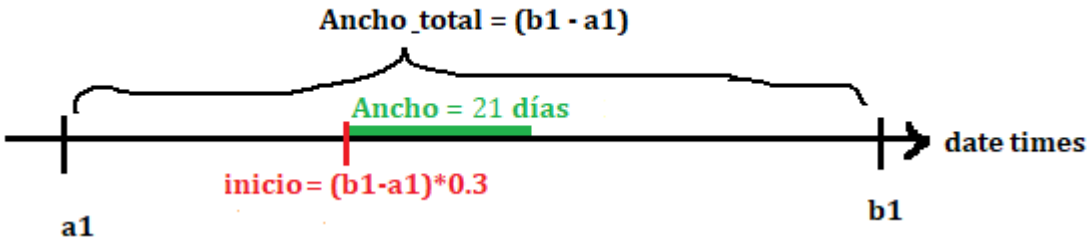

Figura 6: En esta imagen puede verse un ejemplo de un conjunto de datos usados para las predicciones de esta sección.

Cuadro 3: En esta tabla se ven las distancias por horas y distancias relativas entre los datos de calibración lineal y los de AMB para seis subconjuntos de datos. Note que todos ellos tienen una distancia relativa menor que la tolerancia.

Sub-intervalo	Distancia $[\mu g/m^3]$	Distancia relativa $[\mu g/m^3]$
1	130.12	7.71
2	133.43	7.92
3	115.07	6.82
4	129.25	7.67
5	148.57	8.80
6	125.18	7.43

Cuadro 4: En este cuadro pueden verse las distancias relativas medidas en $\mu g/m^3$ para cada uno de los parámetros usados en esta sección. Note que el valor de cada uno se mantiene dentro de la tolerancia.

Ancho / Inicios	Inicio	0.3 veces el Ancho Total	0.5 veces el Ancho Total
71 días	6.623587	6.577129	11.083333
56 días	6.976803	6.430222	6.006215
42 días	6.890205	6.636378	6.577129
21 días	6.905637	6.842152	6.45863

casos al aumentar el ancho la distancia relativa aumentaba, pero en otros, disminuía.

El mejor valor de la distancia correspondió al conjunto de datos de 56 días iniciando desde 0.3 veces el ancho total. Es de resaltar que para todos los datos, exceptuando uno, se obtuvieron valores de distancia dentro de la tolerancia.

Los valores de distancia relativa estuvieron entre 6 y 11 $\mu g/m^3$, siendo principalmente valores cercanos a 6. Al considerar cual sería el mejor conjunto de datos para una tolerancia dada, se encontró que si se busca una buena exactitud, se debe aumentar el inicio de la toma de datos a una donde las mediciones de PM 2.5 sean parecidas o tengan una tendencia similar. Igualmente, el tamaño de la ventana no fue un valor concluyente, pero si se busca reducir la exactitud al mínimo se debería usar el mayor ancho que se pueda. En general, el ancho a usar va a depender de la tendencia de los datos, pues si estos son muy variados se debería intentar un ancho grande para captar esa variabilidad; pero si los datos no varían mucho, entonces se debería usar un ancho pequeño para reducir el tiempo de cómputo. Para una tolerancia especifica mayor que 6 $\mu g/m^3$ se puede observar la tabla 4 para tener una idea del conjunto de datos apropiado.

6.1. Alcance del mejor conjunto mínimo de datos

Igual a como se mencionó en secciones anteriores, se tomaron los datos que no fueron tomados en cuenta en el ajuste para analizar su distancia relativa con respecto a los datos de AMB. De esta forma, se usaron cuatro sub-intervalos para comparar. Cada intervalo tuvo una duración de 4 días.

Los resultados se muestran a continuación:

Cuadro 5: Esta tabla muestra los valores de distancia y distancia relativa para los datos calibrados con el mejor conjunto mínimo de datos que se encontró, Note que en todas las partes del sub-intervalo la distancia relativa estuvo dentro de la tolerancia.

Sub-intervalo	Distancia $[\mu g/m^3]$	Distancia relativa $[\mu g/m^3]$
1	72.626383	7.877437
2	75.859962	8.180190
3	47.983343	5.204524
4	78.167690	8.478476

7. Conclusiones y Recomendaciones

Con base en los resultados obtenidos y el análisis realizado en la sección anterior se llegó a las siguientes conclusiones:

- Primero que cuantificar el error de medición de un sensor y hallar una correspondiente calibración, supone el uso del concepto de métrica en espacios vectoriales, el cual en este contexto toma un sentido distinto al usual de distancia pero igual de relevante por su aplicabilidad.
- Segundo que la ventana y pasos óptimos según lo planteado corresponden a los valores de 1 hora y media hora correspondientemente o 1 hora y 1 hora.
- Tercero, que el valor máximo de distancia entre los datos (según el valor de tolerancia correspondiente al orden de los datos de AMB 10 $\mu g/m^3$) es el de 256.47 $\mu g/m^3$ aproximadamente, y el mínimo valor de distancia relativa fue el de 4.07 $\mu g/m^3$.
- Cuarto, el alcance de la mejor calibración realizada sobre todo el conjunto de datos consistió a todos los sub-intervalos analizados. Es decir, la calibración continuó siendo buena al analizar cada sub-intervalo.
- Finalmente el mínimo conjunto de datos que permiten mantener el valor de tolerancia al extrapolar la calibración, es el de 56 días iniciando desde 0.5 veces el ancho total de los datos. Igualmente, esta predicción fue buena en cada sub-intervalo del intervalo completo posterior al de la calibración, pero fue notablemente buena en solo uno de los sub-intervalos.

8. Referencias

Referencias

[1] Núñez L (2020) Métricas, datos y calibración inteligente URL https://github.com/nunezluis/MisCursos/blob/main/MisMateriales/Asignaciones/TallerDistancias/TallerDistancias.pdf.