Error de Medición y Calibración de Sensores IoT

Juan Diego Duarte, Wilmar Alejandro Sotelo *
Universidad Industrial de Santander
Calle 9 con cra 27

Versión 1 - 12 de septiembre del 2021

Índice

Resumen

Actualmente, la contaminación del aire es un tema de importancia a nivel general, conocer el estado del aire a nuestro alrededor se puede medir y cuantificar, por eso en este artículo se presenta el desarrollo para el ajuste de la curva de las lecturas de material particulado $PM_{2,5}$ realizadas por medidores de baja calidad IoT con respecto a las mediciones entregadas por sensores de alta calidad de las estaciones del Área Metropolitana de Bucaramanga (AMB) ubicados en las instalaciones de la Escuela Normal Superior.
El problema surge debido al desarrollo de dispositivos de medición de bajo costo con menor calidad que hoy en día se producen en mayor cantidad. El objetivo estimar el error de medición y construir un modelo de ajuste lineal de curvas para disminuir (en cuanto sea posible) la diferencia (medida como distancia) entre los datos de alta y baja calidad.

Para trabajar con las mediciones de los sensores IoT, se ordenan los datos generados por los mismos; luego se cuantifica una base de error mediante distancia euclídea con respecto a los sensores AMB y finalmente se construyen dos modelos lineales para ajuste de curvas considerando todos los datos y utilizando una fracción de los datos (80 %).

De esta manera, para el segundo caso se cuantifica la calidad del ajuste lineal aplicándolo al 20 % de la información restante y comparándolo con 20 % final del primer caso. La diferencia entre estos no es considerable; sin embargo, el ajuste lineal no permite una correcta aproximación, dejando un error relativamente grande.

*e-mail: Juandduartec@gmail.com, wilmar2218418@correo.uis.edu.co

1. Introducción

Se dice que existe contaminación del aire cuando contiene gases tóxicos, compuestos orgánicos volátiles y partículas sólidas en suspensión. La Organización Mundial de la Salud (OMS) informa sobre las muertes prematuras ocurridas en el mundo debido a la contaminación del aire y también presenta la relación que existe entre la concentración de los contaminantes y las enfermedades que ocurren como consecuencia directa de ellos [1], esta es una de las razones por las cuales se debe monitorear la pureza del aire; y esto se hace mediante sensores de material particulado. Actualmente existen sensores de bajo costo que forman parte de los dispositivos de la revolución del Internet de las cosas (IoT); estos no son altamente precisos, por lo que deben ser sometidos regularmente a procesos de calibración.

De acuerdo con lo anterior, para el caso de la escuela Normal Superior de Bucaramanga, el problema radica en cuantificar el error de medición de los sensores de bajo costo; para luego realizar el proceso de calibración y obtener finalmente una aproximación más cercana a las medidas proporcionadas por los sensores de las estaciones del Área Metropolitana de Bucaramanga *(AMB)*. Las medidas en cuestión se hacen para material particulado $PM_{2,5}$ de dimensiones $\leq 2,5\,[\mu m]$. Dentro del estado del arte de este tipo de problemas no se hace ningún aporte innovador, sólo se presenta un caso en particular en que se cuantifica un error de medición.

En la Sección 2 se presenta la estrategia utilizada para realizar el tratamiento de datos, cuyos resultados se encuentran en la sección 3. En la sección 4 se presentan las conclusiones del trabajo.

2. Metodología

Los datos para este trabajo fueron proporcionados por el profesor[1]. Se tomaron específicamente las medidas tomadas en el colegio Normal Superior, para el cual existen archivos tanto de sensores de alta calidad (Estaciones AMB) como sensores de bajo costo (IoT). Se cuenta con datos de material particulado $PM_{2,5}$ tomados cada hora desde el 01/10/2018 hasta el 31/08/2019 por los sensores de las estaciones AMB; sin embargo, las medidas tomadas por los sensores de bajo costo no están a la misma hora de las de alta calidad; por lo cual, para el procesamiento de datos, sólo se consideraron los puntos en los que los medidores de baja calidad realizaron mediciones.
Para el tratamiento de datos se utilizó la herramienta **MATLAB**. Debido a que la información proporcionada por los dos tipos de sensores no se encuentra emparejada sino en desorden, la información leída por las estaciones de bajo costo se reordenó como se describe a continuación:

1. Los datos más antiguos se dejaron al principio.
2. Se verificó que a cada día le correspondieran 24 puntos, así:
   - En los días con muchas más de 24 lecturas se promediaron las mediciones para obtener 24 intervalos, correspondientes cada uno, en orden, a una hora del día.
   - En los días con lecturas repetidas se realizó un recorte para despreciar esta información.

[1] https://github.com/nunezluis/MisCursos/tree/main/MisMateriales/Asignaciones/TallerDistancias/DatosDistancias

- Se despreciaron los días con información faltante.

3. Los días fueron reordenados internamente, ya que al principio de cada uno de ellos existe una hora diferente o muy lejana a la hora cero. Nota: la hora cero se refiere al intervalo comprendido entre las 00:00 h y las 00:59 h.

Habiendo realizado este proceso, los datos de los sensores IoT quedan lo más cercanos en tiempo posible a los datos proporcionados por los sensores de alta calidad.

Para el cálculo de la distancia se debe tener en cuenta que, como la información no se encuenta a horas exactas y existen picos producidos por la baja calidad de los medidores, primero se realizan promedios por ventana móvil[2]. Esto nos permite suavizar la curva y trabajar con un resultado aproximado para grupos de un cierto número de horas; en este caso se decidió definir ventanas de 3 horas porque permiten colapsar datos cercanos sin despreciar mucha información.

Habiendo realizado los promedios por ventanas móviles, se procedió a calcular la distancia considerando la definición euclídea (1).

Nota: se debe tener en cuenta que para verificar el cambio en la distancia luego de realizar la calibración, el tamaño de la ventana móvil no debe variar; ya que una modificación de n implicaría una reducción o aumento en D debido a que conforme se incrementa el número de datos que no son iguales entre sí (para un mismo i), se aumenta D.

$$D(\widehat{\mathbb{D}}_{\widehat{i}}, \mathbb{D}_i) = \sqrt{\Sigma_{\widehat{i},i}^{n} (\widehat{\mathbb{D}}_{\widehat{i}} - \mathbb{D}_i)^2} \tag{1}$$

$$n = \text{Número de datos.}$$

Calibración: Una manera de determinar el valor de α para conseguir la distancia mínima consiste en generar un vector de $\alpha = 0{,}5 : paso : 1$; con un *paso* del tamaño de la exactitud que se desee; por inspección de la figura 1, la mayor parte del tiempo los datos tomados por los sensores de bajo costo tienen mayor magnitud que los de referencia, pero no alcanzan a ser del doble de tamaño; así, los límites para esta constante proporcional se fijaron entre 0.5 y 1.

Dicho vector de α se evaluó en la ecuación 2, de la que se obtuvo como resultado un vector de distancias D y del que se tomó el mínimo valor, el procedimiento se realizó en primer lugar para el total del grupo de datos. Esto último con el objetivo de considerar todas las variaciones y no sólo un conjunto de mediciones aisladas; es decir, si sólo se considera la segunda mitad del total de los datos, la constante α será mejor para ajustar específicamente esa parte; mientras que cuando se ajusta con todo el historial, se asegura una mejor aproximación en circunstancias futuras.

Finalmente, se realizó una comparación entre los valores de la constante α obtenidos para cuando se dividen los datos en subgrupos; así, mediante la comparación de este factor de ajuste, se puede determinar aproximadamente qué tan apropiada es la utilización de un modelo para cierto número de datos de entrada.

$$D(\alpha\widehat{\mathbb{D}}_{\widehat{i}}, \mathbb{D}_i) = \sqrt{\Sigma_{\widehat{i},i} (\alpha\widehat{\mathbb{D}}_{\widehat{i}} - \mathbb{D}_i)^2} \tag{2}$$

[2] https://en.wikipedia.org/wiki/Moving_average

Ya con el valor de α determinado, se realiza el ajuste por medio de los mínimos cuadrados (LMS) ($|y\rangle = c|x\rangle + |b\rangle$), y así determinar un modelo lineal $f(\xi_j) = \alpha\widehat{f}(\xi_j)$, es decir, $\alpha|y\rangle \simeq f(\xi_j)$. Donde $c = \dfrac{\langle x|y\rangle}{\langle x|x\rangle}$ y $|b\rangle = |y\rangle - c|x\rangle$.

Cálculo del error: Para cuantificar la mejoría hecha por la aproximación lineal, definiremos el error como se presenta en la siguiente ecuación:

$$\%Error = 100 * \frac{D_{corregida}}{D_{inicial}} \tag{3}$$

Donde $D_{inicial}$ corresponde con la distancia entre las muestras de los sensores de alta y baja calidad luego de realizar el promedio por ventana móvil y $D_{corregida}$ se refiere a la distancia obtenida entre el mismo conjunto de datos anterior, escalando la magnitud del vector de las medidas de las estaciones IoT por el factor α.
La determinación del alcance del modelo se realizó midiendo el error propuesto en la ecuación 3 para los siguientes dos escenarios:

1. Cálculo de α con todos los datos trabajados. Aquí, tanto la $D_{inicial}$ como la $D_{corregida}$ tienen en cuenta todas las muestras.

2. Cálculo de α para el primer 80 % de los datos trabajados. En este caso, la $D_{inicial}$ y la $D_{corregida}$ se calcularon para el último 20 % de las muestras.

Para determinar el mínimo conjunto que replica el modelo realizamos lo anteriormente dicho, es decir, tomando conjunto de datos cada vez más pequeños para generar el modelo hasta que el error sea el menor posible.

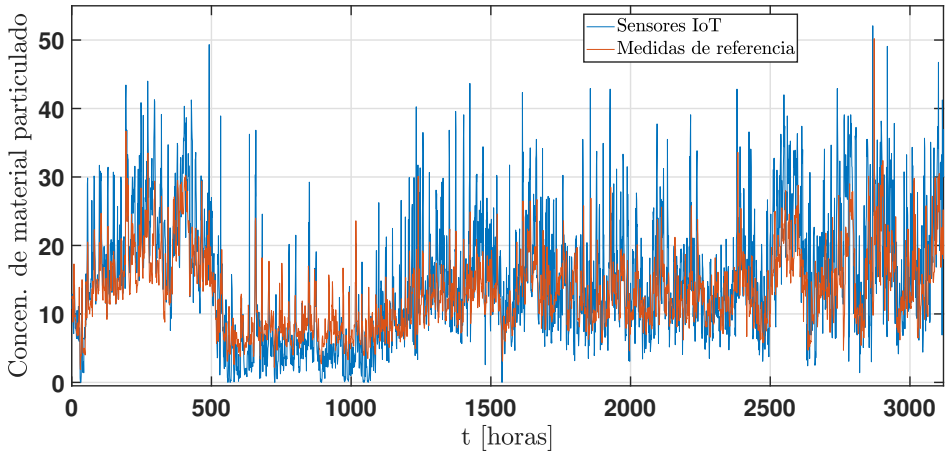

Figura 1: Concentración de material particulado para todas las lecturas

3. Resultados

En el cuadro 1 se presentan los resultados de los cálculos de distancia euclidea, el error asociado y el valor de α para la calibración.

Cuadro 1: Tabla Resultados de Error y Calibracion

		Iterando	LMS
Para todos los datos	$D_{inicial}$	331.4666	331.4666
	$D_{corregida}$	253.2304	253.3847
	α	0.779	0.788
	Error	76.4 %	76.44 %
Para el 80 % inicial	α	0.8180	0.818
Para el 20 % final	$D_{inicial}$	166.7828	215.4965
	$D_{corregida}$	124.7924	161.5898
	Error	74.8 %	74.98 %

Como podemos observar en la tabla 1, se tienen los resultados de distancia entre todos los datos manejados ($D_{inicial}$), y la distancia para las dos calibraciones realizando iteraciones y aplicando el método de mínimos cuadrados (LMS) ($D_{corregida}$); de acuerdo con esto, es claro que ambas formas de aproximar arrojan resultados semejantes pero no muy precisos; es decir, después de la calibración se mantiene un error considerablemente grande (76,4 %), este error sirve para definir el alcance que tiene un modelo para la toma del 80 % de datos y así evaluar el mismo para predecir el 20 % final de la medida de referencia.

De acuerdo con lo anterior y considerando los resultados del cuadro 1, para el modelo del 80 % de datos iniciales, se tiene un error de 1.6 puntos porcentuales menos que para el modelo con todos los datos; esto nos indica que el modelo que se realiza con una fracción de la información total, entrega resultados cercanos a los esperados; el cambio en las aproximaciones se debe a la variación del comportamiento de los datos; si nos fijamos en la ventana de tiempo de 500 a 1000 horas de la fig.1 esta no sigue la tendencia del resto de información; por lo que, si utilizáramos sólo esta ventana para construir el modelo y luego verificarlo con los datos restantes, tendríamos una distancia y error mayores.
El alcance de predicción del modelo se puede cuantificar con el error mostrado en el cuadro 1, el cual no permite una mayor disminución de las distancias debido a su componente lineal.

Por otro lado, el valor de α para la calibración por LMS surge de su comportamiendo al calcular todas las distancias entre los datos totales, esto se muestra en la Fig.2; el intervalo para obtener una menor distancia entre las lecturas esta entre $[3/4, 4/5]$. Además, el valor de α para evaluar su alcance en las predicciones se obtiene de la misma manera pero para el 80 % de los datos de los promedios móviles.

Para visualizar qué tanto se acercan las aproximaciones a los datos de referencia se tiene la Fig.3, esta ventana de tiempo corresponde con distancias promedio para el intervalo entre 3/oct/18 hasta 8/oct/18 y muestra el comportamiendo de las lecturas de los sensores y la aproximación realizada; de acuerdo con los resultados de la tabla 1 para el total de los datos, se observa que la

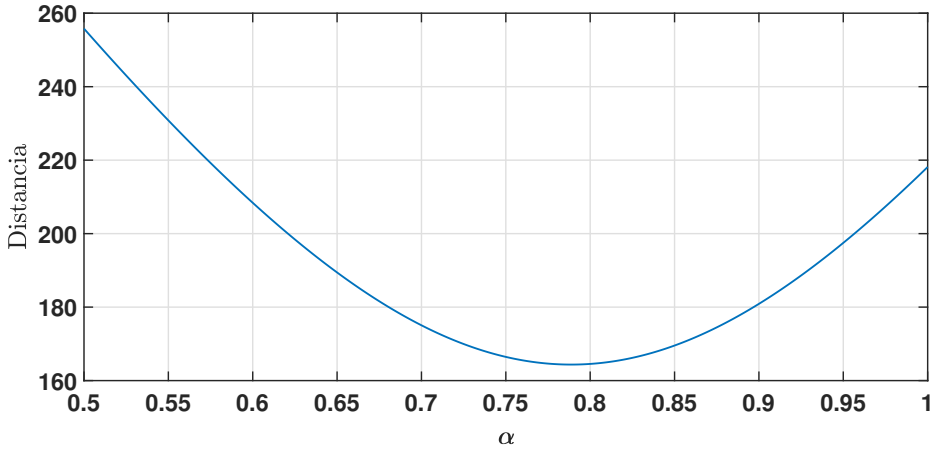

Figura 2: Distancia entre las dos medidas para diferentes α

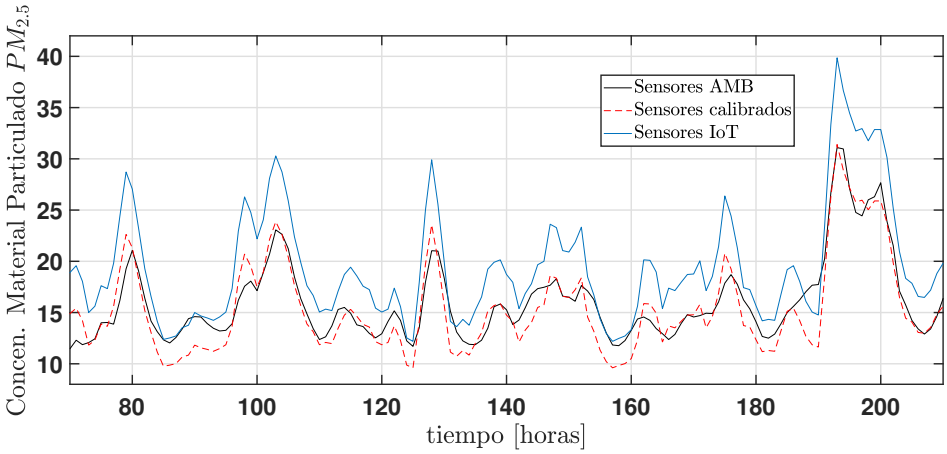

Figura 3: Comparación antes y después de la calibración

aproximación no es muy funcional. Nota: se debe mencionar que esta ventana es mostrada para ilustrar el comportamiendo en detalle de la aproximación en un intervalo de tiempo en particular.

En la Fig.4 se muestra, para el rango final del 20 %, tanto los datos de referencia como los entregados por los sensores de baja calidad luego de haber pasado por el proceso de ventana móvil; la predicción de la aproximación aparece como "sensores calibrados".
Debido que estas lecturas tienen un comportamiento semejante a los datos con los que se genera el modelo, los resultados son similares; esto se puede corroborar con la tabla 1, el error con base en la distancia no varía drásticamente (1.6 %).

El menor conjunto de datos que puede reproducir el modelo es precisamente la ventana de datos

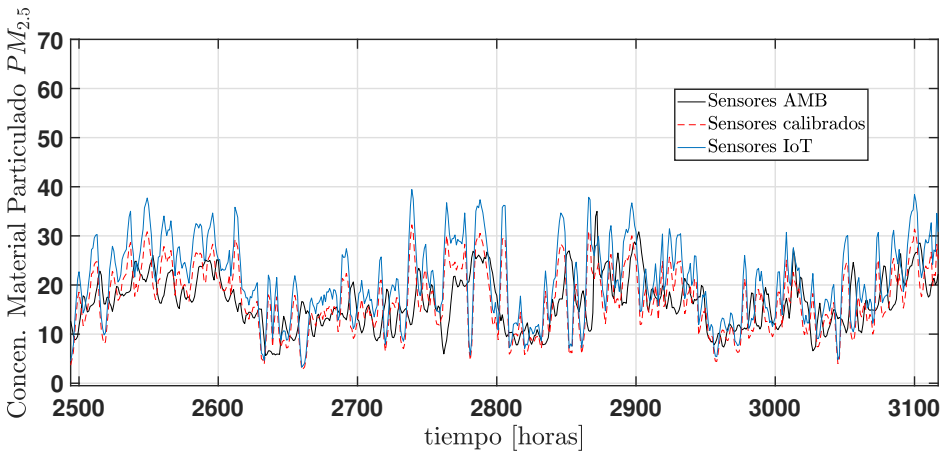

Figura 4: Predicción de la calibración

de la Fig.3 ya que posee un error del 41.798 %. Su máxima tolerancia respecto al error total de acuerdo con la tabla 1 es de 58.2 %.

4. Conclusiones y Recomendaciones

La calibración lineal se puede considerar como una opción para cuando se tenga mucha información a calibrar con el fin de reducir costo computacional; se pueden establecer intervalos típicos de comportamiento para determinar qué modelo utilizar (esto puede hacerse para varios años en donde los datos varíen con las estaciones).
Para esta aplicación, específicamente, el modelo lineal no realiza un correcto ajuste de datos; de allí un error tan grande (74 %) como el que se obtuvo al trabajar con toda la información. Adicionalmente, para la predicción de datos de los sensores, no se tienen en cuenta factores adicionales, como lo son los efectos asociados a la velocidad del viento, condiciones de humedad y temperatura; los cuales pueden influir en gran medida.
Dependiendo de la tendencia del comportamiento de las medidas de los sensores, se pueden hacer aproximaciones del modelo con más o menos ventanas; es decir, si los datos mantienen su comportamiento uniformemente, se puede trabajar con menos del 80 % de la información y obtener un modelo con una calidad de aproximación similar.

Para trabajos futuros se propone la implementación del ajuste mediante un modelo no lineal y el uso de datos para más de un año (esto con el fin de considerar el comportamiento dependiendo del cambio de clima).

5. Referencias

Referencias

[1] Esteban D. Volentini, Carlos Albaca Paraván, José Younes, Sergio D. Saade, Luis A. Tek, and María de los A. Gómez López. Descripción de un sistema iot para la medición y registro de la calidad del aire. In *2020 IEEE Congreso Bienal de Argentina (ARGENCON)*, pages 1–7, 2020.