Taller 1: Métricas, datos y calibración inteligente

Yesid Alfonso Gutiérrez & Michael Andrés Tapias[*]
Universidad Industrial de Santander
Carrera 27 - calle 9na

11/09/2020

Índice

1. Introducción **2**

2. Metodología **3**
2.1. Error de la cuantificación del material particulado 3
2.2. Reducción de ruido de la serie temporal 4
2.3. Función $\hat{f}$ de calibración 4

3. El experimento y los resultados **6**
3.1. Conjunto de datos 6
3.2. Preprocesamiento de datos 6
3.3. Resultados 7

4. Conclusiones y Recomendaciones **10**

5. Referencias **11**

[*]e-mail: yesid.gutierrez@saber.uis.edu.co, michael2208457@correo.uis.edu.co

Resumen

Una de las mayores preocupaciones que se puede observar en la actualidad son los componentes que generan contaminación en el aire, dentro de estos componentes se encuentra uno que ha sido investigado a nivel mundial con bastante interés, este es el material particulado PM2.5, el cual posee un tamaño inferior a 2.5 micras y genera un impacto negativo en el sistema cardiorespiratorio, esto ocurre debido al diminuto tamaño de las partículas, las cuales pueden acceder a las vías respiratorias y permanecen en el sistema respiratorio, generando un aumento en el riesgo de enfermedades cardiorespiratorias. Generalmente, para cuantificar la presencia del material particulado se suelen utilizar sensores certificados, los cuales vienen previamente calibrados desde fábrica. Sin embargo, el costo elevado de estos dispositivos representa una limitación para el despliegue de estos dispositivos en producción. Por otra parte, existe una familia de sensores de bajo costo pertenecientes al Internet de las cosas (IoT, por sus siglas en inglés), los cuales carecen de una calibración de fábrica. En este trabajo, nosotros proponemos una estrategia inteligente para calibrar dispositivos IoT de captura de material particulado PM2.5 , a partir de un conjunto de muestras de un sensor IoT y un sensor calibrado de referencia utilizando métodos tradicionales de médias moviles y regresión lineal. [1]

1. Introducción

En la actualidad la presencia de partículas que generan contaminación en el aire está relacionada con diferentes problemas de morbilidad y mortalidad [1]. Dentro de estos contaminantes es de especial interés el material particulado, el cual es una mezcla de partículas líquidas y sólidas, de sustancias orgánicas e inorgánicas, que se encuentran en suspensión en el aire. Su composición es muy variada y se puede encontrar, entre sus principales componentes: sulfatos, nitratos, el amoníaco, el cloruro sódico, el carbón, el polvo de minerales, hollín, cenizas metálicas y agua [2], habitualmente estas partículas pueden provenir de los automóviles, camiones, fábricas, quema de madera y otras actividades[3].

El material particulado presente en la atmósfera, se clasifica en fracciones que tiene que ver con el tamaño de cada una de las partículas que lo constituyen [4]. Normalmente, son tres los grupos de clasificación más habituales. Material con tamaño de partícula superior a 10 micras, material con tamaño inferior o igual a 10 micras conocido como *PM10*, y material con tamaño de partícula inferior a 2.5 micras conocido como *PM2.5*. Éstas últimas son las que más relevancia poseen en la contaminación urbana, ya que pueden penetrar profundamente en los pulmones y poseen riesgos potenciales significativos para la salud.
Las partículas de tamaño comprendido entre las 2.5 y las 10 micras, no son realmente inhaladas hasta las vías profundas y se expulsan de manera relativamente eficaz a través de las mucosidades o de la tos, o sedimentan directamente sin llegar a penetrar en el árbol respiratorio. De hecho, debido a los efectos adversos que las partículas finas pueden infligir a un gran número de personas y ambiente [5], en la actualidad el monitoreo y medición de las partículas PM2.5 es uno de los mayores retos que afrontan las autoridades sanitarias de todos los países del mundo por tres motivos principales: primero, este tipo de partículas resultan más nocivas para la salud humana que cualquier otro

[1] Código y experimentación disponible aquí: https://github.com/yesid08/MAvanzadas-Taller1

contaminante, ya que la exposición crónica aumenta el riesgo de enfermedades cardiovasculares, respiratorias y cáncer de pulmón[6]. Segundo, sólo se necesita exposición con el ambiente para que el material particulado afecte la salud de las personas, y finalmente, resulta prácticamente imposible evitar la interacción con este tipo de partículas que afecta tanto personas de zonas rurales como urbanas, en países desarrollados y en vía de desarrollo[7]. Por esta razón, es de vital interés capturar datos del material particulado en tiempo real a partir de una muestra de aire, donde por medio modelos matemáticos se cuantifica la relación y concentración de estas emisiones, permitiendo tener una visión de comportamiento y estrategias de disminución de las mismas. Todos los equipos de medición de partículas trabajan según las normativas vigentes y se envían calibrados de fábrica (certificado de calibración ISO opcional) [8], pero, debido a su alto costo se convierte en una metodología poco aplicable en las diferentes entidades o espacios requeridos. Por esto surge como alternativa de solución los sensores de bajo costo que forma parte de la revolución generada por el Internet of Things (IoT, por sus siglas en inglés).
En este trabajo, nosotros proponemos una estrategia para calibrar dispositivos de captura del material particulado de bajo costo utilizando medias móviles. Adicionalmente, en este trabajo proponemos un marco de trabajo para combinar métodos de regresiones lineales clásicos (como los mínimos cuadrados), para realizar una calibración de dichos dispositivos de bajo costo respecto al material particulado *PM2.5*.

2. Metodología

En este trabajo, nosotros proponemos una estrategia para calibrar un sensor IoT que cuantifica el material particulado PM2.5 con respecto a un dispositivo certificado. Es decir, buscamos encontrar una función $\hat{f}$, tal que $P \approx \hat{f}(\hat{P})$, donde P corresponde a una medición de material particulado en el dispositivo de captura certificado y $\hat{P}$ corresponde a una medición del dispositivo IoT. Tal como se muestra en la figura xxx, nuestro marco de trabajo consiste en realizar una captura de datos para la calibración en ambos dispositivos. Luego, utilizamos una media móvil con diferentes tamaños de ventana, para reducir las fluctuaciones bruscas que se producen en el tiempo durante la toma de las muestras. Finalmente, estimamos una función $\hat{f}$ minimizando la diferencia de la medición ξ entre las muestras de material particulado en ambos dispositivos. De esta manera, $\hat{f}(\hat{P})$ será nuestra función de calibración para el dispositivo IoT, y así podremos cuantificar el material particulado PM2.5 presente en el ambiente para futuras mediciones.

2.1. Error de la cuantificación del material particulado

Generalmente, los dispositivos IoT, no cuentan con el estándar ISO de calibración para la captura de datos de material particulado. En este trabajo, nosotros buscamos que dichos dispositivos, puedan capturar datos del material particulado con una medición similar a la de un equipo de captura certificado. Por esta razón, para una medición P_i de material particulado utilizando un equipo de captura certificado y una medición $\hat{P}_i$ de material particulado utilizando un equipo IoT. Definiremos al error de medición entre el equipo certificado y el equipo IoT como la distancia euclidiana $d(P_i, \hat{P}_i) = \xi = \sqrt{\sum_i (P_i - \hat{P}_i)^2}$. Esta distancia, nos permite cuantificar la variación

existente entre un equipo de medición certificado y un equipo de bajo coste IoT.

2.2. Reducción de ruido de la serie temporal

En este trabajo, para reducir el ruido generado debido a la existencia de fluctuaciones obtenidas de los sensores de bajo costo soportados por tecnologías IoT en la medición del material particulado PM2.5, con respecto a los datos de referencia brindados por el sensor certificado de AMB, proyectamos como una solución viable a la media móvil. Cabe resaltar que estas fluctuaciones pueden ser ocasiondas por múltiples eventos, dentro de los que se destacan: la ubicación del sensor, humo generado por las horas pico en el transito de vehículos o hasta el hollín generado por algún tipo de quema. Con base en esta premisa, utilizamos la media móvil debido a que esta nos permite suavizar y disminuir las fluctuaciones existentes en la medición de material particulado en ambiente. Dicha media movil, se puede definir por su expresión matemática como:

$$\bar{P}_{SM} = \frac{P_M + P_{M-1} + \cdots + P_{M-(n-1)}}{n} = \frac{1}{n}\sum_{i=0}^{n-1} P_{M-i}$$

Donde $\bar{P}_{SM}$ es el valor resultante de la ventana para el material particulado, $P_M, P_{M-1}, \cdots, P_{M-(n-1)}$ son las mediciones de material particulado que componen la ventana, y n es la cantidad de valores de material particulado P_{M_i} que componen el subconjunto de muestras que pertenecen a la venta móvil. Finalmente, esta ventana se desplaza sobre el conjunto de mediciones de material particulado, suavizando las fluctuaciones marcadas debido al ruido presente en nuestra fuente de datos.

2.3. Función $\hat{f}$ de calibración

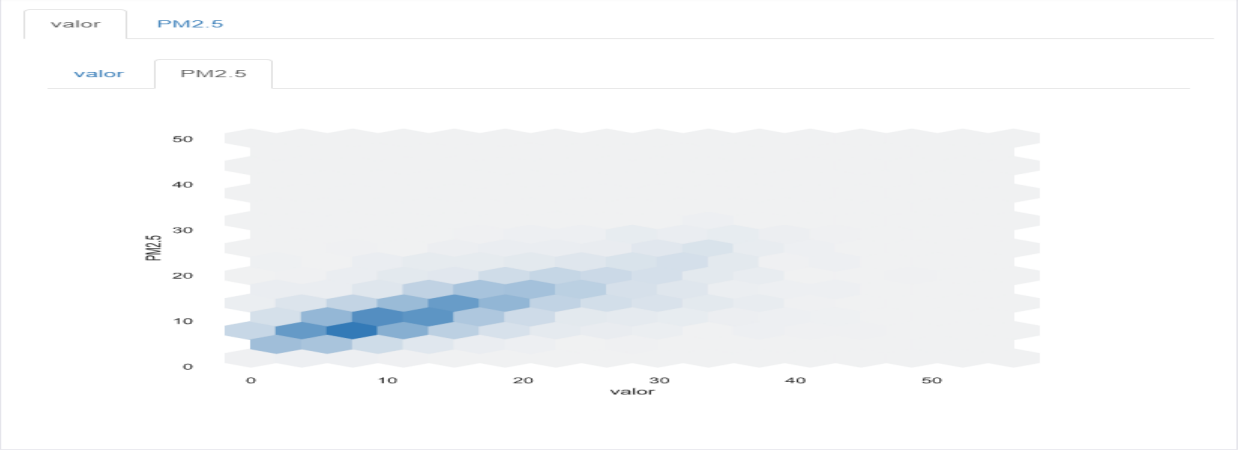

Figura 1: Relación lineal entre el equipo de captura AMB (PM2.5) certificado y el dispositivo IoT (valor)

Además de la media móvil anteriormente descrita, nosotros observamos a través de un análisis estadístico, que el comportamiento entre las mediciones del material particulado PM2.5 brindado por el sensor AMB y el sensor IoT, poseen un comportamiento lineal positivo (ver figura 2), en donde la gráfica del coeficiente de correlación de Pearson nos muestra que el color azul es lineal pero con una pendiente positiva. Por otra parte, el color rojo de la barra lateral indica que es lineal pero con una pendiente negativa. Por otra parte, al representar gráficamente las mediciones del sensor AMB de referencia calibrado y el sensor IoT de bajo coste, se puede apreciar un comportamiento lineal (ver figura 1).

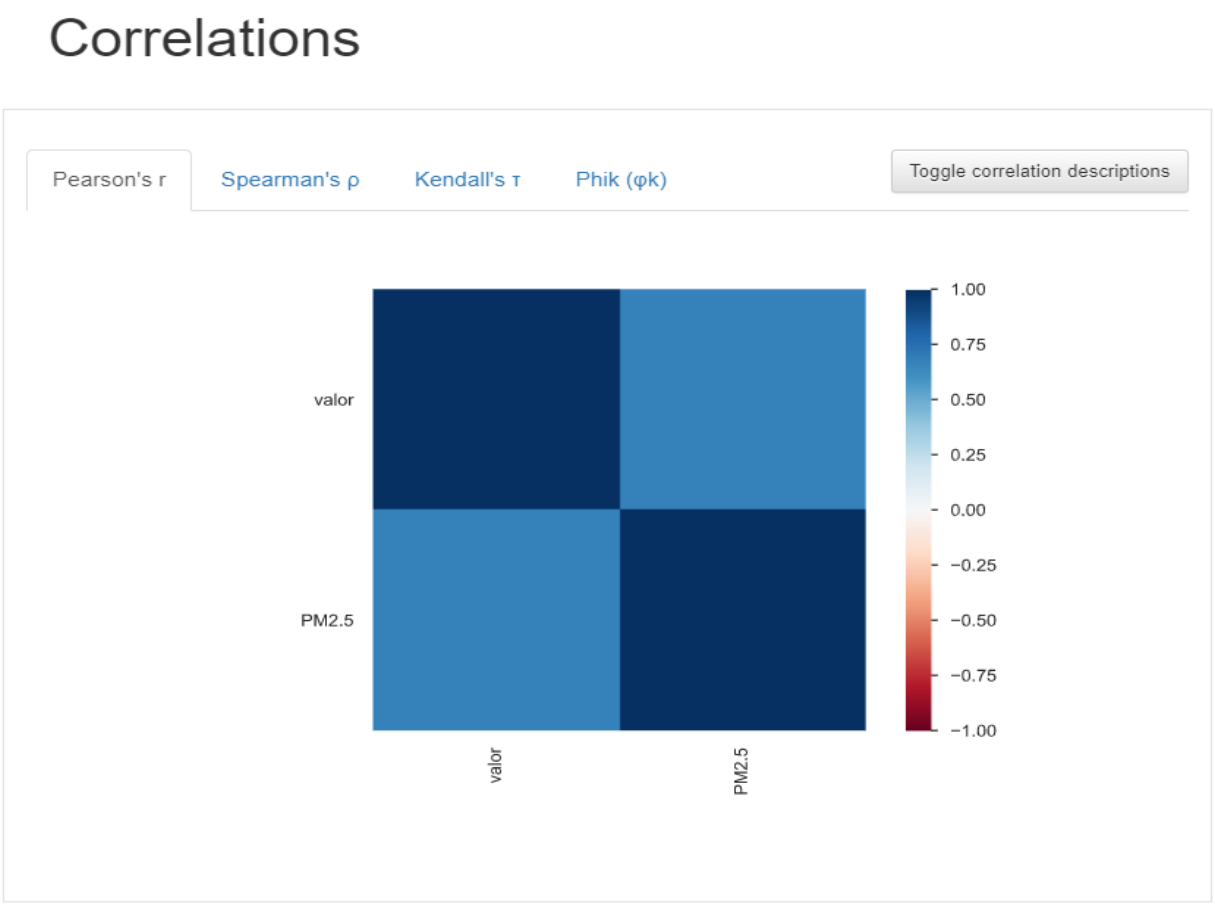

Figura 2: Coeficiente de correlación de Pearson entre el sensor de IoT

Teniendo en cuenta la relación lineal anteriormente descrita, nosotros proponemos estimar la función $\hat{f}$ como una función de regresión lineal, la cual puede calcularse utilizando el método de mínimos cuadrados, el cual a partir de un conjunto de medidas de material particulado PM2.5 de un sensor certificado $P : \{P_1, P_2, P_3, \ldots, P_n\}$ y un conjunto de medidas de material particulado de un sensor de IoT $\hat{P} : \{\hat{P}_1, \hat{P}_2, \hat{P}_3, \ldots, \hat{P}_n\}$, aprende una función $\hat{f}(\hat{P}) = \beta\hat{P} + b \approx P$ donde P

es una lectura sensorial del dispositivo certificado de la AMB, $\hat{P}$ es una lectura sensorial IoT del PM2.5 y donde β y b son las constantes que representan la pendiente de la recta y el intersecto respectivamente. Por otra parte, dado que la relación existente entre las lecturas del material particulado de ambos sensores es de tipo lineal, en este trabajo no utilizamos ninguna estrategia de machine learning para aprender la función $\hat{f}$, ni tampoco barajamos la posibilidad de utilizar estrategias de deep learning, ya que estas requieren un gran número de datos para su entrenamiento, son computacionalmente costosas y además, los resultados de este tipo de estrategias son difíciles de analizar e interpretar.

3. El experimento y los resultados

3.1. Conjunto de datos

En este trabajo, se utilizó un conjunto de datos perteneciente al Acueducto Metropolitano de Bucaramanga (AMB), el cual cuenta con un total de 8040 muestras que corresponden a registros de material particulado PM10, PM2.5, temperatura del aire, lluvia, humedad relativa, hora, fecha, entre otras características que fueron capturadas utilizando un equipo con estándar ISO previamente calibrado desde fábrica. Adicionalmente, se provee un conjunto de datos correspondiente a un sensor IoT de bajo coste el cual carece de calibración. Dicho conjunto de datos posee un total de 5009 muestras de material particulado PM2.5.

3.2. Preprocesamiento de datos

Los sensores de bajo costo soportados por tecnologías IoT normalmente ofrecen una cuantificación de material particulado a un coste bajo. Sin embargo, estos dispositivos generan una diferencia significativa con respecto a las mediciones de un equipo certificado. Adicionalmente, en un ambiente real de captura de datos existen diferentes limitaciones que afectan a los dispositivos IoT, tales como: la batería de los sensores, fallos de red, entre otros; los cuales pueden ocasionar una interrupción en la captura del flujo de datos en tiempo real, resultando en la pérdida de algunas muestras del conjunto de datos utilizado en este trabajo. Por esta razón, nosotros mitigamos la interrupción de datos sustituyendo las muestras corrompidas (NaN) por μ_p y $\hat{\mu_p}$, los cuales hacen referencia a la media del material particulado PM2.5 del sensor de la AMB de referencia, y la media de las lecturas del dispositivo IoT de bajo coste.

Por otra parte, debido a que el dispositivo IoT realiza la captura de datos durante diferentes periodos de tiempo, nosotros estandarizamos los datos en un término de frecuencias por horas. De esta manera, dado un intervalo de tiempo $T_a = [a, b)$, la medición de material particulado PM2.5 en el periodo T_a corresponde a $P_{T_a} = \frac{\sum_i p_i}{n_{T_a}} \forall p_i \in T_a$, siendo n_{T_a} el número total de mediciones de material particulado en el intervalo de tiempo T_a dado. Finalmente, debido a que los conjuntos de datos del AMB y del sensor IoT están desfasados en el tiempo (ya que el conjunto de datos AMB posee algunas muestras que no se encuentran en el dispositivo IoT y viceversa), nosotros calculamos la intersección de los dos conjuntos de datos con respecto al tiempo. De esta manera, para cualquier medición de material particulado P del AMB definida en un tiempo t, también existirá una medición

de material particulado $\hat{P}$ para el dispositivo de IoT, resultando en un total de 3797 mediciones de material particulado PM2.5.

3.3. Resultados

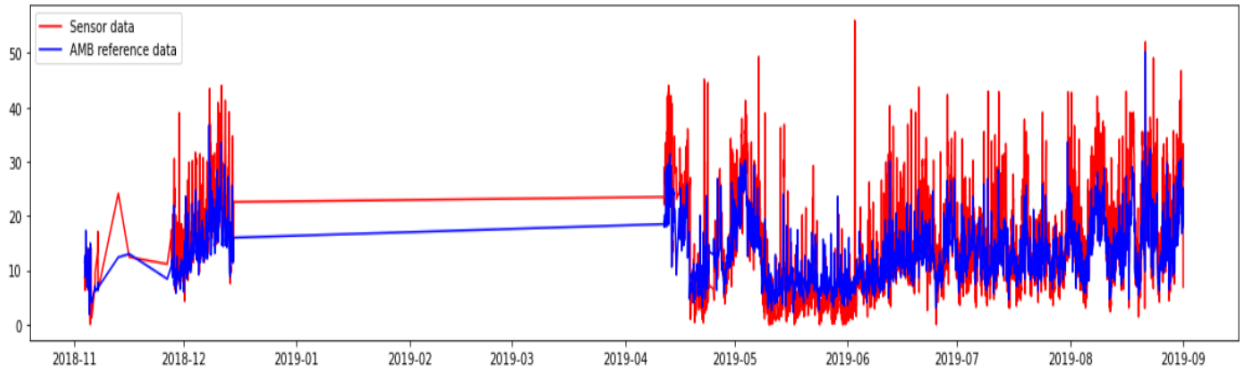

Figura 3: Serie temporal de la cuantificación de material particulado, la serie temporal de color azul corresponde al dispositivo de captura certificado de la AMB y la serie temporal de color rojo corresponde al dispositivo de captura IoT.

Después de preprocesar nuestro conjunto de datos, procedimos a representar gráficamente nuestra serie temporal de mediciones de material particulado PM2.5 en ambos dispositivos. Tal como se muestra en la gráfica 3, a pesar de que ambos dispositivos miden niveles de material particulado similares en el ambiente, estas muestras poseen fluctuaciones bastante marcadas producto del ruido que se produce en el entorno de captura de los datos, tales como el alto tráfico vehicular o la quema de productos inorgánicos en zonas industriales. Por esta razón, utilizamos una técnica de medias móviles para reducir la variabilidad de las mediciones. Debido a que en este trabajo buscamos encontrar la mejor configuración para reducir el error de medición ξ entre los dispositivos de captura, nosotros calculamos el error ξ utilizando las primeras 10 potencias de 2 como tamaños de ventana. En la figura 4 se pueden apreciar algunas de las variaciones de las series temporales de ambos dispositivos de medición del material particulado en función del tamaño de ventana de la media movil.

Por otra parte, debido a que existe una relación lineal entre las mediciones de los dispositivos (ver figuras 2 y 1), nosotros estimamos la función $\hat{f}$ como la regresión lineal entre las mediciones de los dispositivos. En la figura 5, se puede observar la serie temporal de mediciones de material particulado en el dispositivo certificado de referencia y un dispositivo IoT, el cual fue calibrado artificialmente de manera inteligente utilizando la función $\hat{f}$.

Tal como se muestra en la figura, la función $\hat{f}$ logró reducir la diferencia en la medición del material particulado entre los dispostivos con resultados interesantes. Sin embargo, las series originales poseen fluctuaciones marcadas, las cuales introducen ruido a la estimación de dicha función $\hat{f}$. Por esta razón, decidimos aplicar una estrategia de medias móviles utilizando las primeras diez

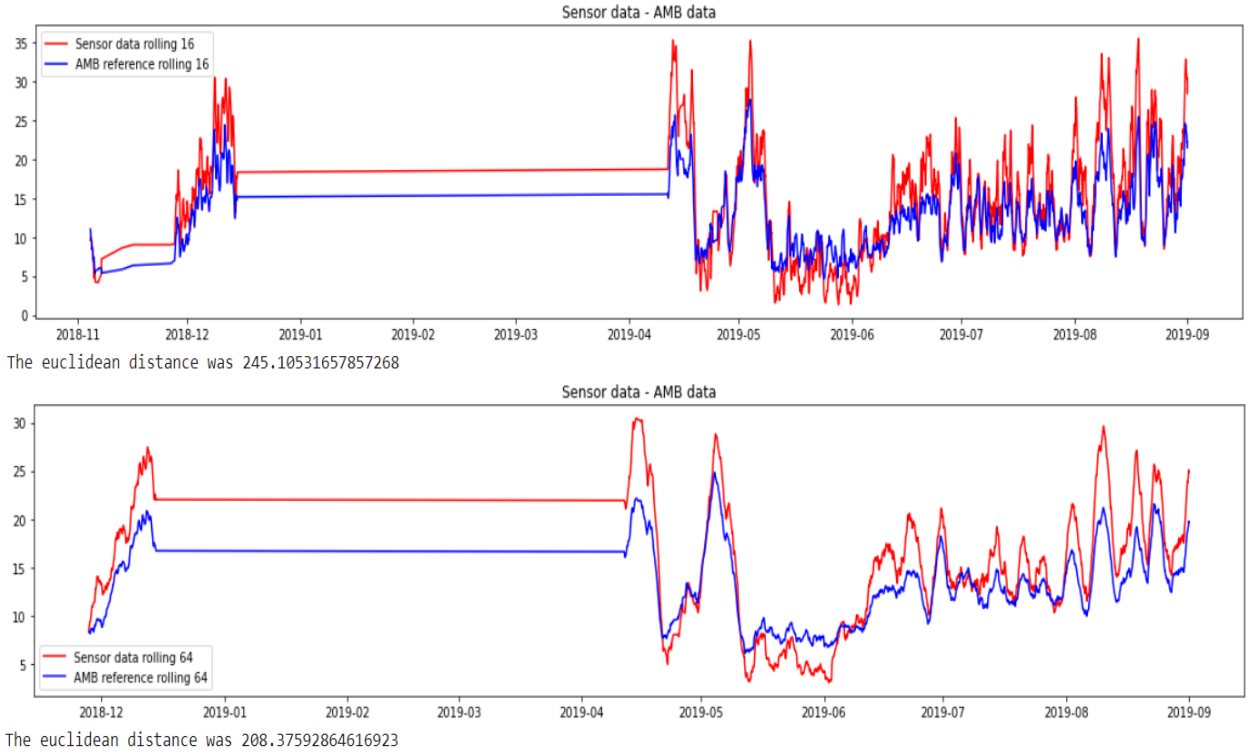

Figura 4: Comparativa de series temporales de la cuantificación de material particulado de un dispositivo de captura certificado y un dispositivo de captura IoT, en función de diferentes tamaños de ventana: la gráfica superior corresponde a un tamaño de ventana de 16 muestras y la gráfica inferior, representa las series temporales de ambos dispositivos utilizando un tamaño de ventana de 64 muestras.

potencias de dos como ventana para reducir las fluctuaciones fuertes en las series temporales. Posteriormente, estimamos la función $\hat{f}$ sobre las mediciones suavizadas de material particulado PM2.5 de ambos dispositivos, tal como se muestra en la figura 6.

Adicionalmente, decidimos calcular la diferencia o error de medición ξ, para cada uno de los experimentos anteriormente mencionados, con la finalidad de encontrar la mejor configuración para minimizar el error de medición ξ entre los dispositivos. Tal como se ilustra en la figura 7, A medida que se aumenta el tamaño de la ventana de las medias móviles, se reduce el error de las mediciones, esto puede deberse a que la media móvil suaviza las series temporales, reduciendo las fluctuaciones de las series temporales originales y por ende, la variabilidad entre las mediciones de ambos dispositivos. Finalmente, la figura 7, muestra que cuando se combina esta técnica de medias móviles con la función de calibración $\hat{f}$, el error de medición entre dispositivos se reduce considerablemente.

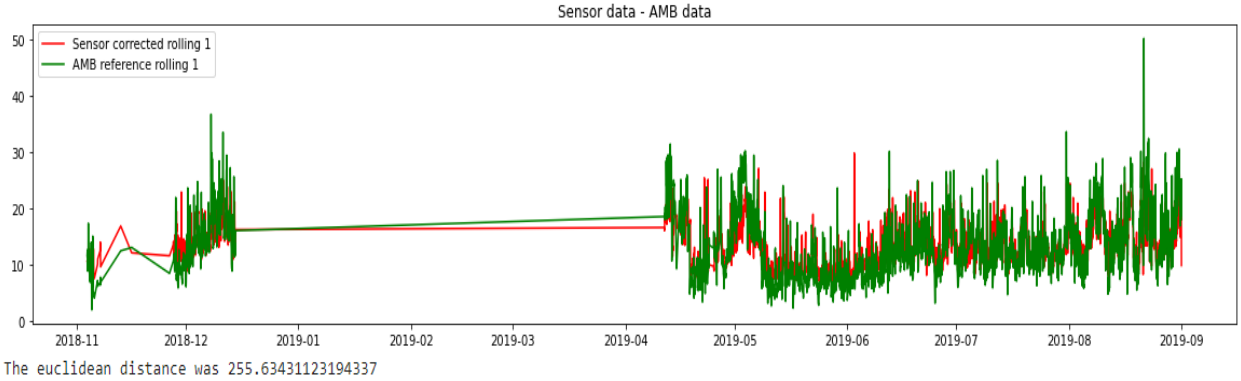

Figura 5: Series temporales de la cuantifiación de material particulado utilizando un sensor certificado y un sensor IoT calibrado artificialmente. La serie verde corresponde a las mediciones realizadas por el dispositivo AMB certificado y la serie de color rojo corresponde a las mediciones realizadas por el dispositivo IoT calibrado artificialmente.

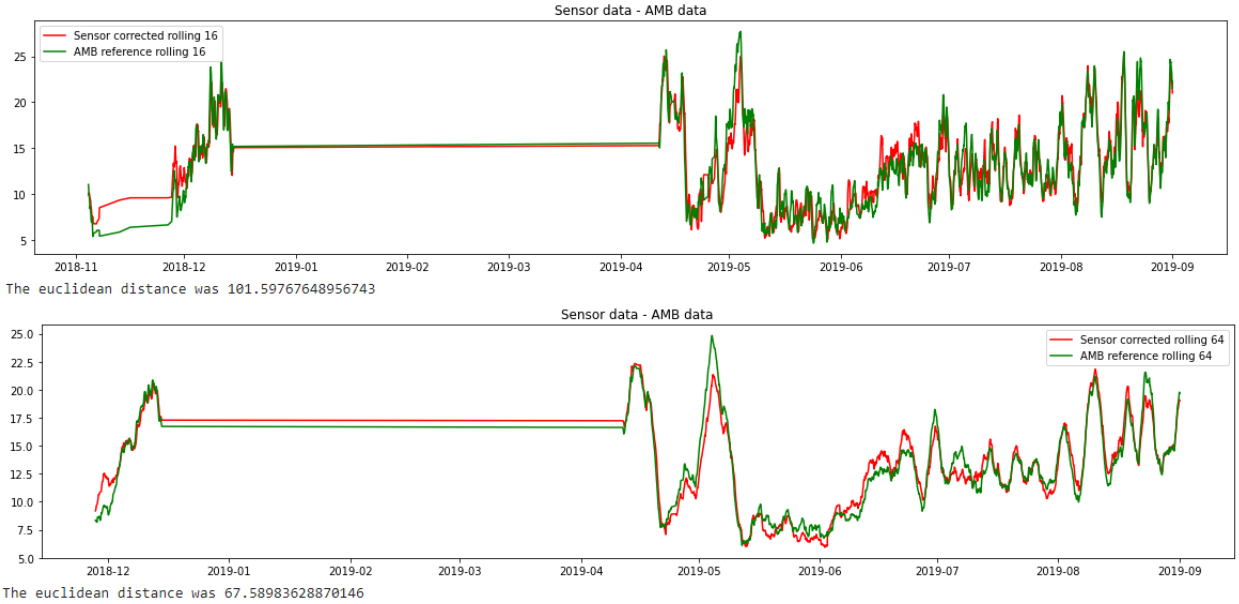

Figura 6: Comparativa de las series temporales de un dispositivo de captura certificado y un dispositivo IoT calibrado artificialmente, en función de diferentes tamaños de ventana. La gráfica superior ilustra las series temporales para un tamaño de ventana de 16 elementos y la gráfica inferior, muestra a las series temporales en función de un tamaño de ventana de 64 elementos.

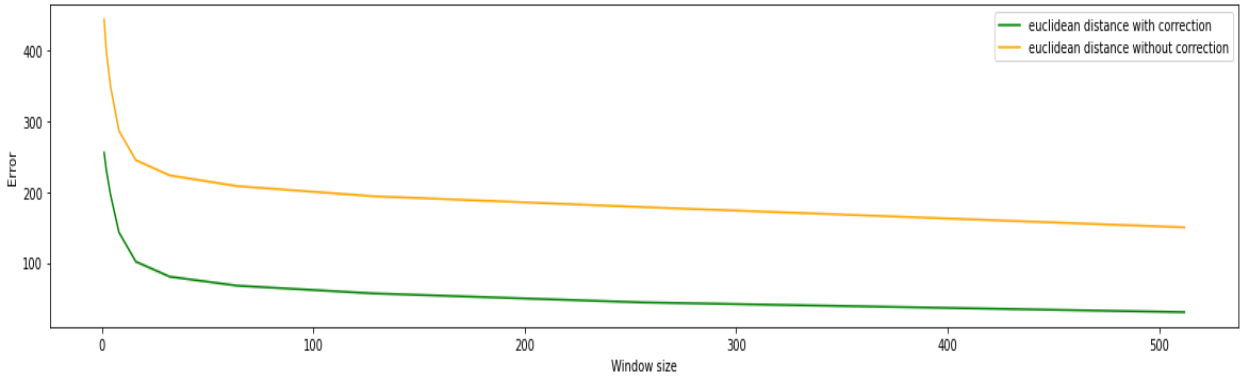

Figura 7: Comportamiento del error de la cuantificación del material particulado en ambos dispositivos, en función del número de muestras que conforman la ventana móvil. La gráfica naranja representa la diferencia de medición entre dispositivos sin utilizar una corrección, mientras que la gráfica verde ilustra la diferencia de medición entre los dispositivos, utilizando una corrección artificial en el dispositivo IoT.

4. Conclusiones y Recomendaciones

En este trabajo, se presentó una estrategia de calibración para dispositivos IoT, los cuales pueden medir el material particulado PM2.5 presente en el ambiente a un bajo coste. Los resultados sugieren que utilizar medias móviles entre las lecturas de un sensor certificado y un sensor IoT, ayuda a reducir el error o diferencia de la medición del material particulado. Por otra parte, los resultados sugieren que si combinamos esta estrategia con una regresión lineal entre las lecturas de dichos dispositivos, obtendremos una mejora considerable respecto a las mediciones del material particulado PM2.5 del dispositivo IoT. Finalmente, la estrategia presentada muestra resultados muy prometedores para realizar una calibración artificial e inteligente de dispositivos basados en tecnologías IoT, esta estrategia no es computacionalmente costosa y podría ser utilizada por diferentes industrias o entes gubernamentales, para cuantificar el material particulado a un coste más bajo sin pérdidas en la precisión de las lecturas de los dispositivos. Para el trabajo futuro, implementaremos esta estrategia en diferentes conjuntos de datos para validar estadísticamente los resultados del trabajo y utilizaremos un modelo de tipo ensemble de diversos modelos de aprendizaje de máquina.

5. Referencias

Referencias

[1] Environmental Protection Agency (EPA). Particulate matter (pm) pollution, 2015.

[2] Organización Mundial de la Salud. Calidad del aire y salud, 2018.

[3] California office of environmental health hazzard assessment (OEHHA). Air quality: Pm2.5, 2020.

[4] Ayuntamiento de Valladolid. Material particulado pm10/pm2,5, 2020.

[5] Bliss Air. What is pm2.5 and why you should care, 2019.

[6] Instituto para la salud Geoambiental. Material particulado, 2013.

[7] Siber Ventilación. Partículas pm2.5, ¿las más contaminantes del aire?, 2018.

[8] PCE Instruments Colombia. Medidor de particulas, 2019.