Métricas, datos y calibración inteligente

Juan Sebastian Carrillo Rodríguez[*]
Jonathan Stiven Gómez Zuluaga[**]
Universidad Industrial de Santander
Bucaramanga

8 de septiembre de 2021

Índice

Resumen

En la actualidad vivimos en un mundo donde las diferentes tecnologías de sensores hacen parte fundamental de la cotidianidad de la mayoría de las personas, incluso aun sin ser notorio en algunos casos. Los diferentes sensores pueden variar en calidad y precio, así mismo con diferencias en la precisión de sus medidas. Estos sensores de bajo costo no son tan confiables como unos de mayor calidad, permiten obtener mediciones de forma mas económica a costa de una menor precisión en sus resultados. Este problema puede ser abordado de forma computacional para mejorar los resultados por medio de un ajuste realizado previo a la recolección de los datos a modo de calibración, esto permite mejorar la calidad de los datos recogidos posteriormente. Con el fin de sacar un mayor provecho a los sensores mencionados anteriormente y así simular comportamientos de sensores de mayor calidad.

Para esto se llevan a cabo diferentes métodos de ajuste como una aproximación a una curva polinomial o métodos de inteligencia artificial para reconocimiento de patrones, donde de los datos del sensor son comparados con un conjunto de datos de referencia para realizar una calibración adecuada. Utilizando estos métodos se consigue mejorar la precisión de los sensores en 4 %.

[*]e-mail: juan.carrillo3@correo.uis.edu.co
[**]e-mail: jonathan2218424@correo.uis.edu.co

1. Introducción

El auge de las comunicaciones IOT (Internet of Things) en los últimos años ha introducido en el mercado diversos sensores de bajo costo que buscan apoyar el muestreo de múltiples variables físicas y permitir la realización de desarrollos e investigaciones basadas en los datos que se pueden recoger con estos[1]. Estos datos pueden muchas veces presentar una gran diferencia de resultados obtenidos con sensores y dispositivos especializados, es en ese punto donde toma gran importancia el papel de la calibración con un patrón de referencia que permita subsanar las debilidades de estos sensores, con esto no se quiere decir que un sensor de bajo costo sea malo o sus datos no permitan realizar investigación o análisis sobre una variable. Lo que se quiere recalcar es que calibrando estos sensores se puede sacar máximo provecho de su bajo costo para obtener datos fiables y con un grado alto de precisión que sin dudarlo llevara a mejores resultados de los distintos proyectos realizados[2].

En este caso particular se mostrará un procedimiento de calibración para sensores que miden la concentración de material particulado de 2.5 micrómetros o menos de diámetro (PM2.5) partiendo de un patrón o referencia basado en una estación meteorológica del área metropolitana de Bucaramanga (AMB) ubicada en el colegio normal superior sede C. Para cumplir con el objetivo primero se lleva a cabo un proceso de homogeneización entre los datos de referencia y los captados por el sensor de bajo costo ya que estos han sido muestreados en cantidades y frecuencias diferentes, esto se realiza con el método de la ventana dinámica, este método promedia valores en un determinado espacio de tiempo para cada conjunto de datos, seguidamente estos datos homogéneos son utilizados para aplicar métodos de ajuste que permiten mejorar las mediciones obtenidas por los sensores de bajo costo. En este articulo se hace una comparación entre el método de ajuste de mínimos cuadrados y una red neuronal profunda entrenada para realizar regresión lineal[3],[4]. A continuación se presentan de forma mas detallada los algoritmos utilizados y los resultados obtenidos.

2. Metodología

Los datos obtenidos por medio de la estación meteorológica para la calibración y los datos experimentales facilitados por los sensores IOT fueron analizados en primera instancia, donde se identifican distintas lineas temporales de muestreado o distintos tiempos de toma de muestra, lo cual, no permite una comparación directa entre los conjuntos de datos debido a la inviabilidad de comparar 2 muestras tomadas en distintos intervalos. Adicionalmente, se puede evidenciar la falta de datos experimentales en muchos intervalos de tiempo, algo común y natural debido a diversos factores que modifican o alteran una toma de muestras experimentales. Dados estos problemas expuestos anteriormente se hace necesario el procesamiento de las muestras ejecutando un algoritmo de promediado móvil, esto permite disminuir el efecto de datos erróneos que causan picos o incrementos inesperados en los datos y adicionalmente permite encajar los datos resultantes en un mismo espacio temporal facilitando la comparación entre estos.

Para el diseño de la ventana móvil se deben establecer 2 parámetros importantes. El tamaño de la misma que implica la cantidad de muestras que se van a usar para hallar el promedio y el numero de muestras superpuestas entre una ventana y otra 1.

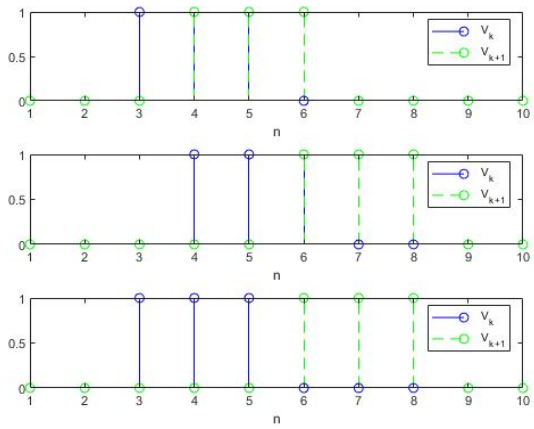

Figura 1: Esta figura incorpora 3 imágenes donde se evidencian la ventana de promediado móvil V_k y V_{k+1} . La imagen superior muestra la opción de superponer una ventana sobre otra avanzando una sola muestra. La imagen intermedia señala la opción de superponer simplemente la última muestra. Finalmente, la imagen inferior detalla la posibilidad de no superponer las ventanas de promediado móvil

La elección de estos 2 parámetros mencionados tiene un gran impacto en los resultados, un mayor tamaño de la ventana implica un suavizado robusto a los datos perdiendo posiblemente información, por su parte, 2 ventanas superpuestas de la forma presentada en la imagen superior de 1 implican un costo computacional bastante elevado.

Para encontrar el tamaño apropiado de la ventana es necesario tomar en cuenta que un aumento en el tamaño de la ventana se ve reflejado en un suavizado en la curva natural de los datos, esto pude tener repercusiones en los resultados finales y afectar que tan apegados a la realidad son. Por esto es necesario tomar una medida que si bien suaviza la curva, no afecte en gran medida la representación de los datos en el problema. También es necesario tomar en consideración el tiempo de computo que requiere una ventana grande, al tener que promediar mas datos por cada iteración se aumenta el coste computacional.

La ventana móvil busca acoplar valores cercanos en cada conjunto de datos y promediarlos para obtener los arreglos $f(\varepsilon_j)$ y $\hat{f}(\varepsilon_i)$ que si pueden comparar. Para generar cada ventana se toman intervalos de tiempo donde los limites están dados por $a_j \leq x_i, \hat{x}_i \leq bj$, en este caso particular se define un tamaño de ventana de n elementos para el conjunto de datos teórico y a partir de esa ventana teórica se calculan los respectivos limites a_j y b_j para generar la ventana apropiada en el conjunto de datos experimentales, con esto se busca ajustar la ventana en el conjunto de datos experimental al intervalo temporal de la ventana de datos de referencia. Este procedimiento se puede evidenciar de mejor manera en la gráfica 2 donde se realiza un ejemplo con una ventana de promediado de tamaño n = 3 en los datos de referencia, para definir la ventana de promediado en los datos experimentales se toman todos los datos obtenidos en ese mismo intervalo de tiempo. Llegado el caso en el que la ventana experimental no tenga muestras esta no será tenida en cuenta.

Cabe mencionar que en esta caso particular se toman ventanas en el conjunto teórico que se solapan avanzando en un elemento, es decir que la ventana $V_(k+1)$ va a repetir $n-1$ elementos de la ventana V_k agregando un nuevo elemento que corresponde al siguiente valor siguiendo el orden del arreglo al que se le esté aplicando el algoritmo.

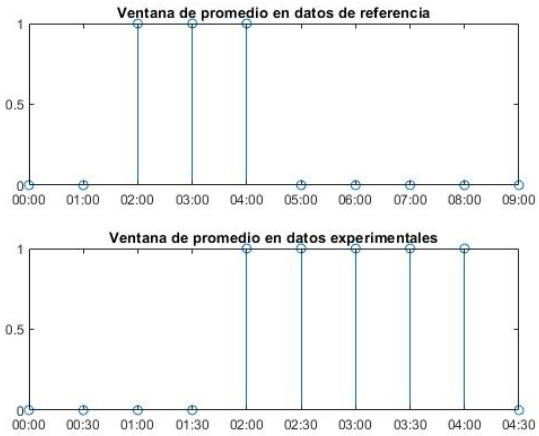

Figura 2: Esta figura ilustra un ejemplo de la diferencia de tamaño entre la ventana de promediado para los datos de referencia y los experimentales

En este punto se deben realizar algunas aclaraciones involucradas con lo mencionado anteriormente. La distancia entre la muestras experimentales y las muestras de referencia se encuentra haciendo uso del concepto de distancia euclidea $D(\mathbb{D}_i, \hat{\mathbb{D}}_i)$:

$$D(D_i, \hat{D}_{\hat{i}}) = \sqrt{\sum_{i,\hat{i}} (D_i - \hat{D}_{\hat{i}})^2} \tag{1}$$

cuando ya se tienen los conjuntos de datos comparables entre si $f(\varepsilon_j)$ y $\hat{f}(\varepsilon_i)$, se puede medir la distancia que hay entre ellos utilizando la ecuación 1

Finalmente, después de realizar el promediado de los datos de referencia y experimentales se obtienen 2 conjuntos de datos con igual cantidad de valores que se pueden comparar. A estos conjuntos de datos se les realiza un proceso de ajuste por mínimos cuadrados que permite realizar la calibración de los datos experimentales a partir de los datos de referencia. Esto se ilustra a través de la figura 3.

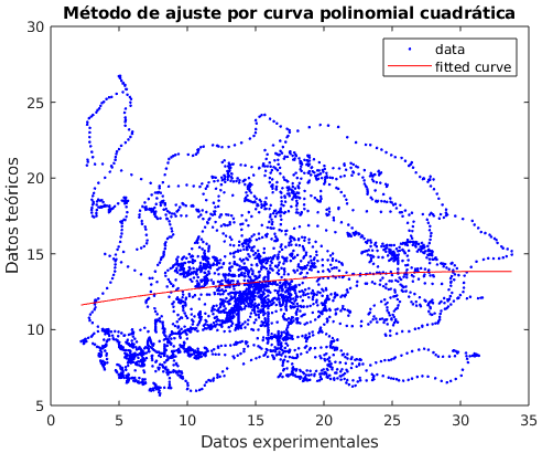

Figura 3: Se presenta la gráfica de datos experimentales vs datos de referencia, adicionalmente, el modelo obtenido a través del ajuste por mínimos cuadrados.

De esta misma forma buscando un ajuste a los datos se opta por métodos de inteligencia artificial que son ampliamente usados en problemas de reconocimiento de patrones, partiendo de que se tiene un conjunto de datos teórico para realizar un entrenamiento de un algoritmo supervisado. En este caso figura 4, se utilizan redes neuronales profundas adaptadas a problemas de regresión[3]. Las redes neuronales constan de algoritmos que permiten realizar un entrenamiento previo de la red utilizando conjuntos de datos A, B donde A son los datos que se tienen y B los datos que se esperarían obtener, en nuestro caso los conjuntos de datos teórico y experimental.

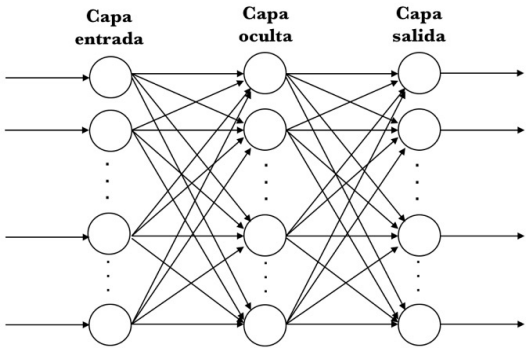

Figura 4: Gráfico de una red neuronal

Las capas de entrada y salida varían según los datos, estos dictaminan la cantidad de neuronas a utilizar en esas capas. Por otra parte la cantidad de neuronas en cada capa oculta y así mismo el numero de capas ocultas es un son definidas por la complejidad del problema.

3. El experimento y los resultados

Para determinar el tamaño de la ventana se debe tomar en cuenta como afecta este a la distancia planteada en la ecuación 1, por esto se realizaron pruebas variando el tamaño y calculando esta misma, los resultados se evidencian en la Fig. 5. Como se puede apreciar en la gráfica al aumentar el tamaño de la ventana, la distancia entre los dos conjuntos de datos disminuye.

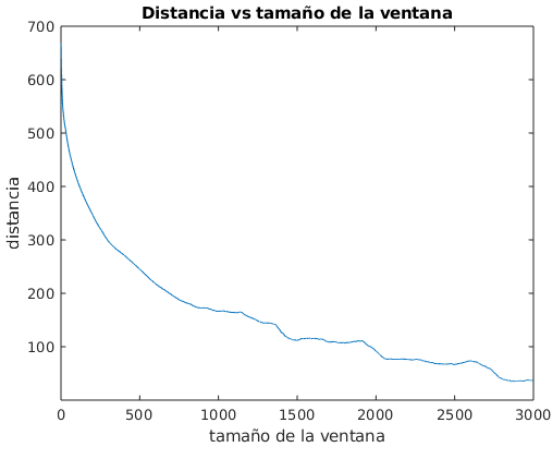

Figura 5: Comportamiento de la distancia entre los conjuntos de muestras experimentales y de referencia al cambiar el tamaño de la ventana móvil para realizar el promediado.

Esta relación que se puede intuir de la gráfica no indica que la ventana de mayor tamaño sea la mas óptima, ya que esta modifica distribución de los datos al aplicar un suavizado y puede afectar el desempeño final del ajuste realizado a los datos, ver Fig. 6. Además las ventanas grandes requieren un mayor costo computacional para ejecutar el algoritmo.

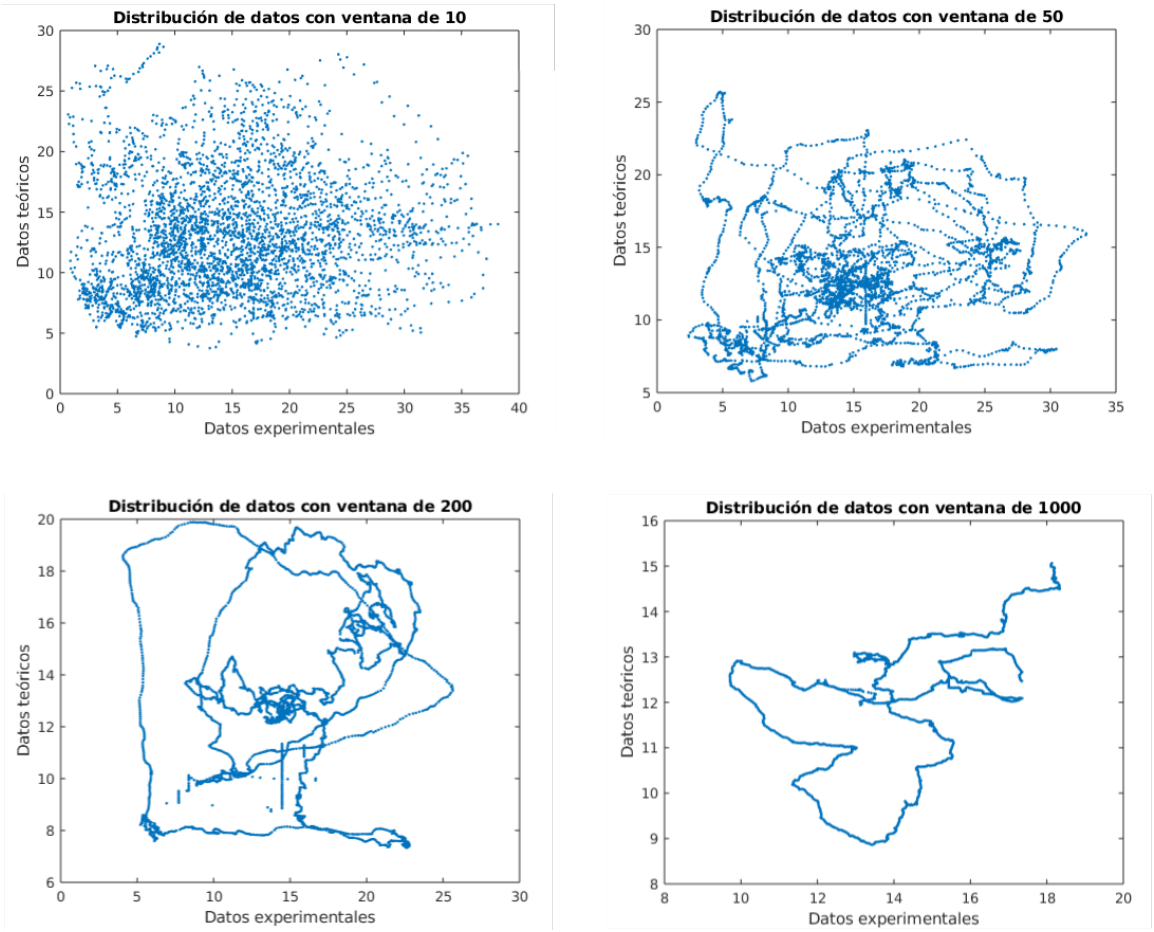

Figura 6: Comportamiento de la distribución de datos variando el tamaño de la ventana utilizada.

Basado en lo anterior se escoge una ventana de tamaño 40 que nos permite disminuir la distancia obtenida como se muestra en Fig.5, sin afectar de sobremanera la distribución de los datos para seguir manejando datos lo mas acercados a la realidad posible.

Seguidamente con los conjuntos generados al aplicar el algoritmo de la ventana utilizado, se procede a métodos para ajuste de datos. En primera instancia se realiza un ajuste utilizando un ajuste por Curva polinomial cuadrática, para la ventana planteada de tamaño cuarenta se encuentra el polinomio mostrado en la ecuación 2 como se puede observar en la Fig. 3.

$$f(x) = -0{,}002499 * x^2 + 0{,}1599 * x + 11{,}29 \tag{2}$$

Con este método se encontró una mejora en la distancia que se muestra en la tabla de distancias, tabla 2, para la validación del método se utilizó el 70 % de los datos para obtener el polinomio del ajuste y 30 % para validar los resultados y calcular la distancia obtenida por el modelo con ajuste.

Para el ajuste de los datos también se utilizo una red neuronal, la arquitectura utilizada se muestra en la tabla 1,

Cuadro 1: Arquitectura de la red neuronal.

Capa	Descripción
1	Entrada
2	Capa profunda de 15 neuronas
3	Capa de salida 1 neurona
4	Salida

Al ser un problema de tipo regresión lo que se busca es que la red entregue un único valor, por lo tanto la capa de salida cuenta con una neurona, como se puede ver en la figura 7.

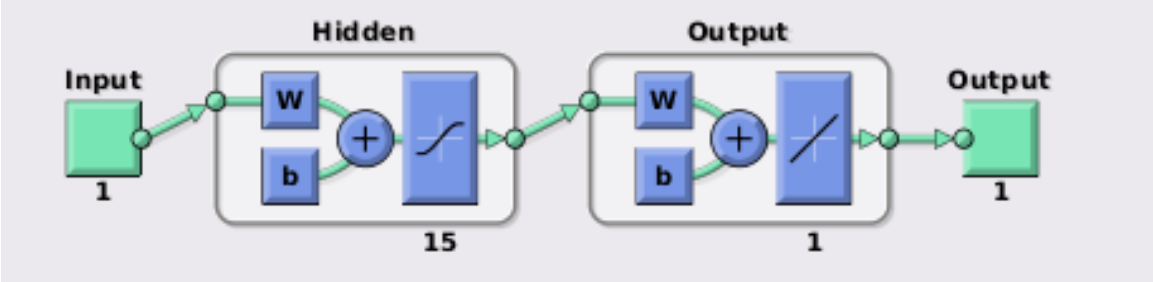

Figura 7: Gráfica de la red neuronal utilizada.

Los valores utilizados para el entrenamiento de la red corresponden a una tasa inicial de aprendizaje de $lr = 8e - 3$, una métrica de validación *meansquarederror* y los datos utilizados apara el entrenamiento equivalen al 70 % de los datos y el otro 30 % restante se utiliza para la validación. Los resultados obtenidos por cada método son mostrados en la tabla 2 donde se aprecia que la red neuronal obtiene una mejora del 4 % con respecto a la curva polinomial.

Cuadro 2: Tabla de comparación de distancias.

Método	Distancia
Sin ajuste	486.017
Curva polinomial	221.13
Red neuronal	212.82

Si utilizamos otras métricas para validar los resultados, como podría ser el error medio cuadrático tenemos los datos que se ven en la tabla 3 o la métrica del error relativo que se muestra en la tabla 4.

Cuadro 3: Tabla de comparación de distancias con error medio cuadrático.

Método	Distancia
Sin ajuste	57.0839
Curva polinomial	39.6267
Red neuronal	36.7067

Cuadro 4: Tabla de comparación de error relativo.

Método	Error
Sin ajuste	48.85 %
Curva polinomial	47.89 %
Red neuronal	44.50 %

Los resultados son comparados en la gráfica 8, donde se puede ver que la gráfica generada por la red neuronal tiende a imitar mejor el comportamiento de los datos teóricos.

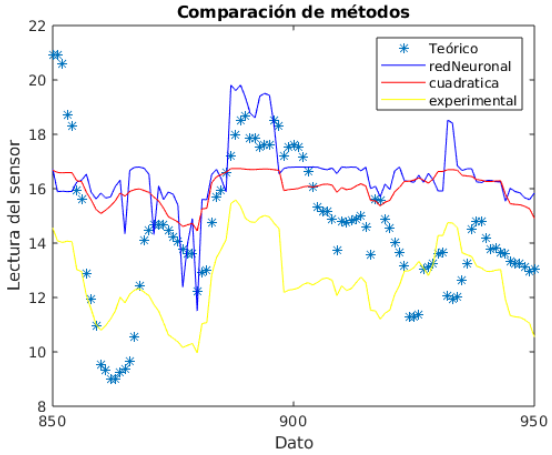

Figura 8: .

4. Conclusiones y Recomendaciones

Es claro que en al momento de utilizar un sensor se depende de la calidad del mismo para la confianza en las medidas entregadas, aun así con sensores de baja calidad es posible obtener resultados aceptables, mejorar estas mediciones por medio de un ajuste de los mismos utilizando algoritmos computacionales y así simular el comportamiento de un sensor de mayor calidad al

disponible. La importancia de esto recae en que por medio de estos algoritmos se pueden obtener mejores resultados en pruebas o ensayos posteriores a la recolección de los datos realizada por el sensor.

En este articulo se proponen dos métodos para realizar el ajuste de los datos entregados por el sensor. primero aplicando un promediado de ventanas a los datos para posibilitar su comparación y luego con métodos de ajuste, Estos métodos muestran una mejora significativa con respecto a los datos sin un procesamiento previo, los dos permiten al usuario del sensor tener una mayor confianza en los datos obtenidos por el sensor que esta siendo empleado. Con las tres tablas comparativas de distancia y error propuestas en la sección anterior podemos verificar el mejor resultado lo obtenemos al aplicar una red neuronal al problema presentado.

Referencias

[1] Alejandro Cama-Pinto, Emiro De la Hoz, and Dora Cama-Pinto. Las redes de sensores inalámbricos y el internet de las cosas. *Inge Cuc*, 2012.

[2] Naomi Zimmerman, Albert A Presto, Sriniwasa PN Kumar, Jason Gu, Aliaksei Hauryliuk, Ellis S Robinson, Allen L Robinson, and Ramachandran Subramanian. A machine learning calibration model using random forests to improve sensor performance for lower-cost air quality monitoring. *Atmospheric Measurement Techniques*, 11(1):291–313, 2018.

[3] Ana Luna, Álvaro Talavera, and Luis Cano. Uso de sensores electroquímicos de bajo costo para el monitoreo de la calidad del aire en el distrito de san isidro-lima-perú. 2017.

[4] Elena Esposito, Saverio De Vito, Maria Salvato, V Bright, Roderic Lewis Jones, and Olalekan Popoola. Dynamic neural network architectures for on field stochastic calibration of indicative low cost air quality sensing systems. *Sensors and Actuators B: Chemical*, 231:701–713, 2016.