Métricas, datos y calibración inteligente

Brayan Monroy* **Emmanuel Martinez**
Universidad Industrial de Santander
Cra 27 Calle 9 Ciudad Universitaria

25 de marzo de 2022

Índice

Resumen

En el presente informe se hace desarrollo del taller de distancias, partiendo del procesamiento de los datos por medio del calculo de las medias moviles locales y el posterior calculo de la distancia siguiendo la definición de la distancia euclidiana. Finalmente se realizó la calibración de las medidas obtenidas por sensores de bajo costo dadas unas medidas de referencia mediante un ajuste lineal por medio del método de mínimos cuadrados. El ajuste lineal siguió dos estrategias: la primera sobre todo el conjunto de datos y posteriormente se hizo un análisis de la porción mínima del conjunto de datos para realizar una óptima calibración teniendo en cuenta una tolerancia para validar dicha calibración.

1. Calculo de distancia

Dado el conjunto de datos de referencia $\mathbb{D}_i = \{(x_1, y_1), (x_2, y_2)...(x_n, y_n)\}$ y el conjunto de datos a calibrar $\hat{\mathbb{D}}_i = \{(\hat{x}_1, \hat{y}_1), (\hat{x}, \hat{y}_2)...(\hat{x}_m, \hat{y}_m)\}$, donde x representa las fechas con horas de cuando se adquirieron los datos y de concentración de material particulado $PM_{2,5}$. Adicionalmente, se observa que $n \neq m$, siendo $n, m \in \mathbb{Z}^{++}$, por lo tanto, es necesario realizar un ventaneo de los conjuntos de datos considerando las fechas de ambos conjuntos de datos.

El ventaneo o media movil se encuentra expresado por

$$\epsilon_i = \frac{1}{w - c_i} \sum_{k=i-w}^{i} a_k, \qquad c_i = \sum_{k=i-w}^{i} \delta(a_k), \qquad \delta(a) = \begin{cases} 1, & a = 0 \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

*e-mail: brayan2180032@correo.uis.edu.co

donde se promedian los w últimos elementos a partir del elemento actual i de algún conjunto de datos considerando que c_i la cantidad de elementos $a_k = 0$ en dicha ventana.

Antes de aplicar el ventaneo se realizó un reordenamiento de los conjuntos de datos para hacerlos de la misma longitud. Más especificamente, si se genera un conjunto de referencia que considere solo la fecha x tal que $\mathbb{D}_i^0 = \{(x_1, 0), (x_2, 0)...(x_n, 0)\}$ y para los datos a calibrar $\hat{\mathbb{D}}_i^0 = \{(\hat{x}_1, 0), (\hat{x}, 0)...(\hat{x}_m, 0)\}$. Entonces se puede entrelazar la información de tal manera que $\mathbb{S}_i = \{(x_i, y_i)|(x_i, y_i) \in \mathbb{D}_i \cup \hat{\mathbb{D}}_i^0 \wedge \neg((x_i, y_i) \in \hat{\mathbb{D}}_i^0 : x_j \in \mathbb{D}_j = x_i)\}$ es la unión del conjunto de datos de referencia $\mathbb{D}$ con el conjunto $\hat{\mathbb{D}}$ de tal manera que se omiten del conjunto $\mathbb{S}$ todos los pares $(x_i, y_i) \in \hat{\mathbb{D}}_i^0$ tales que $x_i \in \mathbb{D}$ ya existan, e $i = 1, \ldots$ son los indices del conjunto de datos tales que $x_1 < x_2 < \cdots$. Igualmente se genera un conjunto de datos para las medidas a calibrar $\hat{\mathbb{S}}_i = \{(x_i, y_i)|(x_i, y_i) \in \hat{\mathbb{D}}_i \cup \mathbb{D}_i^0 \wedge \neg((x_i, y_i) \in \mathbb{D}_i^0 : x_j \in \hat{\mathbb{D}}_j = x_i)\}$.

Debido a que los conjuntos de datos $\mathbb{S}_i$ y $\hat{\mathbb{S}}_i$ pueden contener elementos $y_i = 0$, se aplicaron dos tipos de ventaneo. El primer tipo de ventaneo consistió en aplicar una interpolación lineal a los elementos $y_i = 0$ de ambos conjuntos de datos con una ventana de 3 datos, por lo tanto $c = 0$ para cada ventaneo. El segundo tipo de ventaneo consistió en ignorar los elementos $y_i = 0$, durante el ventaneo, de tal manera que $c =$ número de datos en cero. Por lo tanto, podemos expresar los conjuntos de datos de referencia y de medidas a calibrar de igual cantidad de datos como

$$\mathbb{D}_i = \{z_i | z_i \in \mathbb{S}_i \wedge z_i = \epsilon_i\}$$

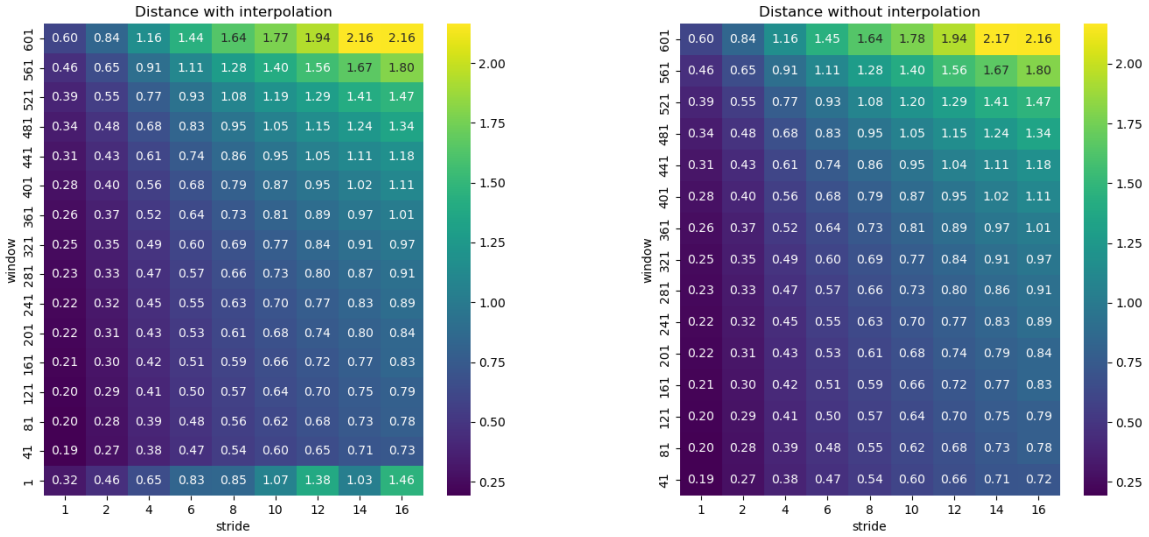
y

$$\hat{\mathbb{D}}_i = \{z_i | z_i \in \hat{\mathbb{S}}_i \wedge z_i = \epsilon_i\},$$

respectivamente. Finalmente, se puede calcular la precisión del conjunto de medidas a calibrar respecto al conjunto de datos de referencia mediante la ecuación de distancia

$$\mathcal{D}(\mathbb{D}_i, \hat{\mathbb{D}}_i) = \frac{1}{N}\sqrt{\sum_i^N (\mathbb{D}_i - \hat{\mathbb{D}}_i)^2} \tag{2}$$

Para la estimación de estos nuevos conjuntos de datos se probó con ventanas $w_l = \{1, 41, 81, 121, \ldots, 601\}$ y espaciados o 'strides' $s_q = \{1, 2, 4, 6, 8, 10, 12, 14, 16\}$. Para una región local del conjunto de datos se pueden observar los resultados generales, tanto para los datos con interpolación como sin interpolación en las Figuras 1a y 1b, respectivamente. De estas figuras se observa que los mejores resultados ocurren cuando $w = 41$ y $s = 1$. En general se observa que entre menor sea el tamaño del ventaneo y del espacio, mejores resultados se van a obtener. Además, cabe resaltar que entre mayor sea el tamaño del ventaneo y/o el espaciado, mayor información se va a perder en el ventaneo del conjunto de datos, por lo que entre más pequeño se mantengan ambos valores, se mantedra con mayor consistencia el comportamiento original del conjunto de datos.



(a) Distancia entre datos interpolados.

(b) Distancia entre datos no interpolados.

Figura 1: Compararación de la distancias calculadas entre los conjuntos de datos para diferentes tamaños de ventana y espaciados. (a) Considerando la interpolación, (b) Sin considerar la interpolación.

2. Calibración

Considerando que $f(x_i) = y_i$ sería la función que define la concentración de material $PM_{2,5}$ de acuerdo a su fecha de adquisición para los datos de referencia, de la misma manera que $\hat{f}(x_i) = \hat{y}_i$ lo define para los datos a calibrar, entonces para la calibración de las muestras promediadas $(\hat{f}(\epsilon_j), f(\epsilon_j))$, se hizo uso de un ajuste de mínimos cuadrados para determinar un modelo de ajuste lineal, $f(\epsilon_j) = \alpha\hat{f}(\epsilon_j)$. De esta forma, el calculo de α fue determinado como,

$$\alpha = \frac{\langle f(\epsilon)|\hat{f}(\epsilon)\rangle}{\langle f(\epsilon)|f(\epsilon)\rangle} \tag{3}$$

Como se muestra en la Figura 2(b), se presenta una gran dispersión entre los datos de referencia y los datos a calibrar al no ser evidente un comportamiento lineal de los mismos. Sin embargo, una vez se realiza el ajuste lineal. como podemos observar en la Figura 2(a), los datos de las estaciones de bajo costo se ajusta correctamente a los datos de referencia, presentando un comportamiento similar y una menor distancia en comparación con los dato datos sin calibrar, diferenciándose por un desfase temporal posiblemente debido a un error en la documentación de la captura de los datos. Adicionalmente, se realizo el calculo de un valor para la tolerancia y de esta forma analizar eel porcentajee de datos que podrian considerarse .ªceptables"dentro de la estimaciám realizada por el modelo lineal. Para esto, se estimo el valor de la media para los errores absolutos entre los datos

predichos por el modelo lineal y los datos de referencia, posteriormente se utilizo dicha media como valor de tolerancia o umbral para obtener el porcentaje de datos calibrados los cuales consideramos aceptables dentro del margen de error, dicho calculo se puede definir matematicamente como,

$$tol = \mathbb{E}[|f(\epsilon_j) - \alpha\hat{f}(\epsilon_j)|] \tag{4}$$

en donde $\mathbb{E}[\cdot]$ denota el valor esperado del error absoluto sobre la calibración de las mediciones de bajo costo, para esta caso siendo la misma media aritmetica de los errores absolutios $|f(\epsilon_j) - \alpha\hat{f}(\epsilon_j)|$.

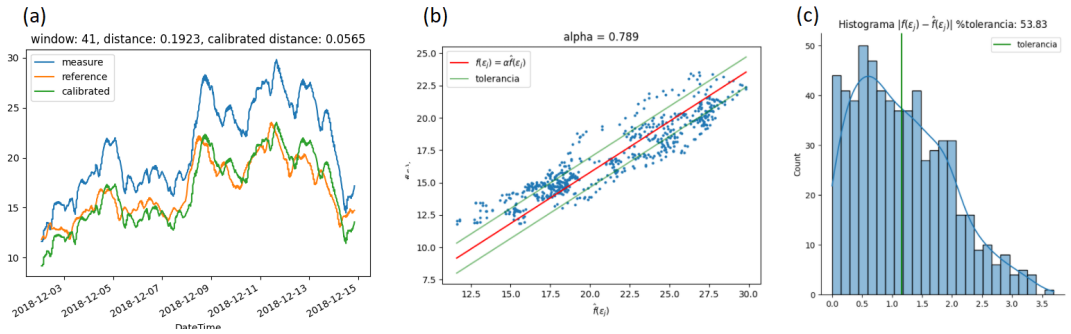


Figura 2: (a) Ajuste lineal por el método de mínimos cuadrados, (b) Calibración de las muestras dado el modelo lineal obtenido.

Posteriormente, se siguió otra alternativa mediante la división de los conjuntos de datos, en un principio se realizo una división equitativa con el 50 % de las muestras como conjunto de calibración y el otro 50 % como conjunto de validación, los resultados se pueden observar en la Figura 3. En esta estapa tambien se realizo el calculo del valor de tolerancia y porcentaje de datos aceptados siguiendo el procedimiento descrito anteriormente, solo que en este caso se obtuvo el calculo de tolerancia sobre lo datos utilizados para calibración y se obtuvo el porcentaje de tolerancia con sobre los conjuntos de datos utilizados para validación. Al comparar lo resultados de la Figura (2) y Figura (3), podemos oberser como el porcentaje de datos aceptados para nuestro modelo por ajuste lineal disminuye del 53.83 % al 18.28 % al utilizar solo el 50 % de las muestras del conjuntos de datos para la estimación del parametro α.

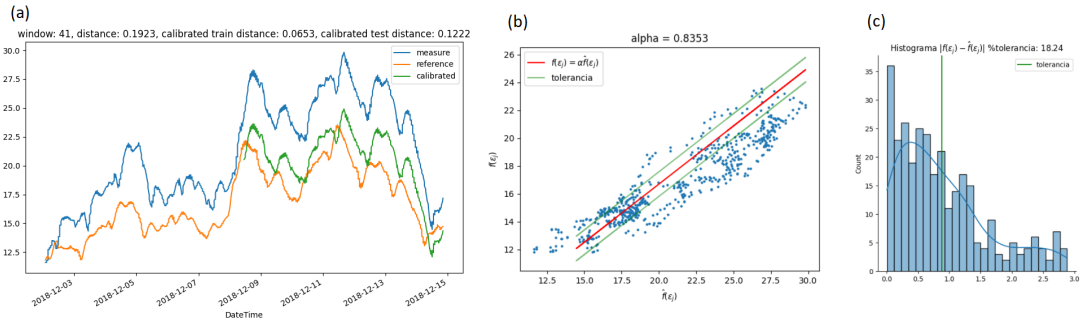


Figura 3: Ajuste lineal empleando el 50 % del conjunto de muestras de referencia y calibración.

porcentaje	**D calibración**	**D validación**	**alpha**	**% tolerancia**
20	0.0858	0.0939	0.8494	23.83
30	0.0873	0.0802	0.8185	33.26
50	0.0653	0.1222	0.8353	18.24
70	0.0617	0.1325	0.8074	22.28
100	0.1923	-	0.7890	53.83

Cuadro 1: Primera columna, variación del porcentaje de conjuntos para realizar el ajuste lineal de calibración. Segunda y tercera columna, distancia del conjunto de calibración y conjunto de validación. Ultima columna, factor de ajuste lineal obtenido.

Finalmente, se evalúa la estrategia por división de conjunto de datos para distintos porcentajes de división los cuales fueron $\{10, 20, 30, 40, 50, 60, 70, 100\}$ para el conjunto de calibración y su respectivo complemento para el conjunto de validación. Los resultados obtenidos se muestran en el Cuadro (1). Como se puede observar, conforme mayor es el porcentaje del conjunto de calibración, menor es la distancia entre el conjunto de validación y de referencia, por lo que podría indicar una mayor precisión en la calibración de las muestras de las estaciones de bajo costo. Adicionalmente encontramos como a mayor porcentaje de datos utilizados para la estimación del modelo de ajuste lineal, el parametro α tiende al valor obtenido al utilizar todo el conjunto de datos, sin embargo encontramos un comportamiento particular en dondeal utilizar solo el 30 % de los datos para calibrar, se obtiene un alpha muy cercano al esperado incluso siendo mejor que la estimación obttenido al usar el 50 % de los datos. Por ultimo, analizando el porcentaje de los datos aceptados para cada caso, podemos encontrar como utilizar un porcentaje del 30 % de los datos para calibrar, no ofrece el mayor rango de datos aceptados como el menor error de calibración, lo cual se corrobora el analisis realizado sobre la estimación del parametro α.

3. Conclusiones y Recomendaciones

En el presente informe se emplearon los conceptos aprendidos de aproximación de funciones y bases ortogonales en el problema de calibración de sensores dados unas muestras de referencia. Durante el desarrollo del informe, se observo un comportamiento inusual al analizar el calculo de la distancia de los conjuntos de datos en función del tamaño de la ventana móvil, en donde a un mayor tamaño de ventana la distancia era menor. Sin embargo, dada la definición de distancia euclídea utilizada, podemos evidenciar como esta no contempla el tamaño del conjunto de muestras, por lo que los valores de distancia obtenidos pueden estar sesgados a obtener una menor distancia conforme en conjunto de datos disminuye sin poder evidenciar correctamente el error general, por este motivo y con el fin de ofrece un mejor análisis se realizo la dividisión de la distancia obtenida por la cantidad de muestras evaluadas. Adicionalmente, en la estapa de calibración se observo como con simplemente una estimación de un factor de escalado sobre los datos a calibrar se disminuye de forma significativa el error entre estos datos y los datos de referencia, sin embargo se requiere de un modelo lineal que incluya un termino independiente el cual nos permite realizar correciones sobre el desfase de los datos los cuales se pueden evidenciar claramente en las graficas analizadas.