Variables Climáticas e Interpolación

Nicolás Pérez Fonseca   
*Departamento de Ingeniería de Sistemas*  
*Pontificia Universidad Javeriana*Bogotá, Colombia  
[n\_perez@javeriana.edu.co](mailto:n_perez@javeriana.edu.co)

José Alejandro López García   
*Departamento de Ingeniería de Sistemas*  
*Pontificia Universidad Javeriana*Bogotá, Colombia  
[lojose@javeriana.edu.co](mailto:lojose@javeriana.edu.co)

Juan Pablo Amorocho Manjarres  
*Departamento de Ingeniería de Sistemas*  
*Pontificia Universidad Javeriana*Bogotá, Colombia  
jp.amorochom@javeriana.edu.co

Diego Fernando Trujillo Álvarez  
*Departamento de Ingeniería de Sistemas*  
*Pontificia Universidad Javeriana*Bogotá, Colombia  
diego\_trujillo@javeriana.edu.co

***Abstract*— By using the numerical method known as interpolation and cubic spline predict the inner temperature given the day and hour in the year 2013, such our model predicts the temperature variable based on data from a nearby region in the Fortaleza (Brazil) area.**

***Keywords— numerical-methods, interpolation, spline***

1. Introducción

En la medición y predicción de variables climáticas hay implicadas varias disciplinas y campos del conocimiento que se complementan para desarrollar los modelos climáticos usados en institutos meteorológicos.

Uno de los campos con aplicación en este medio son los métodos numéricos para predicción y análisis de datos, ya que basados en las lecturas de variables climáticas en locaciones geográficamente cercanas y con condiciones similares de terreno se pueden aproximar variables basados en otras lecturas [1]

Esto es de gran utilidad ya que ayuda a complementar o intuir con un error información de regiones donde no se encuentra una estación o en la que se imposibilita realizar mediciones, como puede ser una región de difícil acceso como selva, o regiones inestables como una zona de conflicto.

En el caso que concierne al presente artículo, basados en los datos de temperatura interna de la región de Quixadá en la región de Fortaleza, Brasil, tomados cada hora desde el martes 31 de marzo del 2013 al miércoles 29 de abril del mismo año usando interpolación para crear un modelo que permitiera completar y aproximar la información faltante de temperatura interna en la región cercana de Quixeramobim.

El método usado para interpolar usado es spline cúbico, debido al comportamiento oscilatorio de los puntos. Basados en el modelo de aproximación polinomial por tramos [2]

1. METODOLOGÍA
   1. *Selección de Datos*

Para la construcción del modelo se contaba con mediciones de 17 estaciones cada una con 22 parámetros de los cuales 18 correspondían a mediciones del de la estación climática, y un número variable de registros que iba desde los 720 y 163.

Se debían elegir dos estaciones, una sería la fuente del modelo y la otra sería el objetivo de la aplicación del modelo. Para ello se debían elegir dos estaciones con cercanía geográfica suficiente como para que tuviera sentido la aplicación del modelo. Entonces basados en las coordenadas cartesianas de cada una provistas previamente se eligieron Quixadá y Quixeramobim.

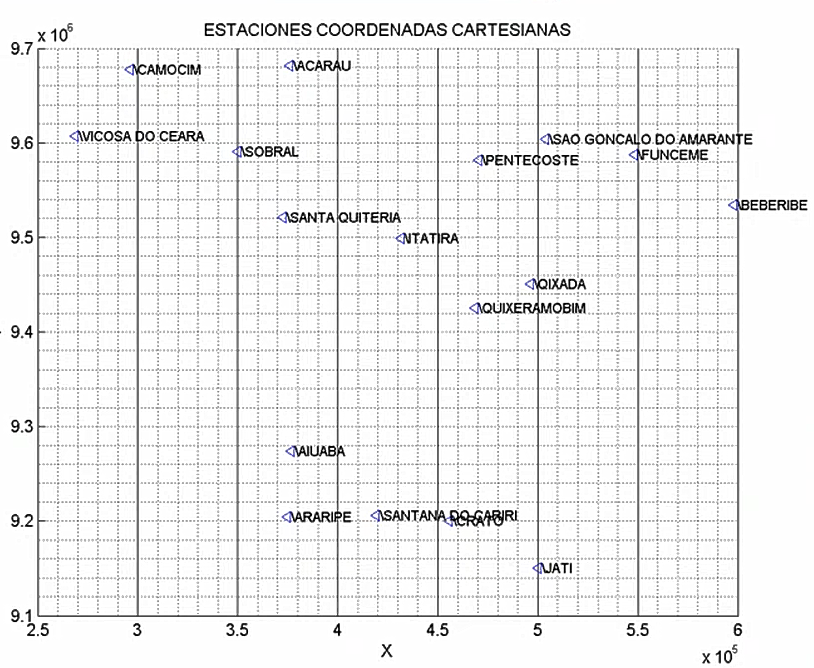


Ilustración 1 Estaciones Climáticas en coordenadas cartesianas

Arbitrariamente se eligió Quixadá como estación para el elaborar el modelo y a Quixeramobim como objetivo de aplicación.

Una vez se elegidas las estaciones la elección la elección de la variable se a modelar se hizo bajo los siguientes criterios: evitar variables con ceros ya que su tratamiento puede ser conflictivo, las unidades ya que magnitudes vectoriales como dirección del viento añaden dificultad al proceso, y finalmente variables que no posean valores nulos o no existentes.

Con esos criterios se eligió la variable de Temperatura Interior ya que estaban los datos completos en las dos estaciones elegidas, y sus unidades era grados Celsius, una magnitud escalar lo cual favorecía el proceso de modelado.

Los datos de Quixeramobim no estaban tomados con la misma frecuencia que los de Quixadá, lo cual hacía esta estación el objetivo perfecto ya que tenía momentos sin datos donde podíamos aplicar el modelo. Como comparación Quixadá posee 720 registros, mientras que Quixeramobim posee 313.

* 1. *Filtrado de Datos*

Una vez elegidas las estaciones y las variables, se debían filtrar los datos.

De los datos de Quixadá se debían tomar el 80% correspondiente a 576 registros, los cuales serían usados para construir el modelo, y los restantes 144 correspondientes al 20% se usarían para calcular el error del modelo resúltate.

La elección de qué registros iban en qué grupo, se hizo tomando muestras aleatorias.

Los datos de temperatura interna se correspondían con un tiempo dado en días julianos y horas en formato de 24 horas. Entonces, se tomó como el momento cero el momento en el que ocurre la primera medición y desde ahí se convierten las fechas a horas desde ese primer momento, con lo cual tenemos variables de tiempo dado en horas y no días y horas.

* 1. *Construcción de los archivos de datos*

Con la información que sería usada de la que inicialmente fue prevista, se crea un libro de Excel con las variables de interés para el modelo y se crean archivos independientes separados por comas para listar los datos destinados para error y modelo en el caso de Quixadá. Mientras que en el caso de Quixeramobim. Sólo se listan la totalidad de los datos, sin discriminación.

* 1. *Comportamiento de la temperatura interna en el tiempo según los datos*

Tomando los pares ordenados resultantes del filtrado de datos y construcción de archivos, el comportamiento de la temperatura interna en el tiempo en Quixadá se comporta de la siguiente manera:

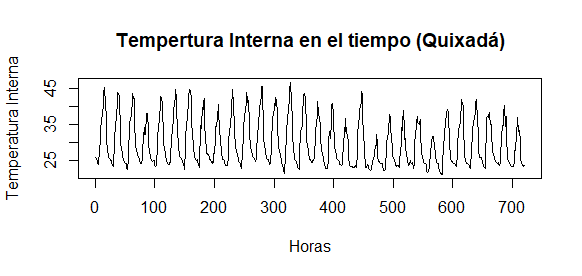


Ilustración 2 Temperatura Interna en el tiempo para Quixadá

Mientras que en caso de Quixeramobim se comporta de la siguiente manera:

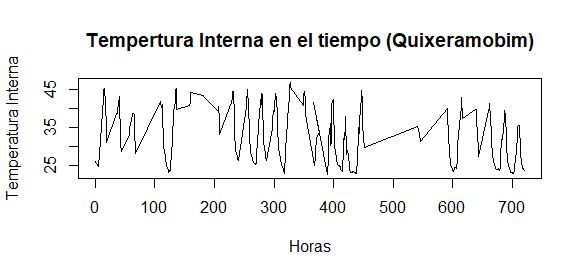


Ilustración 3Temperatura interna en el tiempo para Quixeramobim

El comportamiento de se puede explicar por la ausencia de datos en ciertos momentos, por ejemplo, hay días donde no se toman mediciones.

* 1. *Interpolación*

El comportamiento de los polinomios de grado alto oscila lo cual se ajusta al comportamiento de los datos que tomamos para elaborar la interpolación por lo que una interpolación polinómica es pertinente en este caso.

Debido a los polinomios de grados muy altos tienden a oscilar, decidimos partir el intervalo atendiendo así a un modelo de aproximación polinomial por tramos [2] para elaborar el modelo que representara el comportamiento de la temperatura en el tiempo.

La representación más simple de este tipo de interpolación sería la lineal ya que esta va uniendo los puntos daros con rectas, pue para generar las ecuaciones de una recta no se requiere más que dos puntos. Al final resulta un modelo con picos y curvas poco suaves, con un aspecto parecido a las gráficas de puntos presentadas previamente. Pero la temperatura no se comporta de esa forma siempre va gradualmente, no existe tal cosa como un cambio bien sea incremental o decremental de la temperatura ocurriendo al mismo tiempo, por lo tanto, su función no puede tener picos porque ello querría decir que hay dos pendientes en un cierto punto.

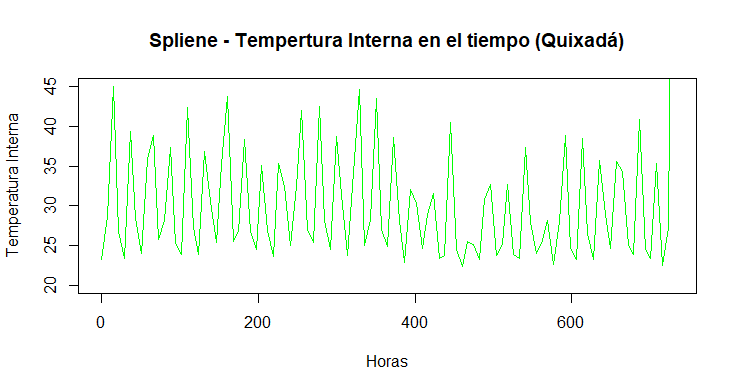
Con lo anterior en mente una interpolación lineal no representaría el comportamiento de la temperatura en el tiempo, entonces se optó por un interpolar usando un spline cúbico, ya que un polinomio cúbico general implica cuatro constantes para garantizar que el interpolante no es sólo continuamente diferenciable en el intervalo, sino también tiene una segunda derivada continua [2].

Entonces, el spline usado se construyó usando la librería *splines* en R, la cual ofrece una gran variedad de formas de construir un spline cúbico dados unos nodos, que no son más que parejas ordenadas, que para nuestro caso son los datos temperatura tomados en unos instantes de tiempo dados [3].

En particular nuestro eje de abscisas se corresponde con el tiempo, mientras que las ordenadas con la temperatura.

Para construir nuestros polinomios, usamos *splinefun* ya que esta elabora la función a partir de los puntos, ya que otras funciones disponibles en la librería *splines* retornan la matriz de valores para construir los polinomios, pero es inmanejable hacer eso con 576 nodos, por lo que se delegó esa tarea de esa manera [3].

El resultado final de la aproximación por splines cúbicos de nuestros nodos se ve como se sigue:

Ilustración 4 Spline cúbico de Temperatura interna en el tiempo para Quixadá

* 1. *El Modelo*

Basados en el resultado del spline cúbico del anterior, se debía contrastar el comportamiento con la totalidad de los puntos, es decir con el 100% de los datos, recordemos que los nodos con los que se elaboró el spline fueron el 80%, es decir 576 registros.

Poniendo lado a lado ambas curvas, se presentan de la siguiente manera:

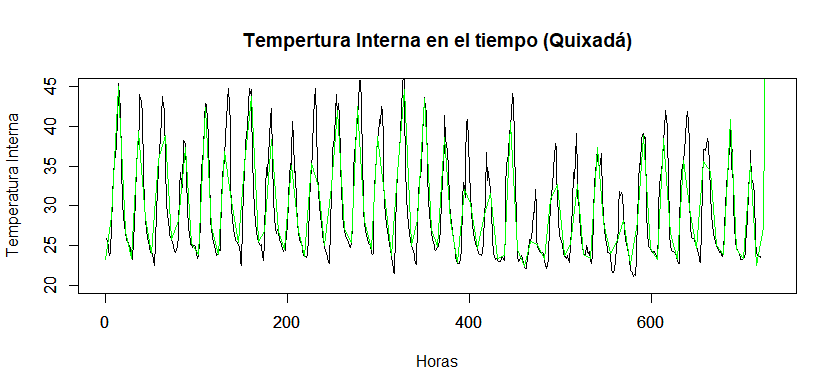


Ilustración 5 Temperatura interna en el timepo para Quixadá y spline cúbico

De la gráfica anterior podemos ver como acierta en los incrementos y descensos de la temperatura, sin embargo, se observa cómo en los máximos y mínimos se aleja del comportamiento real de la temperatura interna en ese instante.

El spline resultante tiende a sobrestimar la temperatura cuando esta alcanza su mínimo del día y a subestimarla cuando es máxima.

* 1. *Errores*

Para encontrar el error entre nuestros datos observados y nuestra estimación polinómica, hicimos una comparación entre los dos puntos por punto. Al ingresar un valor inicial a nuestra función polinómica, en este caso, las horas escogidas en nuestro modelo de tendencia (un 20% de los datos, escogidos aleatoriamente), podemos encontrar la predicción de la temperatura con nuestro polinomio estimador.

A continuación, se mostrará una gráfica con el error encontrado punto por punto, donde el valor Y es el error y el valor X son las horas en la que el error se calculó.

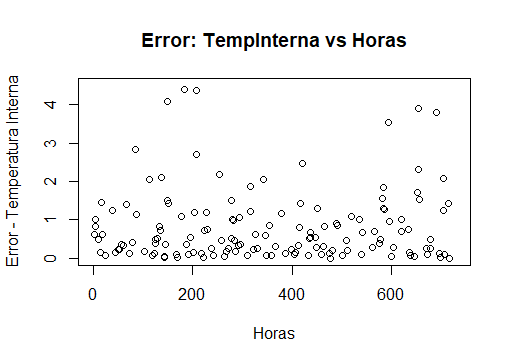


Ilustración 6 Error Ilustrado Punto por Punto

Como esperado, podemos ver ciertos valores atípicos presentes en la gráfica. Esto se debe a la naturaleza de los datos, ya que estos tienen sus propios datos atípicos. Si nos referimos a la *Ilustración 3*, podemos ver que los puntos donde hay más error es en donde hay más espacio entre horas. Para hacer más claro esta comparación, hicimos una interpolación de nuestro error, usando el método de splines cúbicos:

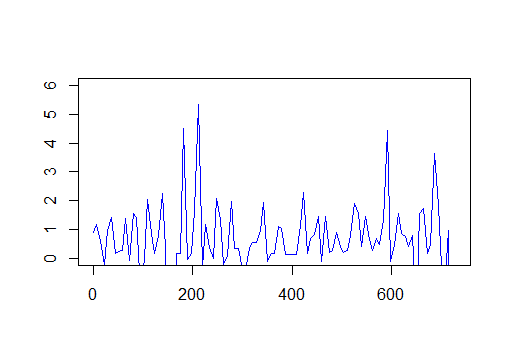


Ilustración 7 Interpolación de Error Punto por Punto

Se puede ver que, claramente, alrededor de las 200 horas y las 600 horas existen los valores máximos. Si la comparamos con la ilustración 3 de desde documento, podemos ver que estas son las áreas en las que faltaban más datos para hacer la interpolación. Ahora, para hacer un análisis completo del error encontrado, hicimos dos cálculos: el Error Cuadrático Medio [5]y un Error Absoluto Medio [6]:

C:\Users\USER\AppData\Local\Microsoft\Windows\INetCache\Content.Word\EstimacionMedia.png

Ilustración 8 Resultados para ECM y EAM.

Acá podemos ver dos resultados distintos sobre nuestra medición de error. El error cuadrático medio resulto con aproximadamente 0.69 más en su valor que en el error absoluto medio. Esto se debe, a que, en general, el error cuadrático medio les da más énfasis a los valores atípicos, o “outliers”, en el error [7]. Esto significa que en general, hay bastantes valore atípicos presente en nuestro error. Para terminar el análisis de errores, haremos un análisis basado en distintos tipos de errores:

* Error máximo: Se encuentra que el error más alto encontrado en la diferencia entre el modelo interpolado y las gráficas normales se encuentra alrededor de los valores de 210 horas y la diferencia se marca con un valor de 35 para los datos interpolados y 45 para el valor de los datos normales dando como resultado 10 de error entre estos dos, aunque la diferencia puede ser también relacionada a que el pico de la temperatura no interpolada puede ser mayor unas pocas horas más adelante del pico de la interpolada aunque la relación es la misma.
* Error mínimo: Este error se encuentra alrededor de los valores de 615 horas donde se encuentra la menor diferencia entre los valores de temperatura interpolada y no interpolada por la parte baja de las gráficas con valores de 23 en la gráfica interpolada y 22 en la no interpolada dando una diferencia de 1 siendo el error mínimo de entre los valores.
* Error medio: Si se suman los valores de error del mínimo (1) y el máximo (10) dividido entre 2 y daría como resultado 5.5
* Error absoluto: Seria el resultado de la suma de todos los valores de error dejando como resultado 16.5 con todos los valores positivos.

1. VECINO MÁS CERCANO
   1. En algunos casos al querer hacer la medición climita de un lugar se vuelve imposible, ya sea porque la zona es de difícil acceso o la maquina de medición se descompuso, para estos casos existe el método del vecino más cercano que es método para clasificar casos basándose en su parecido a otros. Se desarrolló como una forma de reconocer patrones de datos sin la necesidad de una coincidencia exacta con patrones o casos almacenados, donde los casos parecidos están próximos y los que no lo son están alejados entre sí [4].

En este caso se realiza el método entre las estaciones de Quixeramobim y Quixadá, donde se asumirá que los datos de la estación de Quixeramobim son igual al del spline de la estación Quixadá, dando como resultado la siguiente gráfica:

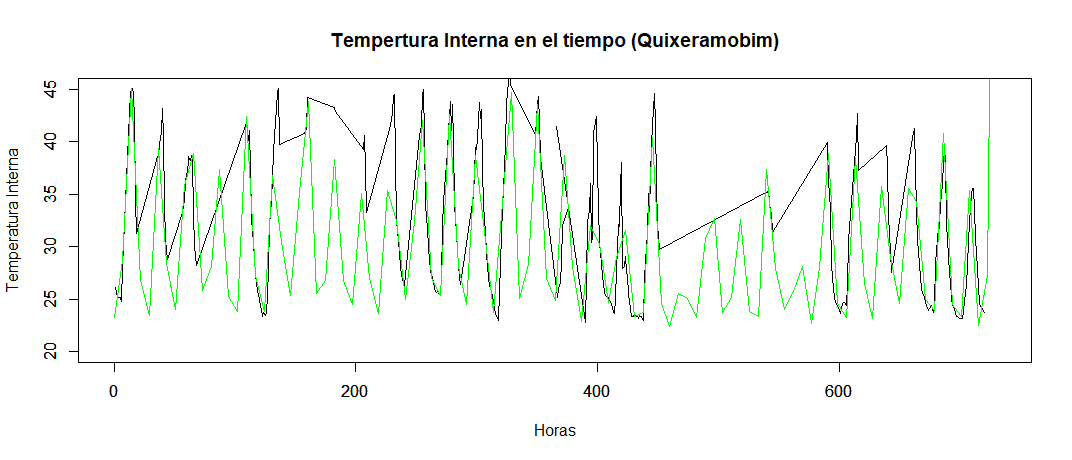


Ilustración 8 Temperatura interna en el tiempo para Quixerambim y spline de Quicadá

De la gráfica podemos observar que en las horas donde se tienen datos de la estación de Quixerambim el spline de la estación de Quicadá acierta en los incrementos y descensos de la temperatura, sin embargo, en algunos puntos máximos y mínimos se equivoca, dando un margen de error. Por otro lado, donde más se diferencian las gráficas son en las horas donde no se tienen datos de la estación de Quixerambim, para esto sirve el método del vecino más cercano, se puede decir que los valores del spline en esas horas son cercanos a los valores reales, claramente esto aún no se puede asumir ya que no se han calculado los errores.

* 1. Error

Para calcular los errores solo se tomaron en cuenta las horas donde se tienen los datos de la estación de Quixerambim y se compararon con el spline de la estación de Quixadá.

Se obtuvo la siguiente gráfica:

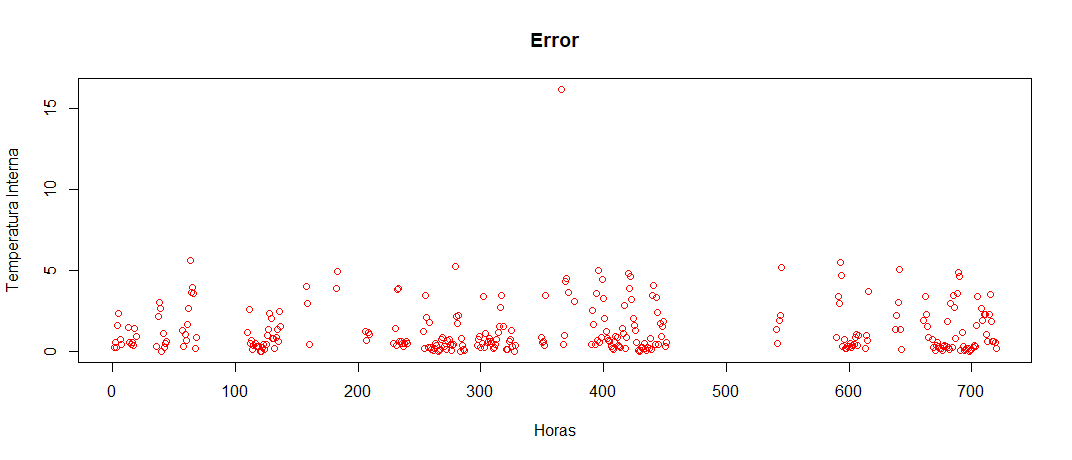


Ilustración 9 Error Vecino más Cercano

Se puede observar que la mayoría de los errores se encuentra entre 0 y 2, unas pocas entre 2 y 6, y un error único que supera el 16, este último puede ser que el spline en ese valor se equivocó al poder ser un valor máximo o mínimo, el cual va a inflar mucho el error medio.

* Error máximo: El error máximo se encuentra alrededor de la hora de 360 y tiene un valor de 16.2.
* Error mínimo: El error mínimo es 0, ya que hubo una hora donde el valor del spline y el valor real son iguales.
* Error medio: Al sumar el valor máximo que es 16.2 y mínimo que es 0, y dividirlo entre 2, da como resultado 8.1.
* Error absoluto: Al sumar todos los errores y dividirlos entre la cantidad de valores da como resultado 1.30, siendo un valor relativamente bajo.

De este resultado se puede decir que el spline de los valores de la estación Quixadá son valores cercanos a los valores reales, con un error promedio de 1.30, por tanto, aplicar el método del vecino más cercano es correcto para este caso.

# Referencias

|  |  |
| --- | --- |
| [1] | J. V. B. GUILLERMO TRINCADO V., «Aplicación de interpolación "spline" cúbica en,» *Instituto de Manejo Forestal, Universidad Austral de Chile, Casilla 567, Valdivia, Chile,* vol. 1, nº 1, pp. 3-27, 1999. |
| [2] | Richard Burden, Douglas Faires, Anette Burden, «Numerical Analysis,» de *Numerical Analysis*, México DF, Cengage Learning, 2017, pp. 105-120. |
| [3] | R-core R-core@R-project.org, «splinefun,» R Documentation, 6 11 2020. [En línea]. Available: https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/splinefun. [Último acceso: 6 11 2020]. |
| [4] | IBM, «Análisis vecino más cercano,» IBM, 11 11 2020. [En línea]. Available: https://www.ibm.com/support/knowledgecenter/es/SSLVMB\_sub/statistics\_mainhelp\_ddita/spss/base/idh\_idd\_knn\_variables.html. [Último acceso: 11 11 2020]. |
| [5] | H. Pishro-Nik, «Mean Squared Error (MSE),» [En línea]. Available: www.probabilitycourse.com. [Último acceso: 11 Noviembre 2020]. |
| [6] | J. O. Berger, «2.4.2 Certain Standard, Loss Functions,» de *Statistical Decision Theory and Bayesian Analysis (2nd ed.)*, New York, Springer-Verlag, 1985. |
| [7] | C. J. Willmott y K. Matsuura, «Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance,» 2005. |