

Reto número 2

Camilo Andrés Buitrago Ladino, David Santiago Meneses Cifuentes
Paula Juliana Rojas Naranjo, Juan Carlos Suárez Motta

Pontificia Universidad Javeriana

Autor correspondiente: Camilo Andrés Buitrago Ladino; bu.camilo@javeriana.edu.co

Autor correspondiente: David Santiago Meneses Cifuentes; @davidsmenenes@javeriana.edu.co

Autor correspondiente: Juan Carlos Suarez Motta; @suarezjuan@javeriana.edu.co

Autor correspondiente: Paula Juliana Rojas Naranjo; @rojasnpaula@javeriana.edu.co

12/05/2021

Para encontrar los repositorios de GitHub respectivos:

Camilo Andrés Buitrago Ladino: <https://github.com/CamiloB1999/AnalisisNumerico>.

David Santiago Meneses Cifuentes: <https://github.com/santiagomeneses/Analisis-numerico-David-Meneses>.

Juan Carlos Suarez Motta: <https://github.com/2qualcuno7/JuanCarlosSuarezMotta>.

Paula Juliana Rojas Naranjo: <https://github.com/ayurojasn/AnalisisNumerico-JulianaRojas>.

1. Resumen

El presente documento describe de manera clara la forma como se emplean los conceptos de interpolación y regresión para solucionar un problema que frecuentemente se da en la vida real, el manejo de los datos obtenidos de sensores y medidores en lo referente al estudio del clima. El desarrollo de esta práctica consistió en tomar un set de datos climatológicos de estaciones de captura de medición cercanas a Fortaleza, Brasil, en el año 2013 y a partir de ellos plantear dos ejercicios puntuales. El primero de ellos consistió en la selección arbitraria de una estación climatológica cuyos datos estuvieron disponibles, en este caso puntual Sobral, y a partir de las mediciones realizar un ejercicio de interpolación para obtener una función que pudiese ser empleada para calcular valores que no estuvieran expresos en los datos de entrada. Para dicho procedimiento se empleó el lenguaje de programación R y las técnicas numéricas de splines e interpolación lineal, comprándolos y evaluando el error de cada uno con respecto al data set original. Un segundo punto contenido en el presente documento consistió en la utilización de datos de la estación más cercana a Sobral, que se determinó empleando métodos geométricos como la estación de Santa Quiteria, para poder interpolar y simular valores cercanos a los realmente obtenidos de la estación de Sobral.

2. Nomenclatura

1. Interpolación Lineal: Método de interpolación empleado para reconstruir una función a partir de puntos dados, en la que se asume que los puntos no conocidos se ubican en una línea recta trazada entre los puntos más cercanos por izquierda y derecha al valor que se desea conocer.[5]
2. Interpolación por Splines: La reconstrucción de una función aproximada a partir de puntos dados empleado splines[3]
3. Spline: Tipo de función definida y diferenciable en un intervalo dado que es posible definir a partir de polinomios [4]
4. Error Absoluto: Es la magnitud de la diferencia entre un valor exacto conocido y una medición a partir de un instrumento, definiéndose de la siguiente manera:

$$\Delta x = x - x_0 \quad (1)$$

[2]

5. Error Relativo: Cociente entre el error absoluto y el valor exacto que permite determinar la proporción de error con respecto al valor real. Se define así:

$$\epsilon = (x - x_0)/x_0 \quad (2)$$

[2]

6. Índice de Jaccard: Valor estadístico empleado para determinar la similitud entre dos conjuntos y calculado a partir de la expresión

$$J_{((A,B))} = (A \cap B)/(A \cup B) \quad (3)$$

[1]

7. Distancia Euclidiana: Fórmula empleada para calcular la distancia entre dos puntos cuyas coordenadas son conocidas. Para coordenadas bidimensionales, esta se define así:

$$d_{xy} = \sqrt{((x_2 - x_1)^2 + (y_2 - y_1)^2)} \quad (4)$$

[7]

3. Introducción

La utilización de métodos de interpolación es una de las herramientas matemáticas más comúnmente empleadas par el estudio de las variables climatológicas, toda vez que es imposible realizar una medición continua en tiempo y espacio [6]La situación antes descrita lleva a que existan espacios temporales o geográficos en los que no hay datos exactos sobre las mediciones climatológicas, por la cual resulta útil emplear métodos numéricos que permitan reconstruir funciones a partir de los datos conocidos capaces de ser empleadas para conocer aproximaciones de los valores no conocidos.

Justamente la situación antes descrita es el foco del presente texto, donde se parte un set de datos climatológicos recolectados en la zona cercana a Fortaleza, Brasil, para poder conocer a partir de los mismos el parte del conjunto de datos no conocidos. Puntualmente, el interés del presente artículo girará en torno a la información de la estación de Sobral. Tal como fue mencionado anteriormente, se trabaja en los dos tipos de discontinuidades en las mediciones, la temporal y la espacial.

A partir de lo anterior, se inicia la exposición de una interpolación con la información interna de Sobral para solucionar la discontinuidad temporal y posteriormente se buscará predecir los datos de Sobral a partir de la información de su estación más cercana, a manera de ejemplo de cómo abordar la discontinuidad espacial.

4. Parte 1

En la primera fase del presente ejercicio se busca interpolar la información disponible de la estación de Sobral para poder llegar a una función capaz de predecir datos de la misma estación en periodos en los que no se tienen datos exactos. Con esto en mente, se opta por probar dos métodos de interpolación para poder compararlos y conocer cuál de ellos ofrece mejores resultados, siendo esta interpolación por splines e interpolación lineal. Específicamente se estarán interpolando los datos referentes a la temperatura interna, toda vez que es una variable apta para este tipo de ejercicios y suele generar curvas suaves.

La estación de Sobral presenta un reto particular, dado que la toma de sus datos no se hizo en intervalos temporales regulares, sino que existen espacios temporales amplios sin muchas mediciones y espacios temporales cortos con una gran cantidad de cantidad de mediciones. Con esto en mente se define que para toda esta sección y la parte dos, la función temperatura se expresará en función de una variable independiente definida por la fórmula $x = \text{día juliano} + \text{hora}/2400$. Esto permite reflejar la discontinuidad temporal de los datos y evitar errores que se generarían si se asumiera que todas las mediciones fueron realizadas en intervalos regulares.

Inicialmente, se muestran un diagrama de puntos que contiene la totalidad de datos sobre la estación de Sobral, evidenciándose la discontinuidad temporal antes descrita.

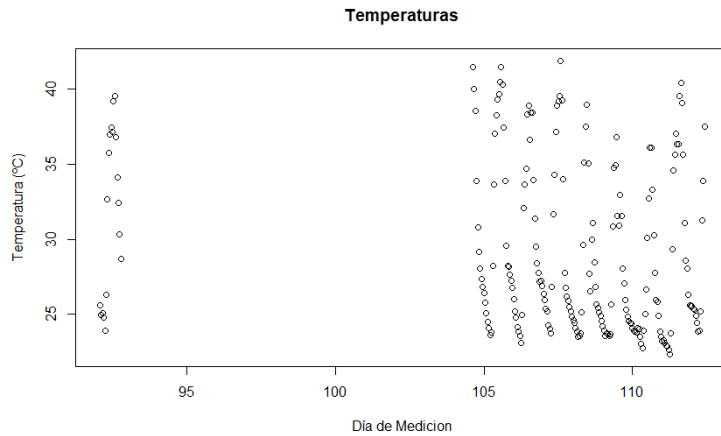


Figura 1: Diagrama con la totalidad de los puntos

Para poder calcular el error de cada método de interpolación utilizados se elimina una sexta parte de los datos disponibles, de tal manera que hubiese un set de puntos no contenidos en el algoritmo de interpolación contra los cuales realizar la comparación final.

A continuación, se expone la gráfica de interpolación por splines en rojo. En dicha gráfica se observa como en el intervalo de tiempo donde no hay mediciones la interpolación no presenta un comportamiento que se esperaría, para solucionar problemas como este se puede tomar la información de estaciones cercanas, como se verá en la parte dos.

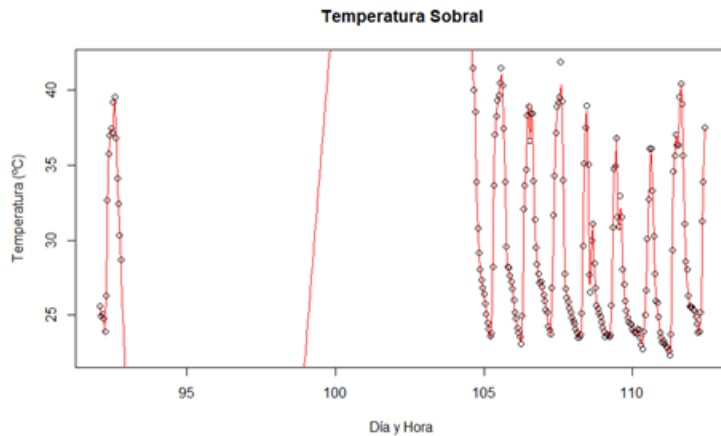


Figura 2: Gráfica de interpolación con Splines

Igualmente, la interpolación lineal, vista en verde, tampoco soluciona este problema, tal como se observa en la siguiente figura:

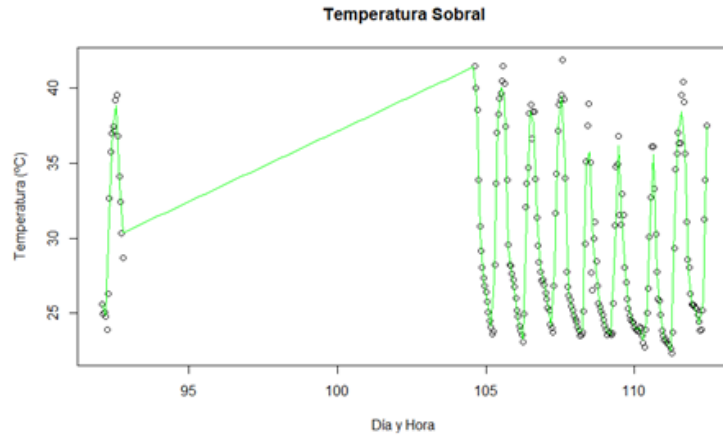


Figura 3: Gráfica con la interpolación lineal

Al realizar el cálculo de los errores de la forma antes descrita y basándose en las fórmulas, se obtiene la siguiente información:

Cuadro 1: Tabla de errores de la primera parte

Metodo	Error	Tipo	Raiz
Interpolación por Splines	ErrorAbsoluto	Minimo	0.019691
Interpolación por Splines	ErrorAbsoluto	Media	0.082803
Interpolación por Splines	ErrorAbsoluto	Máximo	2.318697
Interpolación por Splines	ErrorRelativo	Minimo	0.000646
Interpolación por Splines	ErrorRelativo	Media	0.002666
Interpolación por Splines	ErrorRelativo	Máximo	0.072278
Interpolación por Splines	Indice de Jaccard	Minimo	0.75862
Interpolación Lineal	ErrorAbsoluto	Minimo	0.005
Interpolación Lineal	ErrorAbsoluto	Media	0.13167
Interpolación Lineal	ErrorAbsoluto	Máximo	2.76
Interpolación Lineal	ErrorRelativo	Minimo	0.000646
Interpolación Lineal	ErrorRelativo	Media	0.004176
Interpolación Lineal	ErrorRelativo	Máximo	0.086253
Interpolación Lineal	Indice de Jaccard	Minimo	0.75862

Como se puede observar, los errores calculados indican que el método de Splines cuenta con un mejor desempeño, sin embargo, el análisis a profundidad de estos datos se expondrá en detalle más adelante.

5. Parte 2

Se quieren generar datos de de proyección a partir de otra estación. Para empezar se hizo la lectura de datos segun la entregada. En este caso utilizamos la ciudad de Sobral para nuestro experimento e hicimos una iteración para encontrar el índice de la misma. Se capturaron las coordenadas de la ciudad, Seguido de esto se encontraron las coordenadas euclidianas para calcular la distancia y se obtuvieron los datos de la ciudad más cercana a Sobral.

Se utilizaron los datos de las coordenadas de Sobral para la obtención de los datos de la temperatura en esta ciudad.

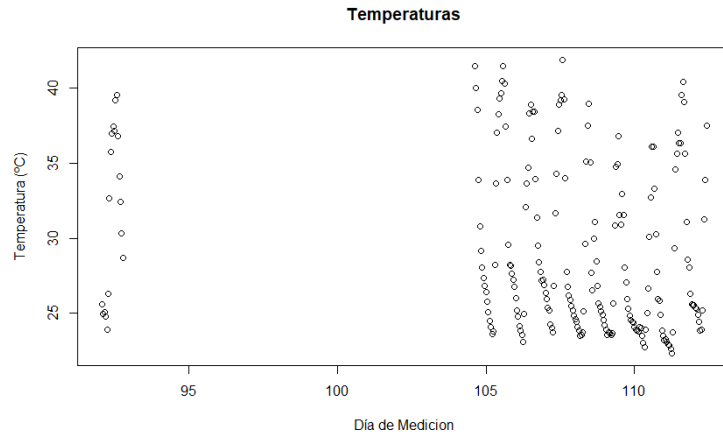


Figura 4: Medición de las temperaturas de la ciudad de Sobral

El tiempo en blanco que se observa en la gráfica quiere decir que no hay datos dentro de ese intervalo de tiempo

Se observa en la figura 1 unos puntos bastante altos y otros bastante bajos en los días cercanos frente a temperatura.

También se utilizaron los datos de la ciudad más cercana para sacar la función Spline de la temperatura de la ciudad más cercana (Santa Quiteria), la cual se muestra a continuación:

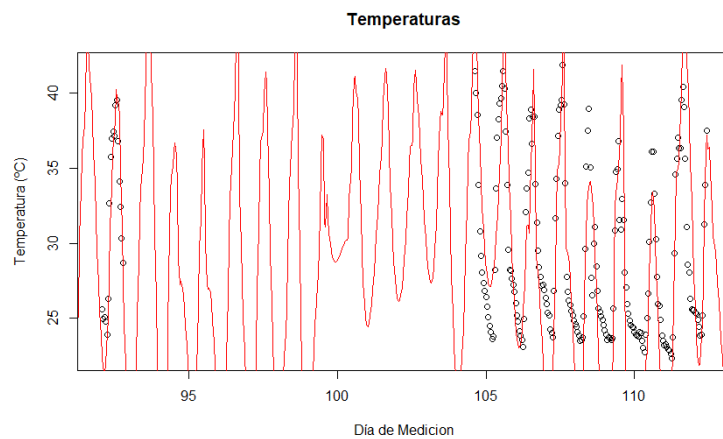


Figura 5: Spline de Quiteria junto con temperaturas de Sobral

Se observa que muchos de los puntos de Sobral se unen con los puntos de Santa Quiteria dando lugar a una conexión en las temperaturas de ambas ciudades.

Por ultimo se compararon los datos de la temperatura de sobral contra la linea de interpolación de Santa Quiteria, se muestra a continuación.

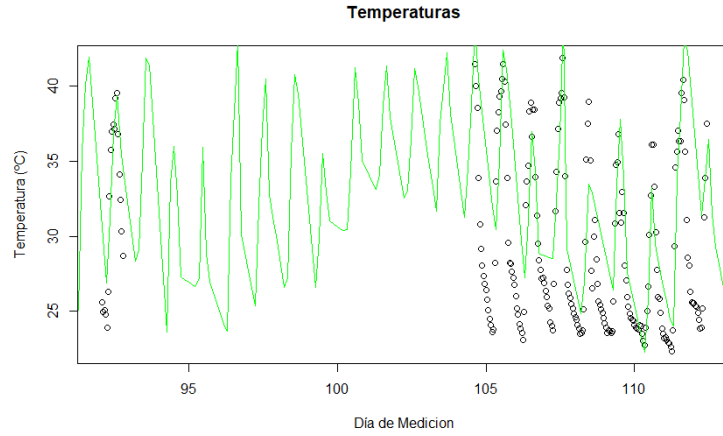


Figura 6: Interpolación con los datos de Santa Quiteria contra las temperaturas de Sobral

Se observa también en la interpolación una intersección en los puntos.

Los errores que se obtuvieron por parte de los datos del Spline y de la interpolación son los siguientes:

Cuadro 2: Tabla de errores de la segunda parte

Metodo	Error	Tipo	Raiz
Interpolación por Splines	Error Absoluto	Minimo	0.06
Interpolación por Splines	Error Absoluto	Media	4.223175
Interpolación por Splines	Error Absoluto	Máximo	20.56
Interpolación por Splines	Error Relativo	Minimo	0.001601281
Interpolación por Splines	Error Relativo	Media	0.1562112
Interpolación por Splines	Error Relativo	Máximo	0.8412466
Interpolación Lineal	Error Absoluto	Minimo	0.06
Interpolación Lineal	Error Absoluto	Media	3.929855
Interpolación Lineal	Error Absoluto	Máximo	13.71
Interpolación Lineal	Error Relativo	Minimo	0.001601281
Interpolación Lineal	Error Relativo	Media	0.1410859
Interpolación Lineal	Error Relativo	Máximo	0.5259664

Los errores son lo suficientemente bajos para concluir que las figuras 1,2 y 3 son confiables pero el método de Interpolación lineal tiende a ser más confiable.

6. Análisis

Gracias a la interpolación, se pueden obtener datos e información necesaria para distintos problemas. Al realizar una interpolación en la información de la estación Sobral, se llegó a la captura de información útil para proceder a una estimación cercana a la estación establecida. Al compararlos con el tipo de interpolación Spline y lineal ya que estos dos disponen de buenos resultados, pero se busca el mejor. Se obtuvo como decisión de interpolación, que Spline es mucho más acertado y con menos error absoluto que el lineal para la segunda realización del reto.

Todo esto con el fin de poder realizar un aceptable acercamiento a nuestro objetivo planteado anteriormente.

Al tener distintos datos del data set y teniendo en cuenta que los dos métodos de interpolación tienen gran efectividad, se encuentran desfases en los resultados obtenidos en la primera parte del reto con la segunda, ya que en la primera no se obtiene como mejor opción interpolación spline sino lineal.

7. Conclusiones

En conclusión el método de interpolación de Spline es mejor al interpolar por disminuir el error en su mayoría

Además se puede concluir que dependiendo de los puntos ambos métodos manejen los mismos errores.

Referencias

- [1] Deep AI. *Jaccard Index*. url<https://deepai.org/machine-learning-glossary-and-terms/jaccard-index>). 2021.
- [2] Antonio Miguel Posadas Chinchilla. *Spline*. url<https://encyclopediaofmath.org/wiki/Spline>. 2021.
- [3] EncyclopediaOfMath. *Spline Interpolation*. urlhttps://encyclopediaofmath.org/wiki/Spline_interpolation. 2021.
- [4] EncyclopediaOfMath1. *Spline*. url<https://encyclopediaofmath.org/wiki/Spline>. 2021.
- [5] ScienceDirect. *Linear Interpolation*. url<https://www.sciencedirect.com/topics/engineering/linear-interpolation>. 2021.
- [6] Raymond Sluiter. *Interpolation Methods for climate data*. url<https://www.researchgate.net/publication/242783501>. 2009.
- [7] Econ UPF. *Measures of distance between samples: Euclidean*. url<http://www.econ.upf.edu/michael/stanford/maeb4.pdf>. 2021.