



Reto 1

Cuestas Samy, Escobar Carlos, Hernández Camilo, Mendieta Juan Camilo

1. Descripción del problema

Dado un conjunto de valores asociados a variables climáticas de estaciones cercanas a la zona de Fortaleza, en Brasil, que están indexados en el tiempo y el espacio, entonces se pide determinar numéricamente:

1. Los valores de la variable Y cada media hora en una estación de monitorero seleccionada, utilizando interpolación ó ajuste de curvas.
2. Los valores de la variable Y cada hora en una estación de monitoreo, utilizando los datos de una estación cercana.

2. Metodología

Partiendo del supuesto de que las variables climáticas admiten interpolación polinómica y que además estas son ajustables a un modelo teórico, lo primero que se hizo fue la selección de una estación, en este caso al tratarse de interpolación, se seleccionó la estación de Itatira. La razón de la selección de Itatira fue porque es una de las estaciones que tiene sus datos completos, es decir tiene 721 celdas donde cada fila representa el paso de una hora. Una vez hecho esto, el paso a seguir fue la selección de la variable cuyos datos servirían para realizar la interpolación, para esto se seleccionó la temperatura interna, esto debido a que es uno de los datos en Itatira que está completo ya que en ninguna de sus casillas existen datos NaN o con valor -6900 que representa lo mismo. Finalmente se seleccionaron los métodos de interpolación, siendo estos la interpolación polinómica de Lagrange y Splines cúbicos. A continuación se muestra el diagrama de flujo correspondiente a Splines cúbicos (principal método de interpolación para lograr el objetivo del reto).

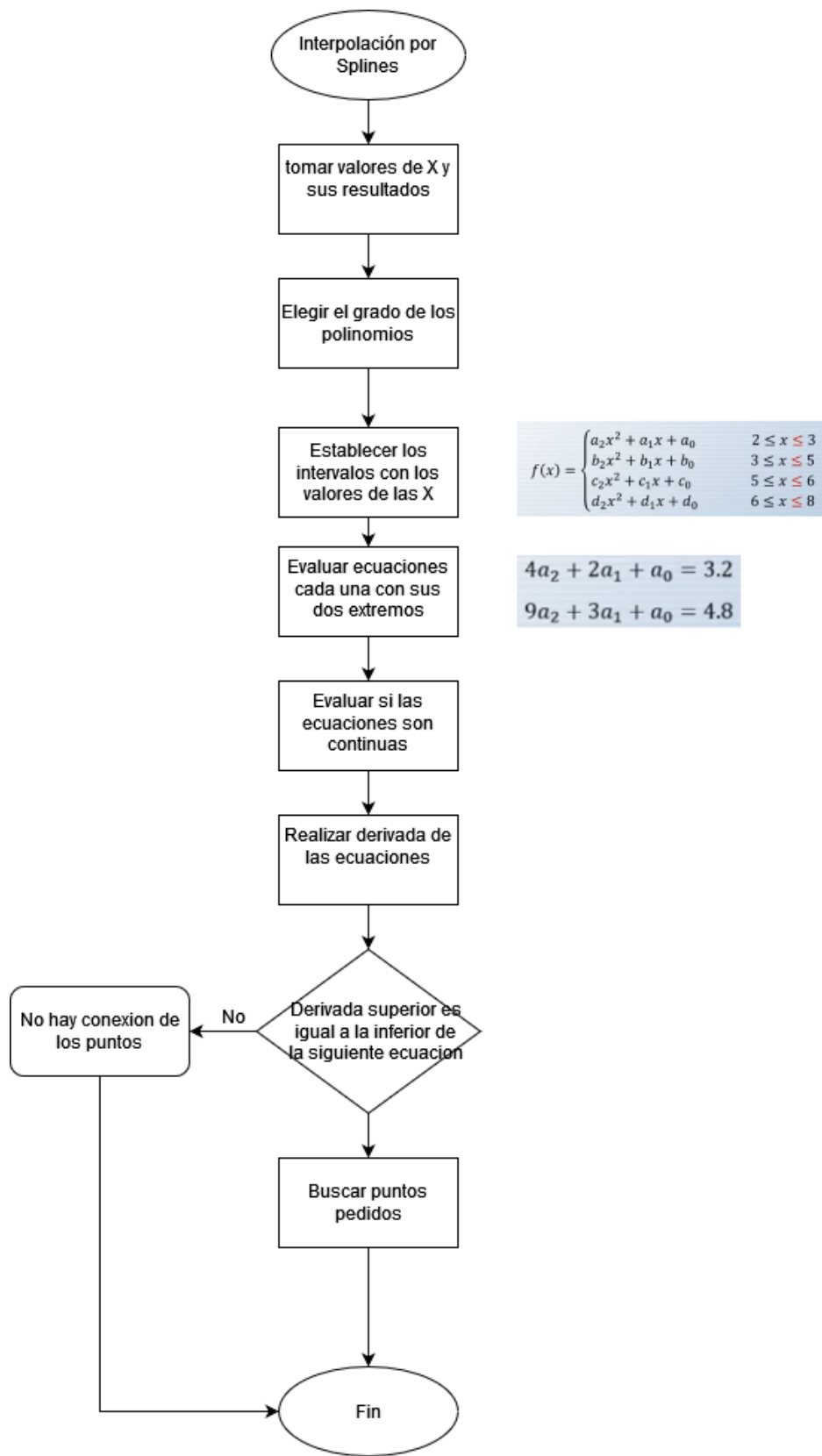


Figura 1: Diagrama de flujo del método de Splines cúbicos

La aproximación polinomial por partes más común utiliza polinomios cúbicos entre cada par sucesivo de nodos y se denomina interpolación spline cúbica. Una interpolación polinomial cúbica general involucra cuatro constantes, por lo que hay suficiente flexibilidad en el procedimiento de spline cúbico para asegurar que el interpolante no solo sea continuamente diferenciable en el intervalo, sino también tiene una segunda derivada continua. La construcción del spline cúbico no lo hace. Sin embargo, se supone que las derivadas del interpolante concuerdan con las de la función que está aproximándose.

Una vez definidos los métodos de interpolación, es necesario tener en consideración que, primero, se debe eliminar el 30 % de los datos disponibles, sin embargo, entre estos datos no debe encontrarse el primer ni el último dato de la tabla. También hay que tener en cuenta la función utilizada en R studio para realizar la interpolación, en este caso Splinefun, la cual recibe como entrada los puntos a interpolar en un arreglo de datos, retorna los puntos interpolados o una función que cumple con las especificaciones de estos puntos. Por último, considerar que el error debe calcularse en el 30 % de los datos que se omitieron, además de que se utilizarán 2 cifras significativas.

Para el segundo punto, tuvimos en cuenta la aplicación real para dar una solución, en este caso con el uso de estaciones cercanas para obtener los datos de una zona, pues la interpolación es utilizada en mapas de isolíneas (por ejemplo, mapas de temperatura) ya que se tienen mediciones obtenidas con estaciones meteorológicas pero se debe estimar con estos datos la temperatura en zonas donde no hayan estaciones posicionadas. Ya sabiendo esto, se procedió a calcular las distancias de las estaciones partiendo del plano dado, esto se hizo tomando como referencia a Itatira, las estaciones seleccionadas para la obtención de los datos fueron Quixada (la más cercana) y Santa Quiteira (la más lejana), posterior a esto se realizaron 2 tipos distintos de interpolación, como se evidencia en los resultados obtenidos en la sección siguiente, esto se debe a que la mitad fue hecha con Quixada y la otra mitad con Santa Quiteira. El paso a seguir fue promediar los datos de ambas estaciones para obtener un mejor y más completo resultado.

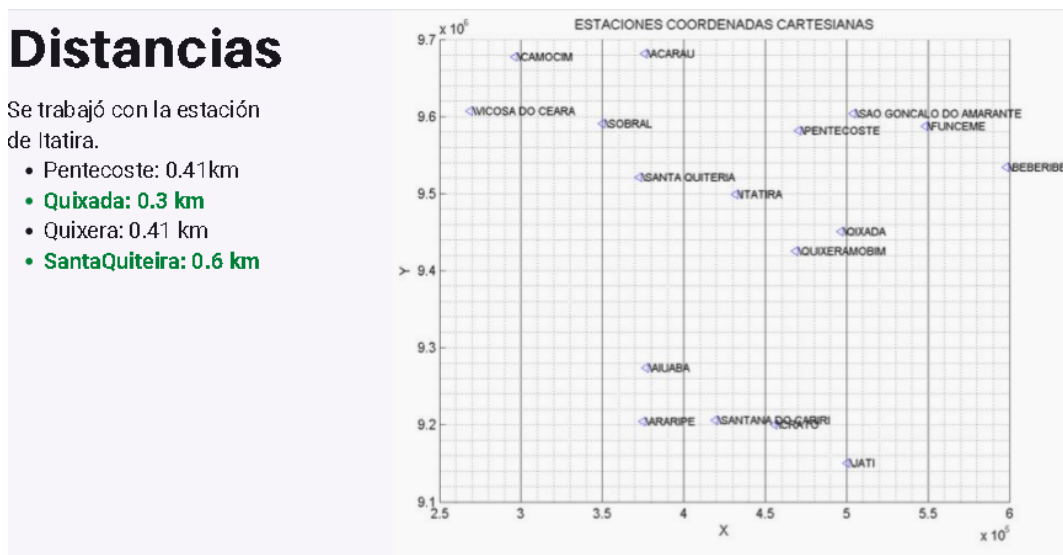


Figura 2: Distancias de las estaciones seleccionadas a la estación de Itatira

3. Implementación

Como se mencionó anteriormente el método a implementar en este caso para realizar la interpolación fue el de los splines cúbicos y el programa utilizado fue R Studio. Para la implementación de Spline cúbico se usó la función `splinefun`, esta realiza una interpolación spline cúbica de puntos de datos dados, devolviendo una lista de puntos obtenidos por la interpolación o una función que realiza la interpolación. para el uso de esta función en el programa solo se hizo uso de 2 parámetros, X e Y, los cuales son los vectores que dan las coordenadas de los puntos a interpolar (Hora y temperatura interna). Por otro lado se utilizó también una librería con el fin de leer el archivo de excel que contenía los datos del clima, se le especifica como parámetro la ruta que contiene el archivo de excel, es necesaria la modificación de esta para su correcto funcionamiento en otro dispositivo, pues la ruta puede ser diferente. Finalmente se hace uso de arreglos con el fin de hacer la parte de la eliminación de los datos y se utiliza la función `plot` para graficar. Una vez ejecutados los programas se obtuvieron distintos resultados. (Las implementaciones se encuentran en sus respectivos repositorios).

4. Resultados y error

Con respecto a los resultados, los errores a calcular fueron el error absoluto, el error relativo, la sumatoria de la diferencia cuadrada de errores, el error absoluto medio, que en este caso en concreto representaría la diferencia de la temperatura o el error con respecto a la temperatura que se podría expresar o entender como los grados de diferencia y, finalmente, el índice de Jaccard, el cual simplemente mide el grado de similitud entre los dos conjuntos de datos.

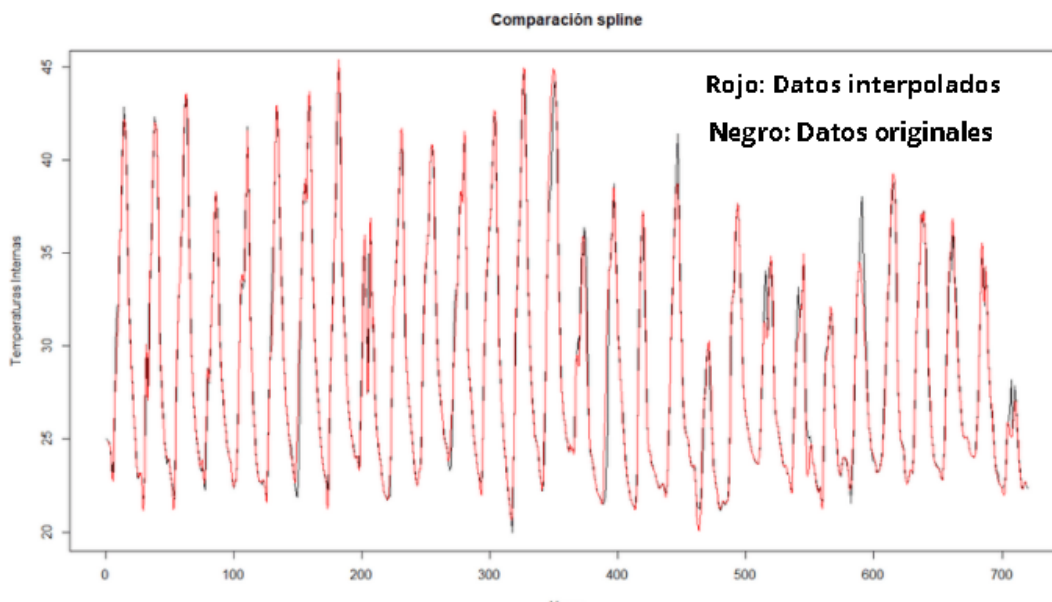


Figura 3: Resultados punto 1: Primer intento de interpolación con splines

Error absoluto: 0.01 / 6.87

Error relativo: 0.01 / 0.3

Sumatoria diferencia
cuadrada errores: 0.47

Error absoluto medio: 0.88

Índice de Jaccard: 0.59

Figura 4: Resultados punto 1, intento 1: Errores

En este caso podemos resaltar que el error absoluto medio es de 0,88 o sea que casi 1 grado de diferencia en lo que refiere a la temperatura.

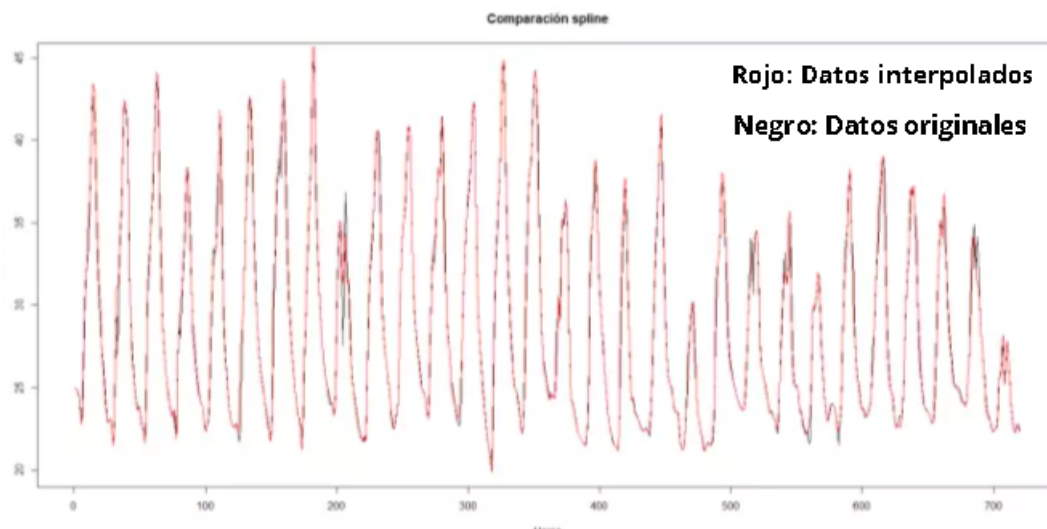


Figura 5: Resultados punto 1: Segundo intento de interpolación con splines

Error absoluto: 0.01 / 4.81

Error relativo: 0.01 / 0.17

Sumatoria diferencia
cuadrada errores: 0.4

Error absoluto medio: 0.64

Índice de Jaccard: 0.59

Figura 6: Resultados punto 1, intento 2: Errores

En este caso podemos resaltar que el error absoluto medio es de 0,64 aún menor que en el anterior intento donde era más cercano a 1 y podría expresarse en grados si lo queremos relacionar con la temperatura.

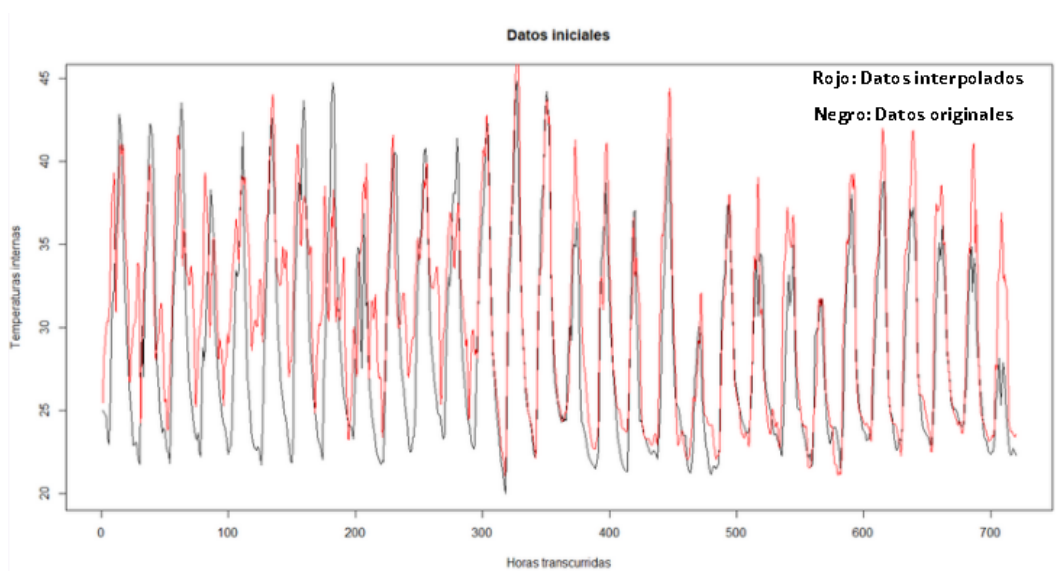


Figura 7: Resultados punto 2: Datos sin promediar (Quixada/Santa Quiteira)

Error absoluto mínimo -> 0.01
Error absoluto máximo -> 11.8
Error relativo máximo -> 0.54
Error relativo mínimo -> 0.01
Error absoluto medio -> 2.8
Indice de Jaccard -> 0.23

Figura 8: Resultados punto 2, Datos sin arreglar: Errores

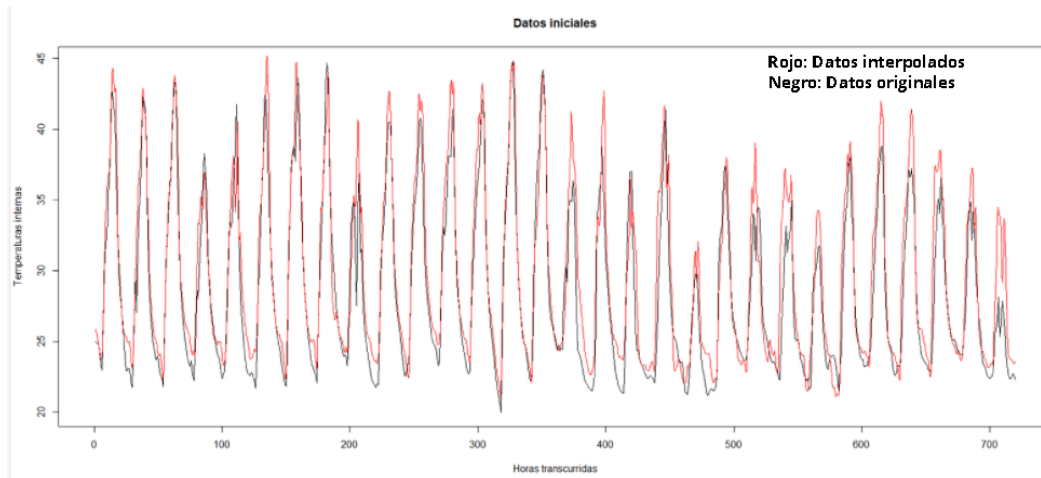


Figura 9: Resultados punto 2: Datos promediados (Quixada/Santa Quiteira)

Error absoluto mínimo [Ⓜ] ⁺ -> 0.01
Error absoluto máximo -> 9.01
Error relativo máximo -> 0.34
Error relativo mínimo -> 0.01
Error absoluto medio -> 1.6
Indice de Jaccard -> 0.24

Figura 10: Resultados punto 2, Datos promediados: Errores

Finalmente, en este caso podemos resaltar que el error absoluto medio es de 1,6, que si bien supera el grado de "diferencia", es menor que en el anterior intento donde era más cercano a 3, teniendo en cuenta que tiene sentido que se arregle esta diferencia, pues ya no se presenta problemas con los datos al haber realizado un promedio para arreglarlos.

5. Bibliografía

[1] Burden Richard L y Faires J. Douglas, Numerical Analysis. Cengage Learning, 9ª. 2012.