Reto 2: Interpolación Clima

María Alejandra López Sandoval, Carlos Erazo, Pedro Guerrero, Valentina Rozo Bernal

***Resumen*—En este escrito se presenta la información de las diferentes estaciones del clima de una zona de Brasil, mediante la interpolación. Así mismo, se da una breve explicación de los métodos utilizados para hallar los valores faltantes, y el cálculo de otros valores, como lo son el índice de Jaccard y el error relativo.**

Adicionalmente todo resultado se sustenta mediante gráficas o tablas que explican su comportamiento. Para finalizar una manera óptima de ver en qué otras situaciones se utilizan los métodos presentados en este documento se ponen en evidencia diferentes aplicaciones que estos pueden tener.

***Palabras clave*— Spline, interpolación, distancia, jaccard, temperatura, estaciones, error relativo**

I. Introducción

Este documento presenta dos tipos de problemas que corresponden a la interpolación, además proporciona la correspondiente información de cómo se resolvió cada uno de ellos. En primer lugar, se presenta la definición de los métodos utilizados para solución del problema así como los datos que se tenían para sacar los resultados.

En segundo lugar, se le presenta al lector el procedimiento para resolver el problema del clima en ciertas estaciones ubicadas en Brasil, los resultados obtenidos, los respectivos errores y la comparación de los diferentes métodos que se implementaron durante el desarrollo, además de las diversas aplicaciones en las que se han utilizado estos temas.

II. Marco teórico

Para resolver los ejercicios se implementaron varios algoritmos como los siguientes:

- **Splines cúbicos** es una función que se usa para la interpolación o el suavizado de curvas. En este caso los polinomios son de grado 3, es decir de la forma:

$$P(x) = ax^3 + bx^2 + cx + d.$$

- **Interpolación lineal:** Es la estimación de una función asumiendo que existe una línea recta entre valores conocidos[5].
- **Distancia:** La distancia euclidiana se define como la distancia que es una línea recta entre dos puntos a y b [4]. Se utiliza la siguiente fórmula para hallar la distancia entre dos puntos:

$$d = \sqrt{(a2 - a1)^2 + (b2 - b1)^2}$$

- **Índice de Jaccard:** el índice de Jaccard, también conocido como coeficiente de similitud de Jaccard, es una estadística que se utiliza para comprender las similitudes entre conjuntos de muestras. La medición enfatiza la similitud entre conjuntos de muestras finitas y se define formalmente como el tamaño de la intersección dividido por el tamaño de la unión de los conjuntos de muestras [1]. La representación matemática del índice se escribe como:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

- **Error relativo:** con el fin de garantizar la calidad de una medida, el error relativo corresponde a el cociente entre el error absoluto y el valor exacto.

$$E_r = \frac{|Valor\ verdadero - Varlo\ aproximado|}{Valor\ verdadero}$$

III. Aplicaciones

Las problemática a tratar dentro de este documento son ejercicios que se han desarrollado varios años atrás, es por eso que dentro de este apartado se van a mostrar ejercicios similares a los ya planteados y cómo por medio de la interpolación y estimación se les pudo dar solución:

- **Regionalisierte Niederschlage (Regine):** es un modelo que se usa para interpolar la lluvia y la temperatura media, también por medio de herramientas de geoprocesamiento se generaron superficies interpoladas.
  Algo que es importante destacar de este modelo es que hace uso de uno de el método de interpolación de Splines, uno de los diferentes métodos que fueron vistos en la clase de análisis numérico.

- **Método Kriging:** es un método de inferencia espacial, mediante este se se pueden estimar los valores de una variable en lugares que no se encuentran muestreados, este método proporciona un estimador lineal con una varianza mínima.

IV. Descripción de los datos

Los datos se presentan en un archivo Excel .xls y un .csv, del primer archivo se obtienen los datos de diecisiete estaciones climáticas brasileras, cada una de ellas tiene datos relacionados con el clima, a continuación se presentan algunos de estos datos los siguientes datos:

- Año
- Día Juliano
- Hora
- Temperatura Interna Cº
- Presión atmosférica
- Temperatura del suelo Cº
- Precipitaciones
- Velocidad del viento

Estos datos están organizados en el documento excel como se presenta en la figura 1. De acá se selecciona la temperatura, los días y las horas para poder hacer las proyecciones y calcular los valores requeridos. Así mismo la información se calcula solo en una de las diecisiete estaciones.

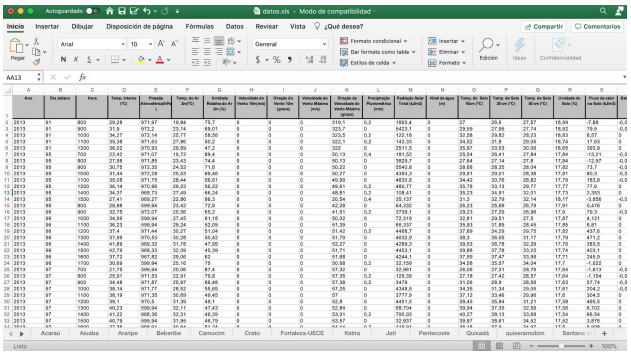

Fig. 1. Tabla excel datos de las 17 estaciones

Del archivo csv se obtienen los datos de las coordenadas de las diecisiete estaciones, y el nombre de cada una, la información presentada se utiliza con el fin de encontrar la estación más cercana a la elegida inicialmente Santa Quiteria para así poder predecir con esta alguna variable de la estación inicial. Se realiza de esta manera pues para proyectar los valores se debe contar con la estación más cercana para obtener resultados más precisos.

V. Validación cruzada

Con el objetivo de garantizar la veracidad de las herramientas que garantizan la interpolación se realiza una comparativa entre dos funciones relacionadas, finalmente se presentan los resultados de forma separada y un resumen conjunto. Se hace uso de la función “spline”[2] que dado una serie de puntos realiza una interpolación spline cúbica, devolviendo la lista de puntos obtenidos por la interpolación, adicionalmente se hace uso de la función “approx”[3] la cual a partir de una serie de puntos, retorna una lista de puntos resultado de una interpolación lineal.

A. Diagrama o pseudocódigo

La comparación se realiza con el mismo conjunto de datos, después de organizar la información se sigue el siguiente paso a paso:

- Graficar los datos iniciales.
- Eliminar el 20% de los datos de forma “aleatoria”.
- Implementar el primer método de interpolación y graficar.
- Implementar el segundo método de interpolación y graficar.
- Realizar una tercera implementación sumando los métodos.
- Obtener el error relativo de cada alternativa.
- Presentar resultados.

B. Gráficas y tablas

En este proceso se detallan un total de cuatro gráficas que corresponden a la presentación de los datos iniciales que para esta y las demás graficas conservan el color de “negro”, la comparación con el primer ajuste que corresponde a la implementación de la funcion “spline”, la comparación con el segundo ajuste que corresponde a la implementación de la función “approx” y una comparación resultante de combinar ambos métodos. Finalmente se presenta una tabla que resume los resultados de la implementación.

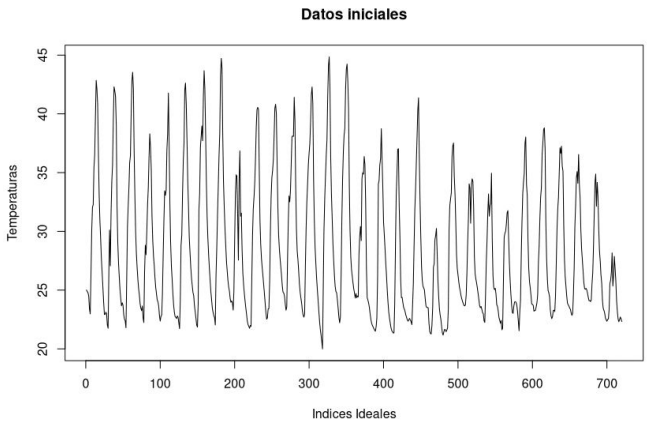

Fig. 2. Datos iniciales

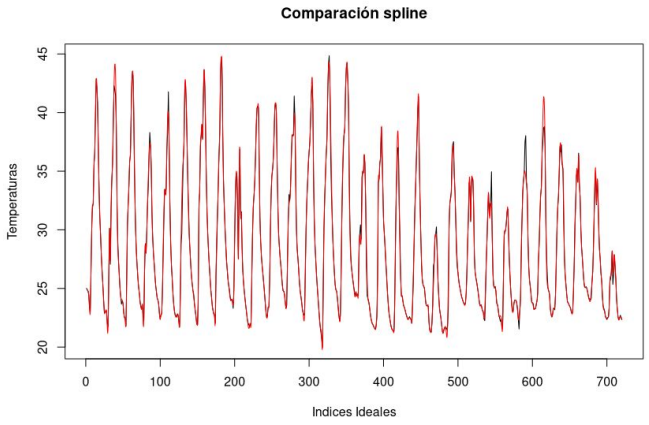

Fig. 3. Comparación con los datos obtenidos gracias a la función “spline”

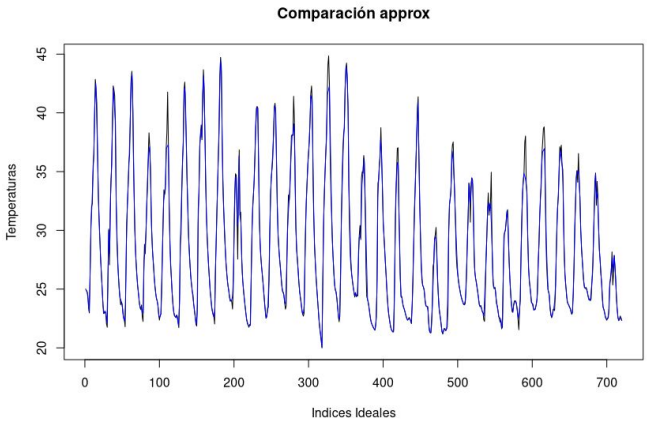

Fig. 4. Comparación con los datos obtenidos gracias a la función “approx”

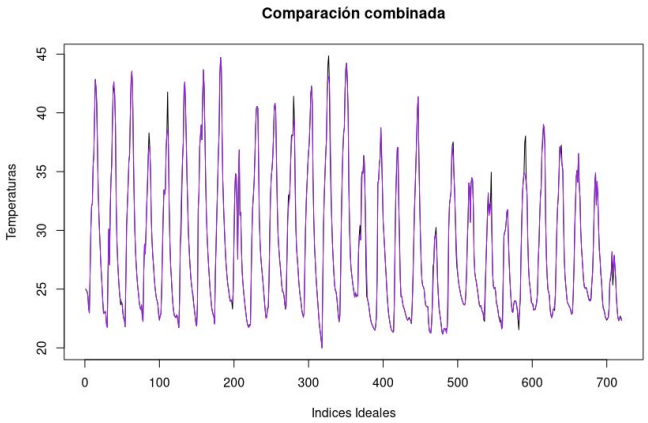

Fig. 5. Comparación con los datos resultantes de sumar los métodos de “spline” y “approx”

	Spline	Approx	Comb
# de errores	102	105	98
Error máximo	0.1	0.12	0.11
Error mínimo	0.01	0.01	0.01
Error medio	0.02	0.03	0.02
Índice de Jaccard	0.8583	0.8542	0.8639

TABLA I (RESULTADOS)

C. Análisis y conclusiones

Tanto el uso de la función “spline” y la función “approx” evidencian fallas notorias en sus respectivas gráficas, pero mantienen un error por debajo del 25% determinando que ambos métodos son confiables. Al unir ambos métodos se reduce el número de errores obtenidos y el índice de Jaccard aumenta, cabe resaltar que en todos los casos el índice supera el 80% confirmando que siempre las funciones de interpolación logran “recuperar” parte de los datos de manera acertada, asimismo el índice al momento de unir los métodos en comparación a los métodos por separado es mayor, si bien no es mucho, frente a un aumento en la cantidad de datos puede llegar a ser representativo.

VI. PROYECCIÓN DE DATOS A PARTIR DE OTRA ESTACIÓN

Este punto consiste en la proyección de datos basados en la estación más cercana a la seleccionada. Los datos que se utilizan se encuentran en un archivo separado por comas (csv), aquí se obtienen las coordenadas donde se encuentran cada una de las estaciones y su respectivo nombre. En este caso se tomó como punto de referencia la estación de Santa Quiteria y se pretende hallar la estación más cercana para así, poder predecir la variable de la temperatura de la estación de Santa Quiteria basado en la información de la estación más cercana.

A. Diagrama o pseudocódigo

El código se inicia de la siguiente manera para definir cuál es la estación más cercana a la seccionada, en este caso Santa Quiteria. Luego se procede a completar los datos de la estación Santa Quiteria basada en la estación más cercana hallada en el punto anterior para interpolar y hallar la temperatura de Santa Quiteria basado en la estación más próxima:

- Leer datos del archivo coordenadas.csv
- Mostrar una gráfica con las coordenadas de cada estación
- Sacar la distancia euclidiana de la estación Santa Quiteria hasta cada una de las otras estaciones
- Obtener el nombre y la distancia de la estación con menor distancia a la de Santa Quiteria.
- Leer los datos de la estación más cercana a Santa Quiteria.
- Elimina el 20% de los datos de forma “aleatoria”.
- Se verifica la información restante.
- Se muestra una gráfica comparando los datos originales con los datos restantes.
- Se interpolan los datos restantes.
- Se calcula el error relativo de los datos interpolados.
- Se estiman los datos de la estación de Santa Quiteria por medio de los datos que fueron interpolados de las estación más cercana.
- Se calcula el error relativo de la estimación de los datos de la nueva estación.
- Se muestra una gráfica comparando los datos originales con los datos interpolados.

A continuación, en la figura 6 se muestra un diagrama de flujo para ver el paso a paso de la programación de manera más simple.

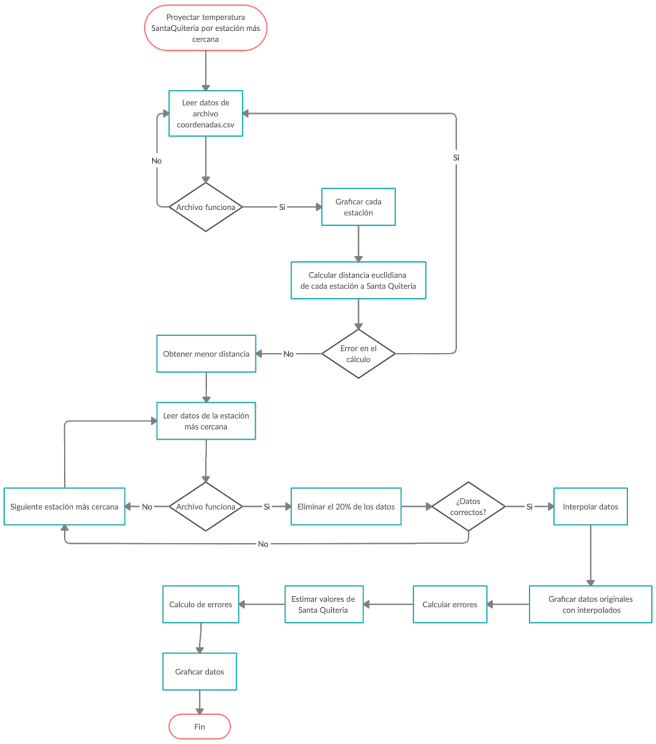

Fig. 6. Diagrama de flujo proyección de datos de Santa Quiteria basado en la estación más cercana

B. Gráficas y tablas

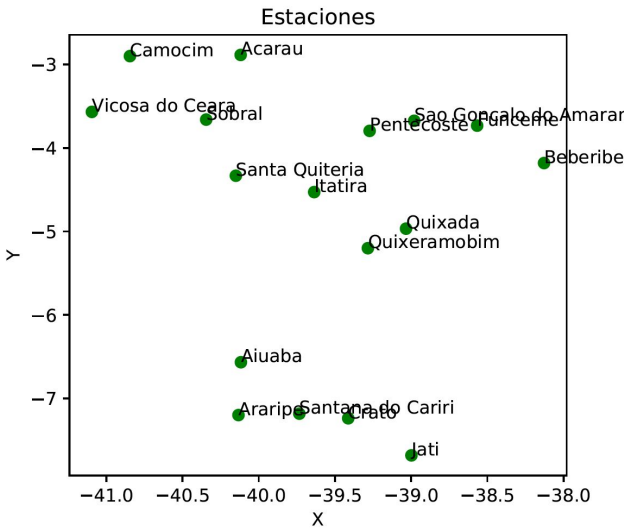

Fig. 7. Estaciones graficadas según sus coordenadas

Nombre Estación	Distancia a Santa Quiteria
Vicosa do Ceara	1.215315240505757
Sobral	0.7000080839500633
Sao Goncalo do Amarante	1.3425411868551271
Santana do Cariri	2.880383881250888
Quixada	1.2840244879343166
Pentecoste	1.0290357841275968
Jati	3.542001878073485
Itatira	0.5500478756414768
Funceme	1.6930797302400145
Crato	2.9974448632698385
Camocim	1.5926572556782543
Acarau	1.4473891886141785
Araripe	2.8668921718210805
Aiuaba	2.234250488785324
Beberibe	2.0278747263272177
Quixeramobim	1.2261613350202576

TABLA II: Distancia de cada estación a Santa Quiteria

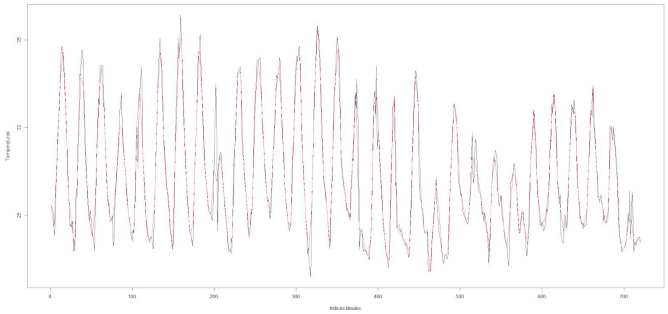

Fig. 8. Datos originales vs datos utilizados

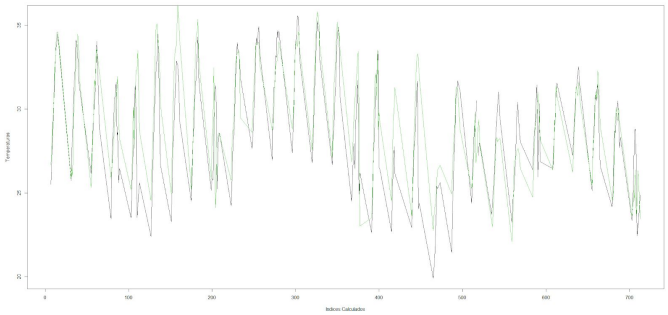

Fig. 9. Datos originales vs datos interpolados

C. Análisis y conclusiones

En la primera parte de este ejercicio se utilizó la distancia euclidiana, el cálculo de esta distancia se realiza con su norma para tener un control sobre sus valores ya que este método es sensible a las unidades y a los valores. Esta distancia se puede utilizar en este caso puesto que las unidades de las variables son iguales en todos los casos. Se obtuvo a la estación Itatira como la más cercana a Santa Quiteria con un valor de 0.5500478756414768, por esta razón se toma esta estación para poder predecir las variables de Santa Quiteria mediante la interpolación.

Para la segunda parte se utilizó el método de los splines cúbicos para realizar la interpolación y hacer la correcta proyección de los datos de la estación de Santa Quiteria a partir de la más cercana, la cual fue la estación de Itatira. Al realizar esta parte se pudo concluir que la reconstrucción del modelo se pudo completar al eliminar 20% de la información real y al realizar la respectiva comparación con los datos como se puede observar en la Figura 8 los datos son bastante aproximados comparados con los datos originales. Después al realizar la interpolación de los nuevos datos y compararlos podemos evidenciar en la Figura 9 que se presenta un error significante comparado con los datos originales, esto se debe a la eliminación del 20% de los datos.

Agradecimientos

A la profesora Eddy Herrera dado el apoyo y conocimiento que brindó a los autores de este trabajo.

Referencias

[1] "Jaccard Index", DeepAI. [En línea]. Disponible en: https://deepai.org/machine-learning-glossary-and-terms/jaccard-index#:~:text=Breaking%20down%20the%20formula%2C%20the,either%20set%2C%20multiplied%20by%20100.&text=Accordingly%2C%20to%20find%20the%20Jaccard,the%20Jaccard%20distance%20(1%20%2D%20. [Accedido: 09- Nov- 2020].

[2] "splinefun function | R Documentation", Rdocumentation.org. [En línea]. Disponible en: https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/splinefun. [Accedido: 09- Nov- 2020].

[3] "approxfun function | R Documentation", Rdocumentation.org. [En línea]. Disponible en: https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/approxfun. [Accedido: 09- Nov- 2020].

[4] NIST, "Euclidean Distance", *Itl.nist.gov*, 2017. [En línea]. Disponible en: https://www.itl.nist.gov/div898/software/dataplot/refman2/auxillar/euclidst.htm.

[5] Merriam-Webster, "Definition of LINEAR INTERPOLATION", *Merriam-webster.com*. [En línea]. Disponible en: https://www.merriam-webster.com/dictionary/linear%20interpolation.

[6] D. Velásquez, G. Carrillo, E. Barbosa, D. Latorre and F. Maldonado, "INTERPOLACION REGNIE PARA LLUVIA Y TEMPERATURA EN LAS REGIONES ANDINA, CARIBE Y PACÍFICA DE COLOMBIA", *Redalyc.org*, 2018. [En línea]. Disponible en: https://www.redalyc.org/jatsRepo/4239/423954588009/html/index.html. [Accessed: 12- Nov- 2020].

[7] A. Porras, "Diplomado en Análisis de Información Geoespacial", *Centrogeo.repositorioinstitucional.mx*. [En línea]. Disponible en: https://centrogeo.repositorioinstitucional.mx/jspui/bitstream/1012/160/1/16-M%C3%A9todo%20Kriging%20de%20Inferencia%20espacial%20-%20%20Diplomado%20en%20An%C3%A1lisis%20de%20Informaci%C3%B3n%20Geoespacial.pdf. [Accessed: 12- Nov- 2020].