Buen trabajo en general, pero mucho cuidado con la redacción y el lenguaje técnico, no podemos escribir como hablamos.. ahí les subrayé un par de cosas que deben revisar.

4.1

Análisis de datos ambientales

Tarea No. 1

Maria Angélica Aguirre López CC. 1 053 852 571

Cristian Alejandro Usma Rojas CC. 1 017 246 409

I. Descripción de serie de datos.

Los datos trabajados fueron obtenidos en formato csv de la estación meteorológica ubicada en la sede del Área Metropolitana del Valle de Aburrá (AMVA) en el barrio Alpujarra. Corresponden a mediciones de temperatura minuto a minuto para el mes de julio, que para efectos prácticos se limitaron a entre las 00:00 horas del día 1 de julio hasta las 23:59 del día 5 de julio del 2018, que suman 7200 entradas.

II. Graficación.

Siguiendo las recomendaciones planteadas, tras la lectura de datos se realizó una inspección visual de los datos. La figura 1 representa la gráfica de los valores de temperatura a través del tiempo para la serie en cuestión.

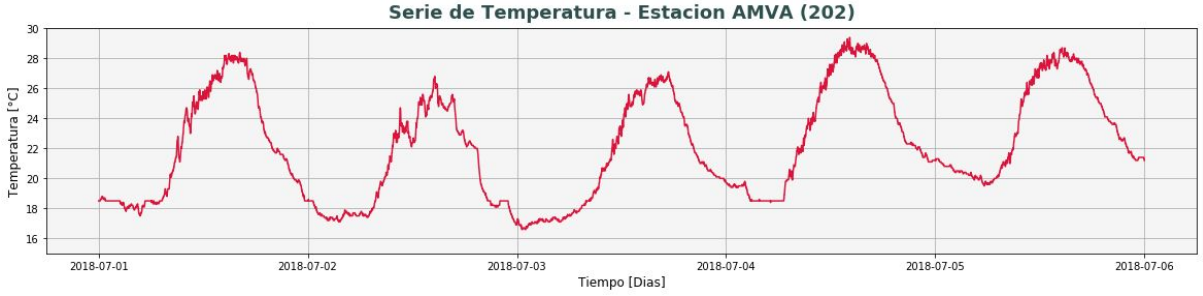

Figura 1. Temperatura vs. Tiempo

Debido a la resolución minutal de los datos, al graficarse se observa "ruido". Aún así, se puede notar la marcada trayectoria del ciclo diurno en el centro de Medellín. Es patente el hecho de que la temperatura asciende a lo largo de las horas de la mañana, hasta alcanzar sus valores máximos diarios en horas posteriores al mediodía, lo cual evidencia la importancia del efecto acumulativo de la radiación solar en la elevación térmica diurna, ya que a pesar de que la radiación incidente es menor en las horas de la tarde que en la mitad del día, este período tiende a presentar una temperatura mayor.

Otra observación, parte de la forma de los diferentes ciclos diurnos para cada uno de los días presentes en la serie de datos analizada. Los días 2 y 3 de junio, correspondientes a los días con las temperaturas más bajas entre los analizados, presentan unas interrupciones particulares (ver figura 4), especialmente en las horas

centrales del día. Esto puede considerarse una señal de presencia de nubosidad, la cual impide la llegada de parte de la radiación solar, afectando de esta manera el desarrollo completo del perfil de temperaturas, perfil que sí logra desarrollarse de una manera más acorde a la teórica en los días 1, 4 y 5 del mismo mes.

A grandes rasgos, puede observarse una tendencia leve a que en cada ciclo diurno las temperaturas aumenten; sin embargo, esto se confirmará al final del trabajo.

III. Características generales de la serie.

A manera de acercamiento inicial a la serie, se calcularon estadísticos paramétricos y no paramétricos, como se muestra a continuación:

- **Extremos**

  Mínimo: 16.6

  Máximo: 29.4

- **Localización**

  Media: 22.12
  Esta medida es coherente con la información disponible sobre Medellín, donde el mes del año con la más alta temperatura media multianual es julio, y cuyo valor es de 22.5°C (IDEAM,1999), muy cercano al presentado anteriormente.

  Mediana: 21.6
  El valor de la mediana, entendida como una medida de localización de la distribución de los valores de la serie de datos, tiene un valor coherente si se compara con la media, ya que se obtiene un valor relativamente cercano. La mitad de los datos de la serie se encuentran en este valor o por debajo.

- **Dispersión**

  Desviación estándar: 3.57
  Según esta medida de dispersión paramétrica, la variación media esperada de la totalidad de la serie de datos con respecto a la media aritmética es de 3.57 °C.

  Rango intercuartil: 6.69
  Análogamente al caso de la desviación estándar, el rango intercuartil es una medida de dispersión, pero en este caso es no paramétrica. Atendiendo a la definición de este estadístico, se puede concluir que el 50% de los datos de la serie se encuentran concentrados dentro de un rango de 6.69 °C, circundantes a la mediana (percentil 50).

- **Asimetría**

  Coeficiente de Yule-Kendall (medida de asimetría no paramétrica): 0.07
  Coeficiente de asimetría (medida paramétrica): 0.29

  Representa una serie cuyos valores están mayormente distribuidos al lado izquierdo de la media. En el caso del estadístico paramétrico, se entiende que gran parte de los datos son menores a 22.12°C.
  Esto puede explicarse bajo el hecho de que la longitud de la serie abarca 6 valles y 5 picos, es decir, más valores en horas de la noche y de la madrugada que del día como tal.

IV. Estimación de función de distribución de probabilidad.

Para ahondar más en el comportamiento del conjunto de datos, se presentan a continuación una serie de gráficas basadas en diferentes aspectos estadísticos de la misma.

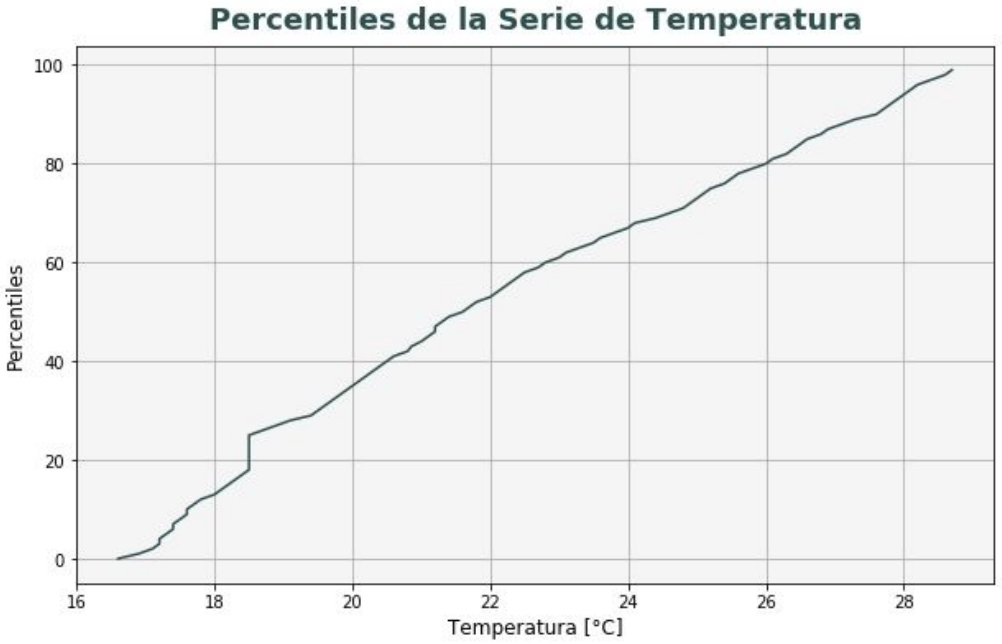

Figura 2. Temperatura vs. Percentiles.
(Función de distribución de probabilidad acumulada - CDF)

Como se observa en la figura 2, la probabilidad acumulada de la serie de datos crece de una manera relativamente uniforme a lo largo de las diferentes temperaturas que contiene. Sin embargo, se encuentra un aumento importante en torno a la temperatura de 18.5 °C, lo cual se puede interpretar como un pico en la función de distribución de probabilidad. El hecho anteriormente mencionado, puede constatarse en la gráfica de la función de distribución de probabilidad (ver figura 3).

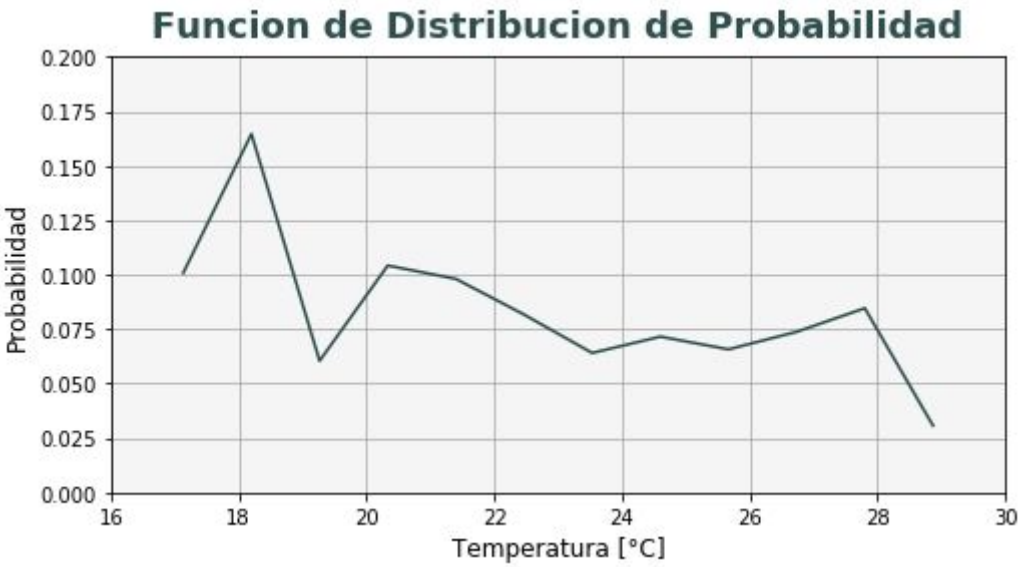

Figura 3. Temperatura vs. Probabilidad (en términos de fracción).
(Función de distribución de probabilidad - PDF)

En cuanto a la función de distribución de probabilidad de los datos, se puede observar que los valores tienden a concentrarse hacia el lado izquierdo del dominio de la función. Desde el punto de vista estadístico, esto es coherente con el hecho de que a lo largo del tiempo, el coeficiente de Yule-Kendall y el coeficiente de asimetría de la serie tengan un valor positivo, aunque con una breve disminución al final del día primero de julio (ver figura 10), lo cual implica que durante este período, el pico de la función de distribución de probabilidad de los datos se desplazó hacia la derecha.

En la figura 4 a continuación, se muestra de nuevo la función de distribución de probabilidad, pero en esta ocasión con marcas para algunos parámetros estadísticos de interés, tales como la media, la mediana, los límites de la desviación estándar, el máximo y el mínimo.

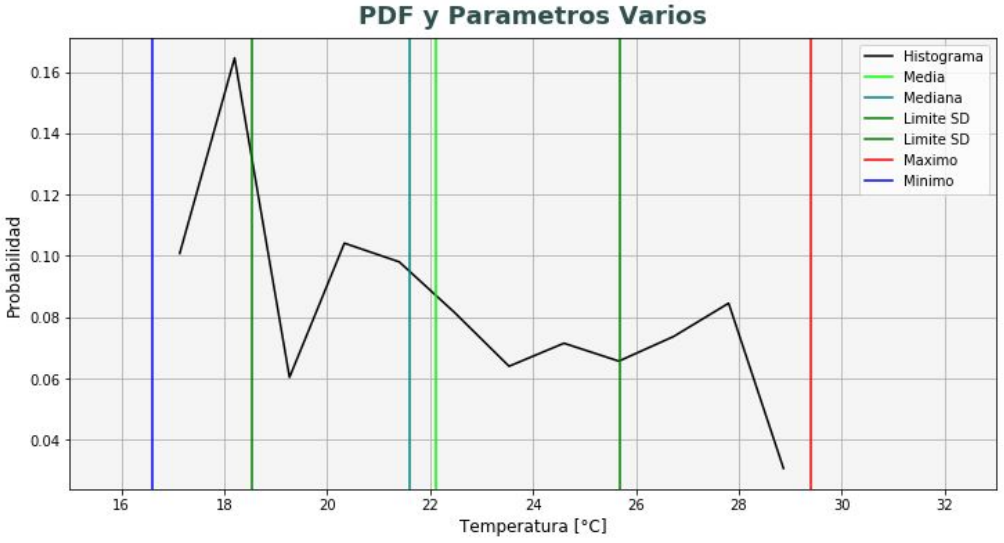

Figura 4. PDF y parámetros de la serie de datos.

V. Evaluación de la estacionariedad.

La estacionariedad se refiere a la continuidad de los estadísticos en el tiempo. Para verificar que en la serie de datos se da esta condición, se debe realizar un análisis por medio de tramos temporales o "ventanas". En esta ocasión se definió una ventana cuya longitud correspondiera a un día, es decir, a 60 datos por minuto multiplicados por las 24 horas del día, obteniendo de esta manera un total de 5760 datos por ventana. En la figura 5 se puede observar de manera gráfica el tamaño de la ventana.

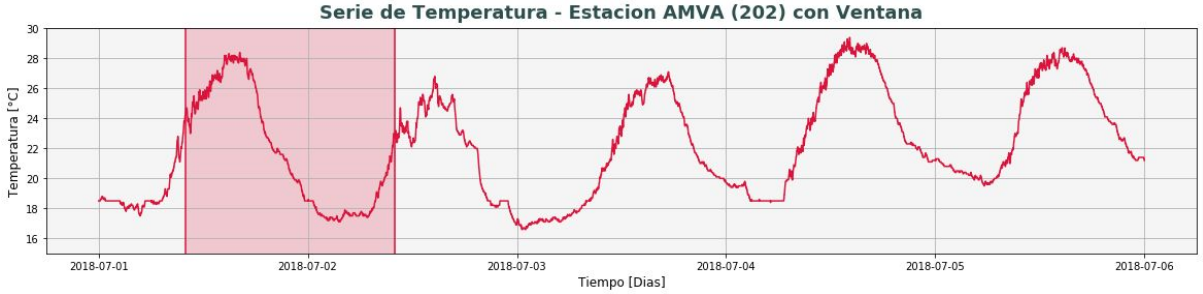

Figura 5. Serie de datos y ventana de análisis.

Dado que la estacionariedad se verifica por medio de la continuidad de los estadísticos (y por tanto, de la función de distribución de probabilidad en el tiempo), se calcularon todos estos elementos para cada una de las ventanas. Las gráficas resultantes se muestran a lo largo de esta sección.

En primer lugar, en la figura 6 se muestra la gráfica del histograma conjunto, el cual consiste en una representación gráfica de la función de distribución de probabilidad para cada ventana. Por lo tanto, puede interpretarse como una "PDF móvil". En esta gráfica, se presentan además las medidas de tendencia central móviles (media y mediana).

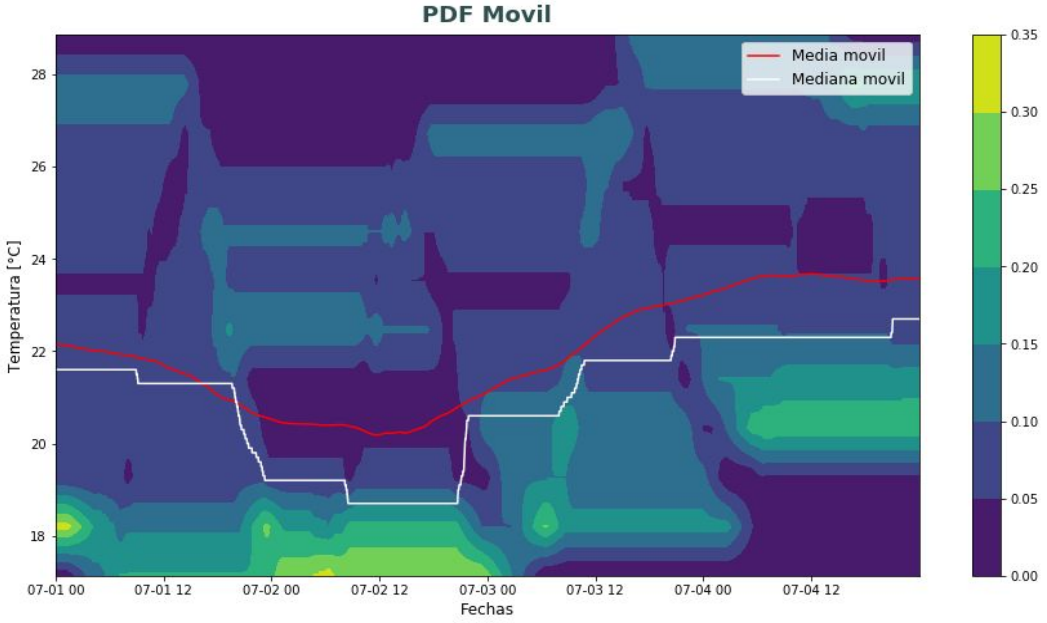

Figura 6. Histograma conjunto, media y medianas móviles.

Enfocándose en el histograma conjunto, puede observarse de manera muy patente que la serie no es estacionaria. Esto por el simple hecho de que la PDF evoluciona en el tiempo, y para que una serie pueda considerarse estacionaria, esta debe conservar absolutamente todos sus estadísticos, sin ninguna clase de tendencia. Es de notarse que, en general, las probabilidades de los histogramas tienden a concentrarse hacia los bordes inferiores del rango de temperatura. Esto es consistente con el hecho de que el coeficiente de asimetría y el de Yule-Kendall sean positivos.

La media tiene un comportamiento consecuente con los valores probabilísticos a lo largo de las diferentes ventanas. Presenta una reducción en los días 2 y 3 de julio, para luego empezar a crecer hasta el final del período de tiempo analizado. Esta tendencia es coherente con las observaciones que se pueden hacer en el histograma conjunto.

En cuanto a la mediana, esta tiene un comportamiento escalonado que, a pesar de diferir en relación a la media, presenta tendencias similares. Este comportamiento escalonado se presenta en los demás estadísticos no paramétricos calculados en este trabajo, y los cuales se tratan en las gráficas posteriores. Los estadísticos paramétricos tienden a ser medidas resistentes, por lo cual a medida que la ventana de análisis avanza, estos se ven poco afectados por la variabilidad (teniendo en cuenta además que la ventana tiene una duración de un día, correspondiente a la totalidad del ciclo que rige las variaciones mayores de temperatura), de manera que sus variaciones principalmente se dan entre días diferentes, permaneciendo relativamente uniformes a lo largo de cada uno de los días.

NO P.

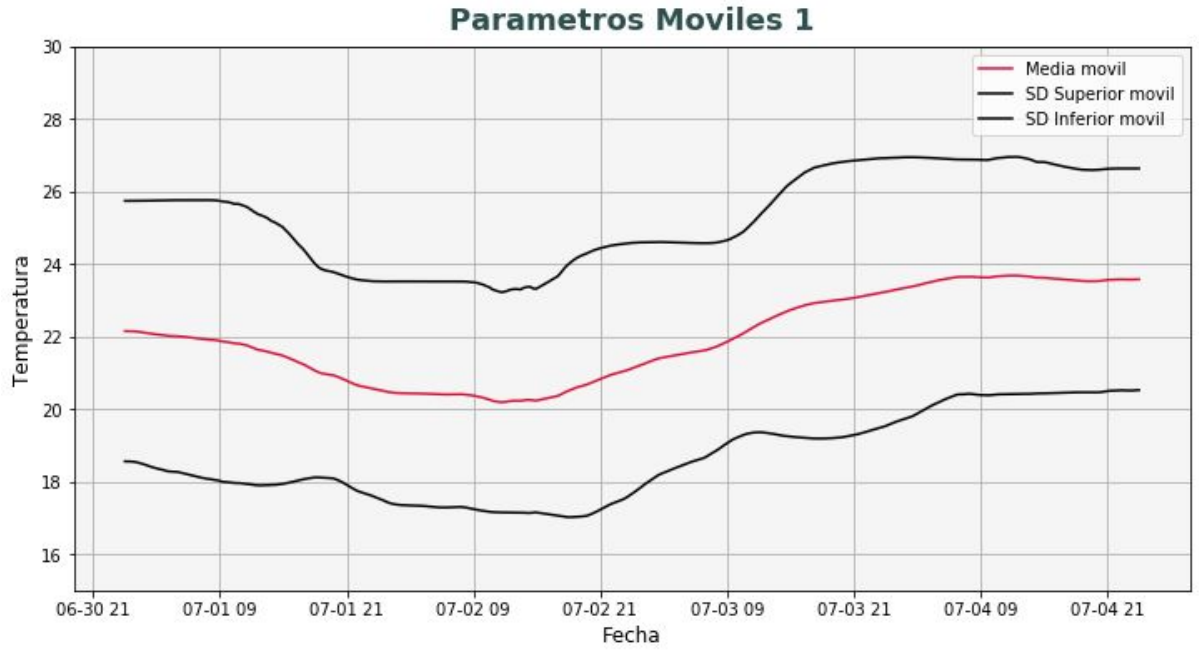

Figura 7. Media y bandas de confianza (desviación estándar sobre y bajo la media) móviles.

El razonamiento anterior se puede observar gráficamente con mayor facilidad en la figura 8. Nótese la marcada división de las líneas para los percentiles 10 y 90 entre los 5 días de datos. Además, esas curvas son mínimas en el día 2, el día identificado por presentar las menores temperaturas dentro de la serie analizada.

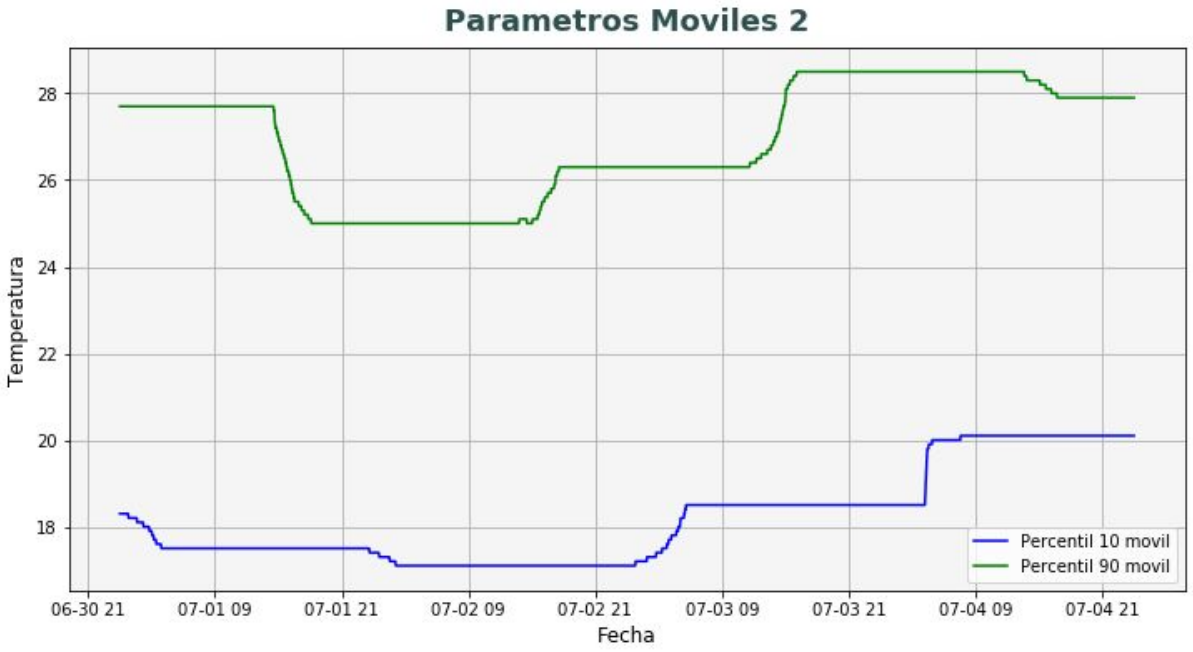

Figura 8. Percentiles 10 y 90 móviles.

En cuanto a las medidas de dispersión, estas se presentan en la figura 9. Siguen el mismo comportamiento mencionado anteriormente. La medida paramétrica (desviación estándar) presenta una curva suavizada, mientras que la medida no paramétrica (rango intercuartil) presenta una curva escalonada. En este caso, se puede observar que los valores máximos de dispersión están ubicados en las ventanas centradas en las horas de la mañana y en las horas del atardecer, en las cuales las temperaturas están en pleno ascenso y descenso. Mientras tanto, la dispersión de los datos tiende a reducirse en los momentos del día en los cuales las temperaturas tienden a estabilizarse (horas de la madrugada y horas del mediodía).

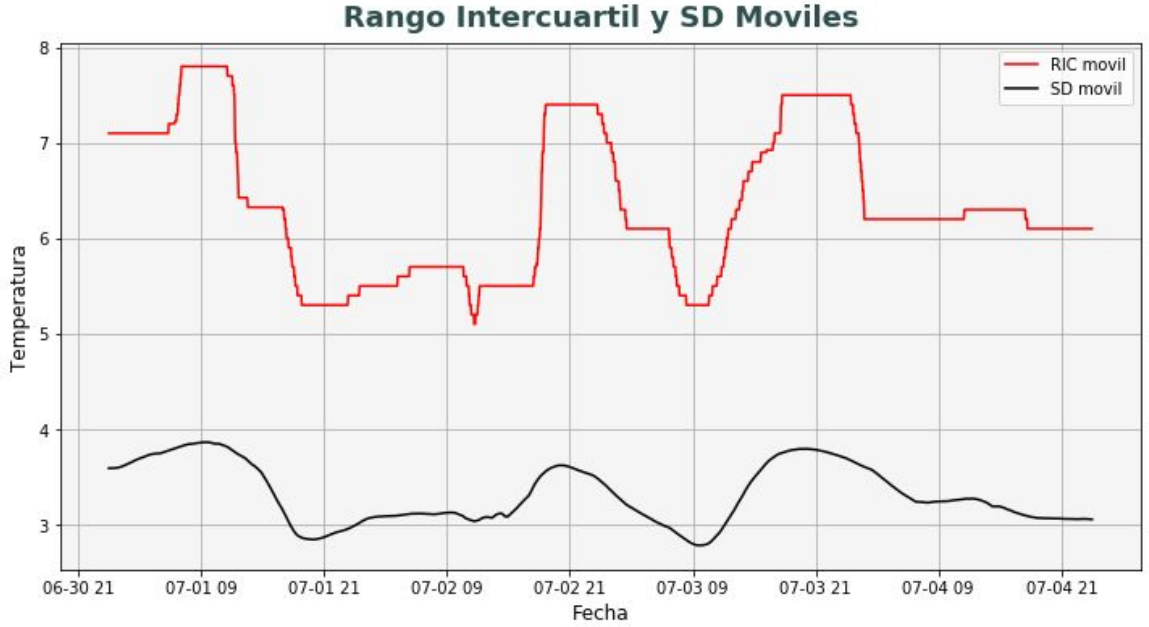

Figura 9. Índice de rango intercuartil y desviación estándar móviles.

Adicionalmente día con menor amplitud térmica (el día 2), presenta medidas de dispersión bajas, esto es coherente con el paradigma planteado.

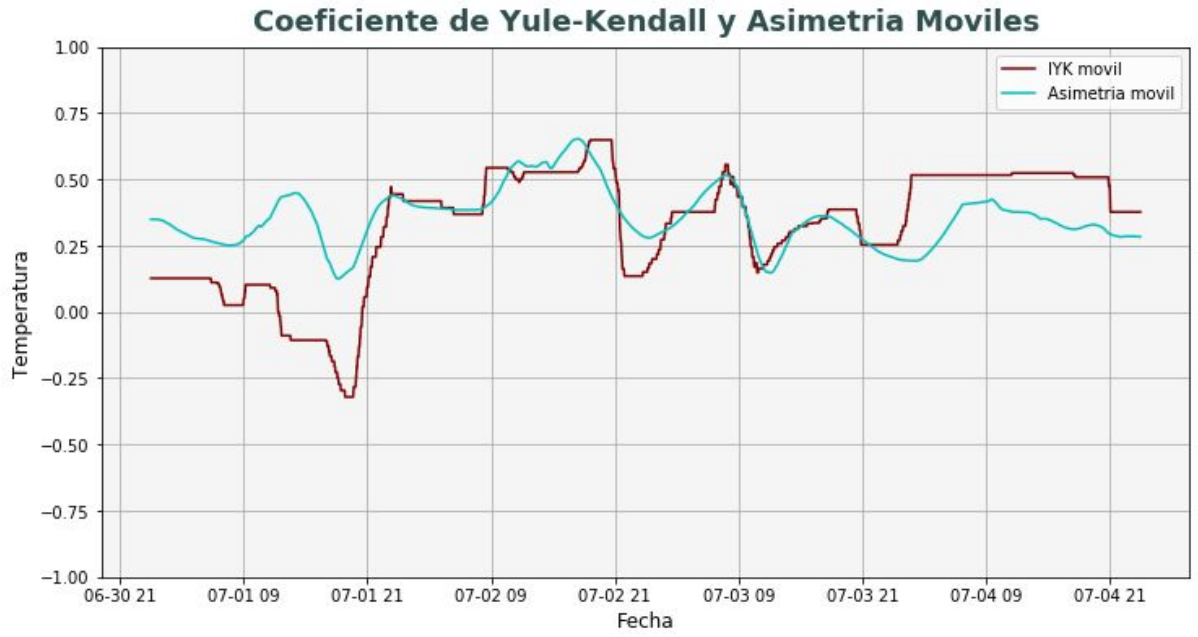

Figura 10. Coeficiente de Yule-Kendall y de asimetría móviles.

Finalmente, en cuanto a medidas de asimetría, estas tienden a ser positivas a lo largo de toda la serie, con excepción del final del día 1. En este día, puede observarse en el histograma conjunto (figura 6) que hay una concentración de probabilidades mayor a la habitual hacia datos con mayor temperatura, lo cual implica un desplazamiento de la masa de probabilidad hacia la derecha, reduciendo el coeficiente de asimetría. Este punto de la serie coincide además con el único tramo en el que la mediana logra superar a la media, lo que implica que en este punto hay mayor cantidad de datos que logran superarla, pero que no son lo suficientemente grandes en magnitud para modificar esta medida de localización.

VI. Evaluación de la tendencia.

Para evaluar correctamente la tendencia de la serie, se procedió de la siguiente manera. Inicialmente, se eliminó la variabilidad de la serie de datos a través de la obtención de la media móvil, variable a la cual se le aplicó el test no paramétrico de Mann Kendall con una certidumbre del 99%. En ésta última, la hipótesis nula consiste en que la distribución de la serie de datos se comporta aleatoriamente o, en otras palabras, carece de tendencia.

Para el caso de estudio se obtuvo que la temperatura entre los días 1 y 5 de julio de 2018 posee una tendencia creciente.

Dónde está el procedimiento, el estadístico Z?
Cálculo de la tendencia para los demás estadísticos...
Punto muy incompleto

VII. Conclusiones.

En primer lugar, los datos analizados tienden a ser mayormente recurrentes abajo de la media y, más específicamente, cerca de los 18°C (como se muestra en la figura 3),

lo cual se debe a que la mayoría de los datos corresponden a mediciones en las horas donde no hay incidencia de radiación térmica.

Es necesario recalcar el carácter periódico de la serie trabajada, aunque esto no implique que se comporte estacionariamente ya que, de serlo los estadísticos analizados en el tiempo se presentarían gráficamente a manera de línea horizontal y es evidente (Figuras 6 a 10) que esto no ocurre .

En cuanto a la tendencia y considerando que estos datos corresponden una porción inicial del mes con la temperatura media multianual más alta del año en Medellín, es entendible que sus valores tiendan a aumentar. De haber analizado la serie completa, seguramente se hubiera encontrado un comportamiento similar hasta alcanzar el máximo del mes (que se presume alrededor del 14 de julio por análisis gráficos realizados antes de cortar la serie), a partir del cual tendería a decrecer.

Como dato curioso, los estadísticos paramétricos se presentan de forma suavizada, mientras que los no paramétricos resultan en una función escalonada. Esto puede relacionarse con sus naturalezas, en el caso de los no paramétricos se trata de un estadístico "resistente" que tiende a analizar la variabilidad de los días por separado.

Bibliografía.

IDEAM. (1999). Información histórica. Obtenido el 7 de agosto de 2018 de: http://bart.ideam.gov.co/cliciu/mede/tabla.htm