ENTREGA: PRIMERA TAREA (07/09/2018)
Universidad Nacional de Colombia Sede Medellín
Curso: Análisis de datos ambientales

4.0
Bien :)
Ojo con la redacción

Estudiantes:
Yesica Paola Quintero Soto, ypquinteros@unal.edu.co
María Fernanda Fuentes Diaz, mffuentesd@unal.edu.co

Encontrar una serie de datos temporal, de más de 1000 datos, y realizar lo siguiente:

1. Lectura de los datos
2. Gráfica
3. Hallar los índices (Localización, Dispersión, Asimetría)
4. Estimar PDF (Histograma, Percentiles, Probabilidades)
5. ¿Son estacionarios los histogramas? Evaluar y graficar
6. ¿Son estacionarios los índices? Evaluar y graficar
7. ¿Existe tendencia en su serie? ¿En los percentiles?
8. Discutir resultados y concluir

Los datos que se analizarán a continuación, representan la medida de la temperatura en grados centígrados (°C) tomados durante cuatro días por una de las estaciones del SIATA, específicamente la que se encuentra ubicada en la sede Volador de la Universidad Nacional Sede Medellín, en la facultad de agronomía. Los días en los que se realizó el análisis corresponde a las fechas comprendidas entre 2016-03-01 y 2016-03-04, y la toma de los datos se realizó durante cada minuto del día (a partir de las 00:00 am hasta las 23:59 pm), obteniendo un total de 5760 datos para analizar. La interpretación de los datos se hizo mediante la herramienta de programación Python 2.7 con el intérprete Jupyter Notebook, apoyados en conocimientos estadísticos.

Como fue mencionado anteriormente, los datos de temperatura fueron obtenidos en la base de datos del SIATA (la cual se puede encontrar en la web y descargar de forma gratuita), el archivo de descarga tiene formato .csv el cual puede visualizarse en microsoft excel. La estructura del archivo es la siguiente:

	A	B	C
1	fecha_hora,Temperatura,Calidad		
2	2016-03-01 00:00:00,21.700000,1		
3	2016-03-01 00:01:00,21.700000,1		
4	2016-03-01 00:02:00,21.700000,1		
5	2016-03-01 00:03:00,21.700000,1		
6	2016-03-01 00:04:00,21.600000,1		
7	2016-03-01 00:05:00,21.600000,1		
8	2016-03-01 00:06:00,21.600000,1		
9	2016-03-01 00:07:00,21.600000,1		

Imagen 1. Visualización de algunos datos de temperatura

De izquierda a derecha, los datos se encuentran organizados así: fecha del registro, hora del registro, dato de temperatura y calidad del dato, donde 1 corresponde a un registro confiable obtenido en tiempo real, y en caso contrario, 1513.
Inicialmente, se desea hacer una visualización no formal de los datos, para tener una idea de la forma como se encuentran distribuidos:

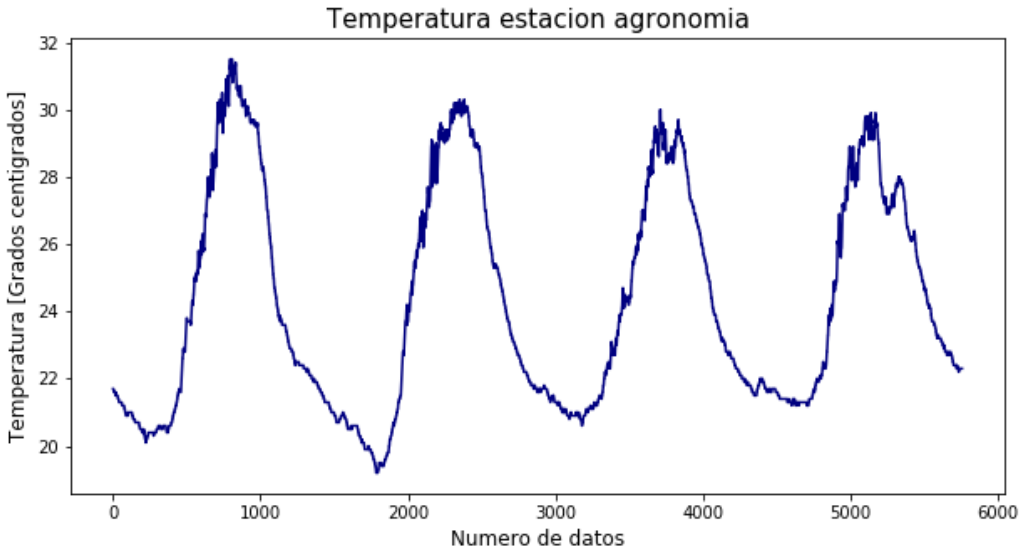

Imagen 2. Gráfica que compara la temperatura y el número de datos

Lo anterior es bastante útil si queremos hacer un análisis rápido de la estructura de los datos, y, por la forma de la gráfica se podría sospechar que presenta tendencia cíclica, pero sin crecimiento o decrecimiento, aunque, siendo de algún modo estrictos, la mejor manera de saber si la gráfica presenta alguna tendencia es a partir de un test no paramétrico, más adelante se entrará en detalles. Se sabe que la temperatura es una variable que va cambiando con el tiempo, así que es difícil que en un tiempo prolongado conserve el mismo valor, que se encuentra condicionado a ciertos factores climáticos. Por esta razón, es que no se tiene una gráfica uniforme, refiriéndose por uniforme a que se tenga una linea mas o menos horizontal, paralela al eje x; en su lugar, aparecen datos que parecieran repetirse cada cierto tiempo, como si se tratara de algo periódico.

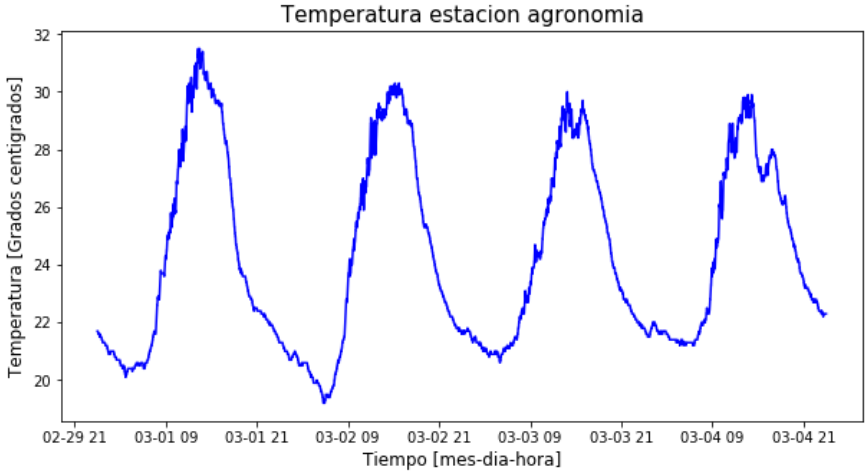

Imagen 3: Gráfica que muestra los cambios de temperatura durante el tiempo analizado.

Una vez son organizados los datos de temperatura con respecto al tiempo en que fueron medidos, en la gráfica anterior se puede observar las variaciones de la temperatura durante los cuatro días analizados. Antes de realizarse un análisis estadístico detallado, se puede observar que el dato de temperatura más alto reportado es de 31 °C y que se presentó en la fecha 2016-03-01 durante 46 minutos alrededor de la 1 de la tarde. De la misma forma, se puede observar el dato de temperatura más bajo de 19 °C el día 2016-03-02, a las 4 de la mañana durante dos horas y media. Se puede observar que alrededor del mediodía se reportaron las temperaturas más altas durante cada respectivo día, y durante las horas de la madrugada se reportaron las temperaturas más bajas en cada respectivo día.

En la siguiente gráfica se encuentra un histograma en el que se grafican las distribuciones de frecuencias, y donde además, se pueden visualizar las medidas de tendencia central más comunes que corresponden a la mediana, la media y la moda. La información exacta de estas medidas de tendencia central se muestra más adelante.

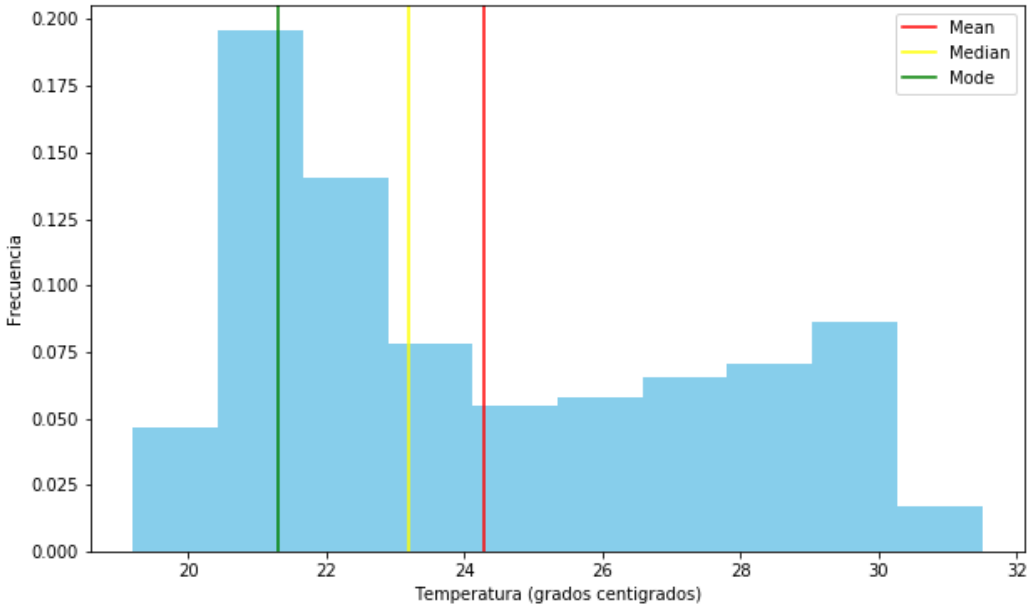

Imagen 4. Histograma de frecuencias

En la tabla a continuación, se exponen las medidas de tendencia central, las medidas de dispersión y las medidas de simetría obtenidas a partir de cálculos. Las medidas de tendencia central son medidas estadísticas cuyo objetivo es resumir la información, con el fin de obtener una primera aproximación del comportamiento de una población (en este caso la población es la temperatura); y las medidas de dispersión son aquellas que entregan información sobre la variabilidad de los datos y pretenden resumir en un valor la dispersión que tienen los datos; y las medidas de simetría tienen que ver con la forma de la distribución. Las medidas encontradas son: la media, la mediana, la moda, la desviación estándar (Std), la varianza, el rango intercuartil (IQR) y el índice de Yule-Kendall (YK). De las medidas encontradas, las que corresponden a tendencia central son: La media o promedio, es el resultado de sumar todos los datos y dividirlo por el número de datos (n) que existen; la mediana corresponde al dato que ocupa la posición central cuando se organizan todos los datos (de menor a mayor magnitud); la moda es el valor de la variable que más se repite. Las medidas que

corresponden a dispersión son: la desviación estándar, que indica cuánto pueden alejarse los valores con respecto a la media; el rango intercuartil que es la diferencia entre el el tercer cuartil y el primer cuartil, que permite obtener una idea de la dispersión de los datos; la varianza es el cuadrado de la desviación estándar; y como medida de simetría se encuentra el índice de Yule-Kendall que resulta al comparar la distancia entre la media y cada uno de los dos cuartiles (el 75 y el 25) siendo una alternativa robusta y resistente. Los resultados de cada una de estas medidas se encuentran en la tabla 1, a continuación:

Media	24.28
Mediana	23.2
Moda [1]	21.3
Varianza	5.90
IQR	10.81
Std	3.28
YK	0.38

Tabla 1. Resultados de los medidas de tendencia central y de dispersión (localización, dispersión, asimetría).

Si se desea analizar el valor de una temperatura en específico, frente a los demás datos que previamente fueron leídos, es necesario ordenar éstos de forma ascendente, y para ello, se utiliza la función percentil, que divide los datos en partes, y a cada una de estas partes les asigna una cantidad de datos iguales. La siguiente gráfica provee información completa de todos los percentiles, y al mismo tiempo, muestra los más importantes señalados por líneas horizontales (paralelas a la abscisa), correspondientes a P10, P25, P50 (mediana), P75 y P90. Los resultados son:

```
P10:  20.7
P25:  21.4
Mediana:  23.2
P75:  27.3
P90:  29.3
```

Tabla 2. Resumen de algunos percentiles.

[1] El resultado de la moda es de 21.3 °C y se encuentra repetido 217 veces.

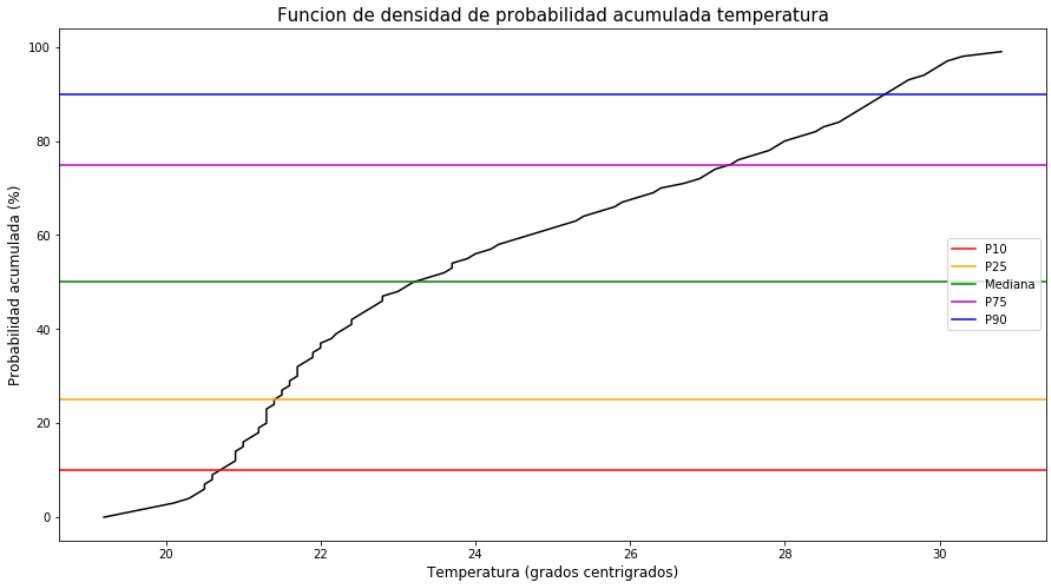

Imagen 5. Probabilidad acumulada - Percentiles.

Lo anterior nos conduce al siguiente análisis: Si la temperatura 20.7 °C es el percentil 10, significa que el 10% de los datos del total de la serie posee valores inferiores de temperatura a éste; también es equivalente a decir que de cada 100 datos de temperatura que existen 90 van a tener una temperatura mayor a la del percentil. Este razonamiento aplica para los demás percentiles, ya que si nos dirigimos a la definición formal de percentil, éste indica el valor de la variable por debajo de la cual se encuentra un porcentaje dado de observaciones en un grupo de datos.

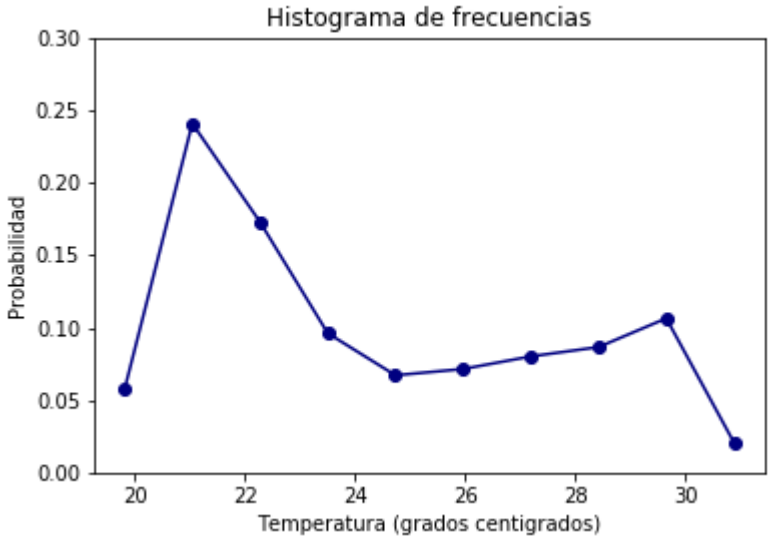

Imagen 6. Histograma de las temperaturas.

La gráfica anterior representa el histograma de frecuencias que se creó para la temperatura, en la que la abscisa representa los valores de temperatura que se dieron durante los cuatro días, y la ordenada representa la probabilidad con la que se presentaron esos datos. Un histograma corresponde a una herramienta muy útil para obtener un panorama de la temperatura, ya que permite observar una tendencia de los datos utilizados de ubicarse en determinados valores. Así es como se puede observar que 23 °C es donde se ubica la mayor

cantidad de datos con una frecuencia de 0.24, y que 31 °C es donde se ubica la menor cantidad de datos, con una frecuencia de 0.01. Además, la suma de las probabilidades debe ser 1.00, teniendo en cuenta que es la suma acumulada de todas las probabilidades de cada temperatura.

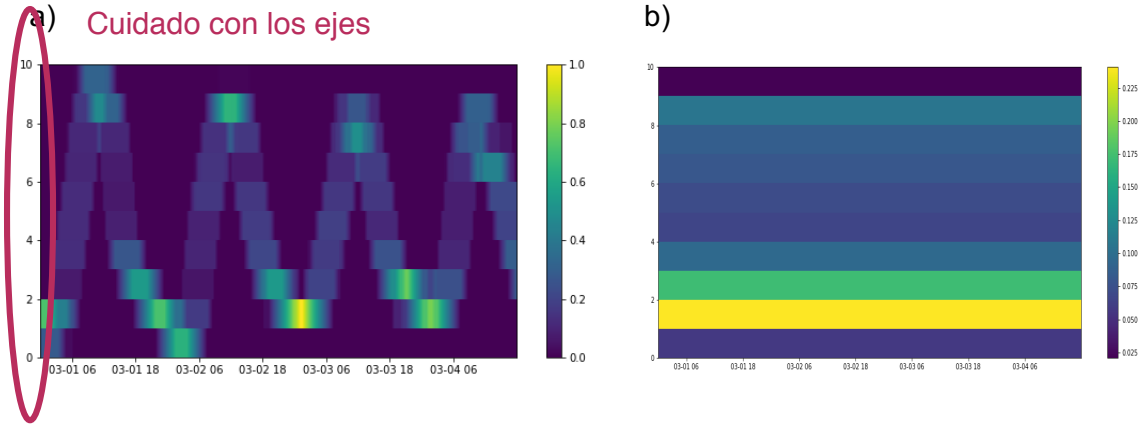

Imagen 7. Gráfica de estacionariedad de histogramas. Imagen a) vista lateral del histograma, b) vista superior del histograma

De cuánto? y por qué?

Se realizó una ventana móvil para analizar con más detalle las medidas de tendencia central a través del tiempo, y también para el caso de los histogramas, donde se pretende evaluar si presentan una condición de estacionariedad. Para el caso de los histogramas (imagen 7), el eje x representa las fechas, y el eje y, los bins o intervalos de clase donde están clasificados los datos. Por un lado, la matriz de colores tiene unas zonas donde hay mayor concentración de un color específico (en la escala de color amarillo), lo que es equivalente a decir que hay una mayor concentración de masa. Esto en definitiva se puede tomar como una condición de no estacionariedad en el histograma, porque de ser lo contrario, tendríamos un color uniforme en todas las zonas de la gráfica. Esto implica que la gran mayoría de los momentos estadísticos tampoco sean estacionarios, o por lo menos la media no lo es, y eso se puede asegurar con cierto grado de certeza, porque también hay unas zonas de tonalidad verde, que podría dar este mismo indicio. La parte **a)** de la imagen 7 también puede ser representada por una gráfica de contornos, como la que se observa en la imagen 8, la cual tiene superpuesta la línea de la media móvil (de color rojo). Esta es una forma un poco más elegante de representar la condición de estacionariedad de los histogramas, y adicional podemos observar que los principales puntos de concentración de masa se encuentra en los picos superiores e inferiores de la gráfica, siendo más marcados en los picos inferiores, donde hay un mayor reporte de datos de temperatura (entre los 21-24°C). La media móvil, aunque siga la misma tendencia cíclica de los histogramas, no es estacionaria.

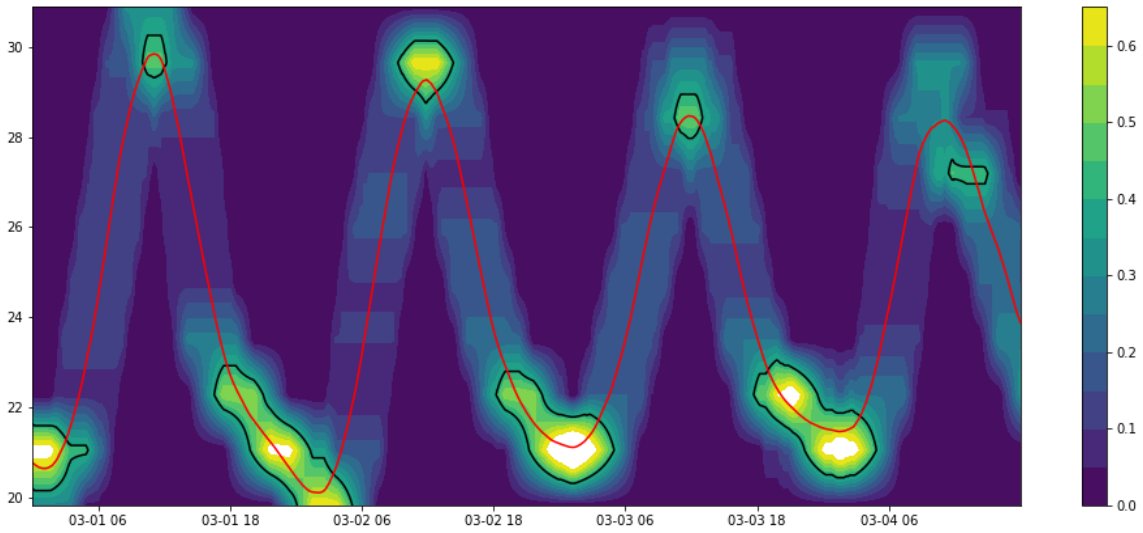

Imagen 8. Gráfica de contorno de estacionariedad de histogramas (la línea roja equivale a la media móvil).

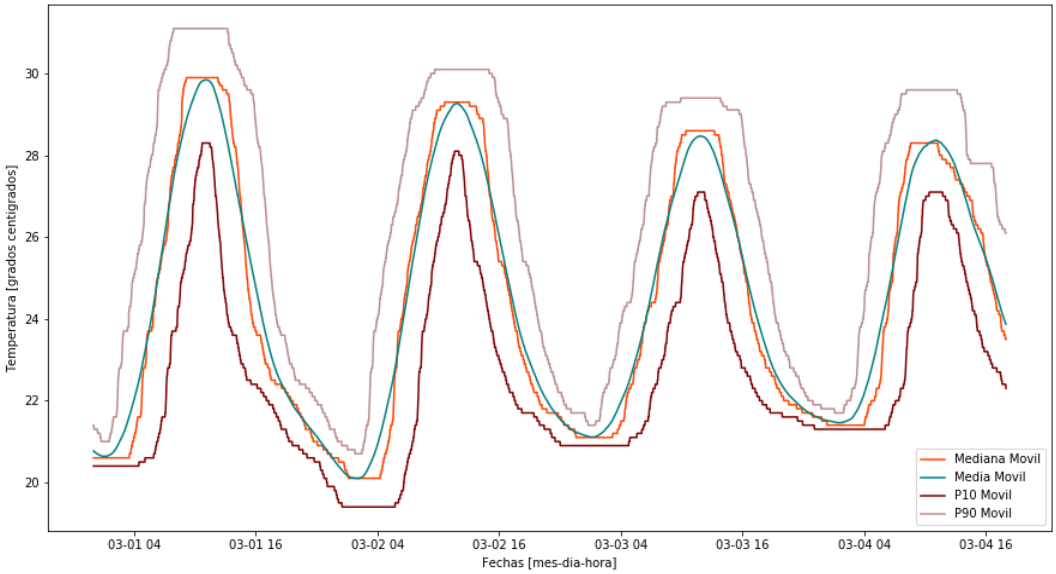

Imagen 9. Estacionariedad de índices como la media, mediana, P10 y P90.

ojo con esto!

En la imagen 9, es posible observar varios índices que poseen las mismas unidades (no fue posible incluir la gráfica de la desviación estándar y el rango intercuartil dado que no cuentan con las mismas unidades al ser adimensionales), los cuales pueden ser comparados. De cierto modo, el percentil 90 siempre permanece superior a los demás índices, al igual que el percentil 10 que permanece en la parte inferior. La media y la mediana móvil casi que están coincidiendo en la forma de sus gráficas, presentan cierta superposición aunque en algunos valores difieran un poco. En general, los índices no siguen una tendencia creciente o decreciente, pero se comportan como si quisieran seguir un ciclo que apunta a una dirección horizontal. Esto puede pasar en una serie de datos de temperatura, que está condicionada a la posición de la tierra con respecto al sol (en ciertos momentos del año), su ángulo de inclinación y la rotación, ello marca las diferentes temperaturas que se puedan dar en un lugar determinado en ciertos meses, donde oscilan valores similares mientras no sean alterados por otros fenómenos o ciclos externos mucho más grandes.

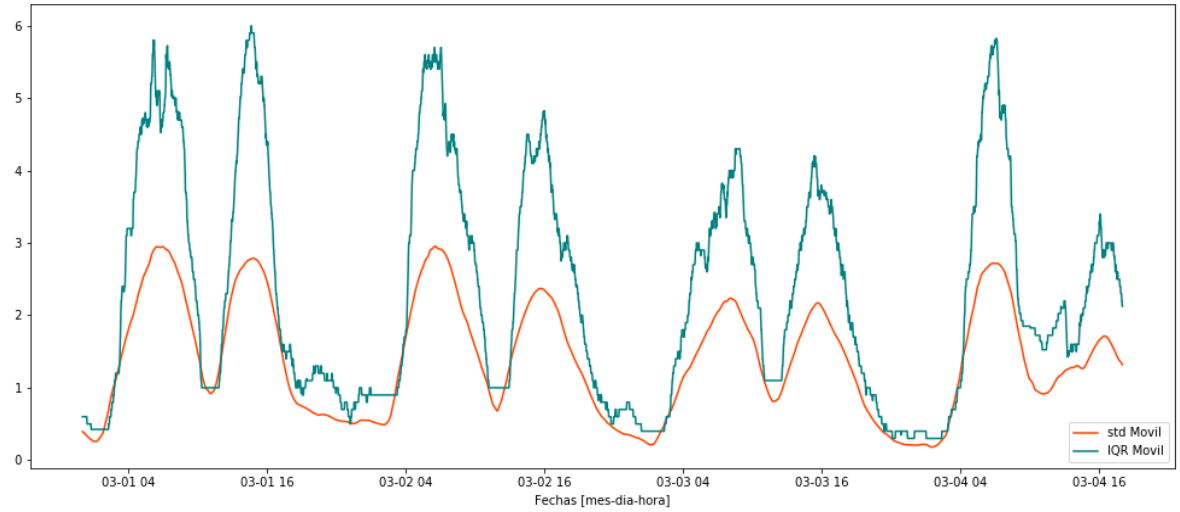

Imagen 10. Gráfica de estacionariedad de la desviación estándar y el rango intercuartil móviles

En la imagen 10, se observa que la desviación estándar no sigue una tendencia creciente o decreciente al igual que el rango intercuartil, lo que sugiere que la dispersión de los datos se mantiene en la misma dirección.

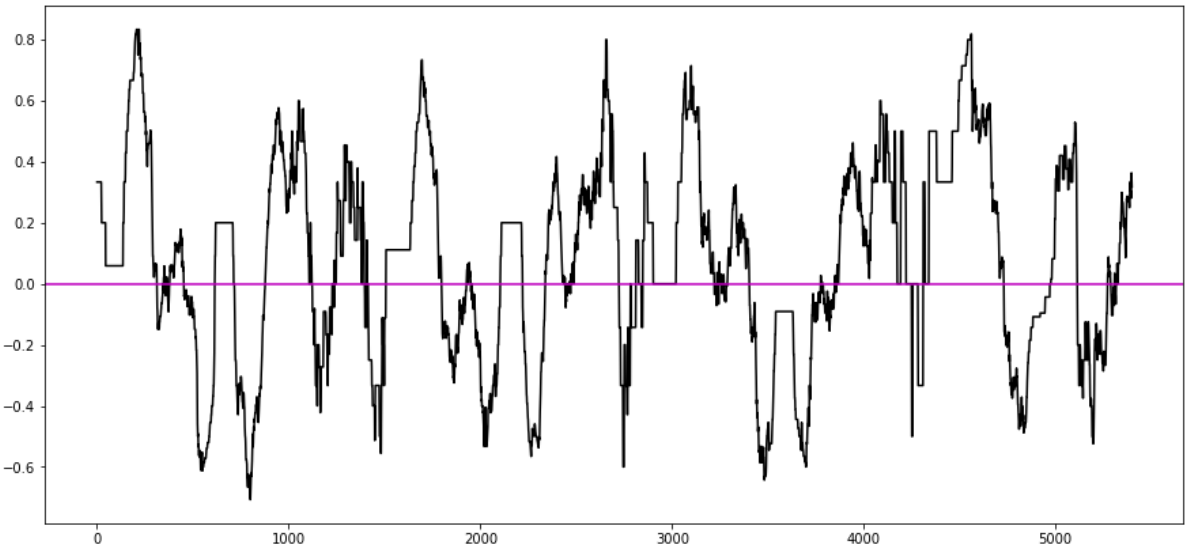

Imagen 11. Gráfica de estacionariedad del coeficiente Yule-Kendall.

En la imagen 11, para el coeficiente YK móvil, no se observa cierta preferencia de la serie a estar concentrada en un solo lugar, ya que se ubica tanto en valores positivos como negativos teniendo como eje de referencia el cero (línea horizontal morada). Esto indica, como en discusiones anteriores, que la tendencia de la serie no es ni creciente ni decreciente, no tiene una preferencia definida, y que los datos tienen comportamiento cíclico.

Para saber si el conjunto de datos presenta tendencia en su serie, se realizó el test de Mann-Kendall, que se utiliza para analizar si la tendencia aumentó o disminuyó a lo largo del tiempo, es un test no paramétrico, lo que significa que funciona para cualquier distribución siempre y cuando tenga una correlación serial. Para este test se corre bajo un hipótesis nula, *Ho* (en esta caso es la condición de aleatoriedad). Al realizar este test existen tres datos que son de

???

interés, estos son: la variabilidad (var), el estadístico de prueba (S) o media y la estadística del test (z) ya que estos son los que nos indican la tendencia de la serie. De esta manera, los resultados obtenidos en este test son:

var	5.943910299017164
S	866247
z	21239192000

Tabla 3. Resultados del test Mann-Kendall

Esto no tiene mucho sentido, el estadístico z no puede dar un valor tan tan grande. La teoría no está bien asimilada, recuerden que debe compararse con la distribución normal. No veo nada además relacionado con la pregunta de alpha/2 ni la tendencia de otros estadísticos

El más importante de los resultados es el dato *S*, que cuenta con 3 posibilidades: si S > 0, significa que la tendencia crece; si S < 0, significa que la tendencia decrece; y si S = 0 significa que no hay tendencia. En este caso S es mayor a 0, por lo que le tendencia es a crecer.

Además del test de Mann-Kendall, se puede mostrar gráficamente la tendencia de una serie comparando la gráfica de temperatura que se muestra en la Imagen 2, con una nueva gráfica de la variación de la temperatura, que trabaja con la ventana móvil. Se hace de esta manera debido a que la variable temperatura contiene tanto variabilidad como tendencia, y por lo tanto, se debe eliminar la variabilidad, que se logra eliminando la media. La serie de tendencias media móvil, no está para todo el dominio debido a la ventana, por eso no se escriben las fechas. Sin la medida de la variabilidad se pueden observar picos más altos, y también valles más bajos. Además, el valor de la media de la temperatura variable es 0.008. De todas formas, en la nueva gráfica no se puede ver una tendencia, pero se pueden observar cambios bruscos.

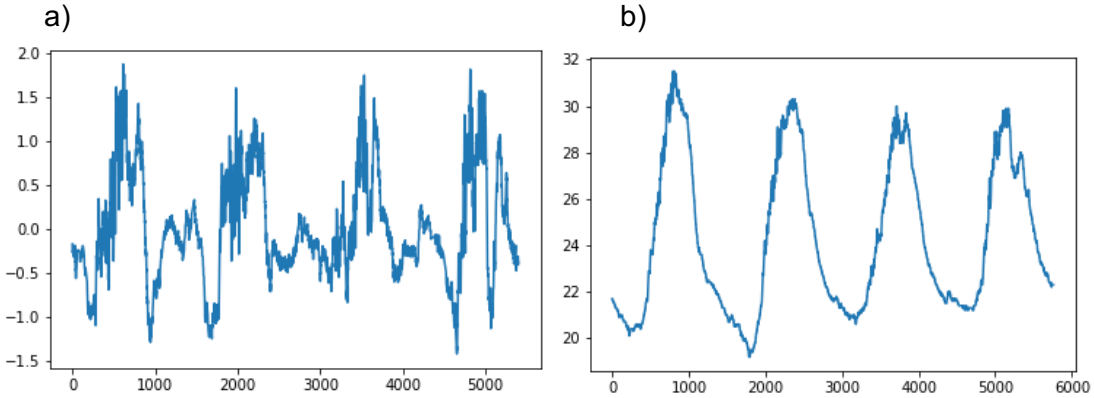

Imagen 11. Comparación entre: a) gráfica de la temperatura variable, y b) gráfica de la temperatura

Como prueba extra, también se le saca la derivada (que es la serie menos el dato anterior) a la variación de la temperatura, en la que se regula la variabilidad de la serie, cambiando el número de la matriz, por esto sirve como una forma de encontrar la ventana más adecuada para trabajar.

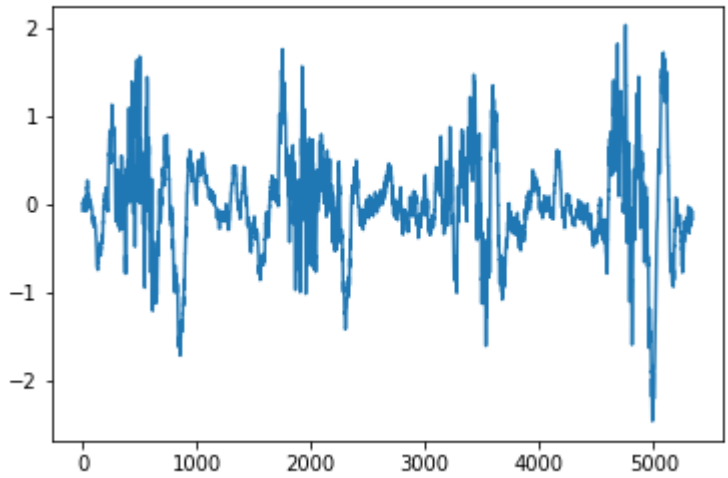

Imagen 12. Derivada de la variación de temperatura

De todo el análisis de datos que se ha hecho hasta el momento se puede decir que se ve un comportamiento muy típico de la temperatura para las zonas tropicales. En la que al medio día se ve el pico más alto de temperatura, y en las horas de la madrugada se ve el registro de temperatura más bajo del día. [1]

Según la distribución temporal de la temperatura en Antioquia, se presentan anualmente dos épocas cálidas y dos épocas frías, cada una alrededor de tres meses. El tiempo que fue evaluado corresponde con la primera época cálida anual (enero-marzo) en la que las temperaturas alcanzan en promedio 1 °C por encima del promedio anual. [2].

Bibliografía:

1. diversidad climática de Colombia. *Rev. Acad. Colomb. Cienc.* https://doi.org/0370-3908
2. Pabón-Caicedo, J. D., & Eslava-Ramírez, J. A. (2001). Generalidades De La Distribución Espacial Y Temporal De La Temperatura Del Aire Y De La Precipitación En Colombia. *Meteorología Colombiana*, *4*, 47–59. https://doi.org/ISSN-0124-6984
3. Poveda, G. (s/f). LA HIDROCLIMATOLOGÍA DE COLOMBIA: UNA SÍNTESIS DESDE LA ESCALA INTER-DECADAL HASTA LA ESCALA DIURNA por CIENCIAS DE LA TIERRA. Recuperado a partir de https://www.uninorte.edu.co/documents/266486/0/Poveda_2004.pdf
4. Poveda, G., & Mesa, O. (1999). Corriente de chorro superficial del Oeste.pdf. *Rev. Acad. Colomb. Cienc.*