TAREA No 1

ANÁLISIS DE DATOS

PRESENTADO A:

CARLOS DAVID HOYOS

PRESENTADO POR:

CARLOS MARIO VALENZUELA ROSAS

UNIVERSIDAD NACIONAL DE COLOMBIA
SEDE MEDELLÍN

MEDELLÍN 2018

Tu informe carece de un análisis detallado de los resultados en función de la física del problema, en tu caso la lluvia. Recuerda que esto no es una clase de programación ni de estadística, es una clase de análisis de datos. No puedes ver los algoritmos como una receta. Te recomiendo que realices nuevamente una revisión de los conceptos que vimos en clase, especialmente la estacionariedad.

3.9

TAREA No 1 ANÁLISIS DE DATOS

1. Datos de la serie
   Los datos para el análisis de la serie son de precipitación en una escala temporal diaria de la estación Villahermosa en el Municipio de Medellín. La estación tiene coordenadas: longitud -75,54613 y latitud 6,25981 con una elevación de 1690 m.s.n.m. y se encuentra dentro de la cuenca Santa Elena (Figura 1).

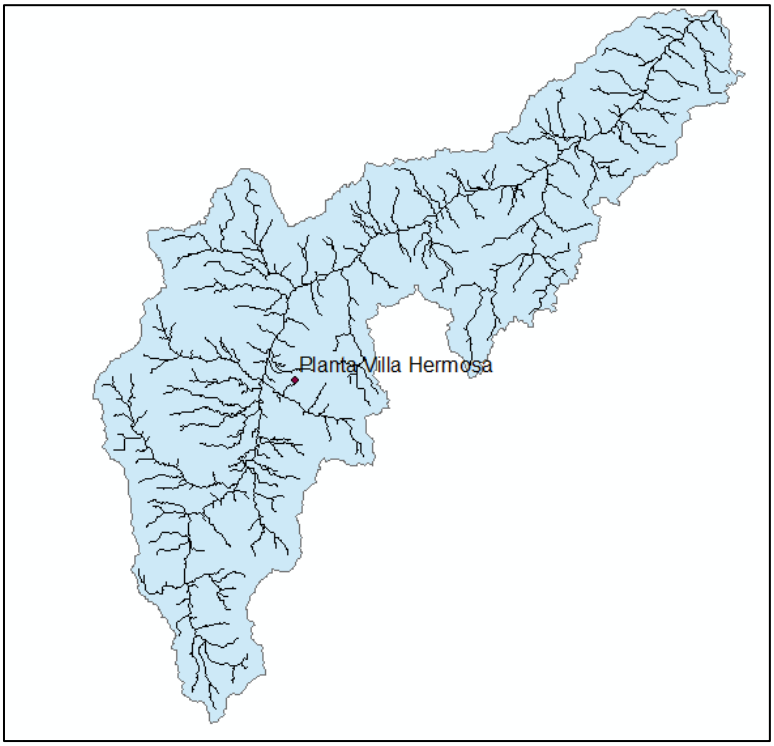

Figura 1. Estación Precipitación Villahermosa

Para la lectura de los datos se contó con información de precipitación desde el 1 de enero de 1995 hasta el 31 de diciembre de 2014, lo que corresponde a 20 años y un vector de longitud de 7300 datos. Se realizó la gráfica de los datos (Figura 2) y se observaron posibles tendencias, valores mínimos y máximos y una posible variabilidad.

Además, se realizó el cálculo de los percentiles y se evaluó en especial los valores del percentil 50 y 75 para conocer los valores de precipitación en dicha ubicación. Se obtuvo valores para el P50 (media) de 0,762 mm y para el P75 de 6,096 mm, lo que indica, por el ejemplo, para el P75 que el 75% de los datos de precipitación se encuentran por debajo de 6,096 mm.

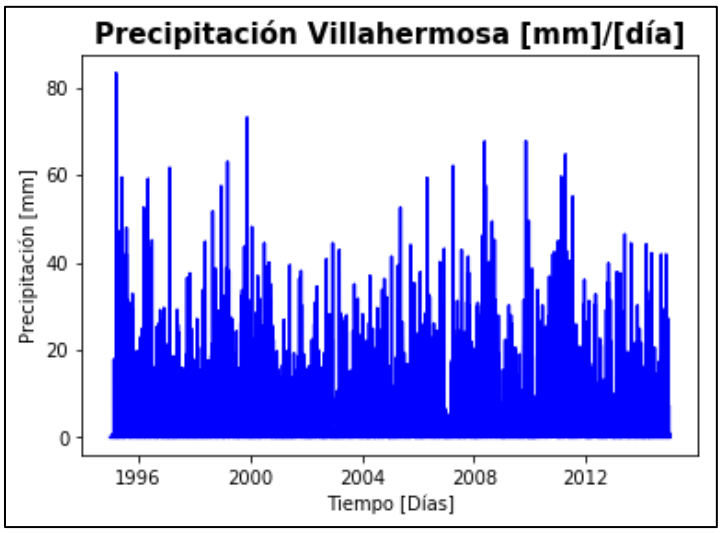

Figura 2. Precipitación diaria estación Villahermosa

Se realizó la gráfica de la probabilidad acumulada y se sombrearon los percentiles P50 y P75 para visualizar en cuál lado se encontraba dicho valor de área acumulada (Figura 3). La función de distribución acumulada (CDF) calcula la probabilidad acumulada hasta el valor de la variable que se especifique

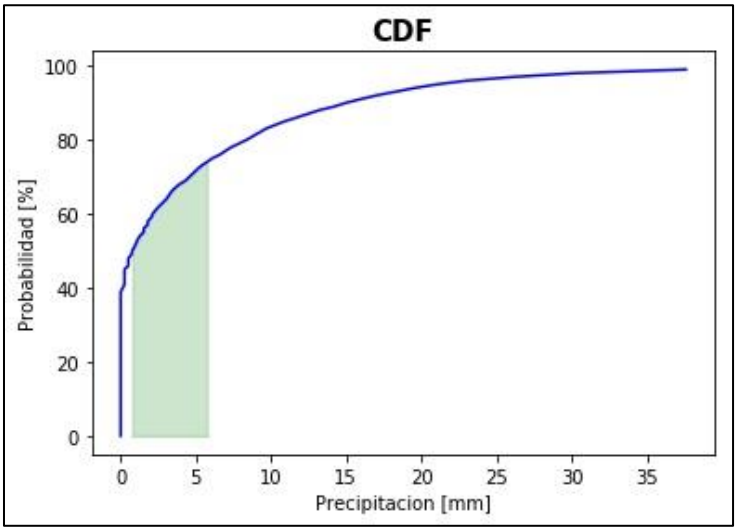

Figura 3. Valor de probabilidad acumulada con P50 y P75

Después, se complementó con un histograma de probabilidad de toda la serie y se observó la curva de distribución de probabilidad (PDF) (Figura 4), la cual indica regiones de mayores y menores probabilidades para los valores de la variable seleccionada.

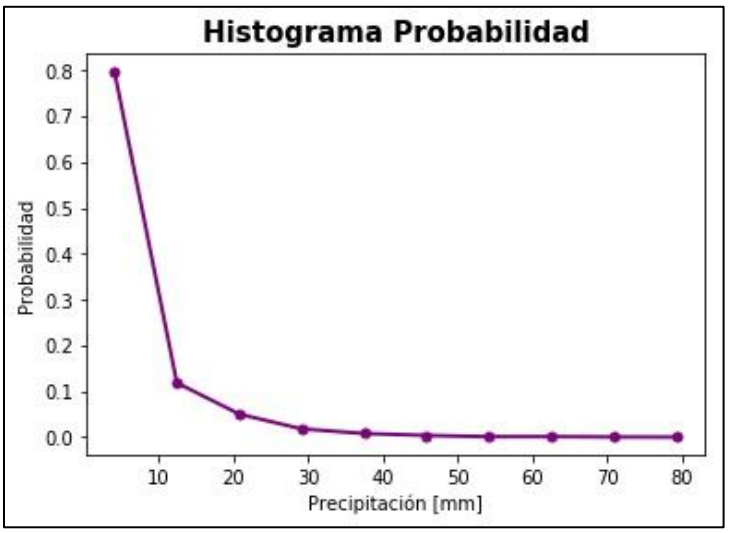

Figura 4. Valor de distribución de la probabilidad

Ahora, para realizar un análisis de datos de forma correcta es necesario realizar ventanas móviles las cuales permitan observa el comportamiento de las propiedades estadísticas de la serie en el tiempo. El valor de ventana seleccionado fue de 3 años y con dichas ventanas se volvió a graficar la curva de distribución de probabilidad para realizar una comparación (**Figura 5**).

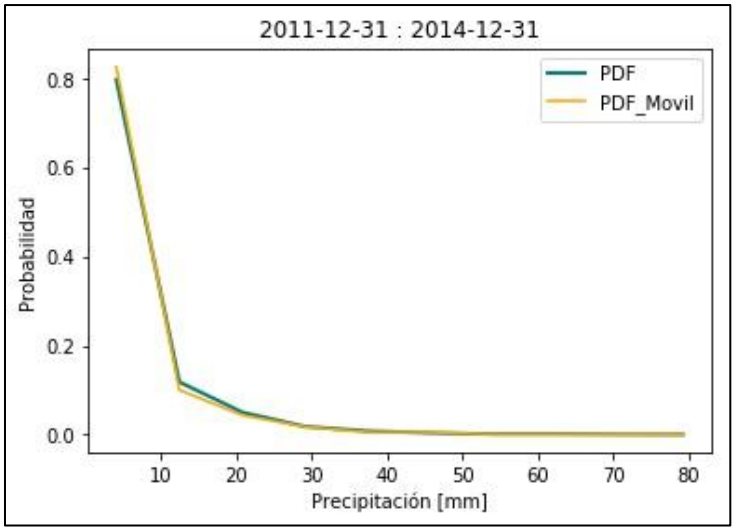

Figura 5. Comparación PDF entre los datos y una ventana móvil

Dicha comparación nos permite observar que los datos presentan cierta estacionalidad con respecto al histograma y que las probabilidades de precipitación tienen valores similares.

El paso siguiente fue calcular los índices de localización, dispersión y simetría vistos en la clase para las ventanas móviles de 3 años. Se crearon los vectores de localización

(Media, Mediana), dispersión (Desviación Estándar, Rango Intercuartil, P10, P90) y simetría (Asimetría, Yule-Kendall). Gráficamente se realizó una gráfica que muestra la matriz de probabilidad para la distribución de probabilidad de los datos (**Figura 6**), la cual muestra que los valores de precipitación con mayor probabilidad se encuentran por debajo de 5 mm.

Alguna conclusión al respecto? con esta gráfica?

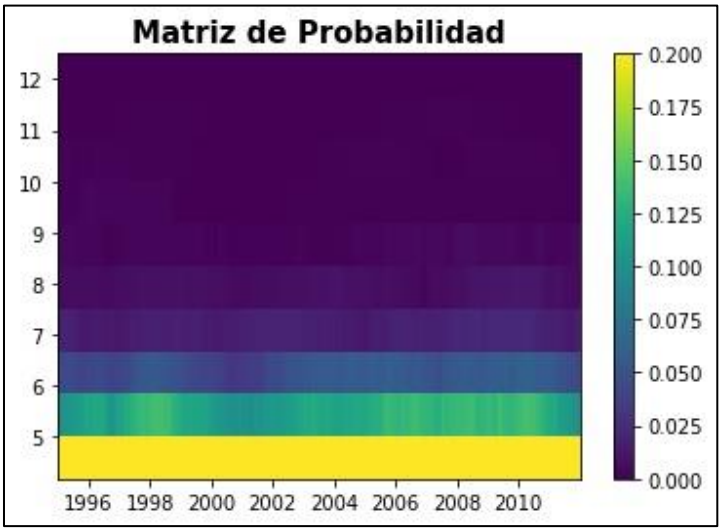

Figura 6. Matriz de probabilidad diaria

En una sola grafica se observó el comportamiento de los percentiles P10 y P90 con la media y la mediana observando que dichos valores no son estacionarios y son diferentes en el tiempo (**Figura 7**). Los valores del P90 se encuentran muy alejados de la media y la mediana, lo cual indica que se podrían estar presentando anomalías en el comportamiento de los datos. Podría mencionarse que se presentan posibles ciclos.

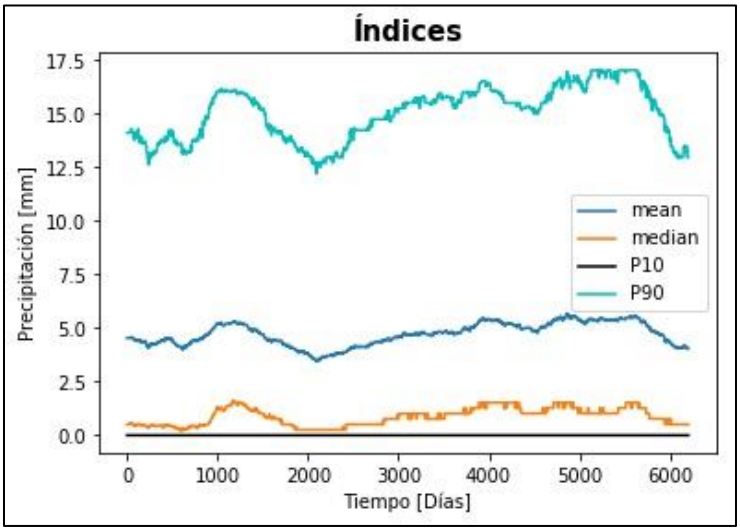

Figura 7. Índices de localización y dispersión

Para comenzar a observar la dispersión de los datos se realizó gráficamente la comparación entre la desviación estándar y el rango Intercuartil, observándose bastante diferencia entre ambos valores (**Figura 8**). Esto indica dispersión entre los datos con valores más altos en los periodos secos.

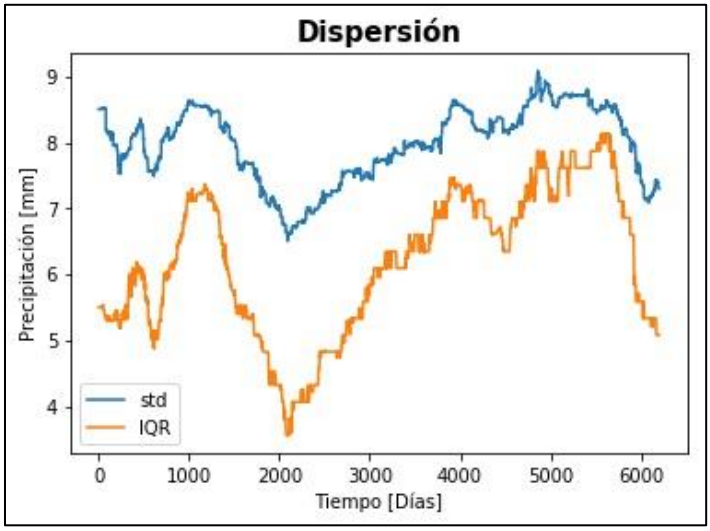

Figura 8. Comparación de la dispersión

Para la simetría de los datos se calculó el índice de Yule-Kendall y se graficó (**Figura 9**), observándose la poca simetría que existe en los datos de precipitación y que dicho valor se encuentra muy por encima de una base igual a 0.

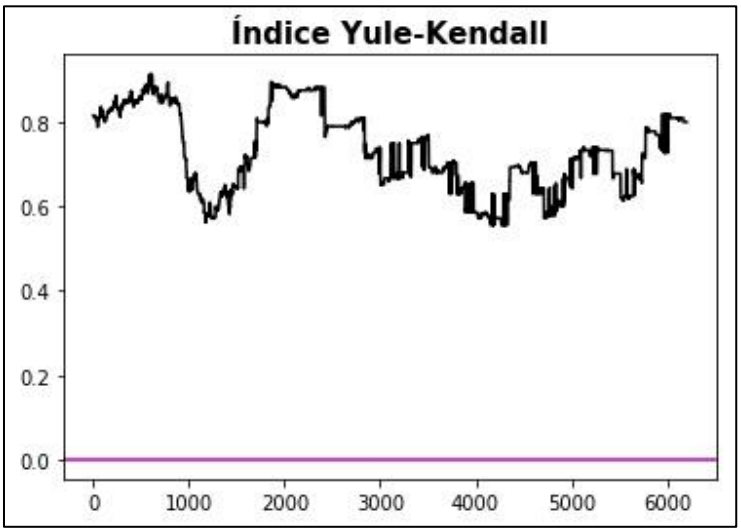

Figura 9. Simetría Yule-Kendall

Para el análisis de tendencia se utilizó el método de Mann-Kendall que es una prueba estadística no paramétrica, que se basa en la correlación entre un dato de una serie y su posición en el tiempo (Kendall and Gibbons, 1990). El estadístico *Sj* se calcula asumiendo la hipótesis nula que las series *xj,k*, *k*=1,…,nj se generan para *nj* de manera

independiente y con variables aleatorias igualmente distribuidas. Un valor positivo del estadístico *Sj* indica una tendencia creciente mientras que un valor negativo del estadístico de prueba indica una tendencia decreciente. Para los datos de precipitación de la serie se obtuvieron los siguientes valores:
Z = 2,987
S = 477947
Var = 43232990750

La hipótesis nula es rechazada si $|Z_j| \geq Z_{j,\alpha/2}$, y se llega a la conclusión que la tendencia es monolítica (no necesariamente lineal), y que no es estadísticamente significativa. Para el caso del test no paramétrico de Mann-Kendall, se considera un test de doble cola porque el resultado puede ser creciente o decreciente, y el nivel de significancia α necesario para poder rechazar la hipótesis nula se va a calcular como $\alpha/2$, con un valor de α = 0,05.

Si Z > 1,96 entonces hay significancia estadística para la serie de datos, es decir una tendencia. Que en el caso de los datos de precipitación es positivo y se podría afirmar que existe tendencia ascendente en la serie de datos analizada.

Se graficó la precipitación y se comparó con la variabilidad separada en otra gráfica (**Figura 10**) observando altos valores de variabilidad y finalmente analizando la información se encontró información de un ciclo que es el anual correspondiente a 365 días de información (**Figura 11**).

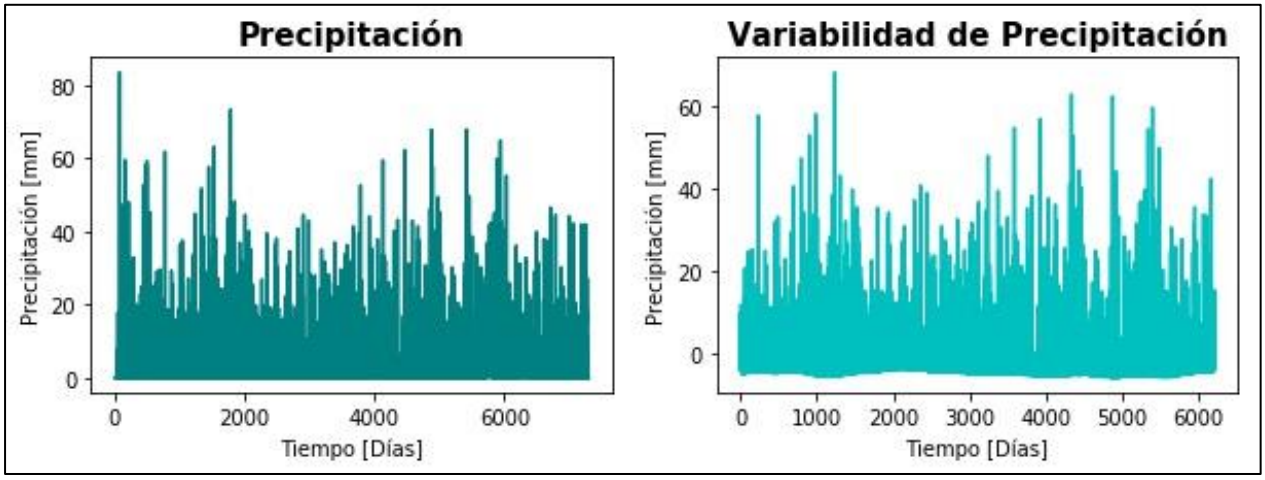

Figura 10. Precipitación vs Variabilidad

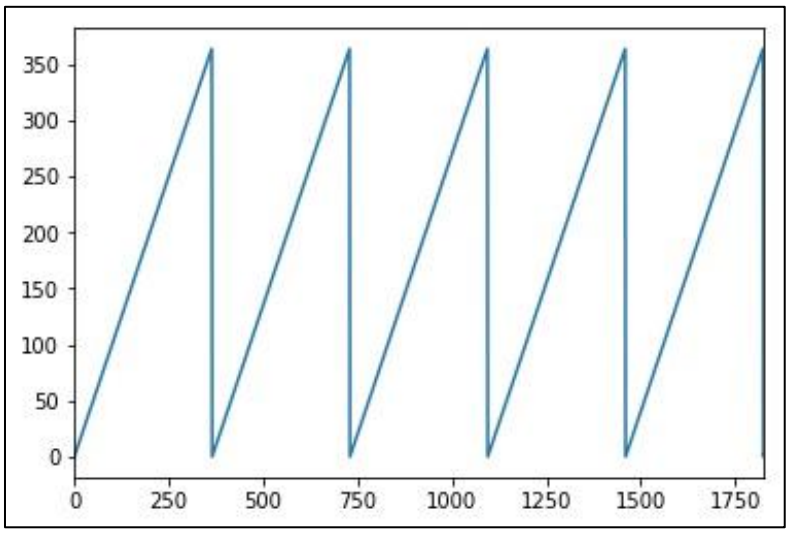

Figura 11. Identificación ciclo anual

Las conclusiones y los análisis fueron realizados dentro de la descripción de cada paso de la tarea. Aunque es importante realizar que los datos de precipitación de la estación Villahermosa en Medellín presentan gran variabilidad, los valores son dispersos y se ven influenciados en gran parte por el ciclo anual.