4.0

UNIVERSIDAD NACIONAL DE COLOMBIA
SEDE MEDELLÍN
FACULTAD DE MINAS

Tarea No. 1 Análisis de Datos Ambientales

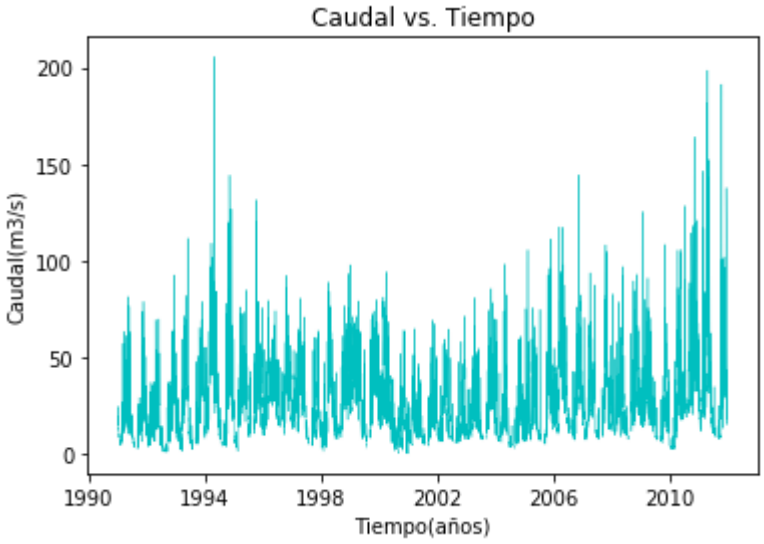
Nicole Elheyn Forero Sacristán C.C. 1094955635, & Daniela Pérez Ortiz C.C. 1017239984

Para el presente trabajo se utilizan datos de caudales, tomados de la estación Nemizaque, correspondientes a la subcuenca del río Pienta. Este desemboca en el río Fonce y su origen proviene de los ríos Negro y Guacha [1]. Se encuentra ubicado en el departamento del Santander y caracterizado por un transporte medio anual multianual de 87,58 (kton/año) y un rendimiento medio anual multianual de 0,05 kton/año*km2 correspondientes al transporte de sedimentos y con un área aferente de 204,10 km2 [2].

La base de datos utilizada en este trabajo cuenta con caudales medidos diariamente desde 1991 hasta 2011 [3]. Sin embargo, cabe resaltar que existe ausencia de datos para todos los años en el mes de agosto, por lo cual el código desarrollado contiene algunas funciones para trabajar con dicha ausencia de datos.

A continuación, se realizará un análisis detallado de la subcuenca Pienta con el análisis de datos obtenido en Python 2.7.

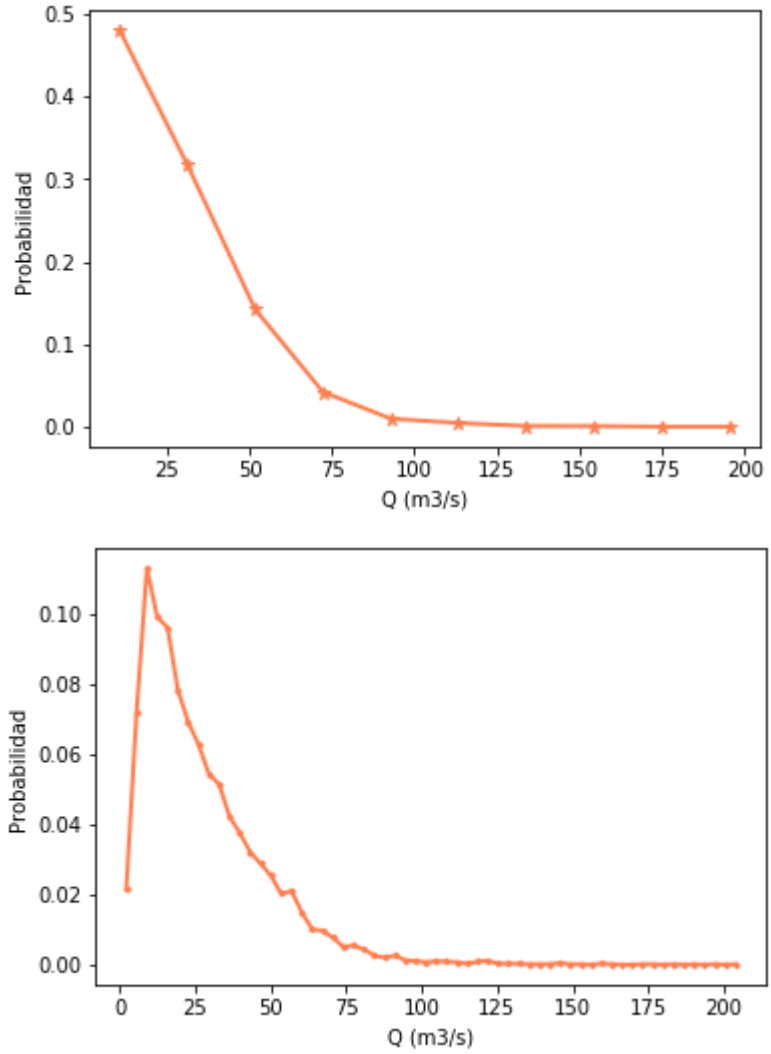
Primeramente, se presenta la gráfica de la serie de caudales (Gráfica 1), en la cual no se logra apreciar muy bien la ausencia de datos, ya que esta presenta un gran volumen de los mismos, haciendo que sean imperceptibles aquellos vacíos. De igual forma se puede evidenciar que los datos tienen la misma tendencia, pues la gráfica muestra una variabilidad estable.



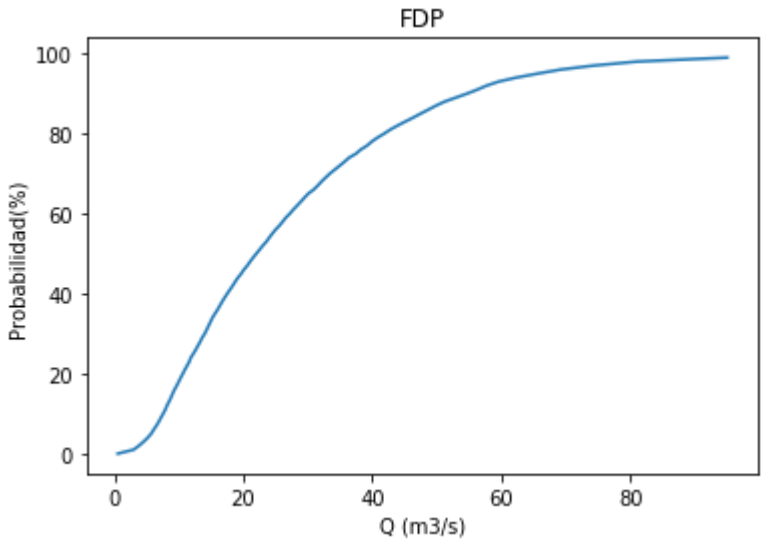
Gráfica 1. Serie de tiempo de caudales.

Luego de esta primera observación de los datos, procedemos a realizar un análisis más profundo. Se grafica un Histograma (Gráfica 2) y la Función de Distribución de Probabilidades (Gráfica 3), ambas gráficas permiten observar las frecuencias y probabilidades de la serie de tiempo.

Si tienen los mismos ejes, cuál es la diferencia entre ambas??



Gráfica 2. Histogramas de Caudales



Gráfica 3. Función de Distribución de Probabilidades de Caudales

A partir de estas gráficas se puede interpretar que la mayoría de los caudales se concentran en valores menores a 50 m3/s y que los valores más extremos son más dispersos y menos frecuentes.

Para concretar las observaciones anteriores y suministrar más información acerca de los datos, se hace el cálculo de algunos estadísticos (índices de localización, dispersión, simetría, etc.), los cuales se aprecian en la siguiente tabla:

Estadísticos de Q1	
mean	27.412.135
std	20.704.530
min	0.500000
25%	12.240.000
50%	22.000.000
75%	37.475.000
max	205.600.000
IQR	25,235
Asimetría	1.855.996
Kurtosis	6.395.774
YK	0.226471

Tabla 1. Medidas estadísticas

El valor de la media y la desviación estándar evidencian lo anteriormente descrito de las gráficas 2 y 3, pues el valor de la media es de 27,41 m3/s, claramente menor a 50 m3/s y el valor de la desviación estándar es de 20,70 m3/s, al ser un valor alto y casi cercano al de la media, hace evidente la gran dispersión de los datos.

Otros valores como los percentiles 25, 50 y 75, o sea los cuartiles, permiten evaluar la dispersión y la tendencia de la serie. El 25% de los caudales son menores o iguales a 12,24 m3/s, el 50% de los datos son menores o iguales a 22 m3/s y el 75% de los datos son menores o iguales a 37,47 m3/s (concluyente con lo analizado anteriormente).

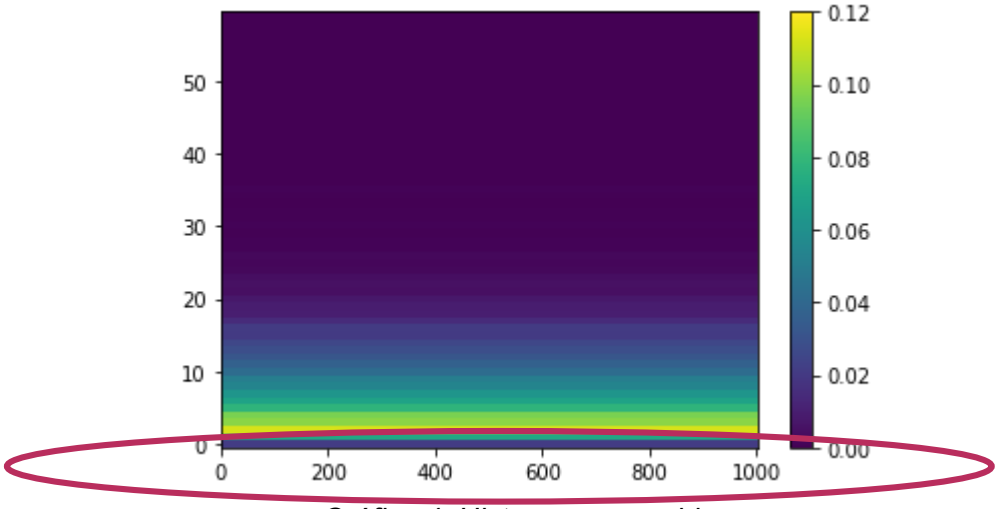
Con respecto al Rango Intercuartil que es un estadístico más robusto, el cual tiene un valor de 25,23 y teniendo en cuenta todos los valores que toman los caudales (con un mínimo de 0,5 m3/s y un máximo de 205,6 m3/s) y los valores de los índices anteriores, se puede decir que es relativamente bajo, así que se interpreta que en este rango los valores son medianamente similares.

Teniendo en cuenta que el valor del coeficiente de asimetría es bueno para indicar la forma de la gráfica, puede confirmarse que la gráfica tiene un sesgo positivo (hacia la derecha) muy pronunciado ya que el valor es positivo y lejano de cero (1,85) y deja más claro la escasa simetría de los datos. El valor del coeficiente de kurtosis (6,39) da indicaciones similares a las del coeficiente de asimetría, en este caso el número también es positivo y alejado del cero, queriendo decir que es una distribución leptocúrtica, o sea, una forma alargada.

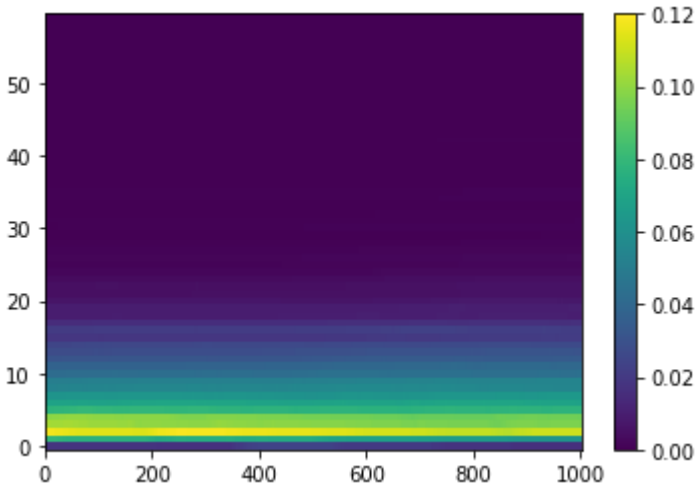
El valor del índice de Yule-Kendall que es de 0,22 confirma la información que destaca el coeficiente de asimetría y este análisis se apoya en la gráfica 8, pues al ser positivo quiere decir que es una asimetría positiva.

Otro índice es la moda, que para esta serie tiene un valor de 8,3 m3/s, queriendo decir que este es el valor que más se repite.

El paso siguiente es responder la pregunta, ¿son estacionarios los histogramas y los percentiles? Para ello se desarrollaron en Python una serie de códigos que permiten obtener un par de imágenes donde se puede apreciar la estacionariedad y se muestran a continuación:

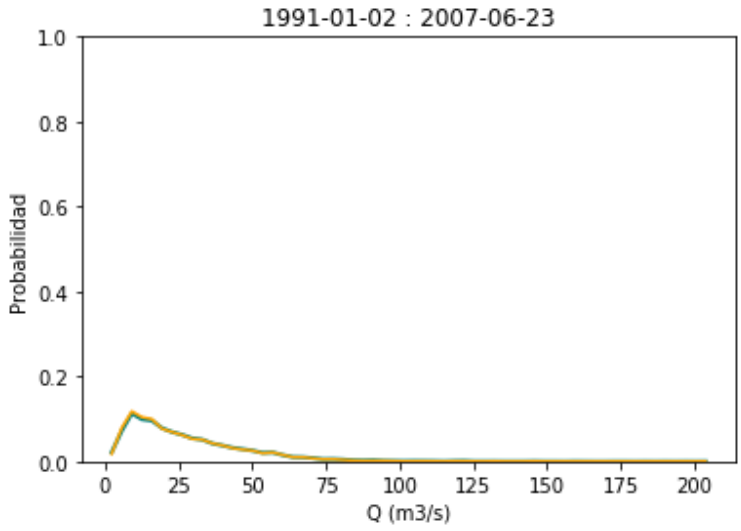


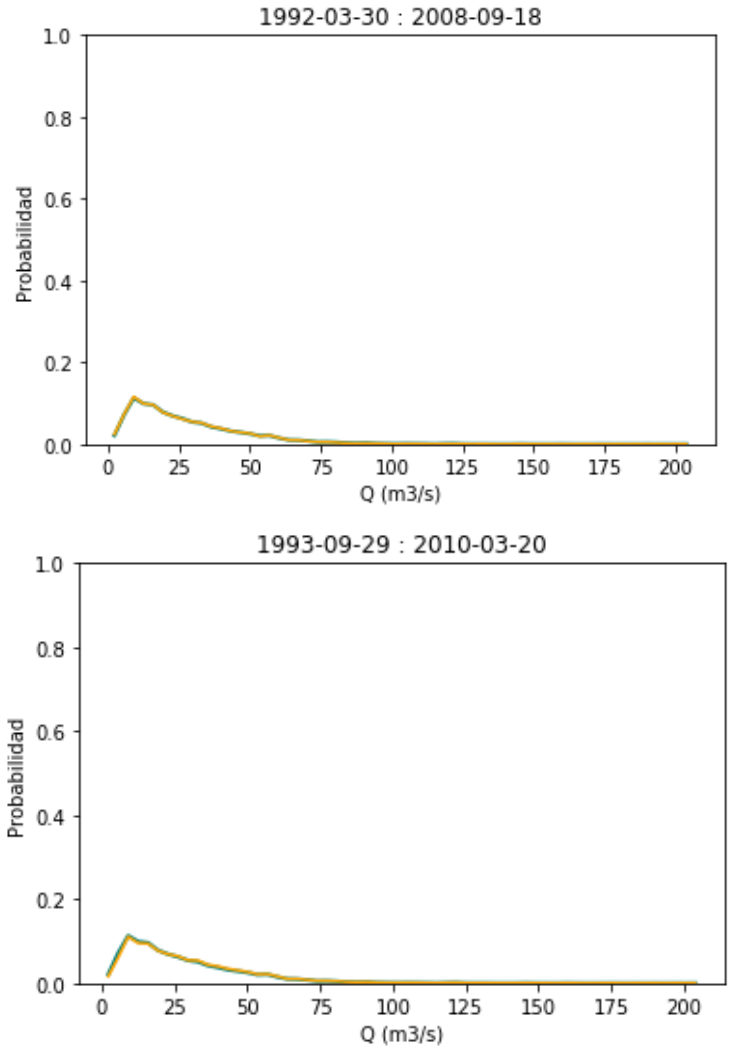
Gráfica 4. Histogramas vs. bins



Gráfica 5. Histogramas vs. bins

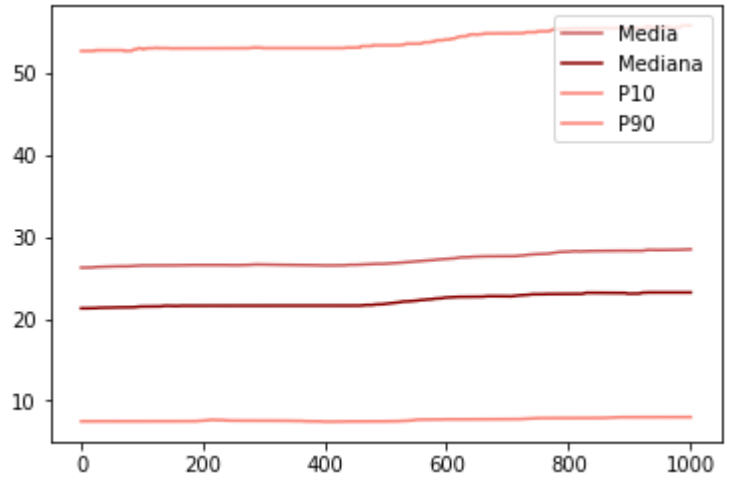
En este caso, las gráficas anteriores ilustran la estacionariedad de los histogramas de la serie de caudales analizada, ya que esta no tiene variabilidad y sus datos se concentran en una franja continua con una mínima dispersión de los mismos. Adicionalmente, las siguientes gráficas 6 de los histogramas a través del tiempo, reafirman todo el análisis anteriormente descrito.



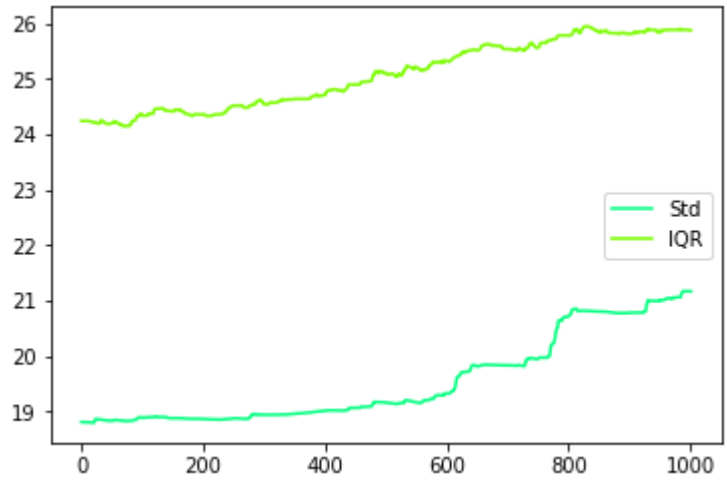


Gráfica 6. Histogramas vs. Bins (por ventanas de tiempo

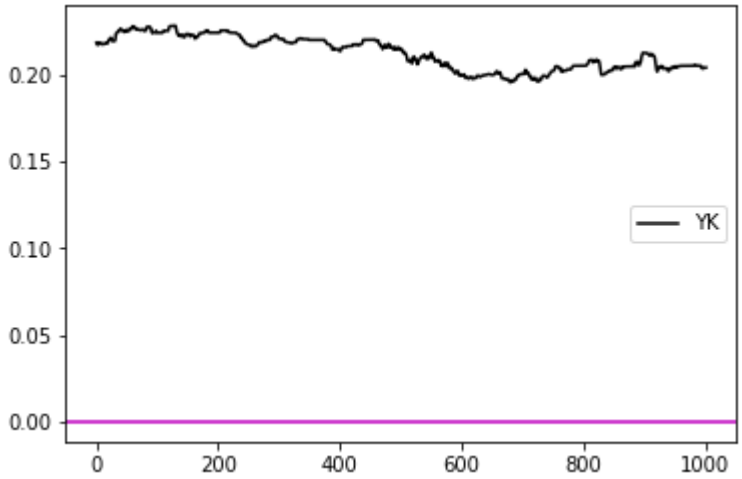
Además, se obtienen otras gráficas que, de manera similar, permiten observar la ausencia o presencia de estacionariedad de los índices (Gráficas 7-9).



Gráfica 7. Índices Media, Mediana, P10 y P90 a través del tiempo



Gráfica 8. Índices IQR y Desviación Estándar a través del tiempo



Gráfica 9. Índice YK a través del tiempo

Al igual que el análisis anterior, la gráfica 7 muestra la estacionariedad de los índices Media, Mediana, Percentil 10 y Percentil 90. Se conservan a lo largo del tiempo en una misma línea horizontal con mínimas perturbaciones.

La gráfica 8 puede ser caracterizada visualmente por la similitud y constancia del IQR y del std en los primeros datos de la secuencia. Esta última tendencia se evidencia en la gráfica 9, es decir que ambas tienen mínimas perturbaciones. En los últimos datos, comparándolas con la gráfica 7, tienen una ligera elevación. Sin embargo, no se logra apreciar una inclinación pronunciada; por lo tanto, se puede decir que, para todos estos índices representados existe una variabilidad regular y que estas perturbaciones se deben a anomalías que afectan los datos.

.

Finalmente, con el test de Mann Kendall se busca rechazar la hipótesis nula de la existencia de aleatoriedad en la serie descrita, para ello se encuentra el estadístico z que nos permite realizar esta acción. Para este caso, el z es igual a 6.83627559 y

con una significancia del 97.5%, dado que nuestro α/2 es 2.5%, se encuentra el estadístico z° de la distribución normal estándar. Lo anterior se debe a que se desea buscar la existencia de una tendencia de la serie, indiferentemente de si esta es ascendente o descendente, es decir la hipótesis alterna. Para ello el valor absoluto de z debe ser mayor o igual a z° de la distribución normal estándar, y ya que z° es 0,065 la hipótesis nula se rechaza [4].

Tendencia para los demás estadísticos? Dónde se evidencia comparación con distribución normal?

Conclusiones

De los análisis anteriores se puede concluir que para este caso donde los datos son tan dispersos y tan sensibles ante valores extremos, es mejor como medida de tendencia central la mediana, ya que es más robusta y resistente; y como medida de dispersión tiene una mejor funcionalidad el rango intercuartil por la misma razón. Es decir que la media y la desviación estándar son medidas poco resistentes y al haber valores tan dispersos son muy modificables y no brindan información tan concisa y confiable.

Otro resultado que cabe resaltar es que la serie de datos de los caudales tienen algún tipo de tendencia, es decir que al ser analizada a ambas colas, esto dado por α/2, se analiza que nuestra serie tenga tendencia ascendente o descendente esto gracias a que se rechazó la hipótesis nula. Además de ello, es asimétrica y sus datos son en su gran mayoría pequeños por su alta concentración en un rango no superior a los 50 m3/s.

Finalmente, todo el análisis de datos realizado representa una cuenca pequeña, con un río de caudales bajos con poca variabilidad y alta concentración de los mismos valores. Por ende, tiene poca ocurrencia de eventos extremos como caudales muy elevados y la serie de datos posee un ciclo anual definido dada la estacionariedad observada. Son datos con una variabilidad regular, ya que son datos que tienen un ciclo anual en este caso por un forzamiento externo radiativo, así que las irregularidades vistas en cualquier caso se deben a anomalías que para estos caudales colombianos se deben principalmente al fenómeno ENSO, generando para esta cuenca caudales menores en fase El Niño y caudales mayores en fase La Niña.

Referencias

[1] Disponible en: http://www.parquesnacionales.gov.co/portal/es/parques-nacionales/santuario-de-flora-y-fauna-guanenta-alto-rio-fonce/hidrografia/

[2] Dispoible en: http://documentacion.ideam.gov.co/openbiblio/bvirtual/023080/ENA_2014.pdf pag 291

[3] Código cuenca: 24027030 (caudales IDEAM)

[4] Tomado de: https://vsp.pnnl.gov/help/vsample/design_trend_mann_kendall.htm