**TAREA N° 1**
**ANÁLISIS DE DATOS AMBIENTALES**

# 4,4
# Muy bien :)

**Por:**
Lorena Andrea Benavides Riaño
Sergio Andrés Ospino Ricardo

**Profesor:**
Carlos Hoyos

**Universidad Nacional de Colombia – Sede Medellín**
**Facultad de Minas**
**2018**

**Tarea # 1. Análisis de Datos Ambientales.**

La serie temporal escogida corresponde a la temperatura mínima diaria registrada en la ciudad de Melbourne (Australia), entre el primero de enero de 1981 y el 31 de diciembre de 1990. Los datos fueron obtenidos del Australian Bureau of Meteorology, y la temperatura fue medida en unidades de grados Celsius (°C). En la Figura 1 se consolida toda esta información.

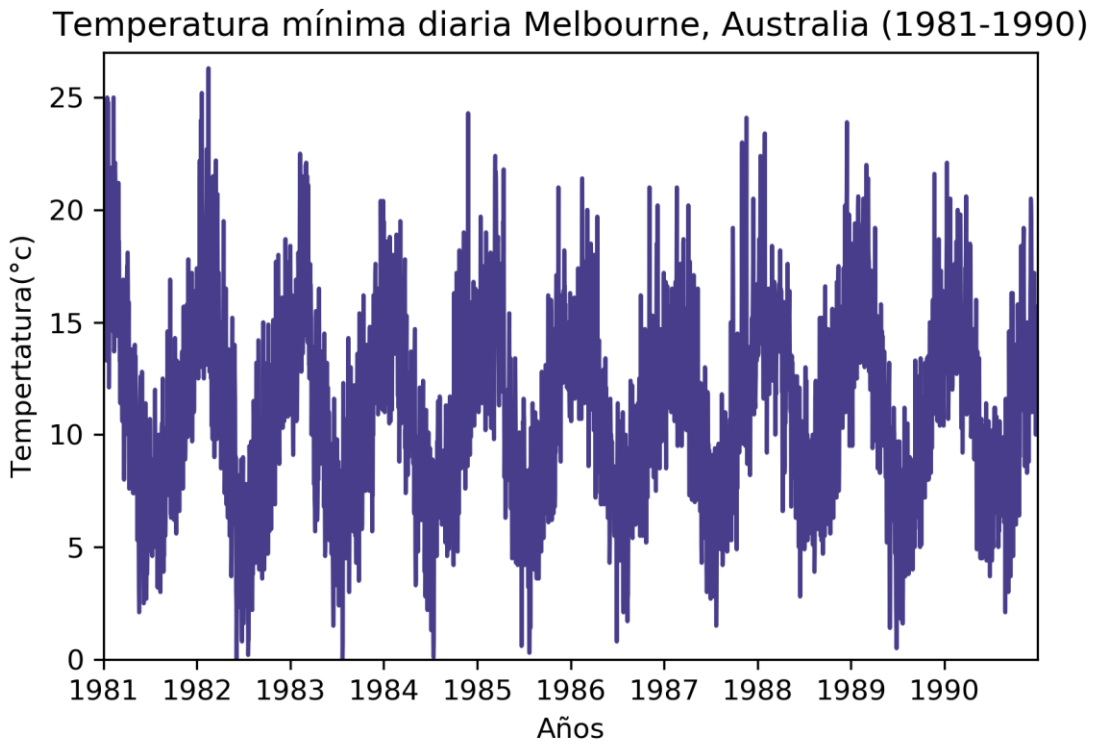


*Figura 1. Serie de tiempo.*

De la gráfica se puede apreciar un patrón que se repite anualmente: entre el último mes de un año y los primeros meses del año consecutivo la temperatura mínima alcanza su máximo valor, y luego tiene un comportamiento decreciente hasta, aproximadamente, la mitad del año, donde registra los valores mínimos. A partir de ahí la gráfica demuestra un comportamiento creciente hasta llegar a diciembre, para dar inicio nuevamente al ciclo. La información suministrada por la gráfica coincide con las estaciones para Australia, que al estar ubicada en el hemisferio sur posee una temporada cálida que dura aproximadamente tres meses, desde la mitad de diciembre hasta mediados de marzo; mientras que su temporada fría tiene una duración aproximada de tres meses y medio, desde finales de mayo hasta inicios de septiembre.

**Ojo con el lenguaje técnico chicos**

Al graficar la curva de distribución de probabilidades (Figura 2), se observa inicialmente cierto grado de simetría en la distribución de los datos respecto a un valor central. Más adelante analizaremos esto con más detalle.

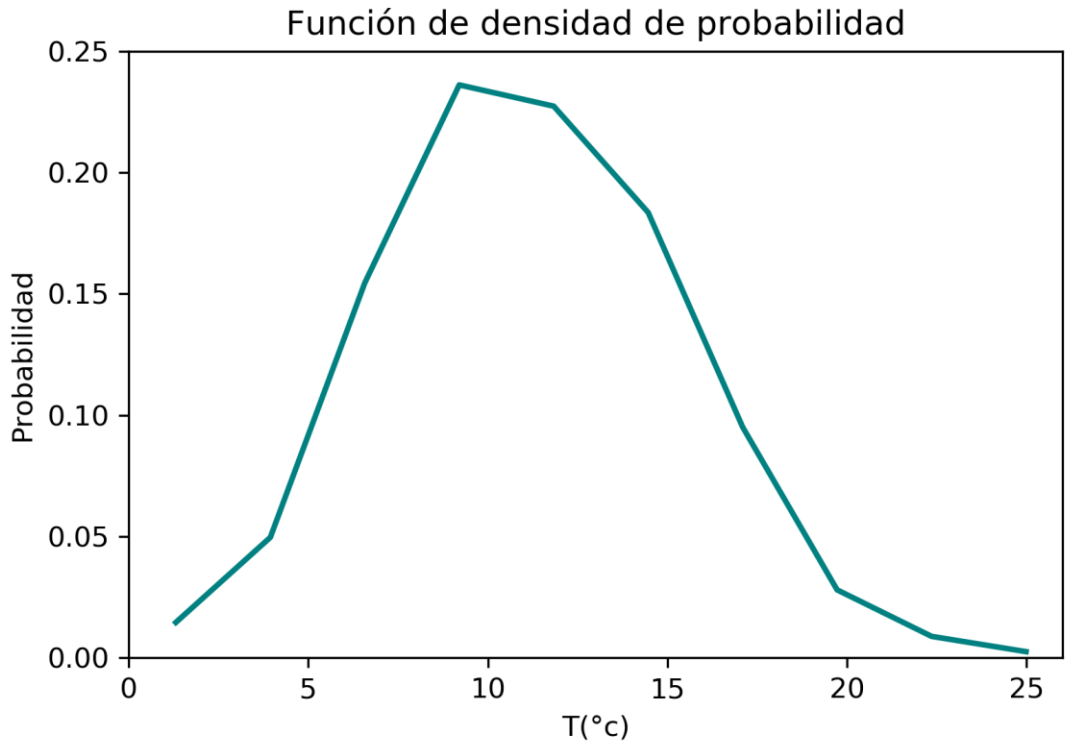


*Figura 2. Función de densidad de probabilidad.*

Respecto a los percentiles, se encontró que el primer cuartil es 8.3°C, la mediana es 11.0°C, y el tercer cuartil es 14.0°C (ver Figura 3).

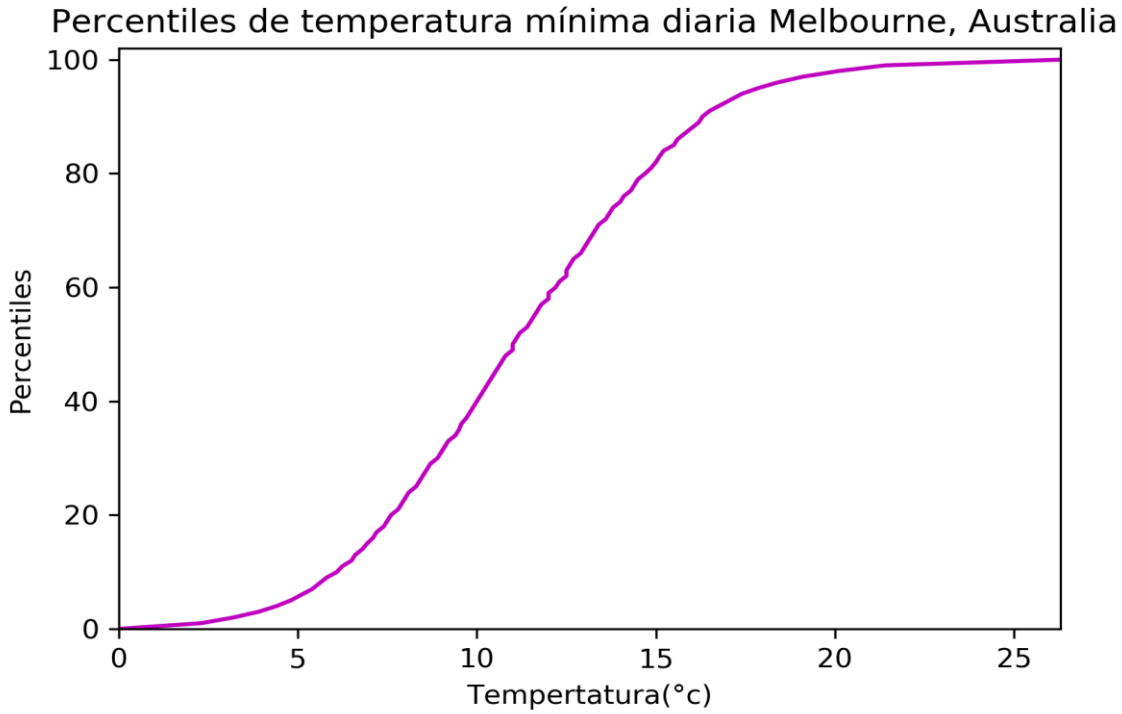


*Figura 3. Gráfica de los 100 percentiles de la serie temporal.*

Ahora bien, esta información en sí misma no es muy diciente. Fue necesario, en primera instancia, hacer una caracterización de los datos en el tiempo, para lo cual se tuvo que establecer unas ventanas móviles con una longitud de noventa días (duración aproximada de cada estación en Australia). Posteriormente, debieron estimarse los respectivos índices paramétricos y no paramétricos para hacer de ellos los análisis pertinentes.

En principio, la media y la mediana (índices de localización) no están muy alejadas la una de la otra, casi solapándose sus curvas (Figura 4). Esto indica la alta resistencia de ambos métodos para esta serie en particular, toda vez que no existen "outliers" o valores atípicos que tengan una influencia significativa en la distribución de los datos.

Los índices de dispersión (Figura 5) —desviación estándar y rango intercuartil—, denotan un comportamiento esperado, a saber, que el primero es menor que el segundo. Si bien las curvas no están muy alejadas entre sí, se reafirma el hecho de que el rango intercuartil es la medida de dispersión más robusta y resistente por defecto.

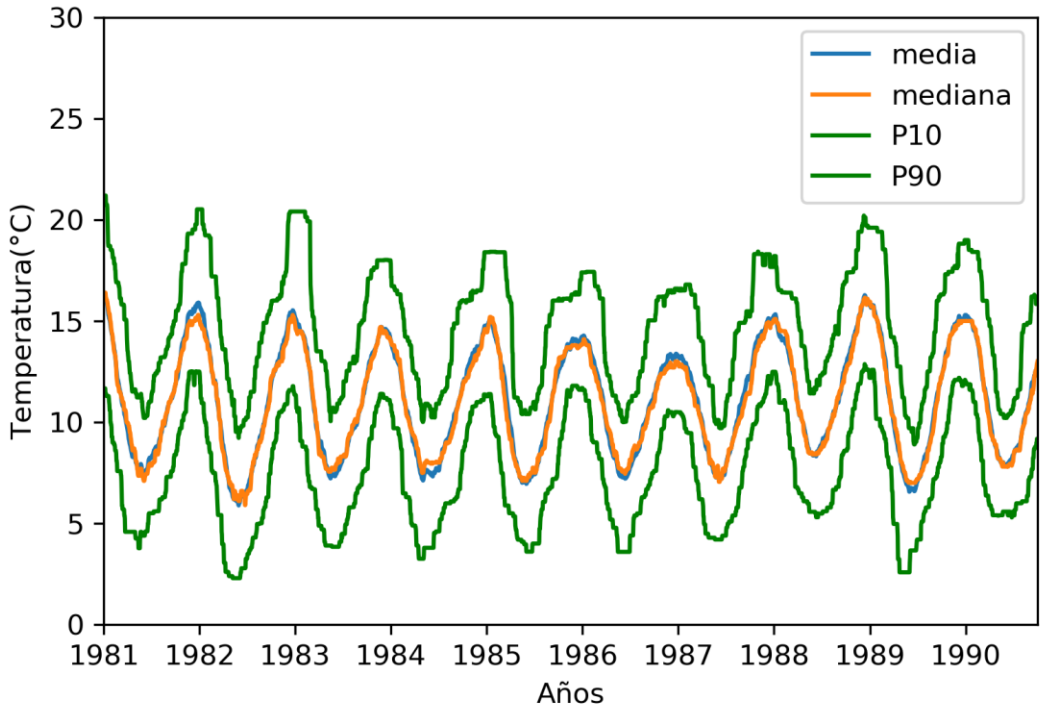


*Figura 4. Índices de localización, con los percentiles $p_{10}$ y $p_{90}$.*

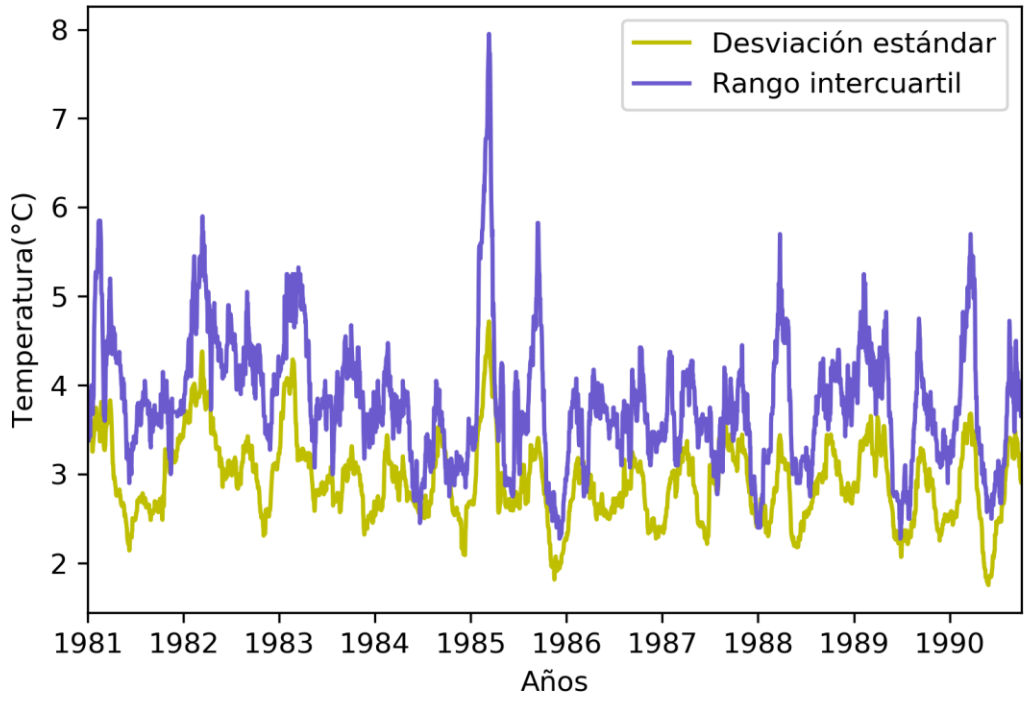


*Figura 5. Índices de dispersión.*

Finalmente, para el índice no paramétrico de simetría (Yule-Kendall) se observa gráficamente (Figura 6) que no existe ninguna preferencia de la distribución por un lado o por el otro; es decir, la serie se distribuye de forma simétrica.

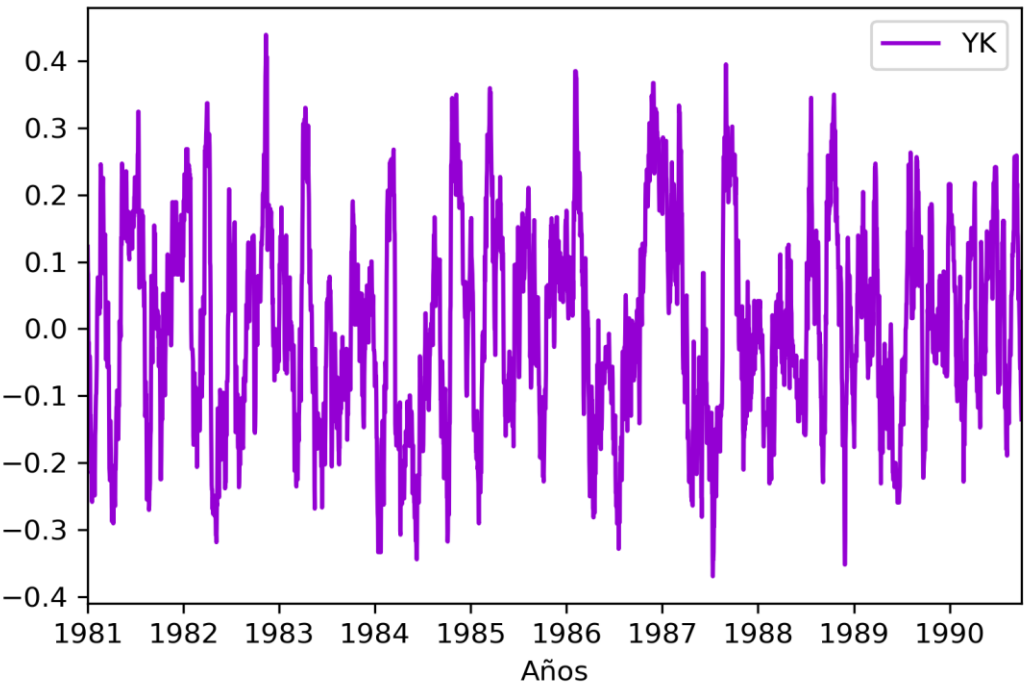


*Figura 6. Índice de simetría (Yule-Kendall).*

Ninguno de los índices mencionados hasta ahora es estacionario. Para el caso de la media y la mediana, existe una variación periódica a lo largo del tiempo; mientras que para la desviación estándar, el IQR, y el índice de Yule-Kendall, los comportamientos son erráticos y, en definitiva, no son constantes.

Así mismo, para analizar la estacionariedad de los histogramas basta con remitirse a la Figura 7, en la cual se observa claramente que la distribución de frecuencias a lo largo del tiempo no es constante. Por lo tanto, los histogramas no son estacionarios.

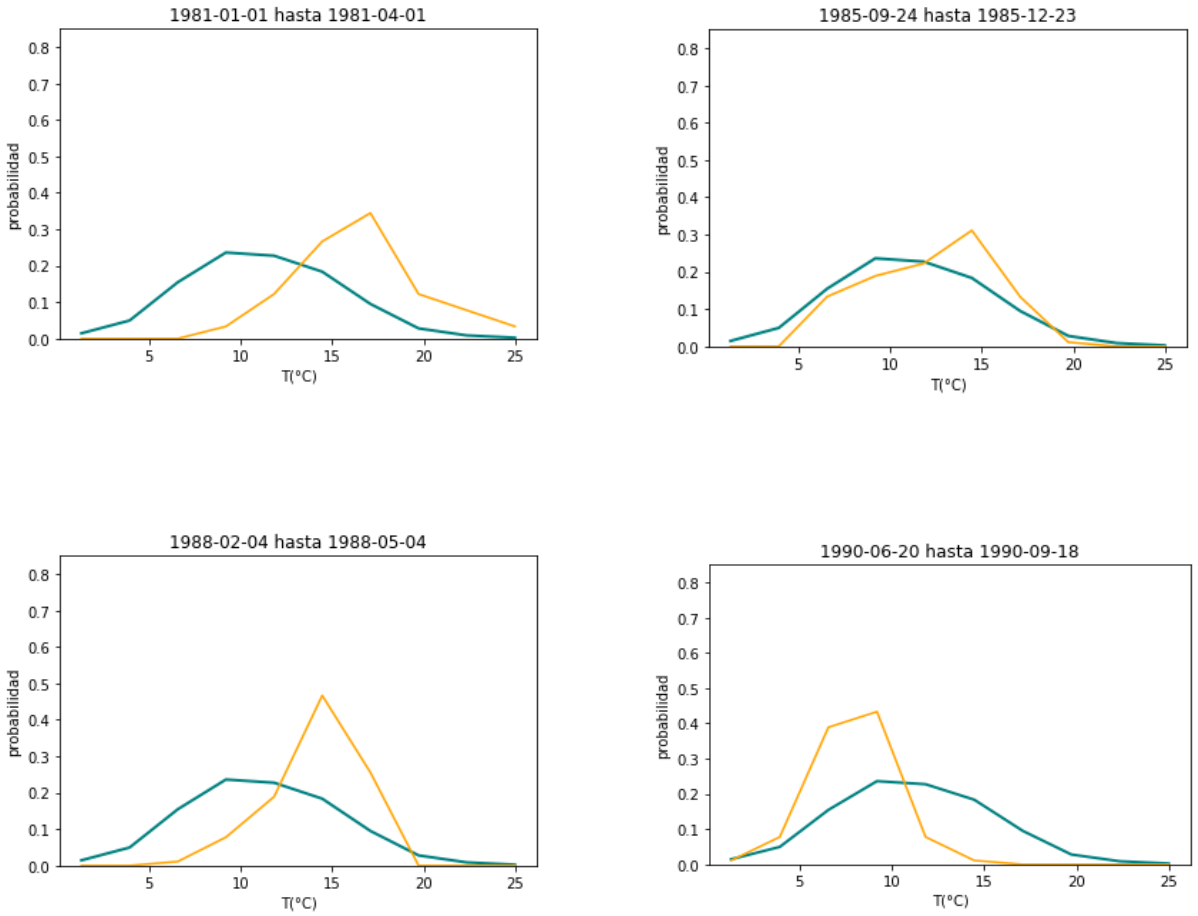


*Figura 7. Histogramas analizados por ventana.*

Otra evidencia de la no-estacionariedad de los histogramas es visible en la Figura 8. Esta gráfica denota las zonas de mayor masa de probabilidad (color amarillo) y las de menor masa (color verde-azulado) a lo largo del período analizado. Evidentemente, las zonas amarillas varían en el transcurso de los años, con lo cual los histogramas no se mantienen constantes.

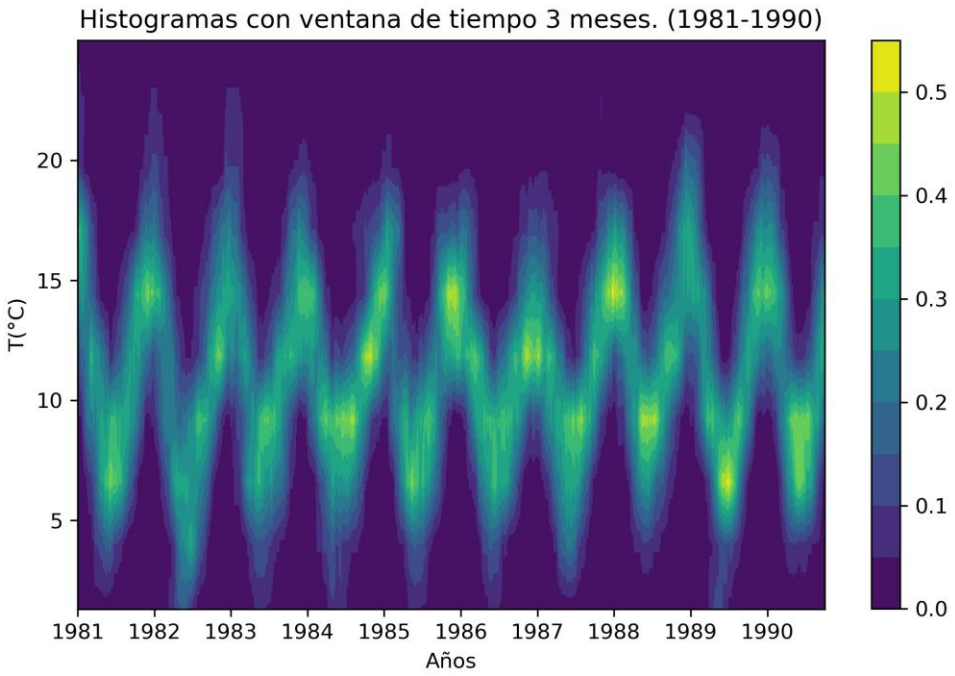


*Figura 8. Histogramas.*

Como parte final del análisis de esta serie de tiempo, se quiso saber si existía alguna tendencia en los datos, para lo cual se empleó un test no paramétrico (Mann-Kendall). A continuación se describe su metodología y los resultados obtenidos.

- **Test de Mann-Kendall**

El test de Mann-Kendall es un test no paramétrico que permite identificar si existe una tendencia en la serie temporal que se desea estudiar. El test busca comparar las magnitudes relativas de los datos. Una de las ventajas de usar este test es que no le impone una distribución a los datos. El test asume que sólo existe un dato asociado a un período de tiempo y si se diera la existencia de múltiples valores para un periodo de tiempo específico, se debe trabajar con el promedio de dichos valores. En el test lo que se busca es comparar cada dato con el dato inmediatamente siguiente. El valor inicial del estadístico de Mann-Kendall es 0, es decir que no presenta tendencia, pero si el dato de un período de tiempo posterior es mayor que el dato de un período de tiempo anterior el estadístico S toma un valor de 1. De lo contrario, si el valor en el período de tiempo posterior es menor que el valor en el periodo anterior el estadístico toma un valor de -1. El resultado final de los crecimientos y decrecimientos de los datos nos dará el valor de S. (Khambhammettu, 2005)

El valor estadístico S del test de Mann-Kendall está dado por la siguiente fórmula, con $x_1$, $x_2$,...$x_n$ los datos de la serie temporal y $x_j$ representa un punto en el tiempo j:

$$S = \sum_{k=1}^{n-1} \sum_{j=k+1}^{n} sign(x_j - x_k)$$

*Ecuación 1. Estadístico S.*

donde:

$$sign(x_j - x_k) = 1 \text{ if } x_j - x_k > 0$$
$$= 0 \text{ if } x_j - x_k = 0$$
$$= -1 \text{ if } x_j - x_k < 0$$

Un alto valor positivo de S es un posible indicador de una tendencia creciente, y un valor muy negativo es un posible indicador de una tendencia decreciente. Sin embargo, es necesario calcular la probabilidad de S para concluir estadísticamente si existe o no tendencia.
Kendall describe una aproximación normal para el test que puede ser usada para muestras con más de diez valores, considerando que no se presenta un gran número de empates en la muestra. A continuación se describe el procedimiento para el test:

- Primero se calcula el valor para el estadístico S (Ecuación 1).
- Luego se calcula el valor de la varianza de S dada por la siguiente fórmula:

$$VAR(S) = \frac{1}{18}\left[ n(n-1)(2n+5) - \sum_{p=1}^{g} t_p(t_p - 1)(2t_p + 5) \right]$$

*Ecuación 2. Varianza de S.*

Donde n es el número de datos, g es el número de grupo de empates (un grupo de empate es donde los datos toman el mismo valor) y $t_p$ es el número de puntos de datos en un grupo. En el siguiente vector {2, 3, 4, 3, 4, 5, 3}, tenemos n=7, g =2, $t_1$=2 para los valores de 4, y un $t_2$ =3 para el empate en el valor 3.

- Calcular el estadístico Z normalizado

$$Z = \frac{S-1}{[VAR(S)]^{1/2}} \quad \text{if S} > 0$$
$$= 0 \qquad \text{if S} = 0$$
$$= \frac{S+1}{[VAR(S)]^{1/2}} \quad \text{if S} < 0$$

*Ecuación 3. Estadístico Z.*

- Calcular la probabilidad del estadístico en la función normal. La función de densidad de la distribución normal tiene una media igual a 0 y una desviación estándar igual a 1.

- Por último, dado un nivel de significancia α, se procede a determinar el punto crítico que es el Z proveniente de la distribución normal y luego se compara los dos valores obtenidos: el Z de la muestra y el Z de la normal. Si $Z_{muestra} > Z_{normal}$ el resultado sugiere una tendencia creciente, sí $Z_{muestra} < -Z_{normal}$ el resultado sugiere una tendencia decreciente. Si ninguna de los dos se cumple la evidencia es insuficiente para demostrar la existencia de una tendencia. (EPA,2009)

En el test de Mann-Kendall se pretende rechazar la hipótesis nula

$H_0$: Existe aleatoriedad en los datos

La hipótesis alternativa en este caso es determinar si existe una tendencia.

- **Resultados**

Realizamos el test de Mann-Kendall para el caso específico de las temperaturas mínimas diarias en Melbourne (Australia) utilizando el lenguaje de programación Python, para lo cual se usaron los métodos de programación aprendidos en clase y así calcular los resultados de los estadísticos anteriormente descritos. Los valores obtenidos de los estadísticos del test fueron:

**Valor S:** 243730.0
**Valor Z:** 0.8112
**Valor Varianza:** 90274034596

El resultado del estadístico S es muy grande y positivo lo que sugiere que puede existir una tendencia, pero al analizar el valor de Z el número es menor que 1, por lo tanto se debe evaluar en la distribución normal para poder calcular su probabilidad y así poderlo comparar con el resultado de Z de distribución normal.

El test de Mann-Kendall es un test de dos colas porque evalúa tanto tendencia creciente como decreciente, es por esto que al comparar con la distribución normal se toma un nivel de confiabilidad de α/2, porque la tendencia puede ser creciente o decreciente.

Una observación importante es que la serie contiene un gran número de grupos de empates, en total se encontraron 65 grupos diferentes de empates, lo que afecta significativamente el resultado de la varianza de S y, por ende, el estadístico Z.

Observando la gráfica de los datos, la evidencia visual no sugiere una tendencia marcada a lo largo del tiempo. Esto se debe a que nuestros datos corresponden a un fenómeno de tipo estacional con ciclo anual, es decir, que los datos se comportan de manera similar cada año porque tienen influencia de un factor externo (en este caso, el Sol). Por esto las hipótesis que definimos para nuestro caso de estudio fueron:

$H_0$: Existe aleatoriedad en los datos
$H_1$: Existe tendencia en los datos

Se escogió un nivel de confianza de α/2 porque queremos evaluar si existe tendencia en los datos sin importar si es creciente o decreciente dado que la naturaleza de los datos hace cuestionar la existencia de la tendencia. El nivel de significancia utilizado fue el 95% con un α/2=0.025.

El valor de Z de la distribución normal con α/2=0.025 fue $Z_{normal}$= 1.959. Al comparar el $Z_{normal}$ con $Z_{serie}$ se concluye que:

$Z_{normal} > Z_{serie}$

**Falta el análisis para los demás estadísticos**

Por lo tanto no es posible rechazar la hipótesis nula y tampoco nos sugiere que se cumple la hipótesis alternativa. Para este caso específico el test de Mann-Kendall no cuenta con evidencia suficiente para sugerir la existencia de una tendencia en la serie.

Una explicación para esto es que debido a la existencia de un gran número de empates la serie temporal afecta significativamente la varianza y por lo tanto reduce la validez de la aproximación normal del test.
Al no existir evidencia de una tendencia es necesario usar un método llamado “Seasonal Kendall Test” donde evalúa cada estación del año por separado; es decir, la primavera de cada año solo será comparada con las primaveras de los años posteriores y así se puede concluir si existe tendencia o no para cada estación del año en específico. Sin embargo, considerando la serie temporal en toda su extensión no es posible evaluar una tendencia. (Salmi, 2002)

## Referencias

Environmental Protection Agency (EPA). (2009). *Statistical Analysis Of Groundwater Monitoring Data At Rcra Facilities.* Estados Unidos.

Khambhammettu, P. (2005). *Mann-Kendall Analysis for the Fort Ord Site.* Retrieved from http://www.statisticshowto.com/wp-content/uploads/2016/08/Mann-Kendall-Analysis-1.pdf

Salmi, T. (2002). *Detecting Trends Of Annual Values Of Atmospheric Pollutants By The Mann-Kendall Test.* Helsinki, Finlandia.