4.3
Bien :)

Análisis exploratorio de datos:
Temperatura superficial del Océano Niño 3.4.

Juan Diego Mantilla Quintero

Introducción

El Niño y La Niña son fases opuestas del ciclo natural que causa mayor variabilidad climática en la zona tropical de Océano Pacífico y que cambia en promedio de una fase a la otra cada 3 a 7 años. El proceso en conjunto es llamado ENSO (El Niño-Southern Oscillation). El ciclo genera un cambio en la temperatura superficial de océano en la región este del Pacífico, de más cálido (El Niño) a más frío (La Niña), así como una variación en la dinámica de los vientos y presiones de la zona que en suma afectan el clima del trópico y el subtrópico.

La Administración Nacional Oceánica y Atmosférica de Estados Unidos (*National Oceanic and Atmospheric Adinistratios, NOAA)* monitorea la temperatura superficial del Océano Pacífico en la zona del Niño 3.4. Esta región está comprendida entre las latitudes 5 Sur y 5 Norte y los meridianos 120 a 170 de longitud oeste *(Figura 1)*. Basados en el promedio mensual de la temperatura, su pronóstico para la temporada precedente y la dinámica de presiones, establecen si el ENSO es El Niño, Neutral o La Niña.

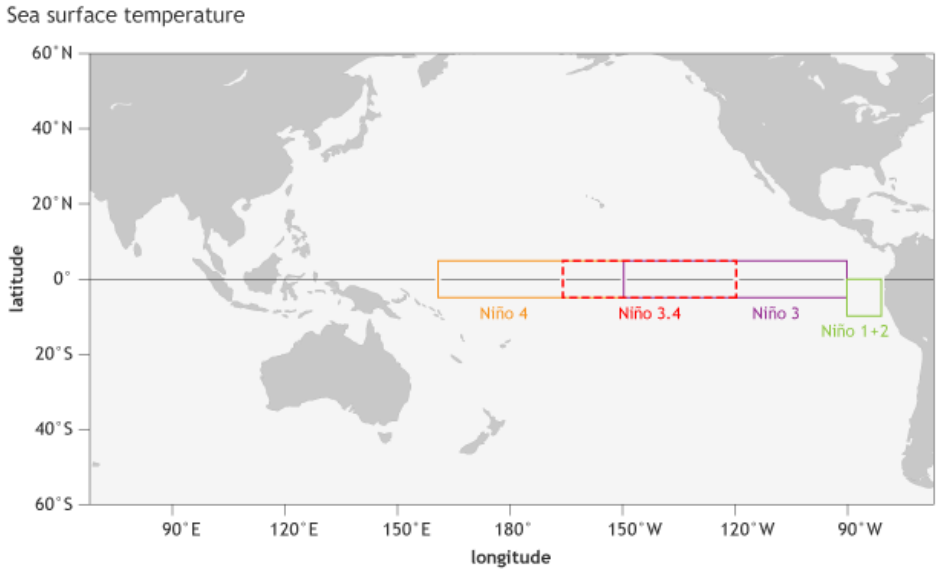

Figura 1: Ubicación de las regiones del Niño para la medición de la temperatura superficial del océano en la zona central y oriental del Pacífico. Fuente NOAA Climate.gov, imagen de Fiona Martin

El propósito del presente trabajo es realizar un análisis exploratorio de datos sobre la variación semanal de la temperatura del Océano Pacífico en la zona Niño 3.4.

Análisis Inicial

Del Centro de Predicción Climática *(Climate Prediction Center, CPC)* se obtuvieron los datos semanales de temperatura superficial promedio de la zona, desde enero de 1990 hasta agosto de 2018 (un total de 1495 datos). Estos fueron procesados y graficados (*Figura 2*) para mirar su variación durante el periodo de tiempo considerado.

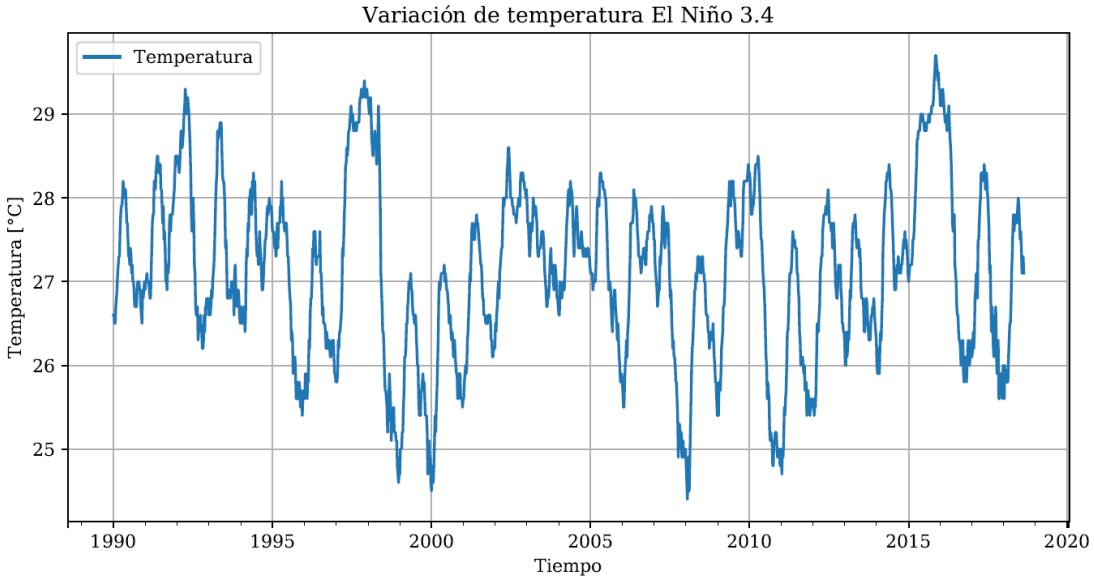

Figura 2: Variación de la temperatura del Niño 3.4.

La *Figura 2* muestra importantes variaciones interanuales en los registros obtenidos, con dos marcados picos de lectura de temperatura cálida, el primero entre 1997 y 1998 y el segundo entre 2015 y 2016. Al rededor del 2000 y 2010, se muestran lecturas de temperatura más fría. Los datos no muestran a simple vista alguna tendencia marcada.

Una característica importante de los datos es su función de distribución de probabilidades (PDF), esta puede mostrar mucho acerca del comportamiento de la variable. Esta función fue graficada en la *Figura 3* haciendo uso los percentiles de la serie. Se observa que las temperaturas entre 25.5 y 28°C agrupan más de la mitad de los datos, al aumentar la pendiente de la curva en esta zona. El inicio y el final de la curva son más tendidos, indicando que temperaturas por debajo de 25°C o por encima de 28.5°C no son usuales. Por otro hay una mayor pendiente en extremo final de los datos comparado con el inicio, lo que da a entender que entre estos dos extremos es más recurrente presentar temperaturas altas.

Otra característica importante de la variable es su función de densidad de probabilidades (pdf). Para determinar su forma se hace necesario establecer las probabilidades por intervalos de clase. En la *Figura 4* se muestra la función de densidad de probabilidades para distintos intervalos de

clase. Se decide utilizar un número de 10 intervalos, luego de concluir que un número más grande únicamente introduce ruido en el ejercicio.

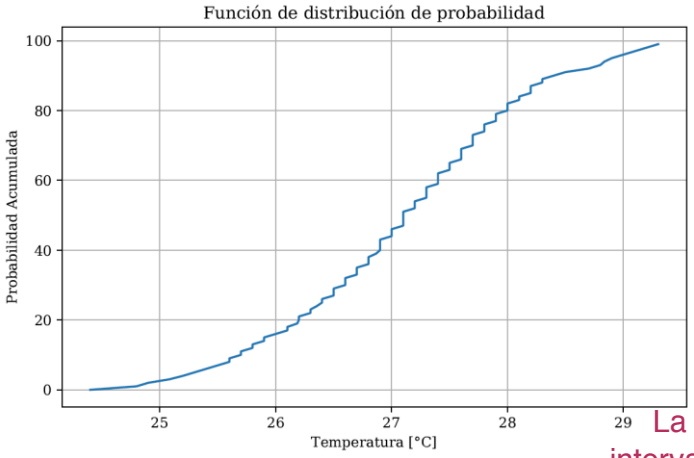

Figura 3: Función de probabilidad.

La exploración del número de intervalos es información irrelevante para el objetivo de la tarea, hace parte del "detrás de cámaras"

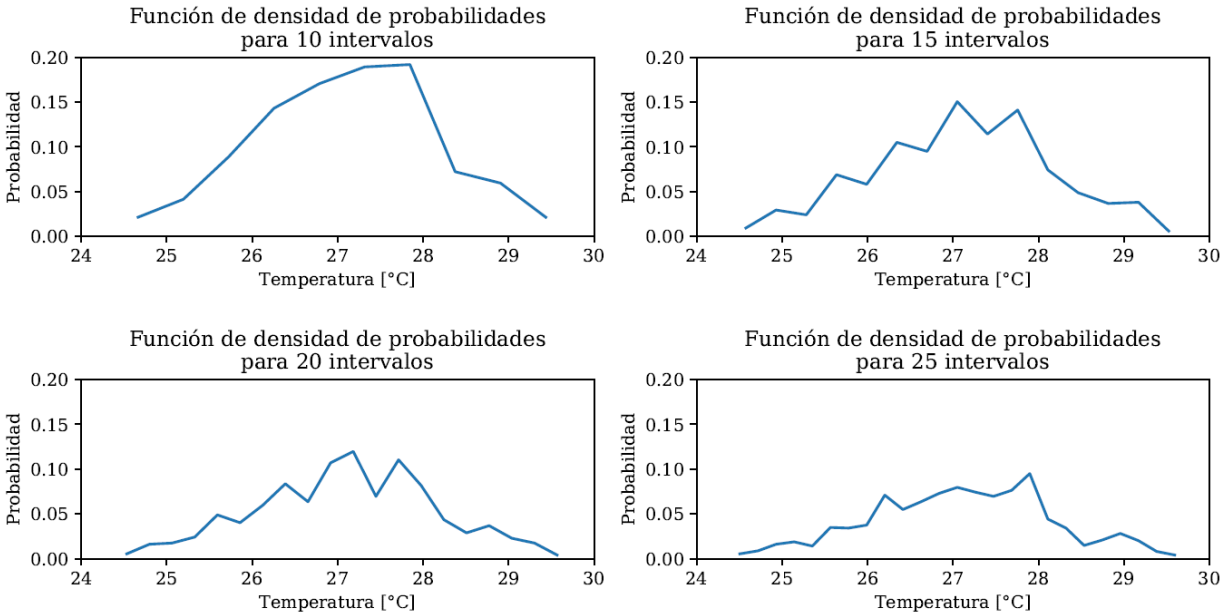

Figura 4: Función de densidad de probabilidades para 10, 15, 20 y 25 intervalos de clase

Utilizando entonces 10 intervalos de clase se graficó la función de densidad de probabilidades y el histograma de la serie *(Figura 5).* Como se intuyó con la función de distribución de probabilidades existe una mayor densidad dentro rango medio de los datos, más exactamente entre 25.7 y 28.3°C.

Ambos dan la misma info

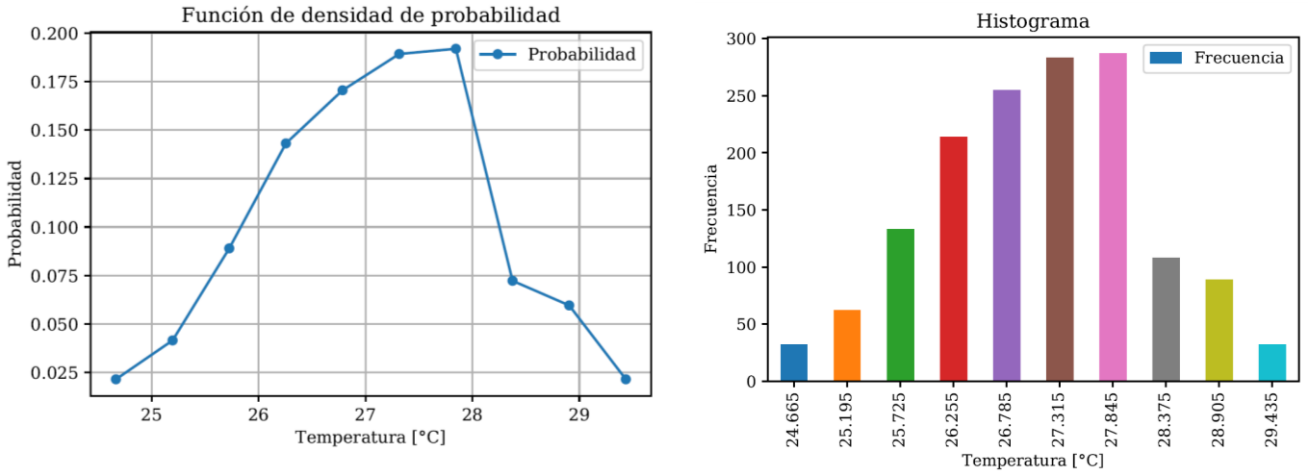

Figura 5: Función de densidad de probabilidad e histograma de la serie calculado para 10 intervalos de clase.

A pesar de que las gráficas de la función de probabilidades dan información importante para caracterizar el comportamiento de la variable de temperatura, se hace interesante calcular otras cantidades más simples que revelen información sobre el dominio de la función. Típicamente para este propósito se utilizan los primeros tres primeros momentos, sin embargo, estos no son propiedades robustas ni resistentes. Por robustez y resistencia hacen referencia a que los métodos sean insensibles a suposiciones que se hagan sobre la naturaleza de los datos (Wilks, 2006).

???

Clásicamente se considera que uno de los momentos más importantes la media. Este es el valor promedio de una variable aleatoria y revela una de las más importantes características de su distribución (Soong, 2004). Sin embargo, la medida es muy susceptible a valores extremos; en contra posición una opción es comprarla con la medida más robusta y resistente de tendencia central, la mediana (Wilks, 2006). De forma análoga los índices clásicos de dispersión y asimetría (desviación estándar y coeficiente de asimetría), fueron comparados con alternativas robustas y resistentes, como el rango intercuartil y el índice de Yule-Kendall, respectivamente. Los resultados pueden ser observados en la siguiente tabla:

Índices de localización	
Media	27.099
Mediana	27.1
Índices de dispersión	
Desviación estándar	1.045
Rango Intercuantil (IQR)	1.4
Índices de asimetría	
Coeficiente de asimetría	-0.0733
Índice Yule-Kendall	-2.54E-15

Qué quiere decir que sean tan cercanos?

Tabla 1: Índices de localización, dispersión y asimetría para la serie de temperatura.

Los índices de localización para este caso en particular muestran que la masa de la distribución se ubica, para ambos casos, en 27.1°C. En comparación con la media, la mediana es algunas veces preferida como medida de tendencia central en el caso de una distribución asimétrica (Soong, 2004). Para este caso, como lo indican los coeficientes de asimetría, se trata de una distribución muy simétrica, con una ligera tendencia hacia temperaturas altas por lo que revelan los signos obtenidos. Mientras que las medidas de dispersión permiten concluir que la serie no es muy diversa y se agrupa alrededor de su media.

Análisis de estacionariedad de la serie

Otro tipo de análisis interesante para la serie es verificar si tanto las distribuciones de probabilidad (pdf) como los índices de localización, distribución y asimetría son estacionarios, es decir que sus propiedades no se ven afectadas por cambios de origen temporal. Para esto se adoptó una ventana móvil para realizar un desplazamiento a lo largo de la serie de tiempo y calcular las mismas propiedades anteriormente obtenidas. Dado que el ENSO cambia de una fase a otra en un rango de 3 a 7 años se adoptó una ventana móvil de 5 años. Al tratarse de datos promedios semanales se definió el ancho de la ventana de 260 elementos.

Para observar el cambio en la distribución de probabilidades se procesaron las iteraciones para la función de densidad de probabilidades (pdf), las cuales que revelaron que la distribución no es estacionaria. En la *Figura 6* se muestran 4 de ellas.

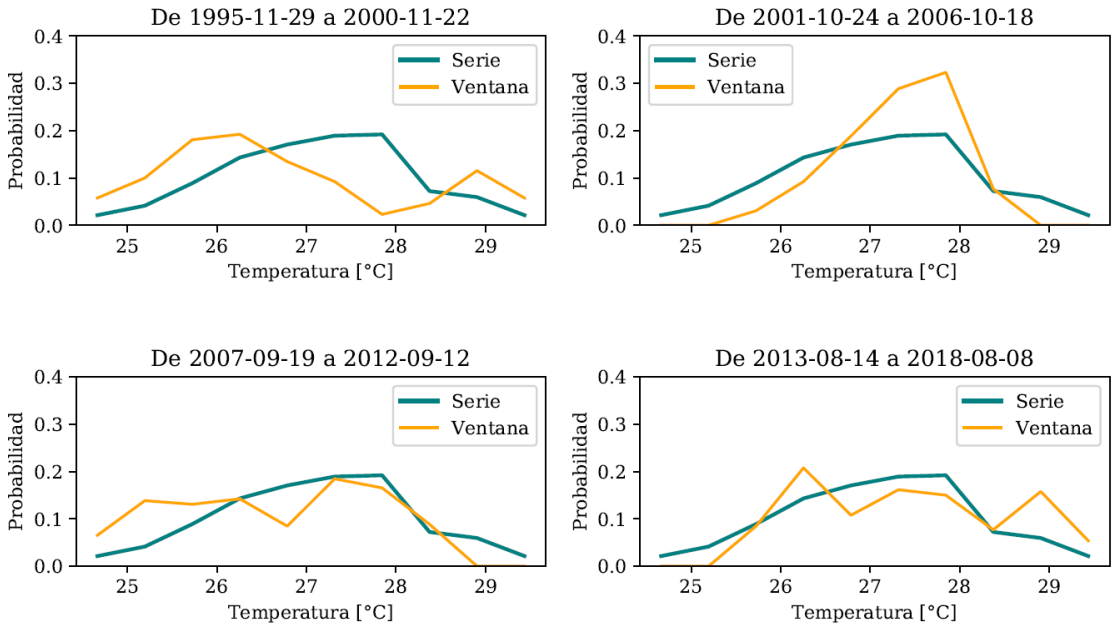

Figura 6: Variación de la función de densidad de probabilidades con la ventana móvil.

Este comportamiento es más contundentemente representado en el *Figura 7* donde se muestra esta variación en toda la serie de tiempo. En la imagen las zonas más claras tienen mayores

probabilidades que las zonas más oscuras. Claramente el comportamiento de la distribución no es estacionario. Se muestran fuertes cambios a lo largo del tiempo, antes de 1994 hay una mayor densidad en valores cercanos a la media, luego las probabilidades de distribuyen un poco hacia los extremos, pero tienden a concentrarse valores entre 25.5 y 27°C. Para el año 2002 se muestra un fuerte pico en temperaturas cercanas a los 28°C, precedido por un periodo donde la densidad se distribuye hace temperaturas más frías. Luego del 2010 se muestra un aumento de densidad para temperaturas más altas, en especial para aquellas cercanas a los 29°C. Sobre la *Figura 7* se graficó la media y la mediana móvil, las cuales muestran un comportamiento similar y acentúan las anteriores conclusiones.

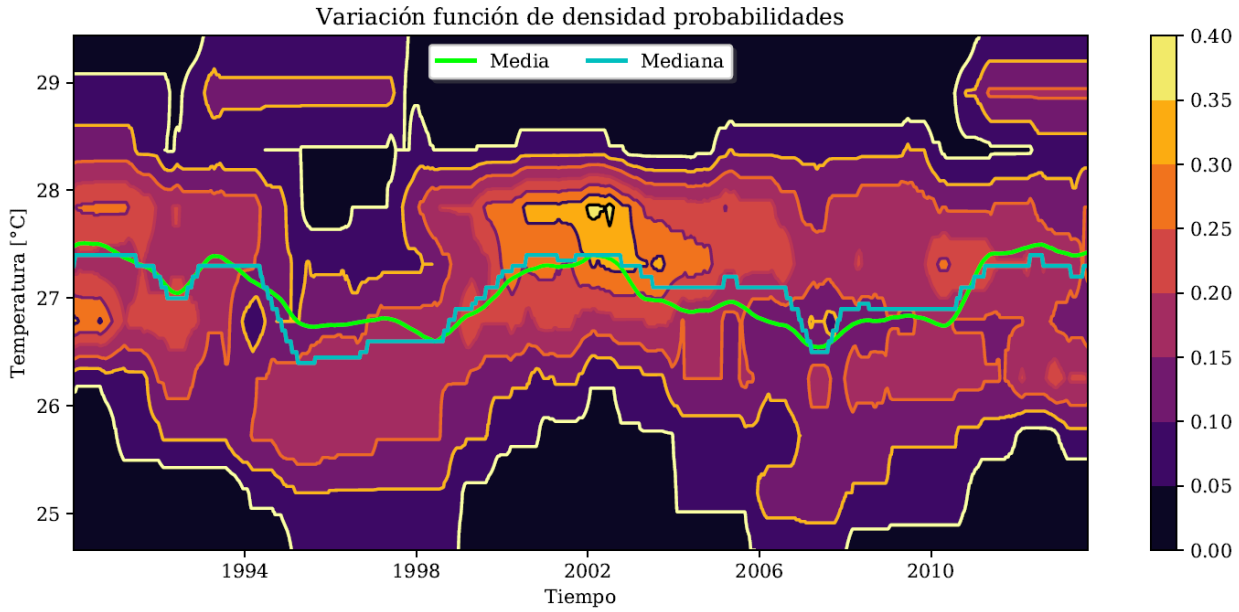

Figura 7: Variación de la función de densidad de probabilidades. En color verde se graficó la media móvil y en azul la mediana móvil.

La *Figura 8* Junto con la gráfica de la *Figura 7* ayudan a identificar más fácilmente las variaciones en la distribución de los datos. Por su parte los índices de localización muestran el mismo comportamiento no estacionario, en la *Figura 8* se muestra su variación. Es interesante ver cómo la mediana, junto con los percentiles 10 y 90 muestran mesetas en diferentes partes de la gráfica, esto se explica al tener en cuenta el método utilizado por la ventana móvil en donde para un rango de iteraciones de la serie se están considerando los mismos picos para calcular los parámetros. Adicionalmente es imposible no notar que los percentiles 10 y 90 siguen un patrón similar que la gráfica de densidad de probabilidades definida para la curva de 0.05. En general se expanden y contraen en la misma zona.

No se recomienda graficar la serie “por debajo”. Añade ruido a la información de la gráfica

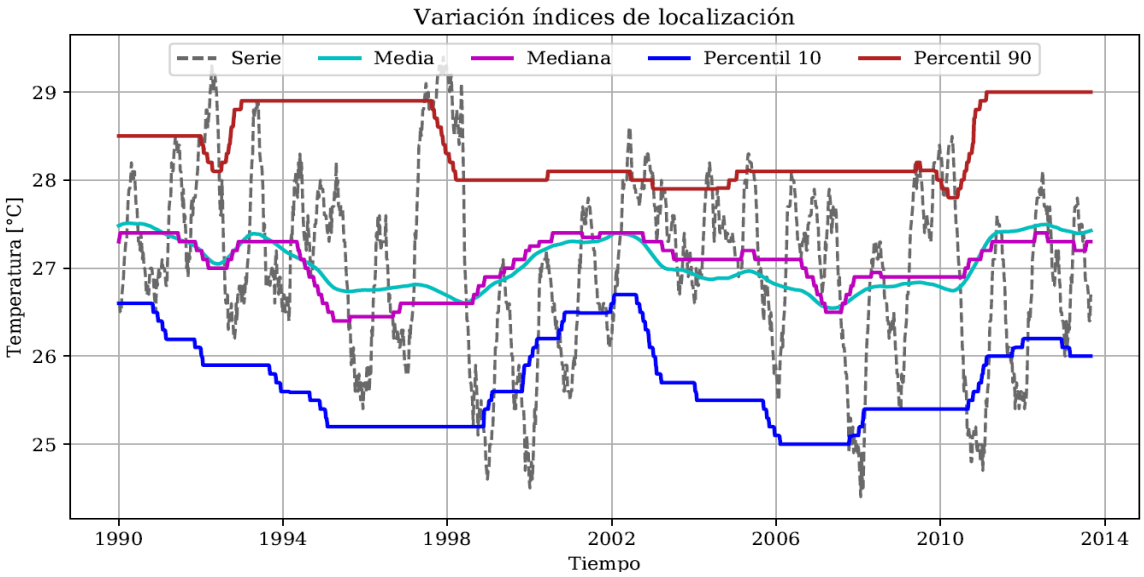

Figura 8: Variación de los índices de localización

Los índices de dispersión graficados en la *Figura 9 y 10* muestran considerables cambios en la distribución de los datos que ocurren con el tiempo. En la primera gráfica se muestran el valor de la media en azul y el valor de la desviación estándar por encima y por debajo de esta media en rojo . Los valores que se encuentran dentro de este margen delimitado se consideran como comunes o valores esperados. Aquí los realmente interesantes son aquellos puntos que se encuentra por fuera de la gráfica. Representan anomalías, dando a entender que es necesario analizar en detalle estos puntos.

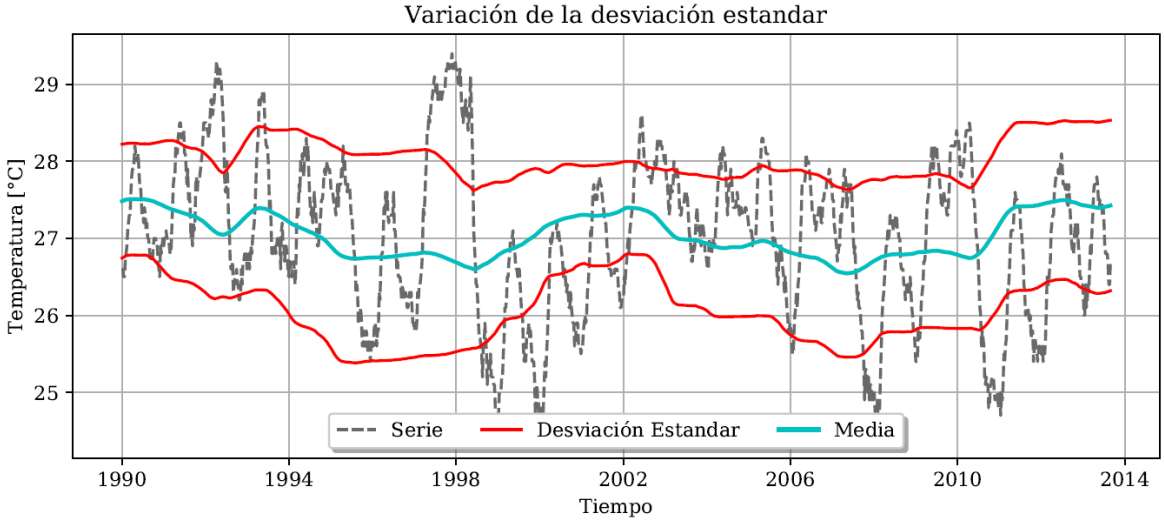

Figura 9: Variación de la desviación estándar.

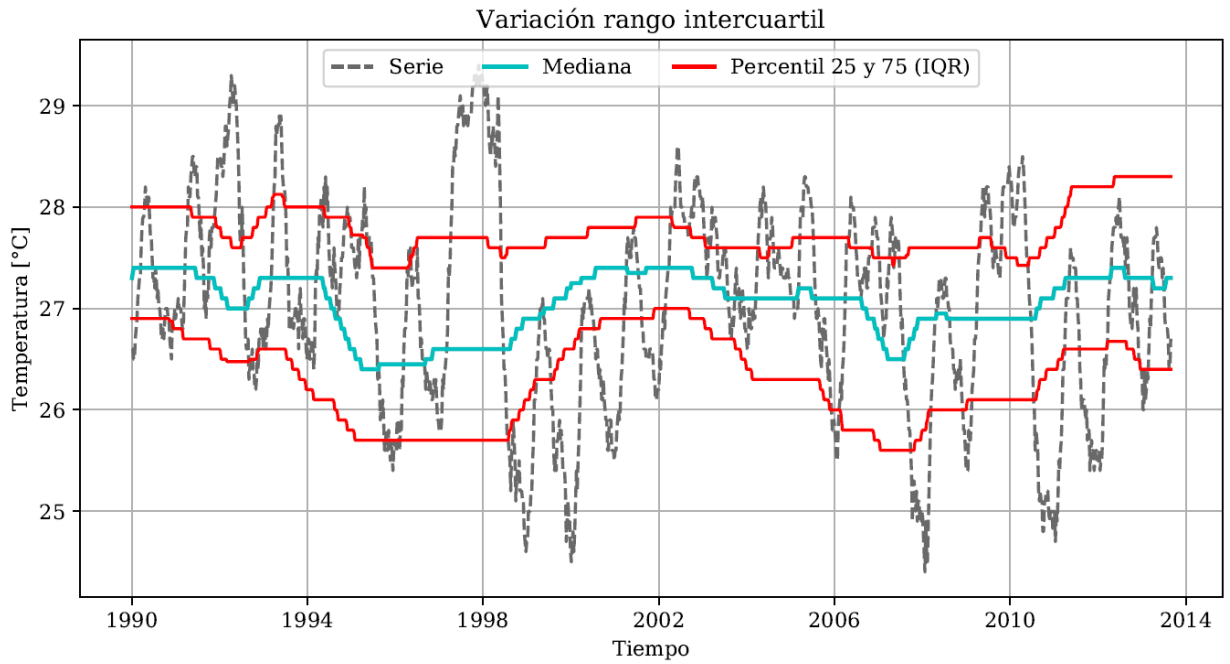

Figura 10: Variación rango intercuartil

Gracias a las gráficas, se puede concluir que los entre 1990 y 1994, hubo periodos con temperaturas particularmente calientes. En 1998 se tuvo un significativo pico de temperatura seguido por un periodo hasta 2002 donde las la variable osciló entre valores cercanos a la media (mediana) y temperaturas más bajas. Hacia el 2002 el rango disperción se estrecha y luego muestra picos de temperatura altas. Entre 2006 y 2014 hubo lecturas baja de temperatura, sin embargo la distribución tiende ubicarse hacia temperaturas más altas.

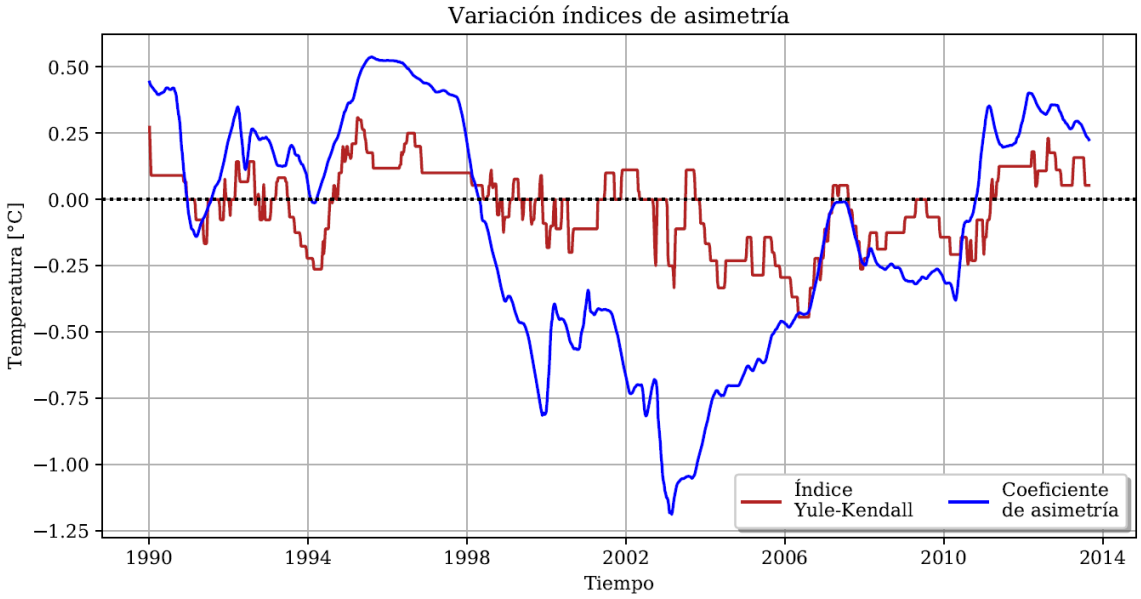

Figura 11: Variación de los índices de asimetría

Y por último la *Figura 11* muestran que tan simétricas es la distribución con respecto a su media (coeficiente de asimetría) o mediana (índice Yule-Kendall). De igual forma esta gráfica indica que la serie no es estacionaria. El área de la curva que se encuentra por encima del cero, indica que la hay una mayor distribución de probabilidad para valores menores de temperatura que su medida de tendencia central (media o mediana). EL contrario ocurre cuando el área de la curva está por debajo de la curva, significando que hay una mayor proporción de probabilidad para valores mayores. Adicionalmente se observa una gran variabilidad del coeficiente de asimetría, el cual se debe a la forma en que es calculada donde la diferencia entre el valor xi y su media es elevado al cubo.

Análisis de tendencias

Para calcular estadísticamente la existencia de una tendencia creciente o decreciente dentro de la serie de datos se utilizó el test no paramétrico de Mann Kendall. El método define como hipótesis inicial que los datos no tienen tendencia y la compara con una hipótesis alternativa utilizando el test estadístico *Z*.

Para el desarrollo de test es importante concretar la confiabilidad con la cual las hipótesis serán comparadas. A esto se le llama grado de significación estadística (1 -α), entre más grande sea en el mismo sentido el resultado es mas fiable. El grado de significación estadística esta definido por el parámetro α, el cual establece el margen de error con el que se va a trabajar.

Las posibles hipótesis alternativas son:

Ha: existe una tendencia positiva o negativa: Si $|Z_{MK}| \geq Z_{(1-\alpha/2)}$ *: la hipótesis Ha es aceptada.* $Z_{(1-\alpha/2)}$ *denota el percentil 100(1 - α) de la distribución normal estandarizada.* Z_{MK} *denota el estadístico calculado por el test Mann-Kendall.*

Ha: existe una tendencia positiva. Si $Z_{MK} \geq Z_{1-\alpha}$ *la hipótesis Ha es aceptada.*

Ha: existe una tendencia negativa. Si $Z_{MK} \leq -Z_{1-\alpha}$ *la hipótesis Ha es aceptada.*

Para este análisis se consideró un α igual a 0.001 para tener un buen grado de confiabilidad. Luego se procedió a determinar la existencias de tendencias en la serie de temperatura, y en las series móviles de media, mediana y percentiles 25 y 75. Los resultados pueden ser observado en la siguiente *Tabla 2*. La metodología que se utilizó fue primero establecer la existencia de alguna tendencia al comparar el valor absoluto de ZMK con Z(1-α/2), luego se estableció que clase de tendencia existía en la serie.

Ojo con la manera de expresar estas conclusiones. Se rechaza la hipótesis nula con x% …..

Serie	Z_{MK}	$Z_{(1-\alpha/2)}$	$-Z_{(1-\alpha)}$	Observaciones
Temperatura (serie completa)	-1.645	3.291		No hay tendencia positiva o negativa
Temperatura (ancho rango móvil)	-5.781	3.291	-3.090	**Existe tendencia negativa**
Media móvil	-4.210	3.291	-3.090	**Existe tendencia negativa**
Mediana móvil	-0.696	3.291		No hay tendencia positiva o negativa
Percentil 25 móvil	-2.979	3.291		No hay tendencia positiva o negativa
Percentil 75 móvil	-6.623	3.291	-3.090	**Existe tendencia negativa**

Tabla 2: Resultados del Test Mann-Kendall, α = 0.001

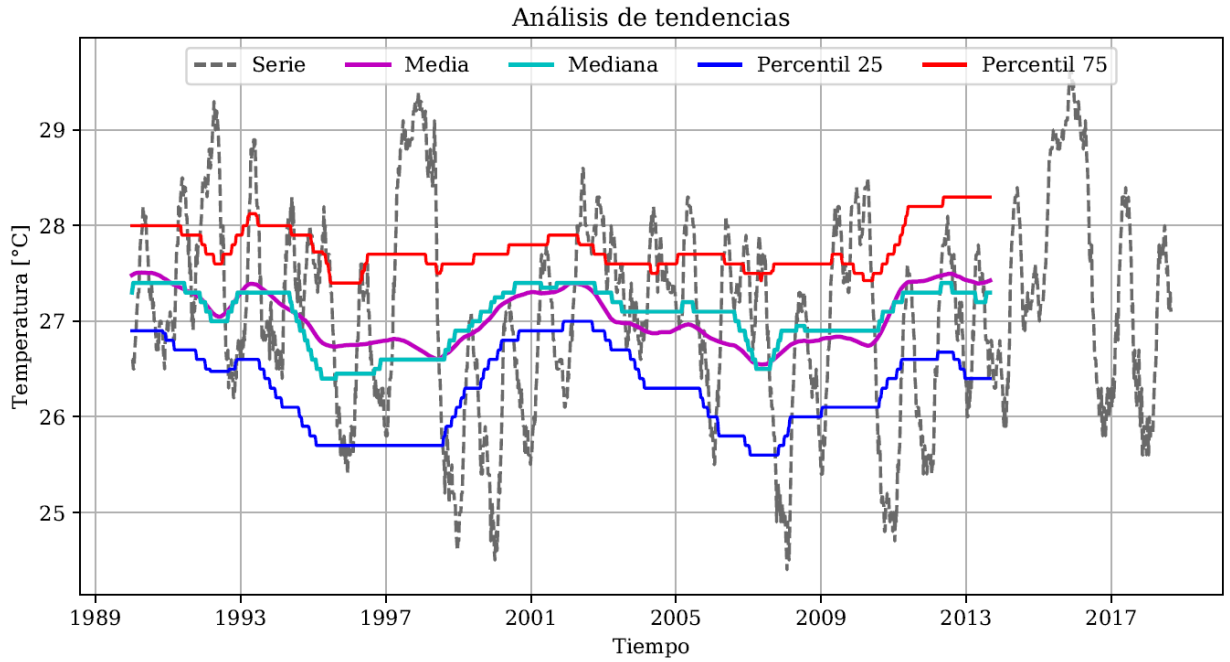

Figura 12: Análisis de tendencias para la serie de temperatura y las series móviles de media, mediana y percentiles 25 y 75.

En *Figura 12* se graficaron las series de datos analizadas. Al analizar los resultados de la *Tabla 2* se observó que para la serie completa el test de Mann-Kendall arrojó que no hay tendencias. Sin embargo cuando se hizo lo mismo considerando el periodo de análisis que las series móviles, el test arrojó que la tendencia es decreciente. Tendencia decreciente fue también obtenida para la media y el percentil 75 móviles. Estas tendencias decreciente pueden ser explicada al considerar que dentro del periodo de análisis (de 1990 a 2013) hubo varios momentos en donde se obtuvieron temperaturas por debajo del rango intercuartil (o de la desviación estándar, considerando la *Figura 9)*. Esta variación en general decreciente, puedo haber influenciado en el resultado del test. Se presume que de haber considerado un periodo de lectura de datos más grandes, los resultados hubiesen dado diferente.

Creo que estos resultados merecen una revisión.

Conclusiones

El análisis exploratorio de datos valiéndose de herramientas simples, probó se una poderoso método para extraer información de series aleatorias por medio de gráficas. Haciendo uso de esta técnica se analizó la serie de temperaturas del Niño 3.4, con datos semanalmente desde enero de 1990 hasta agosto del 2018. Visualmente se estimó una gran variabilidad de los datos, por presencia de picos varios pico de temperaturas muy cálidas y frías que cambiaban constantemente con el tiempo. Esta conclusión fue corroborada al momento de analizar los datos utilizando una ventana móvil de 5 años en donde se calculó distribución de probabilidades e índices de localización, disperción y simetría para la serie. El estudio arrojó que la serie no es estacionaria. Finalmente el test Mann-Kendall arrojó la presencia de una tendencia decreciente para la media y el percentil 75 móviles, así como para la serie de temperatura dentro del rango de ventana móvil. Sin embargo la serie de temperatura completa, no obtuvo tendencia.

Referencias

Soong, T. T. (2004). *Fundamentals of Probability and Statistics for Engineers.* Wiley.

Wilks, D. S. (2006). *Statistical Methods in Astmospheric Science.* Burlington: Elsevier.