Actividad5

Eber Cabrera Campaña

25 de febrero de 2021

1. Sobre la actividad y los recursos para realizarla

En la presente actividad se tiene como objetivo profundizar en análisis de los datos, con ayuda de elementos estadísticos. Entre estos se destaca el uso de las series de tiempo y predicción de comportamientos de las mismas series.

Lo anterior se realiza con el modelo ARIMA y sus co-modelos AR y MA. Dichos modelos utilizan técnicas estadísticas para realizar una predicción de los datos. Para poder aplicarlos, se necesita que la serie de tiempo a tratar sea estacionaria. Esto quiere decir que su promedio y su desviación estándar sean constantes en el tiempo.

Dado que puede no ser el caso, también se tendrán en cuenta técnicas para hacer que una serie sea estacionaria. Hecho esto se aplica el método ARIMA, se toman las predicciones y se comparan con los datos originales.

2. Preparando las series y aplicando el modelo

Al momento de comenzar la actividad lo primero que hay que hacer es probar la estacionaridad de nuestras series (de temperatura máxima y temperatura mínima). Para esto se utiliza la Prueba de Dickey-Fuller, dicha prueba retorna un parámetro de interés p que, al ser menor que 0.05, podemos afirmar con un 95 % de confianza que hablamos de una serie estacionaria.

Poniendo a prueba nuestras series en su forma inicial uno se encuentra con que no son de utilidad pues su valor de p no es menor a 0.05. Dado esto, se procede a aplicar técnicas que bajen el valor de p (hacer constante la tendencia)

Lo primero que se puede hacer es aplicar un logaritmo a la serie en conjunto. Acto seguido se obtiene el promedio móvil con el comando *.rolling* y se le resta a la serie. Si se realiza esto nos encontramos con que, en efecto, el valor p disminuye pero, no lo suficiente. Es por eso que se propone otro método, en vez de utilizar el promedio móvil se utiliza el promedio exponencial con el comando *.ewh, (Exponential Weighted Moving)*. Gracias a lo anterior logramos un valor de p que cumpla con las condiciones, tanto para la serie de temperatura mínima como para la serie de temperatura máxima. Algo que ha de comentarse es que al momento de realizar todo lo anterior para la temperatura mínima, se tuvo que hacer antes un cambio de escala de $°C$ a K pues al momento de tomar el logaritmo base 10 no podría calcularse para valores negativos, cosa que no sucede en escala Kelvin. Pero no se procedió con dicha técnica, con motivo de explorar se realizo una extra que consta de diferenciación gracias al comando *.shift*. Aplicando esto nos encontramos con un valor aún más pequeño para p, por lo que esta técnica fue la definitiva.

Con nuestras series convertidas a estacionarias se puede proceder a aplicar el modelo ARIMA. Un pequeño paso previo a esto es obtener los valores adecuados para p y q (siendo dicho p distinto al de la prueba de Dickey-Fuller). Estos, son los parámetros con los que se correrá el modelo. Para lograr esto aplicamos a nuestra serie estacionara dos funciones: la función de Auto-correlación (para ubicar q) y la función de Auto-correlación Parcial (para ubicar p). Cuando se utilizan dichas funciones obtenemos dos gráficas donde, visualmente, podremos ubicar los parámetros buscados. Una vez obtenidos se procede a realizar primero, de manera independiente, cada uno de los sub-modelos de ARIMA y luego el modelo completo. En cada resultado obtenemos un valor para algo llamado AIC, este será el criterio para elegir el mejor resultado (en base al menor de ellos) que en nuestro caso fue, para la temperatura máxima, el modelo MA y para la temperatura mínima, nuevamente, el modelo MA. Por ultimo nos resta crear una serie de tiempo con las predicciones del modelo y contrastarla con la original.

Antes de comentar los resultados se deben mencionar algunos tropiezos en el procedimiento anterior. Al tener muchos huecos en nuestra serie fue difícil encontrar un periodo amplio para analizar, por tanto lo mejor que se pudo conseguir fueron dos años. Algo malo fue que el periodo más largo en temperaturas máximas estaba muy distante del periodo más largo en temperaturas mínimas lo cual hizo difícil un contraste.

Con respecto a los resultados, vemos que se ajustan relativamente bien, la medida de error **RMSE (Root Mean Square Error)** nos indica esto. Se podrían obtener mejores resultados pero se puede dudar de la calidad de los datos (dada la gran cantidad de huecos).

3. Retroalimentación.

Sin duda alguna fue la actividad que requirió de más dedicación y atención. Fue algo más compleja en cuanto al tratamiento estadístico y el desarrollo del código (dado que se crea una función interesante, donde se aplican comandos y se gráfica). Además el análisis de Series de Tiempo es de gran utilidad para comprender fenómenos a lo largo de los años por tanto el aprendizaje correcto de como utilizar técnicas y análisis y representación con lenguajes tan variados como Python fue fructífero. La curva de dificultad, a lo largo de todo el tema, fue muy buena, con un mayor desafió al final, como ya se comento.