Analisis y predicción de serie temporal
Hielo del polo norte.

Autor: BRUNO SANTOME ANTOLIN
100405889

Indice

1. Datos a trabajar

Podemos observar que la serie temporal a trabajar son datos de la cantidad de hielo por mes en el polo norte desde 1978 a 2022.

Hay un total de 45 filas y 13 columnas.
Cada columna esta asociada a un mes del año, salvo la ultima columna que viene a ser un resumen del indice de hielo anual. Estos datos se representan con un valor numérico.
Vemos que hay una columna llamada Annual, esta columna se puede tratar como una serie temporal diferente. En este analisis solo vamos a trabajar con los datos de cada mes, no vamos a utilizar los índices anuales del deshielo.

2. Tratamiento de Datos

Eliminamos las columnas Years y Annual

Puesto que ya sabemos que es de 1978 a 2022 ya no nos aporta nada esa columna. Como dicho anteriormente no vamos a utilizar la columna anual al ser una serie temporal diferente.
Vemos que hay un total de 15 datos Nulos en todo el dataset que son necesarios tratar para poder seguir con el analisis. Estos valores se sitúan al principio y al final del dataset.

Para tratarlos, he considerado sustituir cada valor nulo por la media de la columna en la que se encuentre. Como hay al menos un valor nulo en cada columna, este proceso se repite 12 veces.

Una vez que el conjunto de datos se encuentra completo, analizamos el tipo de clase que pertenece el conjunto de valores. Tenemos que transformar nuestro conjunto de datos a serie temporal para poder seguir con el analisis.

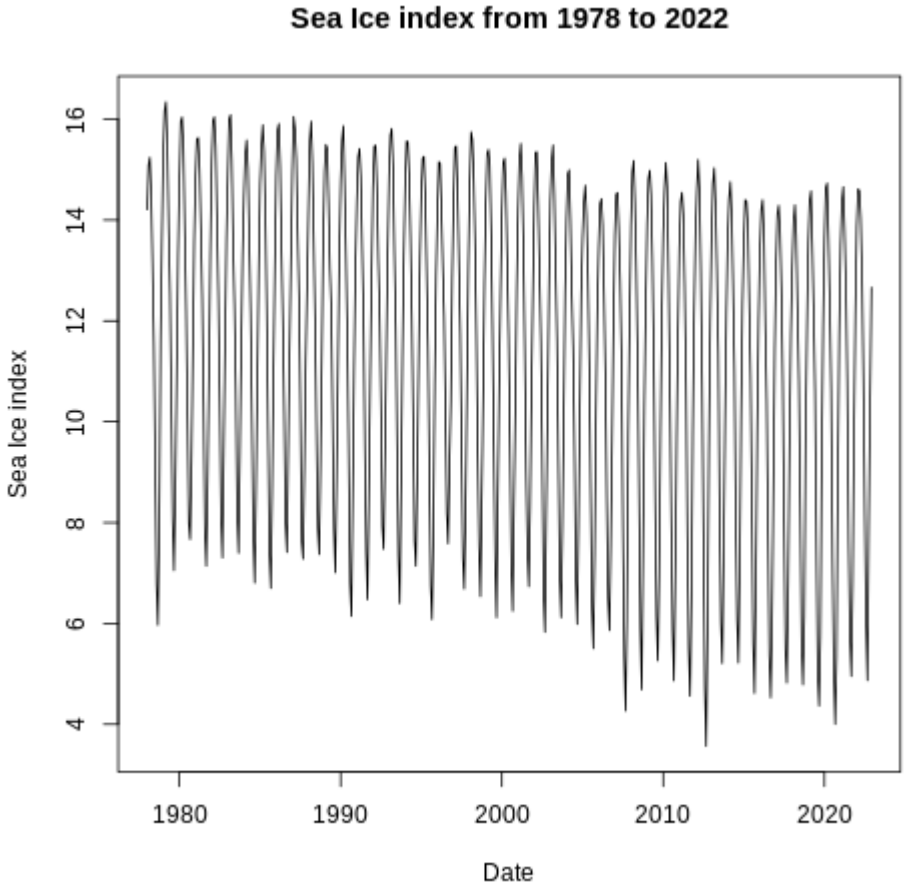
Para ello primero hay que transformar el conjunto a tipo vector para luego transformarlo a tipo "time serie" | "ts". En el resultado de esta transformación vemos que ahora nuestra serie tiene los años a la izquierda como nombre de fila y los meses correspondientes como nombres de columna.

Algunos datos sobre la serie temporal:

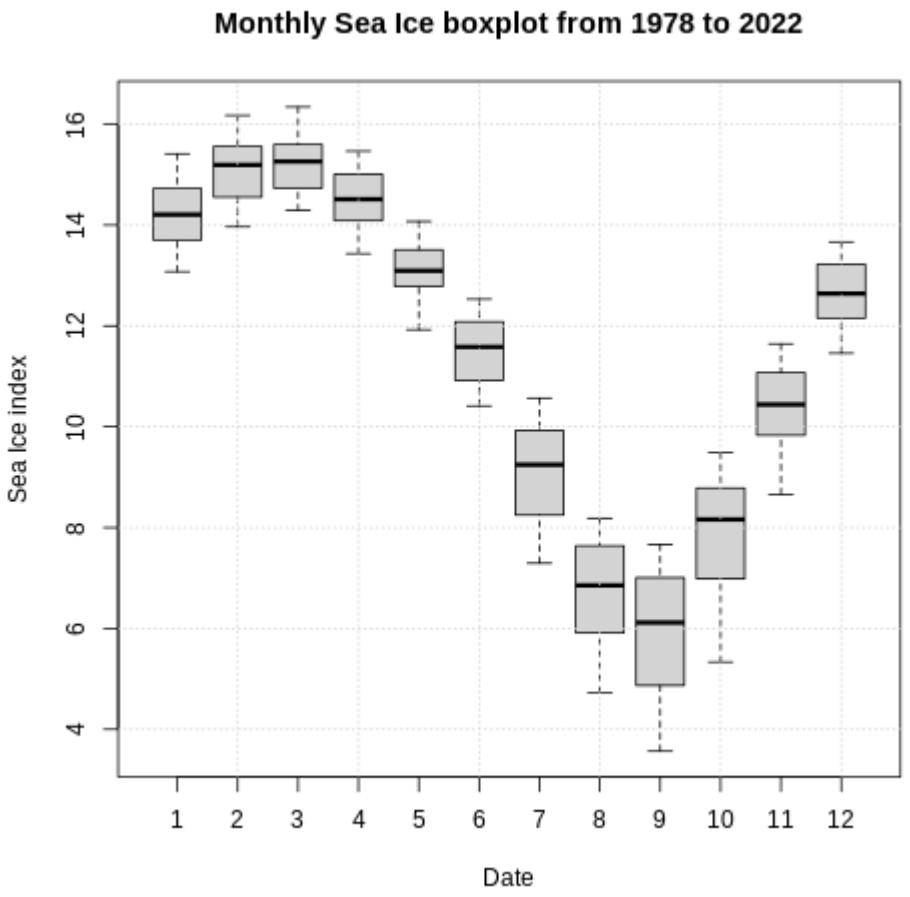
- Empieza en enero de 1978 y termina en diciembre de 2022
- hay tun total de 540 datos entre enero de 1978 y diciembre de 2022
- ⇒Time-Series [1:540] from 1978 to 2023: 14.2 15.1 15.2 14.5 13.1 ...
- El número de ciclo de la serie esta asociado al mes, el ciclo 1 sería enero el ciclo 5 mayo y el ciclo 12 diciembre, Asi con todos los meses.

3. Representación de la serie Temporal.

Al representar la serie temporal nos resulta en:

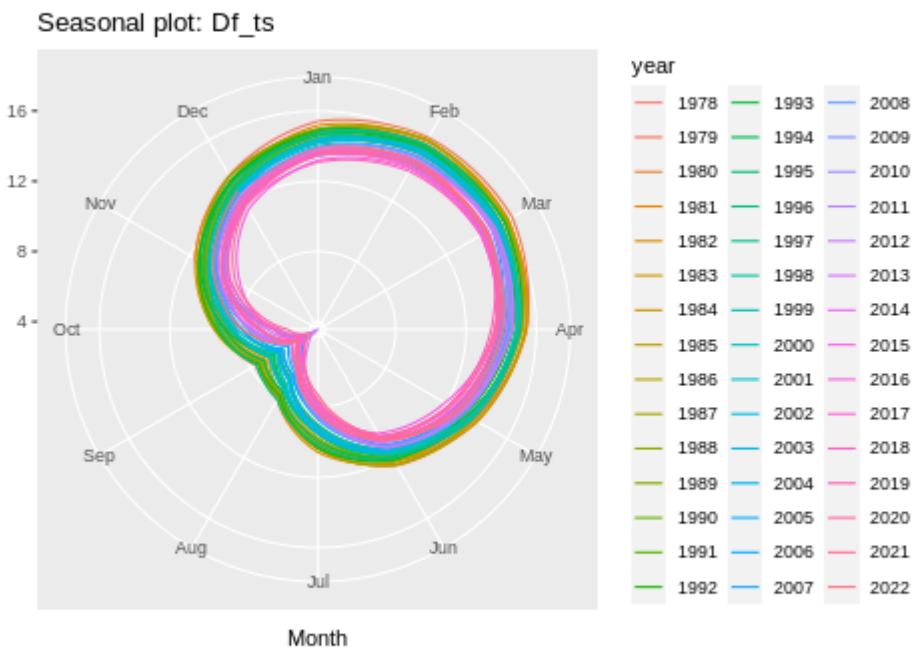


Si comparamos la grafica superior según la clasificación de pegel podemos ver que los datos disminuyen a medida que pasan los años, se puede ver que la estacionalidad es aditiva o ligeramente multiplicativa, ya que los picos varían muy poco a lor largo de los años. Hay que investigar más ya que no nos podemos fiar únicamente de nuestro ojo.



Podemos observar que el hielo alcanza su pico minimo en septiembre, donde la media de valores es más baja. Alcanza su pico máximo entre febrero y marzo.

Hay una relación con las estaciones, en invierno el indice de hielo es muy superior al indice de hielo en verano.



De la información y los gráficos anteriores podemos inferir:

-El indice de hielo disminuye a lo largo del tiempo luego hay indicios de una tendencia decreciente.

-Al comparar lo datos con la clasificación de conducta de Pegel se muestra un comportamiento de estacionalidad aditiva.

-Hay una clara estacionalidad, nivel de hielo más bajo en verano y más alto en invierno.

4. Descomposición de la serie temporal.

Como dicho anteriormente, visualmente parece que tiene un comportamiento aditivo pero hay que verificarlo. Para ellos vamos a descomponer la serie temporal.

Aditivas yt=tt+st+ct+ϵt

Donde tt es la tendencia, st es la componente estacional, ct es la componente de ciclo (o ciclo) y ϵt el error.

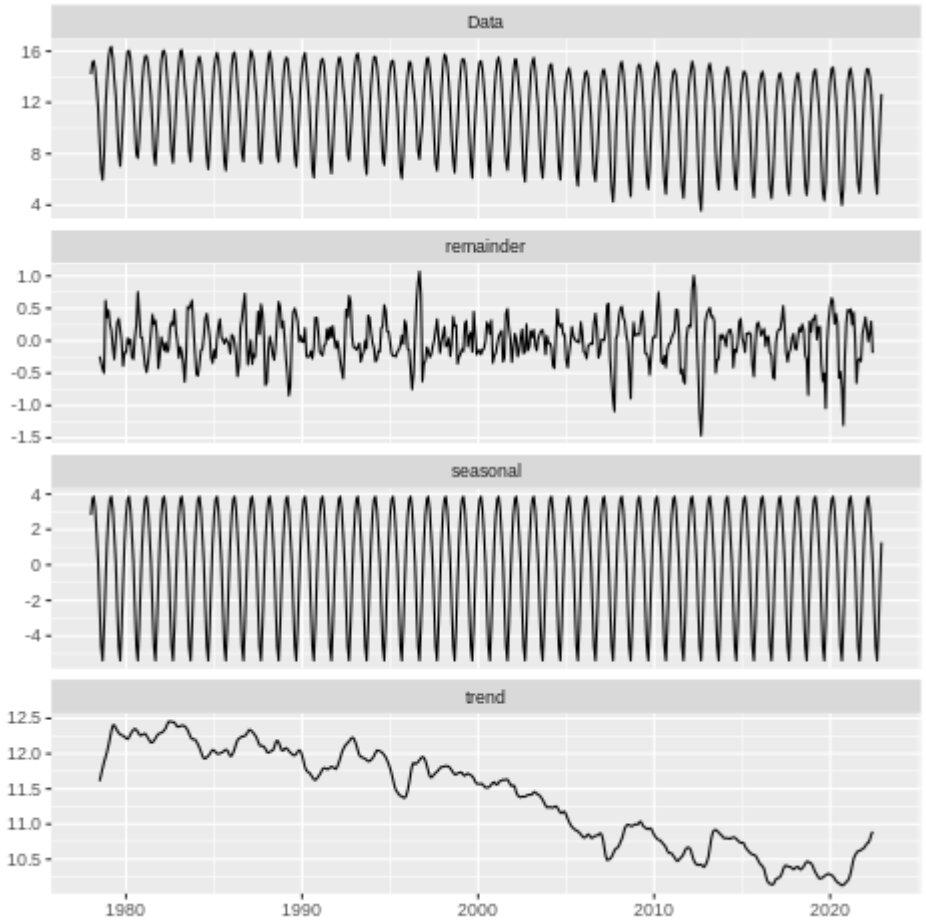
- La tendencia se refiere al comportamiento a largo plazo de la serie.

- La componente estacional se refiere a comportamiento periódicos en la serie.

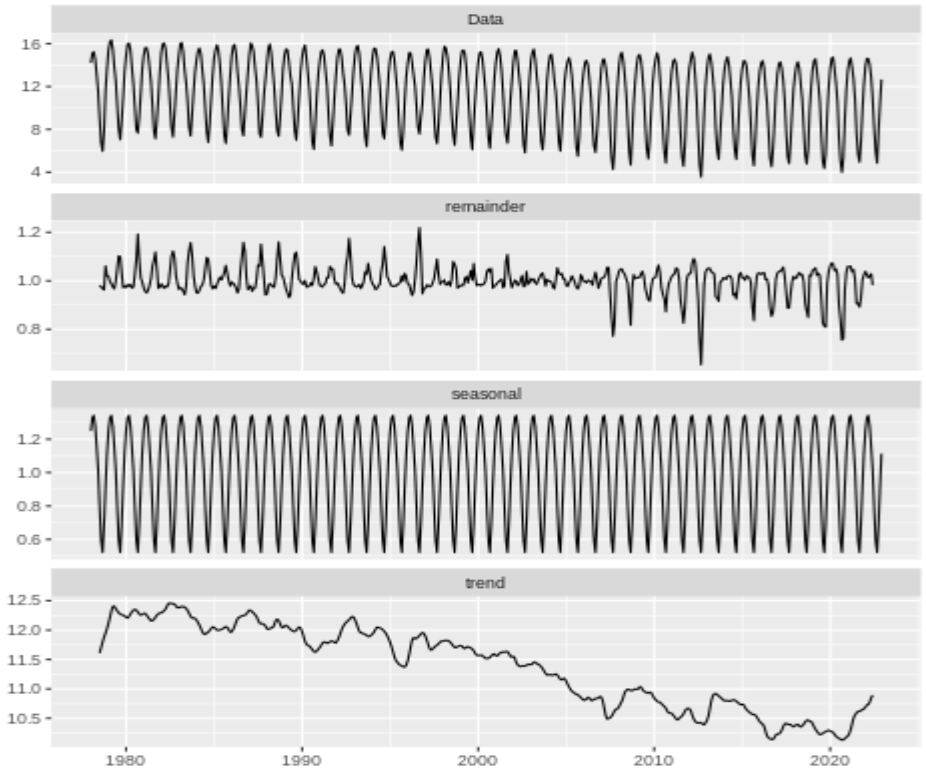
- La componete cíclica se refiere comportamientos periódicos no estacionales.

- El error son movimientos transitorios o irregulares de la serie.

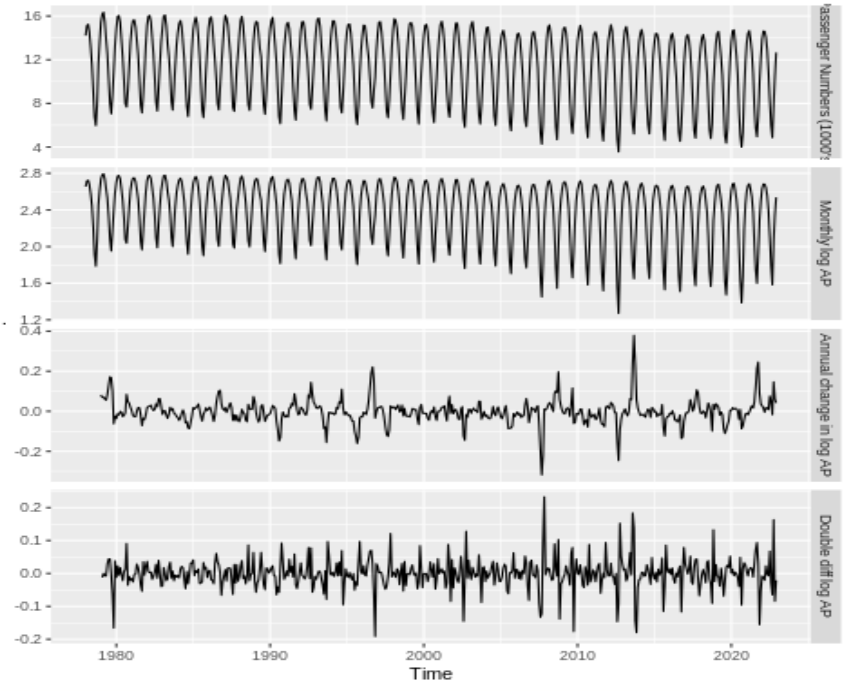
Realizamos primero la descomposición aditiva.



Realizamos la descomposición multiplicativa



Requisito primordial es que la varianza tiene que ser estable en el tiempo. Hacemos la descomposición logarítmica para suavizarla y ver el resultado.



Tras hacer la descomposición de la serie temporal.

Vemos que entre la descomposición aditiva y multiplicativa no hay mucha diferencia. Sin embargo en la grafica de "remainder" que viene a ser el error o residuos vemos que en la descomposición aditiva estos están acotados entre valores más pequeños que el rango de error en la descomposición multiplicativa.

Con esto confirmamos que la serie temporal sigue una estacionalidad aditiva.

5. Comprobación de los requisitos para el modelado

Para poder aplicar los modelos de “Box-Jenkins” ha de cumplirse que la serie sea estacionaria, esto quiere decir que la media, la varianza y la covarianza no cambien con el tiempo. Además no debe haber autocorrelación entre los residuos.

a. Test ADF

Augmented Dickey-Fuller Test (ADF): Este test establece una hipótesis nula H0 en el que la serie no es estacionaria, y como hipótesis alternativa H1 que la serie temporal es estacionaria.

Augmented Dickey-Fuller Test

data: SI

Dickey-Fuller = -12.906, Lag order = 8, p-value = 0.01

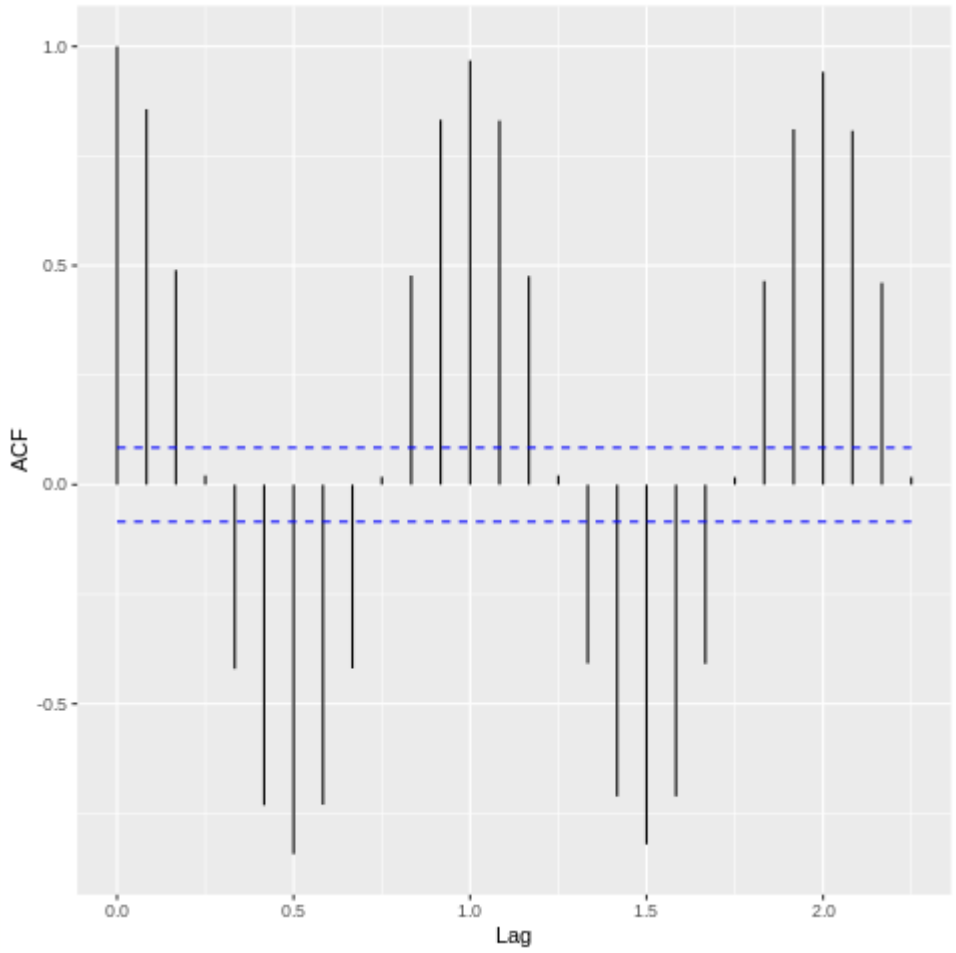
alternative hypothesis: stationary

Según los resultados de la prueba anterior, el valor p es de 0,01, que es <0,05,por lo que rechazamos la hipótesis nula a favor de la hipótesis alternativa de que la serie temporal es estacionaria.

b. Test ACF

Test de autocorrelación: conocido por las siglas ACF del ingles “Autocorrelation and Cross-correlation Function estimation”. Esta función traza la correlación entre una serie y sus retrasos, es decir, observaciones anteriores con un intervalo de confianza del 95 % en azul. Si la autocorrelación cruza las líneas azules

discontinuas, significa que el retraso específico está significativamente correlacionado con la serie actual.



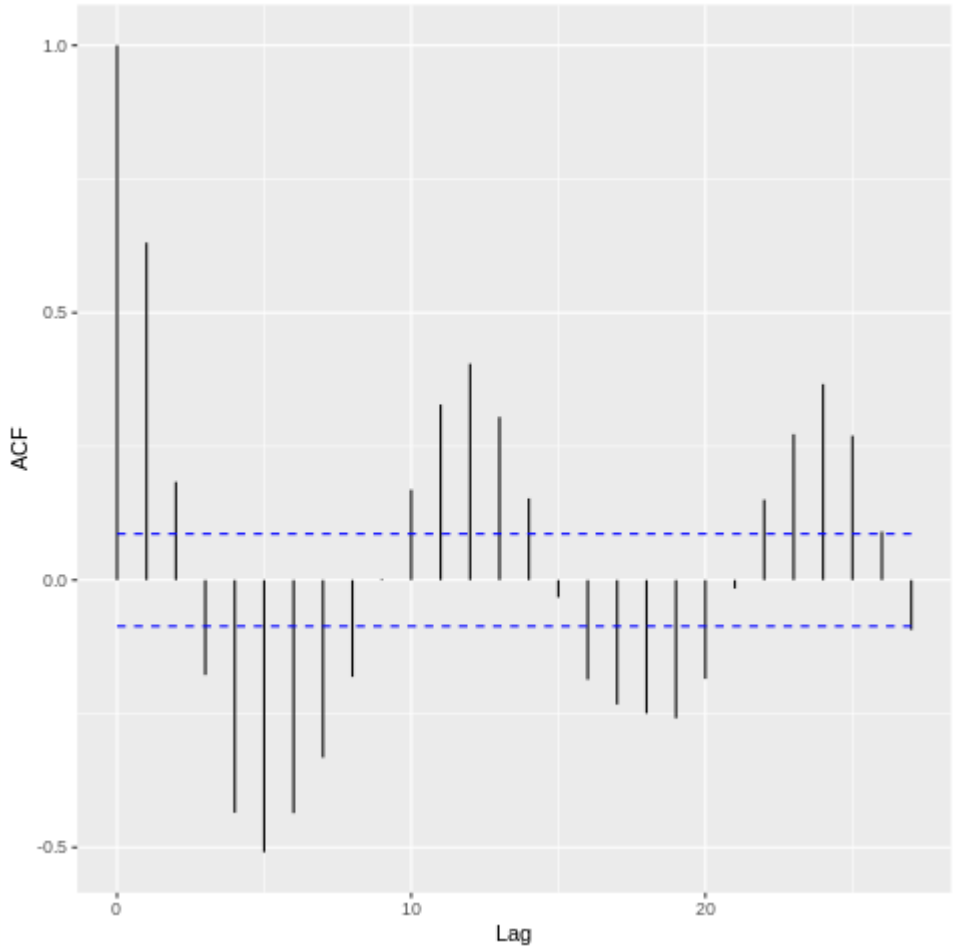
Para disminuir los residuos acotamos el rango de valores ignorando los valores nulos.

los valores nulos se encuentran al principio de la serie y al final, precisamente las celdas con valores empiezan en la celda 7 y terminan en la celda 522

Como aparecen términos que faltan no los contabilizamos.

Estos términos aparecen como NA y se observen entre los meses de Jan-Jun de 1978 y Jul-Dec de 2022

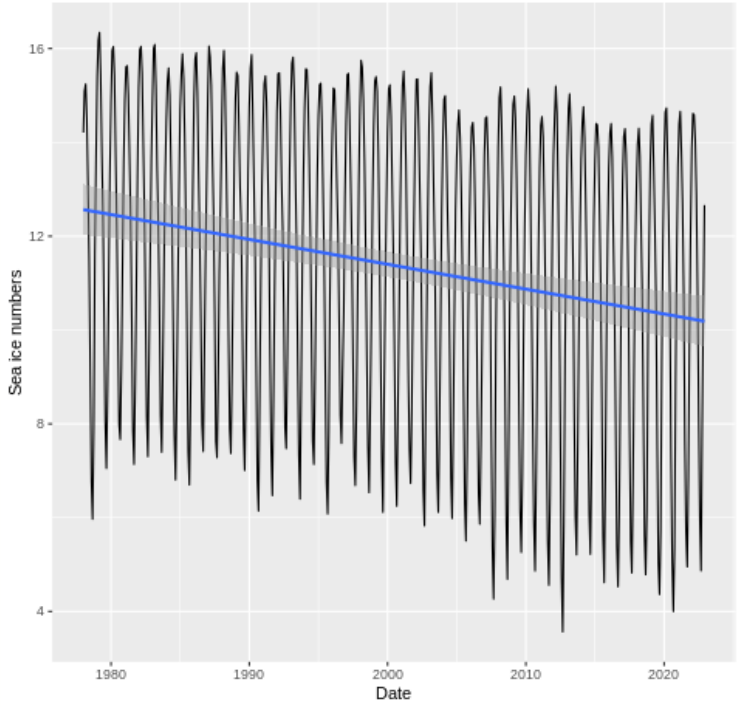
Al repetir el test acotando el rango y después de haber randomizado los valores entre -1 y 1



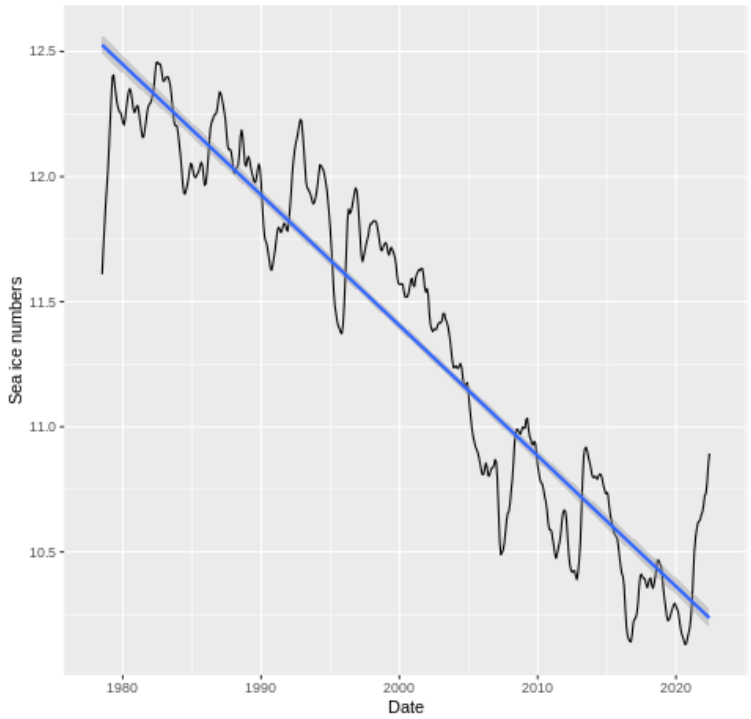
El retraso específico está significativamente correlacionado con la serie actual, ya que en esta última representación se observa que los residuos se encuentran centrados alrededor de cero.

6. Modelo y entrenamiento de la serie temporal

Al aplicar un método linear sobre la representación de la serie temporal obtenemos:



Nos interesa principalmente evaluarlo sobre la tendencia.



El modelo lineal del filtrado se aproxima muy levemente a la tendencia de la serie. Le falta añadir información estacional.

7. Modelo ARIMA

Vamos a analizar los residuos de la aplicación del modelo ARIMA a nuestra serie temporal.

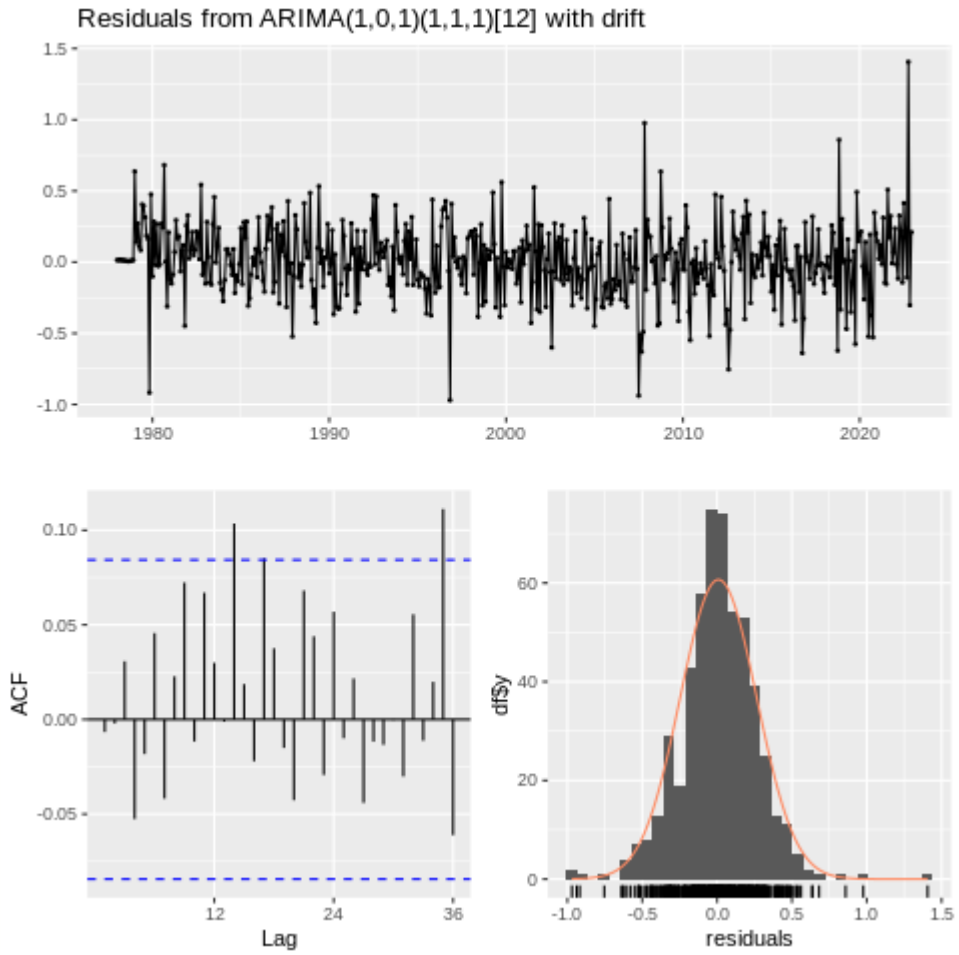
ARIMA (1,0,1)(1,1,1)[12]

Los parámetros de este modelo son los siguientes:

ARIMA(1,0,1): Esto indica que el modelo es un modelo autorregresivo con un valor de retraso y un término de promedio móvil.

(1,1,1): Esto indica que el modelo incluye un término autorregresivo estacional, un término de diferencia estacional y un término de promedio móvil estacional.

[12]: Esto indica que el período estacional es 12.



Realizamos la media de los residuos para ver si esta cercano a zero.

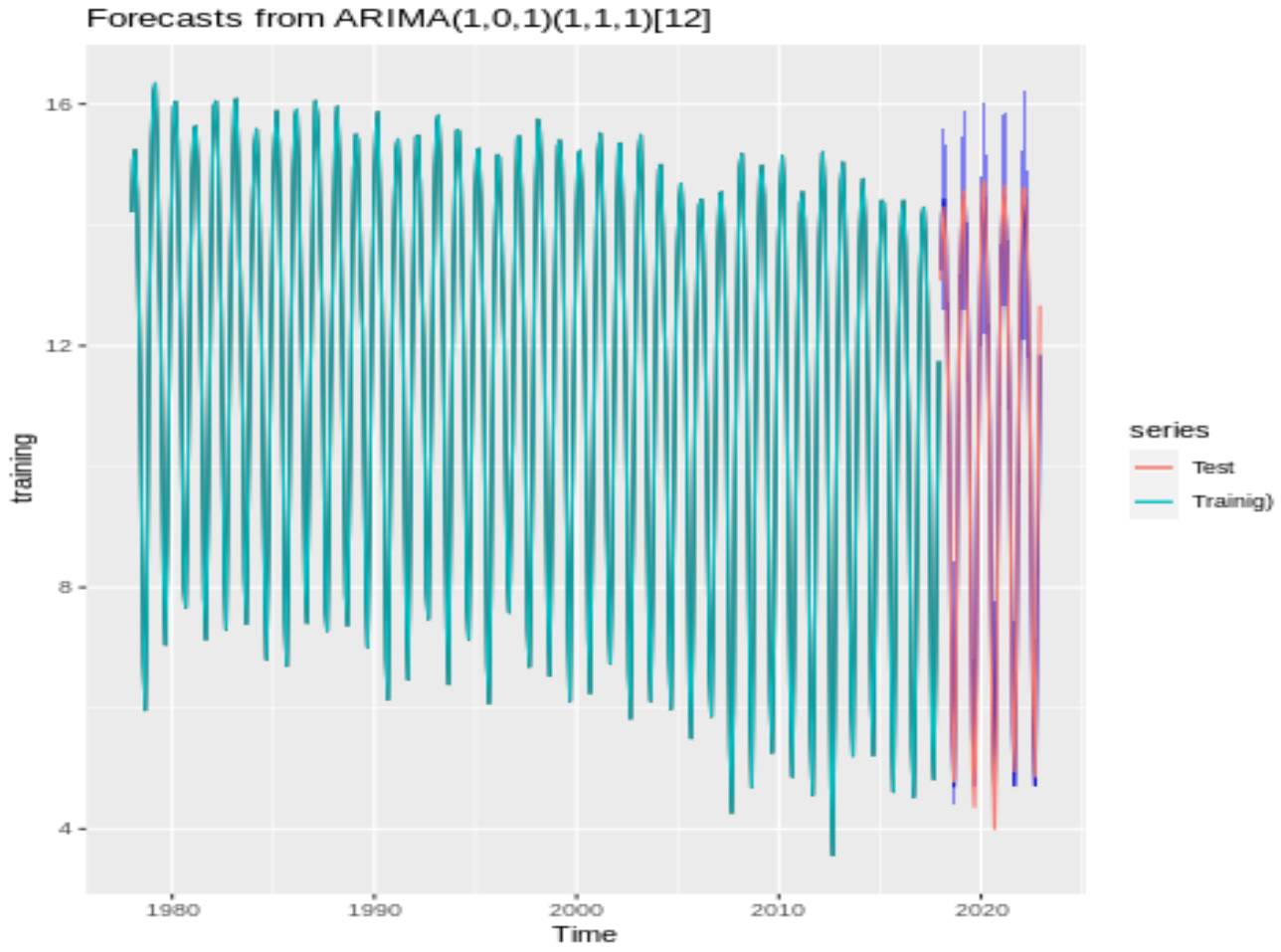
⇒0.008980537

Podemos observar que el valor medio de los residuos es muy cercano a cero lo cual puede ser un indicador importante de un buen modelo predictivo.

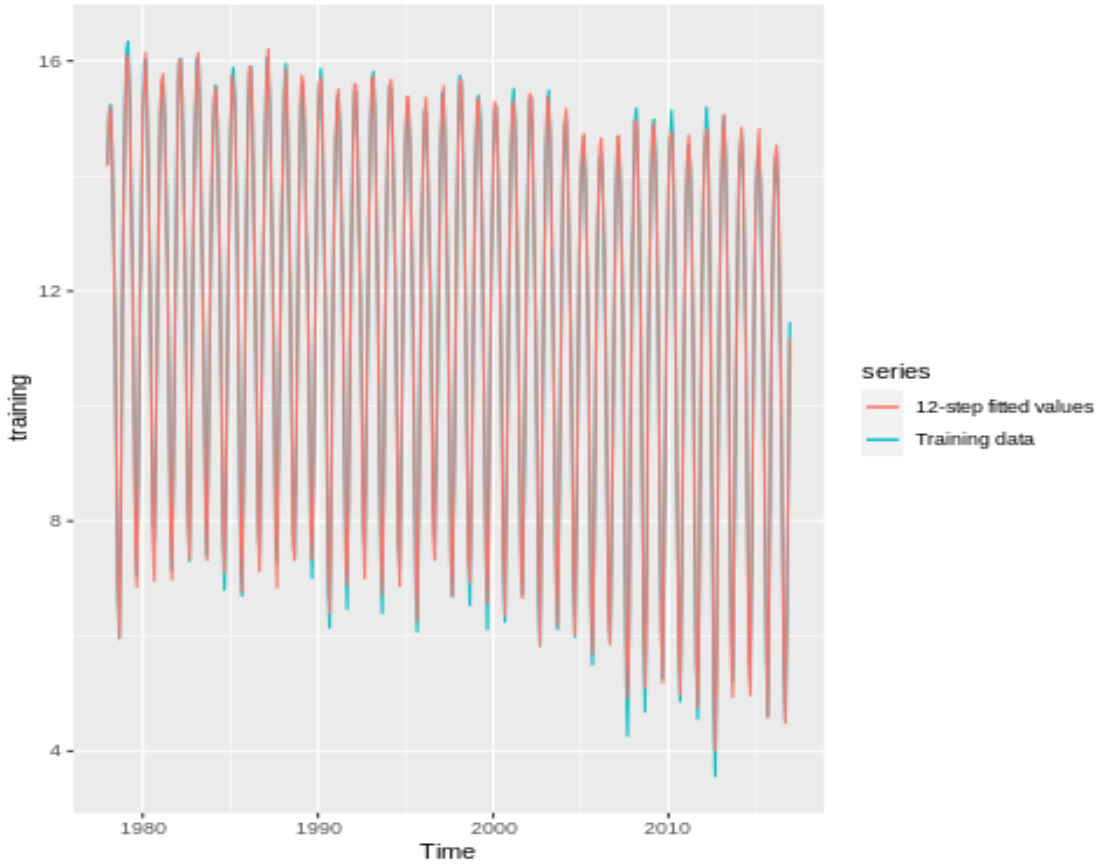
8. Evaluación cuantitativa

En primer lugar, dividimos la serie entre “entrenamiento” y “test”.

En nuestro caso escogemos que los datos de entrenamiento correspondan al conjunto de la serie menos seis ciclos (6 años, 72 meses).



Comprobamos que pasa con los datos de entrenamiento



Métricas evaluadas con los distintos conjuntos.

Metricas	Entrenamiento	Test
ME	-0.02426367	0.02842154
RMSE	0.2713574	0.4109276
MAE	0.2050603	0.245996
MPE	-0.3441623	0.1749743
MAPE	2.09336	2.755223

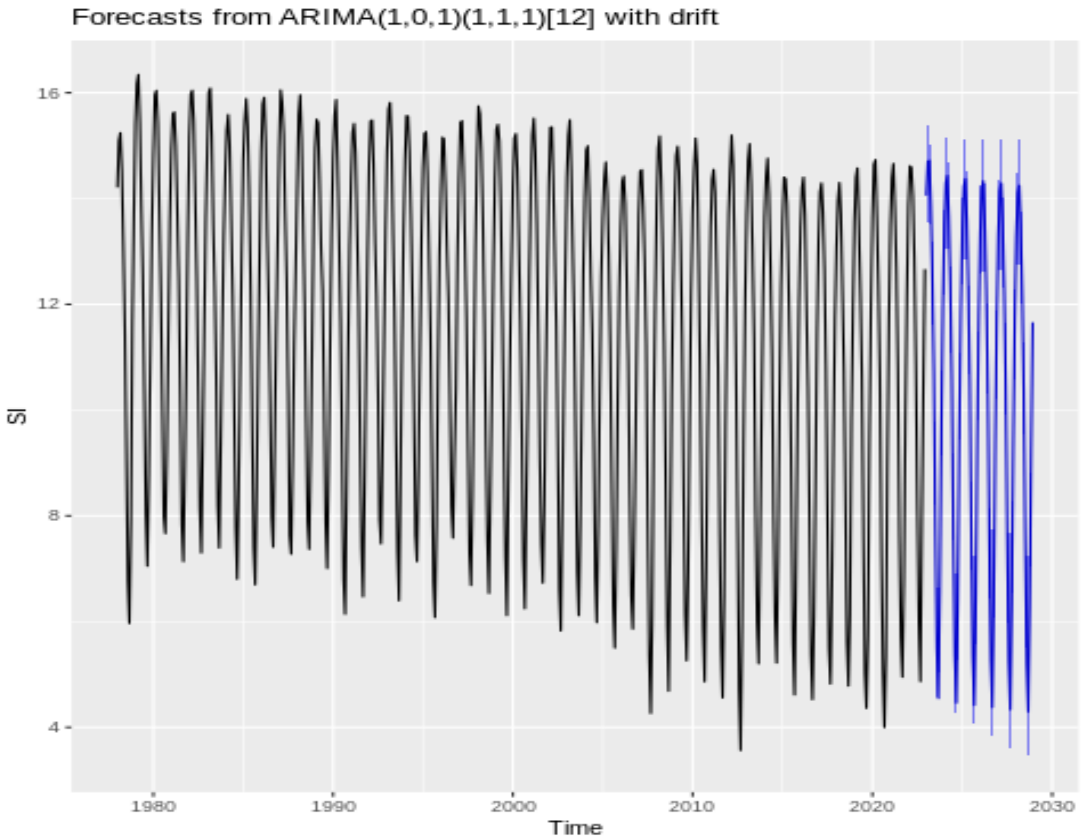
MASE	0.5688931	0.7644358
ACF1	-0.1093358	-0.5497927

Podemos ver que los valores de las métricas de precisión son menores en el conjunto de entrenamiento que en el de test, podría ocurrir que haya un sobreajuste del modelo en el entrenamiento.

9. Predicción con el modelo ARIMA

El objetivo es realizar una previsión con una confianza del 95%.

Hacemos un modelo para la predicción del hielo en los próximos 6 años, desde diciembre de 2022 hasta diciembre de 2028.



10. Conclusiones

Podemos concluir que se consigue una buena predicción con el modelo ARIMA, vemos que el valor del hielo seguirá decreciendo a medida que pasan los años. Podemos ver en la grafica que los picos maximos de cada año predicho están peor ajustados que los picos minimos.

El modelo ARIMA (1,0,1)(1,1,1)[12] es un modelo eficaz para predecir el hielo del polo norte porque tiene en cuenta los componentes estacionales y no estacionales de los datos. El modelo ARIMA utiliza componentes autorregresivos y de promedio móvil para tener en cuenta las tendencias no estacionales en los datos, mientras que el componente estacional permite capturar los efectos de las variaciones estacionales en los datos. Este modelo también utiliza un período de 12 meses para el componente estacional, que captura el ciclo anual de formación y derretimiento del hielo. Además, los términos de retraso y los componentes de diferenciación de este modelo ayudan a reducir la cantidad de ruido en la serie temporal y mejoran la precisión del modelo. Todas estas características hacen de este modelo una herramienta eficaz para predecir el hielo marino.

11. Comentario.

Sería mucho más cómodo compartir el cuaderno de júpiter con estas conclusiones pero por diversos errores de google collab y de las librerías de ggplot2, forecast y ggfortify (al parecer a veces fallan al tener métodos que se superponen). He decido hacer una memoría puesto que en ocasiones el cuaderno compilaba bien y en otras no. Esto ha sido un gran impedimento para profundizar con el analisis, muy frustrante que no compile el codigo de vez en cuando y no se pueda hacer nada.