viu
.es

2023 - 2024

ACTIVIDAD GUIADA 2

Máster en Big Data y Data Science

06MBID – Estadística Avanzada

Nombre: **Rodrigo Sandoval Brito**

Fecha: **Octubre 2023**

Curso 2023 – Ed. Abril

viu | Universidad Internacional de Valencia

INDICE

1. Introducción

• Contexto y motivación

Debido a la cercana llegada del Fenómeno del Niño prevista para finales del año 2023 y su inminente impacto en varios países del mundo, este trabajo se enfocará en analizar los ciclos de este fenómeno climatológico junto con su impacto ecológico reflejado en la métrica del reclutamiento de nuevos peces.

Para la obtención de estos datasets utilizaremos la librería de R llamada "atsa" (Applied Statistical Time Series Analysis), la cual nos bridará acceso a las siguientes series temporales utilizadas en este análisis:

soi (Southern Oscillation Index): El conjunto de datos "soi" contiene información sobre el Índice de Oscilación del Sur, que es un indicador clave utilizado en climatología para realizar un seguimiento al fenómeno climático conocido como El Niño y La Niña. El SOI se basa en las diferencias de presión atmosférica entre la isla Tahití y la ciudad Darwin en Australia. Un SOI positivo indica condiciones de La Niña, mientras que un SOI negativo sugiere condiciones de El Niño.

rec (Recruitment Index): El conjunto de datos "rec" proporcionado por el Dr. Roy Mendelssohn del Grupo de Pesca Ambiental del Pacífico, se relaciona con el índice de reclutamiento de peces, que es un indicador utilizado en ecología para evaluar el éxito reproductivo de ciertas especies de peces en un área particular. Este índice está vinculado a factores climáticos, como la temperatura del agua y las condiciones oceanográficas, que pueden influir en la supervivencia de las crías de peces.

• Objetivos del análisis

El propósito de este estudio es analizar los ciclos del fenómeno climatológico de El Niño, tal como se reflejan en el Índice de Oscilación del Sur (SOI), y comprender en mayor profundidad su relación con el fenómeno ecológico del reclutamiento de peces.

Ya que estas variaciones climáticas pueden ejercer un impacto significativo en las condiciones marinas, incluyendo la temperatura del agua y la circulación oceánica, lo que a su vez puede influir en el éxito reproductivo de las especies de peces, capturado por el índice de reclutamiento.

Por último, emplearemos modelos estadísticos para generar pronósticos para ambas series temporales.

2. Descripción de los datos

• Características estadísticas generales

Para el presente análisis tenemos dos series interconectadas con datos mensuales que abarcan 453 meses desde 1950 hasta 1987. Tal como se aprecia en la Fig.1, se observan patrones recurrentes, con ciclos fácilmente discernibles, compuestos por regularidades caracterizadas por un prominente ciclo anual y una oscilación más lenta de aproximadamente 4 años.

Con respecto al dataset del "soi" tenemos valores normalizados entre -1 y 1. Donde un valor positivo indica condiciones de La Niña, mientras que uno negativo sugiere condiciones de El Niño. Por otro lado, para el dataset del "rec" tenemos prácticamente una normalización de los datos en valores porcentuales, tal como se observa en el siguiente resumen de características estadísticas:

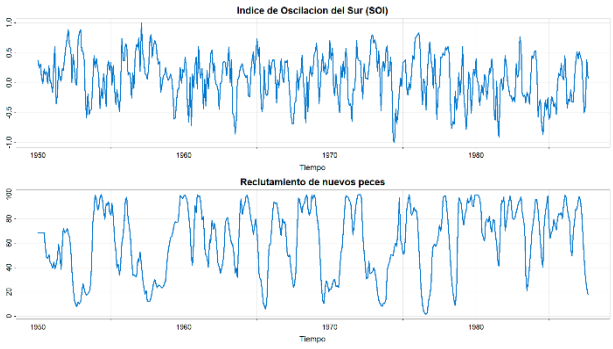

Fig 1. Series temporales del Índice SOI y del Reclutamiento

Dataset	Mínimo	1er Cuartil	Mediana	Media	3er Cuartil	Máximo
soi	-1	-0.18	0.115	0.08004	0.366	1
rec	1.72	39.62	68.63	62.26	86.85	100

- **Descomposición de la serie temporal**

Al realizar la descomposición de las series temporales de nuestro estudio, encontramos que las series no presentan una tendencia clara, sin embargo, se puede apreciar una estacionalidad anual evidente, tal como se observa en la Fig.2.

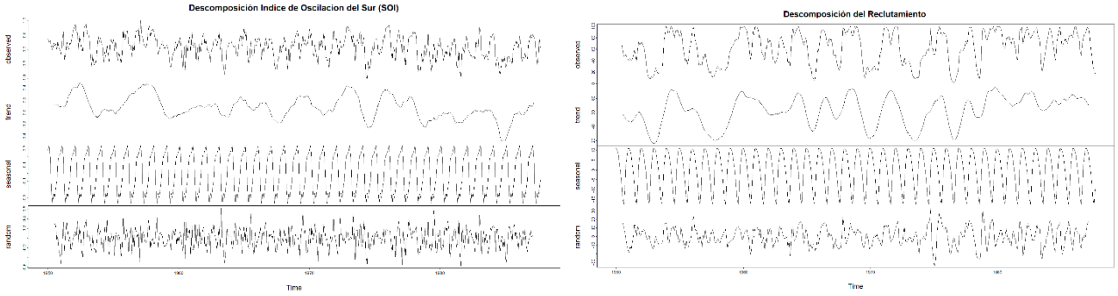

Fig 2. Descomposición de la serie del SOI y del Reclutamiento

Con el propósito de disminuir el ruido o las fluctuaciones aleatorias presentes en la serie del SOI y facilitar así la identificación de sus patrones o ciclos subyacentes, podemos emplear técnicas de suavizado de datos.

Entre las diversas técnicas de suavizado de datos destacan el suavizado Kernel, que suaviza una serie de datos mediante la convolución de cada punto con una función de Kernel. En nuestro caso utilizaremos una función de Kernel normal, es decir una distribución Gaussiana. Al implementar este suavizado definiremos también un ancho de banda, el cual controla la cantidad de suavizado aplicado, donde un ancho de banda más grande dará como resultado un mayor suavizado, dicha implementación se la puede observar en el código R adjunto en el apartado Anexo.

De igual manera, alcanzaremos un efecto de suavizado similar al aplicar la técnica LOWESS (LOcally WEighted Scatterplot Smoothing), la cual realiza una regresión local similar a KNN. De este modo podemos obtener un suavizado de los datos de la serie del SOI que representará los ciclos de El Niño, así como un promedio de dichos ciclos para identificar una tendencia general aproximada, tal como se observa en la Fig 3.

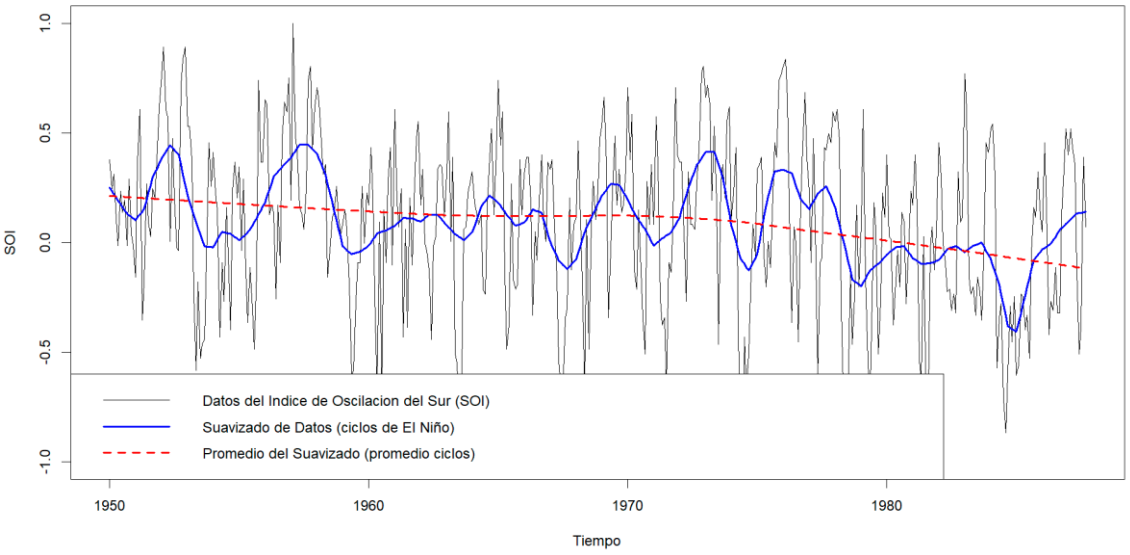

Fig 3. Suavizado de los datos del Índice SOI por medio de técnica LOWESS

3. Análisis

- ## Análisis de autocorrelación

Con el objetivo de definir un modelo adecuado necesitamos realizar un análisis de la función de autocorrelación (ACF). Además, podemos analizar la función de correlación cruzada (CCF) para identificar las relaciones temporales entre nuestras dos series de estudio.

Tal como observamos en la Fig 4, el eje de los *lag* se encuentra en términos de estaciones, es decir de 12 meses. En el ACF del índice SOI observamos existen picos que se repiten de forma periódica, mostrando una fuerte estacionalidad. Mientras que el ACF del Reclutamiento presenta picos periódicos que se van atenuando, mostrando una estacionalidad más débil.

Además, al analizar la relación del índice SOI con el Reclutamiento por medio del CCF podemos notar que las observaciones con 12 meses de diferencia están fuertemente correlacionadas de forma positiva. Mientras que las observaciones separadas por 6 meses se encuentran correlacionadas negativamente. También observamos que el CCF presenta *lags* negativos, lo cual indica que el SOI se encuentra por delante del Reclutamiento.

De esta manera podemos decir que el índice de Oscilación del Sur (SOI) medido en el tiempo *t-6* meses estará asociado con la serie del Reclutamiento en el tiempo *t*, es decir la serie del SOI adelanta en seis meses a la serie del Reclutamiento.

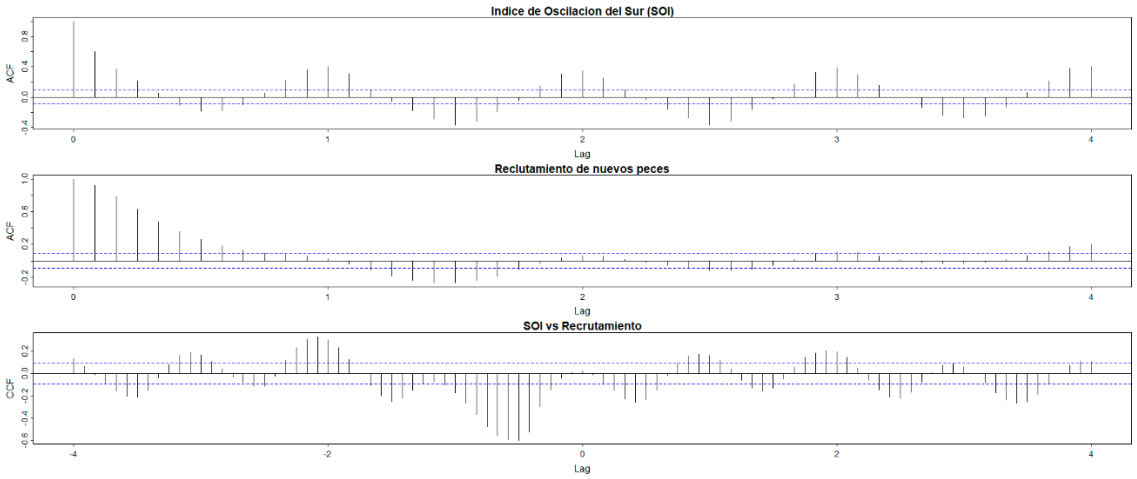

Fig 4. Funciones de autocorrelación (ACF) y de correlación cruzada (CCF) para las series SOI y de Reclutamiento

- ## Aplicación del modelo

Relación entre índice SOI y el Reclutamiento

Como comprobamos al analizar la función de correlación cruzada (CCF), vemos que existe cierta relación entre el índice SOI y el Reclutamiento. Realizar una regresión retardada en R es un poco difícil dado que las series deben estar alineadas antes de ejecutar la regresión. Sin embargo, esta alineación de las series retardadas se puede evitar usando el paquete de R llamado dynlm, utilizado para realizar análisis de modelos dinámicos.

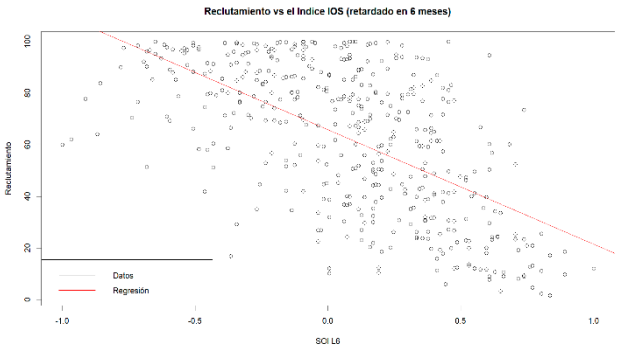

Fig 5. Regresión Lineal entre el Reclutamiento (R_t) y el Índice SOI retardado en 6 meses (S_{t-6})

De este modo asumiendo por simplicidad que entre la serie de Reclutamiento (R_t), y la serie del índice SOI retardada en 6 meses (S_{t-6}), existe una relación lineal:

$$R_t = \beta_0 + \beta_1 S_{t-6}$$

Resolviendo la regresión lineal con el paquete dynlm, obtenemos los siguientes coeficientes:

$$R_t = 65.79 - 44.28 S_{t-6}$$

Donde el modelo presenta un coeficiente de determinación R-squared de 0.36 junto con un valor-p de 2.2e-16, es decir debido a la simplificación de que la relación es lineal el modelo solo explica el 36% de los datos, sin embargo, un valor-p tan bajo nos muestra una fuerte relación estadísticamente significativa entre las variables R_t y S_{t-6}.

Esta aproximación a una relación lineal se la puede apreciar de mejor manera en la Fig. 5, sin embargo es claro que no justifica toda la variabilidad de los datos, por lo que posiblemente exista una relación no lineal entre las variables.

Modelo y predicción de la serie del Reclutamiento

Tal como verificamos en la Fig. 4, la serie del Reclutamiento presenta una estacionalidad más debil que la serie del SOI, por lo tanto en este caso aplicaremos un modelo ARIMA sin estacionalidad: ARIMA (p,d,q).

Para lo cual primero aplicaremos el test ADF para verificar la estacionariedad de la serie, obteniendo así un valor-p de 0.01 por lo cual se comprueba que la serie es estacionaria y no necesita diferenciación por lo que tendremos d=0.

Posteriormente utilizaremos un codigo en loop para encontrar la mejor combinacion entre p y q que obtenga el menor indice AIC. De este modo obtenemos p=1 y q=3, que al aplicarlo a nuestro modelo para predecir los valores futuros, observamos una continuación coherente de la serie, tal se aprecia en la Fig. 6.

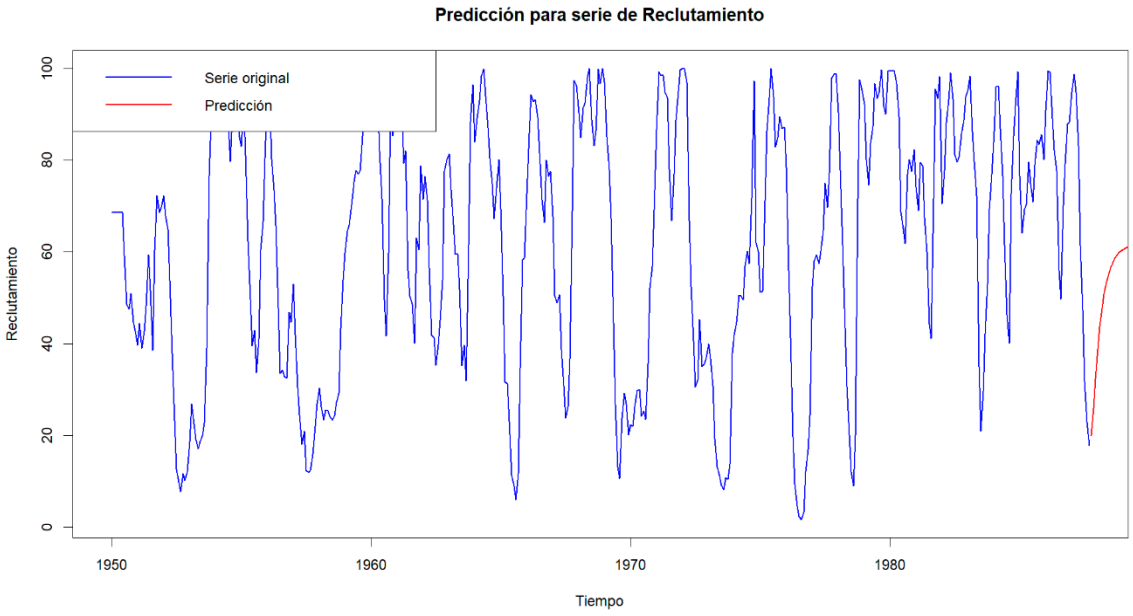

Fig 6. Serie del Reclutamiento (color azul) y predicción del modelo ARIMA (1,0,3) (color rojo)

Modelo y predicción de la serie del Índice SOI

Tal como verificamos en la Fig. 4, la serie de indice SOI presenta una estacionalidad más fuerte que la serie del Reclutamiento, por lo tanto en este caso aplicaremos un modelo ARIMA con estacionalidad anual: SARIMA (p,d,q) (P,D,Q) 12.

Para lo cual primero aplicaremos el test ADF para verificar la estacionariedad de la serie, obteniendo así un valor-p de 0.01 por lo cual se comprueba que la serie es estacionaria y no necesita diferenciación por lo que tendremos d=D=0.

Dado que la serie del Indice SOI, muestra la misma periodicidad de los picos en la funcion de autocorrelacion ACF como en la funcion de autocorrelacion parcial PACF, utilizaremos p=P=1 y q=Q=1. Al aplicar este modelo para predecir los valores futuros, observamos una continuación coherente de la serie, tal se aprecia en la Fig. 7.

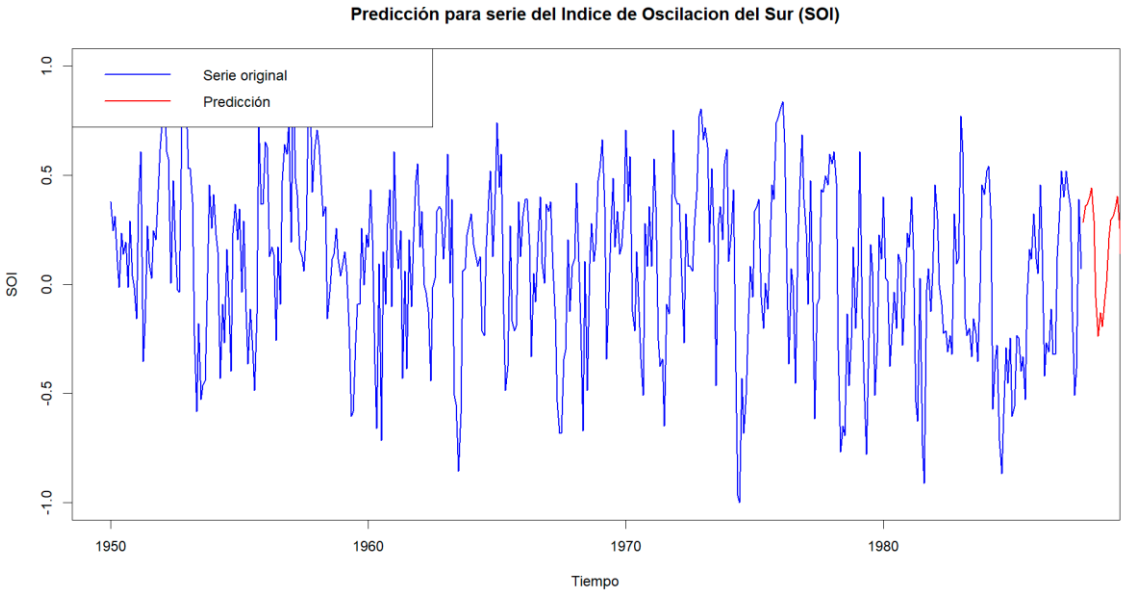

Fig 7. Serie del Índice SOI (color azul) y predicción del modelo SARIMA (1,0,1)(1,0,1) 12 (color rojo)

4. Conclusiones

- **Resultados y limitaciones del trabajo**

En este trabajo verificamos (Fig.1) que las series analizadas están compuestas por regularidades caracterizadas por un prominente ciclo anual y una oscilación más lenta de aproximadamente 4 años. También observamos (Fig. 2) que no poseen una tendencia clara, pero si presentan una marcada estacionalidad anual.

Además, al analizar el promedio de los ciclos de la serie SOI (Fig. 3) observamos muy leves disminuciones a través de un largo periodo de tiempo, lo cual parecería indicar una disminución del índice SOI a través de las ultimas décadas, posiblemente producto del cambio climático, sin embargo, se necesitarían más datos para poder comprobar esta tendencia.

Al analizar la relación del índice SOI con el Reclutamiento por medio del CCF (Fig. 4) podemos notar que la serie del SOI adelanta en seis meses a la serie del Reclutamiento. Al estudiar una relación lineal entre el Reclutamiento y el índice SOI retardado en 6 meses, obtenemos un valor-p muy bajo, lo cual nos muestra una fuerte relación estadísticamente significativa entre las variables.

Pero dado que la regresión lineal solo explica el 36% de los datos, esto sugiere la posible existencia de una relación no lineal entre las series temporales, la cual requeriría un estudio más detenido utilizando técnicas adicionales, como el análisis de la función de transferencia. Este enfoque permite la identificación de patrones no lineales, lo que lo convierte en una herramienta poderosa para comprender y predecir fenómenos interconectados en series temporales, no obstante, este tipo de análisis ya sale del alcance del presente trabajo.

Dado que las series temporales analizadas en este trabajo presentan picos periódicos en sus funciones de autocorrelación ACF y autocorrelación parcial PACF, un modelo

probabilístico SARIMA es una opción adecuada, no obstante, es relevante señalar que los valores predichos no serán completamente precisos, ya que las predicciones meteorológicas son conocidas por ser notoriamente difíciles de proyectar debido a la complejidad de los sistemas atmosféricos y la influencia de múltiples variables.

Además, a menudo, los modelos tradicionales como SARIMA pueden tener dificultades para capturar la variabilidad extrema y las rápidas fluctuaciones climáticas. En estos casos, aplicar un modelo GARCH (Generalized Autoregressive Conditional Heteroskedasticity) podría ser más conveniente. Los modelos GARCH son especialmente adecuados para modelar y predecir la volatilidad en series temporales, lo que puede reflejar mejor las condiciones meteorológicas cambiantes y las fluctuaciones en la variabilidad de los datos.

Al combinar la información de tendencia y estacionalidad de SARIMA con la capacidad de GARCH para capturar la volatilidad, es posible mejorar la precisión de este tipo de predicciones meteorológicas, sin embargo, dicho análisis se encuentra fuera del alcance del presente estudio.

5. Referencias Bibliográficas

• Libros Académicos

- Shumway, R. H., & Stoffer, D. S. (2017). Time Series Analysis and its applications.
- Shumway, R. H., & Stoffer, D. S. (2019). Time Series: A Data Analysis approach using R.

• Repositorios Git-Hub

- Código R de tsa4: https://github.com/nickpoison/tsa4/blob/master/textRcode.md
- Código R de tsda: https://github.com/nickpoison/tsda/blob/main/Rcode.md

6. Anexo

• Script en R utilizado

El siguiente script utilizado para producir las gráficas del presente trabajo se encuentra en la siguiente ruta de Git Hub: https://github.com/rodraxphysics/Time-series-analysis

```r
#Importacion librerias
library(dynlm)
library(atsa)

#Importacion de datasets de atsa
data("soi")
data("rec")

#Obtencion estadisticas principales
summary(rec)
summary(rec)

#Grafica de las series temporales utilizadas
par(mfrow = c(2,1))
tsplot(soi, col=4, ylab="", main="Indice de Oscilacion del Sur (SOI)",xlab="Tiempo", lwd=2)
tsplot(rec, col=4, ylab="", main="Reclutamiento de nuevos peces",xlab="Tiempo", lwd=2)

#Descomposicion de las series temporales
plot(decompose(soi))
title(main = "Descomposición Indice de Oscilacion del Sur (SOI)")

plot(decompose(rec))
title(main = "Descomposición del Reclutamiento")
```

```r
#Analisis de la funcion de autocorrelacion (ACF)
#y analisis de la funcion de correlacion cruzada (CCF)
par(mfrow=c(4,1))
acf(soi, 48)
title(main="Indice de Oscilacion del Sur (SOI)")

acf(rec, 48)
title(main="Reclutamiento de nuevos peces")

ccf(soi, rec, 48, ylab="CCF")
title(main="SOI vs Recrutamiento")

#Analisis de funcion de autocorrelacion parcial (PACF)
par(mfrow=c(3,1))
pacf(soi, 48)
title(main="Indice de Oscilacion del Sur (SOI)")

pacf(rec, 48)
title(main="Reclutamiento de nuevos peces")


#Regresion lineal entre reclutamiento
#y SOI retardado en 6 meses
soi_vector <- as.vector(soi)
lag_soi <- lag(soi_vector, 6)
summary(fit2 <- dynlm(rec~ L(soi,6)))
#Multiple R-squared:  0.3629
#p-value: < 2.2e-16

#Grafico de datos Reclutamiento y regresion lineal del modelo
plot(lag_soi, fish$rec, main="Reclutamiento vs el Indice IOS (retardado en 6 meses)",
xlab="SOI L6", ylab="Reclutamiento")

abline(fit2, col="red")
legend("bottomleft", legend=c("Datos", "Regresión"), col=c("black", "red"), lty=1, lwd
=c(1, 2))


#Suavizado por Kernel (promedio)
plot(soi)
lines(ksmooth(time(soi), soi, "normal", bandwidth=1), lwd=2, col=4)
par(fig = c(.65, 1, .65, 1), new = TRUE) # the insert
gauss = function(x) { 1/sqrt(2*pi) * exp(-(x^2)/2) }
x = seq(from = -3, to = 3, by = 0.001)
plot(x, gauss(x), type ="l", ylim=c(-.02,.45), xaxt='n', yaxt='n', ann=FALSE)


#Suavizado por LOWESS (regresion local)
#similar a regresion KNN
plot(soi)
lines(lowess(soi, f=.05), lwd=2, col=4) # ciclos de El Niño
lines(lowess(soi), lty=2, lwd=2, col=2) # promedio de ciclos

# Grafico de suavizado de datos por LOWESS
plot(soi,ylab="SOI", xlab="Tiempo")
lines(lowess(soi, f = 0.05), lwd = 2, col = "blue", legend.text = "Ciclos de El Niño")
lines(lowess(soi), lty = 2, lwd = 2, col = "red", legend.text = "Suavizado de Datos
(ciclos)")
legend("bottomleft", legend = c("Datos del Indice de Oscilacion del Sur (SOI)",
"Suavizado de Datos (ciclos de El Niño)", "Promedio del Suavizado (promedio ciclos)"),

       col = c("black", "blue", "red"), lty = c(1, 1, 2), lwd = c(1, 2, 2))


# Prueba de estacionariedad ACF para serie de Reclutamiento
adf_result <- adf.test(rec)
print(adf_result)
#p-value = 0.01 (estacionaria)
```

```r
# Prueba del p y q optimo para serie de Reclutamiento
resultados_aic <- matrix(NA, nrow = 5, ncol = 5) # Matriz 5x5
for (p in 1:5) {
  for (q in 1:5) {
    modelo_arima <- arima(rec, order = c(p, 0, q))
    valor_aic <- AIC(modelo_arima)
    resultados_aic[p, q] <- valor_aic
  }
}
min_aic <- min(resultados_aic)
fila_col_min_aic <- which(resultados_aic == min_aic, arr.ind = TRUE)
cat("La combinación con el menor AIC es p =", fila_col_min_aic[1], "y q =",
fila_col_min_aic[2], "\n")
cat("El valor del AIC mínimo es:", min_aic, "\n")

# Modelo ARIMA para serie de Reclutamiento
modelo_arima <- arima(rec, order = c(1 ,0, 3))
pronostico <- forecast(modelo_arima, h = 100)

plot(rec, main = "Predicción para serie de Reclutamiento", ylab = "Reclutamiento",
xlab = "Tiempo", type = "l", col = "blue",lwd=1.7)

lines(pronostico$mean, col = "red",lwd=1.7)

legend("topleft", legend = c("Serie original", "Predicción"), col = c("blue", "red"),
lty = 1,lwd=1.7)


# Prueba de estacionariedad ACF para serie de Indice SOI
adf_result <- adf.test(soi)
print(adf_result)
#p-value = 0.01 (estacionaria)

# Modelo SARIMA para serie de Indice SOI
modelo_arima <- arima(soi, order = c(1, 0, 1), seasonal = list(order = c(1, 0, 1),
period = 12))

pronostico <- forecast(modelo_arima, h = 100)

plot(soi, main = "Predicción para serie del Indice de Oscilacion del Sur (SOI)", ylab
= "SOI", xlab = "Tiempo", type = "l", col = "blue",lwd=1.7)

lines(pronostico$mean, col = "red",lwd=1.7)

legend("topleft", legend = c("Serie original", "Predicción"), col = c("blue", "red"),
lty = 1,lwd=1.7)
```