



# Minería de datos y Modelización predictiva I

## Series Temporales II: Modelos ARIMA



# Tema 2 Series Temporales

2.1.- Introducción.

2.2.- Función de autocorrelación simple y función de autocorrelación parcial.

2.3. El modelo ARMA(p,q).

2.3.1. El modelo autorregresivo AR(p).

2.3.2. El modelo de medias móviles MA(q).

2.4. El modelo ARIMA(p,d,q).

2.5.- El modelo ARIMA estacional..

2.6.- La metodología Box-Jenkins.,

2.7.- Transformaciones para estabilizar la varianza.

2.8.-Identificación y estimación del modelo ARIMA.

2.9.-Diagnóstico del modelo.

2.9.1.- Significación estadística de los parámetros.

2.9.2.- Análisis de los residuos.

2.9.3.-. Medidas de la adecuación del modelo.

2.10. Cálculo de las predicciones



## 2.1.- INTRODUCCIÓN.

**Una serie temporal de  $n$  datos es una muestra extraída del proceso estocástico** que representa dicho fenómeno dinámico, por ello recordamos algunos conceptos básicos de procesos estocásticos.

### Estacionarios

- Toman valores estables en el tiempo alrededor de un valor central, sin mostrar una tendencia o crecer o decrecer a lo largo del tiempo

### No estacionarios

- Pueden mostrar tendencia, estacionalidad y otros efectos evolutivos en el tiempo.



Un **proceso estocástico es estacionario** en sentido estricto cuando **las distribuciones marginales de cualquier conjunto de k variables son idénticas, en distribución y parámetros**

Nos basta con que el proceso **sea estacionario en sentido débil**, es decir

$$\begin{cases} \mu_t = \mu & \forall t \\ \sigma_t^2 = \sigma^2 & \forall t \\ Cov(x_t, x_{t+k}) = E[(x_t - \mu)(x_{t+k} - \mu)] = \gamma_k & \forall k \end{cases}$$

**La media y la varianza permanecen constantes** con el tiempo y la tercera que **la covarianza** entre dos variables de la serie **depende sólo de su separación en el tiempo**

$$\rho_k = \frac{Cov(x_t, x_{t-k})}{\sqrt{Var(x_t)Var(x_{t-k})}} = \frac{\gamma_k}{\gamma_0}$$

$$\gamma_0 = \sigma_{x_t}^2 \quad \forall t$$



## Y la matriz de correlaciones

$$R_k = \begin{bmatrix} \rho_0 & \rho_1 & \cdots & \rho_{k-1} \\ \rho_1 & 1 & \cdots & \rho_{k-2} \\ \cdot & \cdot & \cdots & \cdot \\ \rho_{k-1} & \rho_{k-2} & \cdots & 1 \end{bmatrix}$$



Una propiedad importante de los procesos estacionarios es que son **estables ante las combinaciones lineales**. En particular, **los incrementos de una serie estacionaria son estacionarios**

$$\omega_t = x_t - x_{t-1}$$

Un proceso estocástico es **un ruido blanco** si

$$E[X(t)] = 0 \quad V[X(t)] = \sigma^2 \quad \gamma_k = 0 \quad \forall k = 1, 2, \dots$$

Introduzcamos **el operador de retardo** que nos permitirá escribir de forma más compacta las expresiones anteriores.

$$BX_t = X_{t-1} \quad B^2 X_t = X_{t-2} \quad B^k X_t = X_{t-k}$$

**El operador de retardo es lineal**

$$B(aX_t + bY_t) = aX_{t-1} + bY_{t-1}$$



## 2.2. FUNCIÓN DE AUTOCORRELACIÓN SIMPLE (ACF) Y FUNCIÓN DE AUTOCORRELACIÓN PARCIAL (PACF).

Si de un proceso estacionario observamos  $x_1, x_2, \dots, x_T$

podemos calcular los estimadores de los parámetros anteriores.

El **estimador de la media**  $\hat{\mu} = \bar{x}$

La varianza de este estimador es 
$$Var(\bar{x}) = \frac{1}{T} \left[ \sigma^2 + 2 \sum_{i=1}^{T-1} \left( 1 - \frac{i}{T} \right) \gamma_i \right] = \frac{\gamma_0}{T} \left[ 1 + 2 \sum_{i=1}^{T-1} \left( 1 - \frac{i}{T} \right) \rho_i \right]$$

El **estimador de la autocovarianza de orden k** es 
$$\hat{\gamma}_k = \frac{1}{T} \sum_{t=k+1}^T (x_t - \bar{x})(x_{t-k} - \bar{x})$$

El **estimador del coeficiente de correlación** lo calcularemos 
$$r_k = \frac{\hat{\gamma}_k}{\hat{\gamma}_0}$$



Estos coeficientes de correlación muestrales expresados en función del retardo forman la **función de autocorrelación muestral ACF** y su representación es el **correlograma**.

Estas cantidades **miden la relación lineal entre las variables de la serie separadas por k posiciones**.

El error estándar de este coeficiente es

$$S_{r_k} = \sqrt{\frac{1 + 2 \sum_{j=1}^{k-1} r_j^2}{T}}$$

Dado que **la distribución del coeficiente de correlación es asintóticamente Normal** podemos construir el intervalo de confianza aproximado bajo la hipótesis de que las autocorrelaciones sean cero:

$$\left( \pm 1.96 S_{r_k} \right)$$

**La función ACF nos servirá para identificar si la serie es o no estacionaria.**



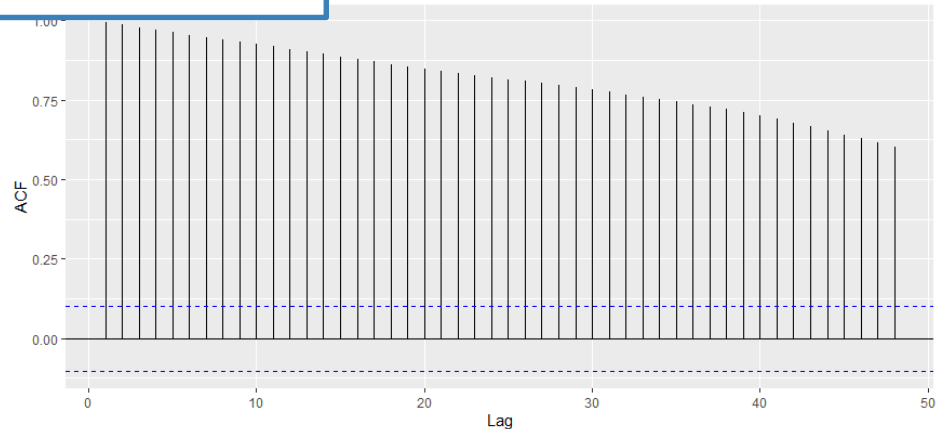


## Ejemplo de correlogramas.

La observación y el análisis de los correlogramas nos sirve para detectar, la estacionariedad, el tipo de modelo y los retardos que son significativamente diferentes de cero.

Con R, obtenemos la representación gráfica de sus autocorrelogramas mediante la siguiente sintaxis:

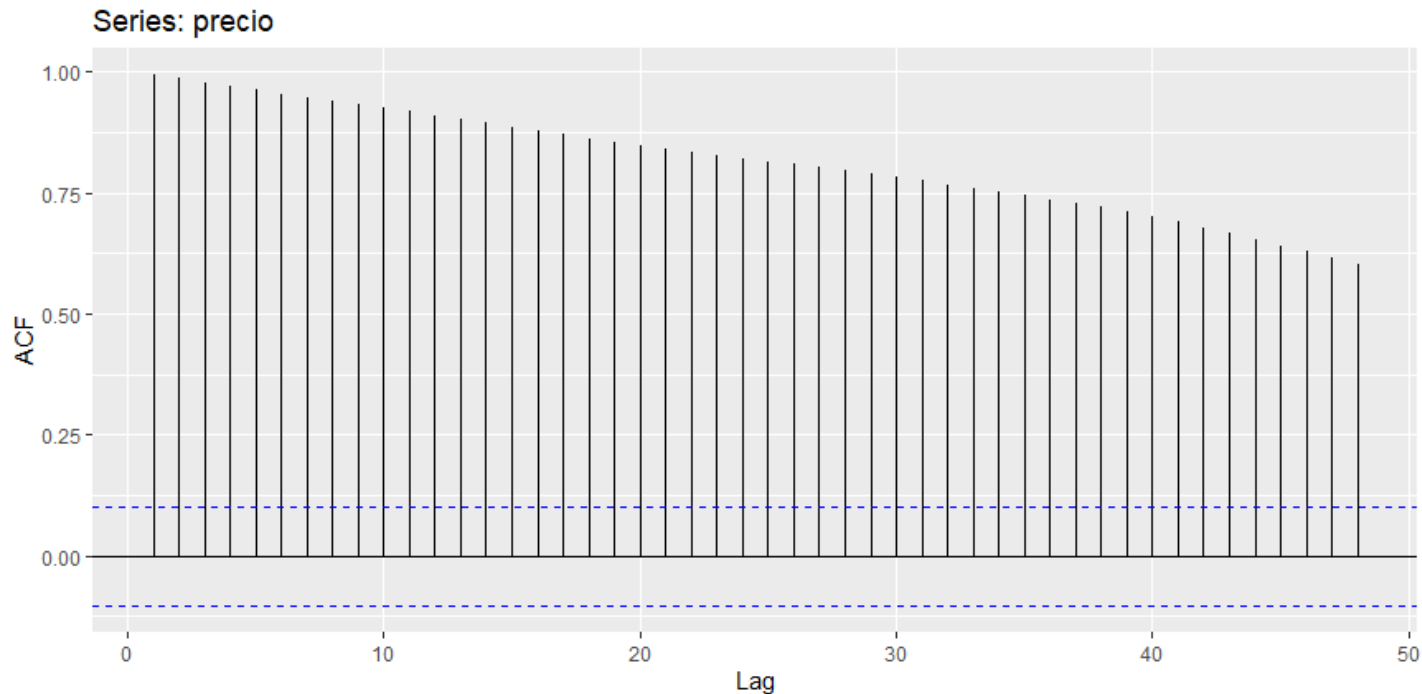
```
#Calculamos las autocorrelaciones simples hasta el retardo 48  
ggAcf(precio, lag=48)  
  
#Calculamos las autocorrelaciones parciales hasta el retardo 48  
ggPacf(precio, lag=48)
```



Si el correlograma ACF **decrece lentamente** la serie es **no estacionaria**,

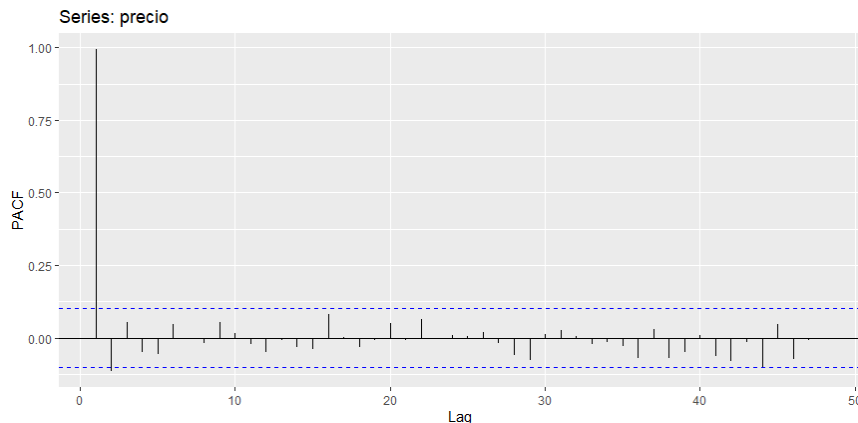
Si el correlograma ACF **se corta o decrece rápidamente** podemos considerar la serie **estacionaria**.

**Esta función también nos va a indicar el tipo de modelo a ajustar y el orden.**



Otro concepto muy útil en el análisis de series temporales es la **función de autocorrelación parcial (PACF)**

Un modelo AR(p) tiene “**solo los p primeros**” coeficientes de correlación parcial distintos de cero y esto nos servirá para identificar el modelo autorregresivo a utilizar.

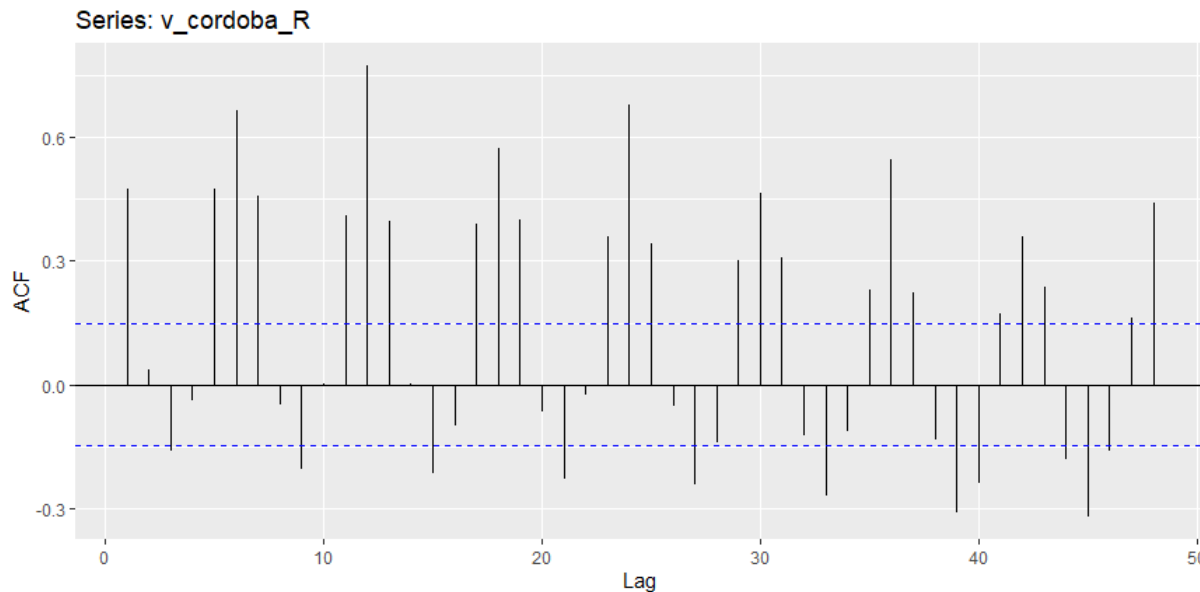


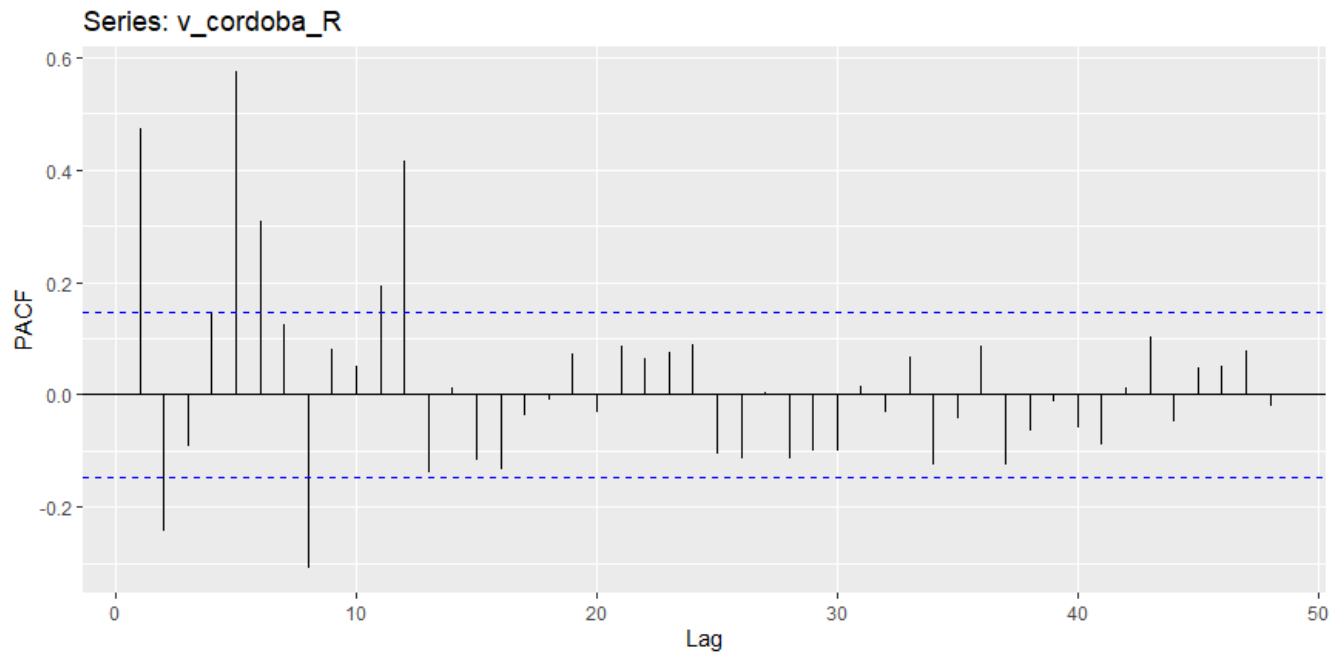
Las autocorrelaciones simples (ACF) y las parciales (PACF) aparecen representadas con las bandas de confianza que se calculan utilizando el error Standard aproximado, como  $\pm \frac{2}{\sqrt{T}}$

Observemos que estas bandas representan el intervalo de confianza en el que podría estar el coeficiente de autocorrelación calculado si el verdadero valor del poblacional fuera cero. Por esto, un coeficiente de correlación dentro de las bandas se considera cero.

**Ejemplo:** Veamos ahora un ejemplo de una serie con comportamiento estacional.

```
#Calculamos las autocorrelaciones simples hasta el retardo 48  
ggAcf(v_cordoba_R, lag=48)  
  
#Calculamos las autocorrelaciones parciales hasta el retardo 48  
ggPacf(v_cordoba_R, lag=48)
```





Se observa un comportamiento repetitivo de las autocorrelaciones cada 12 meses, observando como la autocorrelación más fuerte es el los retardos múltiplos de 12. Esto es debido a que las autocorrelaciones simples tienen un efecto acumulativo de retardos anteriores.

### 2.3. EL MODELO AUTORREGRESIVO AR(P).

Comenzamos con el modelo autoregresivo que **generaliza la idea de regresión para representar la relación entre una variable de la serie y las anteriores:**

$$\text{AR(1)} \quad X_t = c + \phi X_{t-1} + Z_t \quad -1 < \phi < 1$$

$Z_t$  es un proceso de ruido blanco con varianza  $\sigma^2$

Esta dependencia puede también extenderse a los p valores anteriores **AR(p)**

$$X_t = c + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \dots + \phi_P X_{t-p} + Z_t$$

Si centramos la serie restando su media podemos suprimir la constante y el proceso autorregresivo se puede expresar

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \dots + \phi_P X_{t-p} + Z_t$$

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \dots + \phi_P X_{t-P} + Z_t$$

Utilizado el operador de retardo la ecuación anterior se puede escribir como:

$$\phi_p(B) X_t = Z_t$$

$$\phi_p(B) = 1 - \phi_1 B - \phi_2 B^2 - \dots - \phi_p B^p$$

### 2.3.1. El proceso AR(1).

$$X_t = \phi_1 X_{t-1} + Z_t \quad \longleftrightarrow \quad \phi_1(B) X_t = Z_t$$

El modelo AR(1) es un proceso estacionario si

$$|\phi| < 1$$

La función de autocorrelación es

$$\rho_k = \phi^k$$



Va decreciendo de forma geométrica



Esto implica que la ACF de **un proceso AR(1)** pueda tener el siguiente aspecto:

$\phi > 0$   La ACF será una función positiva y decreciente

$\phi < 0$   La ACF será una función alternada, y tendrá palos pares positivos, y palos impares negativos.

En cuanto a la **PACF**, como **sólo existe influencia de primer orden**

$\phi > 0$   La PACF tendrá un único palo, el primero. Este palo será positivo.

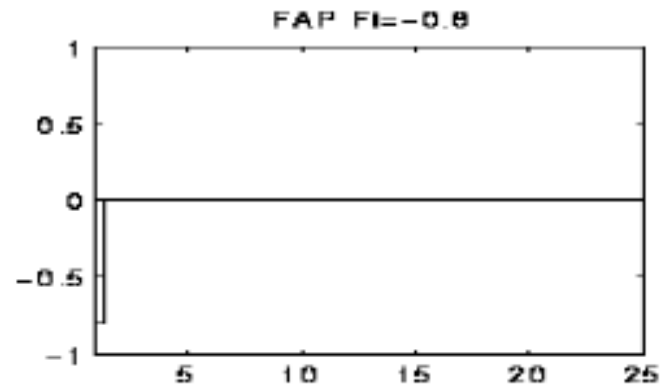
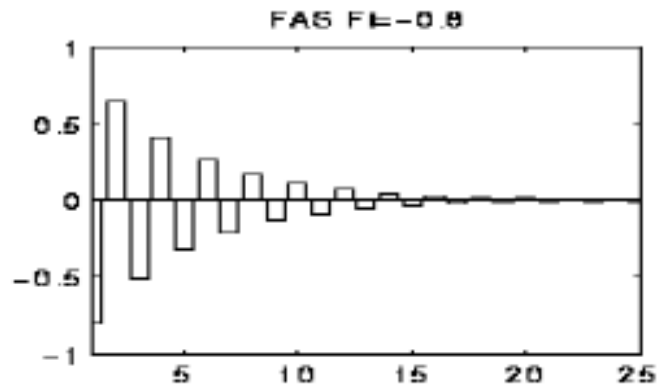
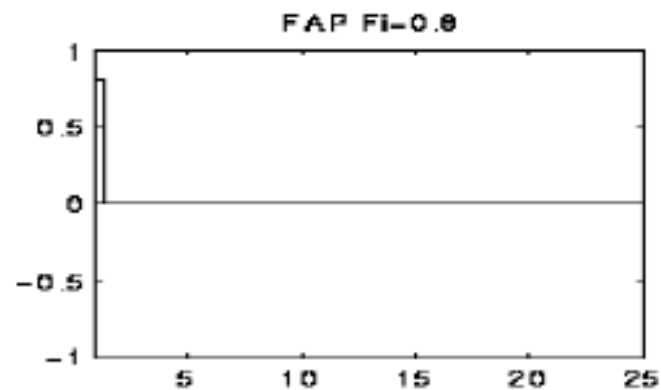
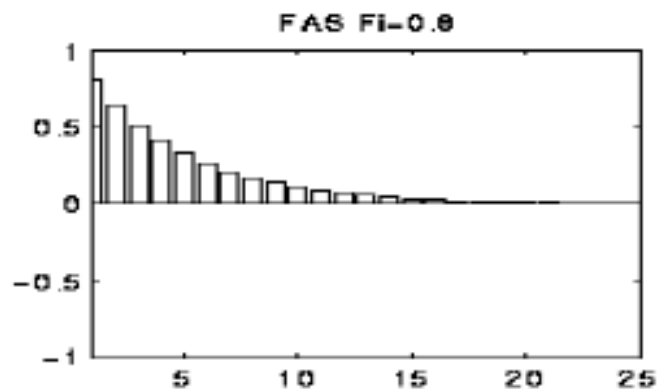
$\phi < 0$   La PACF tendrá un único palo y será negativo.





Ejemplo de funciones de autocorrelación para

$$\phi = \pm 0.8$$



### 2.3.2. El proceso AR(2).

El proceso autorregresivo de segundo orden tiene por ecuación

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + Z_t$$

Se puede demostrar que el proceso AR(2) será estacionario siempre que las raíces del denominado polinomio característico estén fuera del círculo unidad.

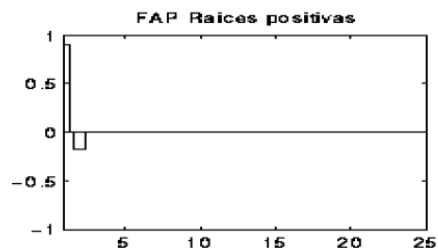
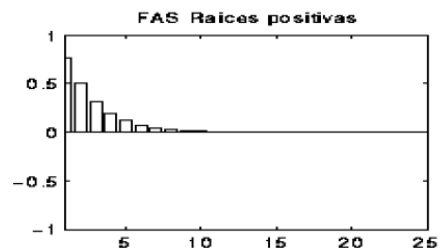
$$1 - \phi_1 B - \phi_2 B^2 = 0$$

•**Dos raíces reales.** En este caso la FAS es la superposición de dos exponenciales decrecientes y la forma depende de que las raíces tengan el mismo signo u opuesto.

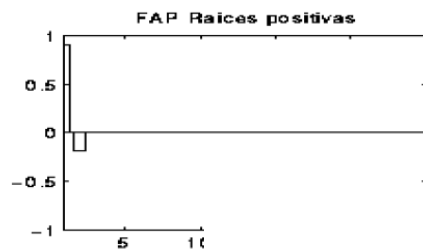
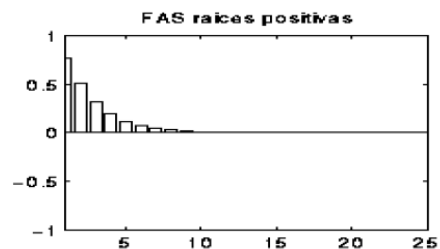
•**Raíces imaginarias.** En este caso la FAS y la serie tienen un aspecto cíclico

La FAP del AR(2), tendrá únicamente dos polos.

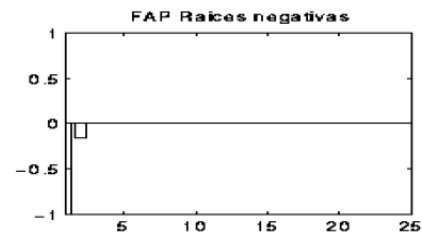
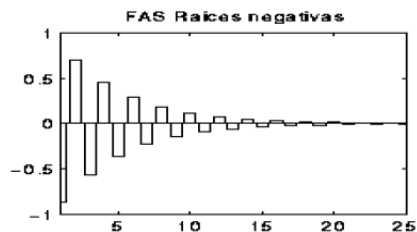
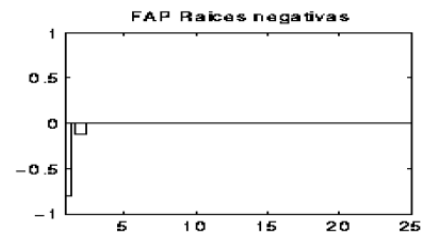
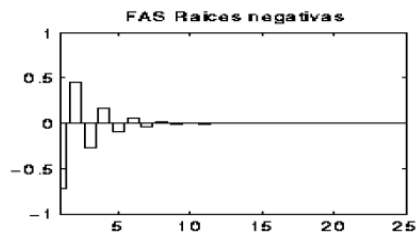
## Raices reales



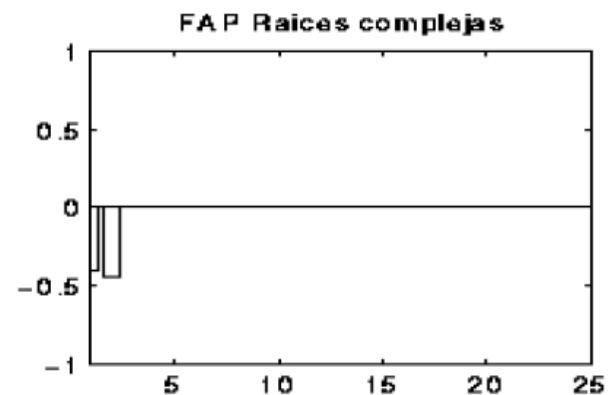
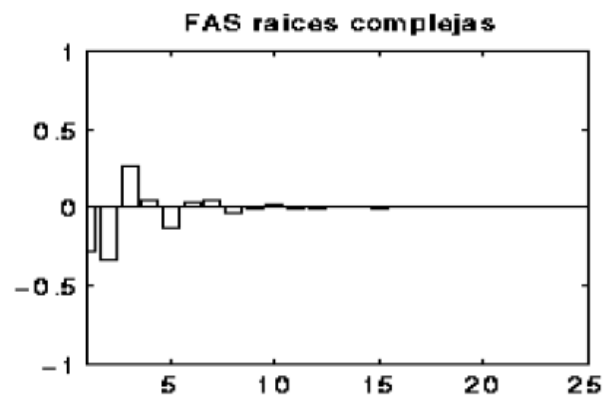
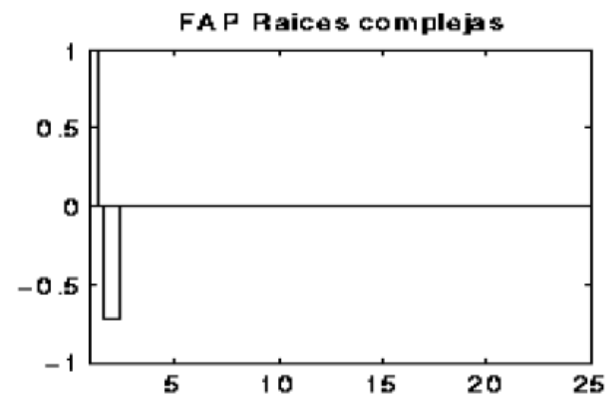
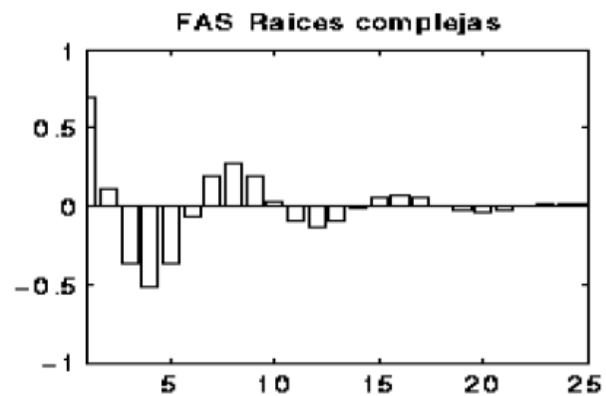
## Raices positivas



## Raices negativas



## Raíces complejas



### 2.3.3. El proceso autoregresivo general AR(p).

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \dots + \phi_p X_{t-p} + Z_t$$

Veamos ahora la **condición de estacionariedad**

$$\phi_p(B) = 1 - \phi_1 B - \phi_2 B^2 - \dots - \phi_p B^p$$

**Así el proceso será estacionario si las raíces de dicho polinomio tienen módulo mayor que 1.**

## 2.4. EL MODELO DE MEDIAS MÓVILES MA(q).

Los procesos autoregresivos representan series que tienen una memoria larga, ya que su función de autocorrelación **FAS decrece de forma exponencial** pero no se corta a partir de un determinado retardo.

Para representar series de memoria corta crearemos los modelos de medias móviles cuya identificación es sencilla porque su función de autocorrelación **FAS se corta** a partir de un determinado retardo. Estos procesos son una media de un número finito de innovaciones pasadas.

### 2.4.1. El proceso de media móvil de orden 1 MA(1).

$$X_t = Z_t + \theta_1 Z_{t-1}$$

Este proceso es la suma de dos procesos estacionarios y por lo tanto sea cual sea el valor del parámetro es estacionario a diferencia de los modelos AR .

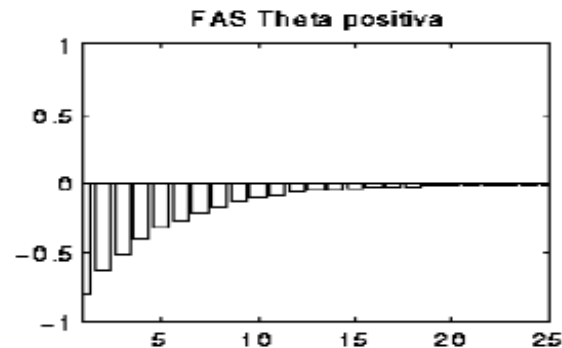
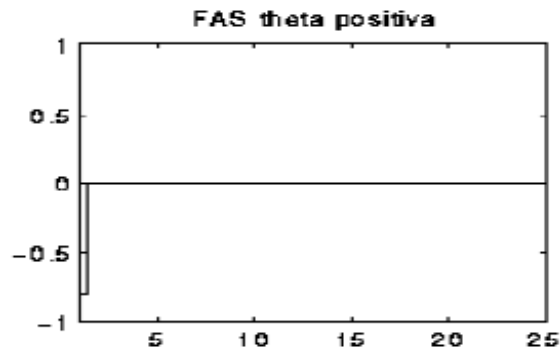
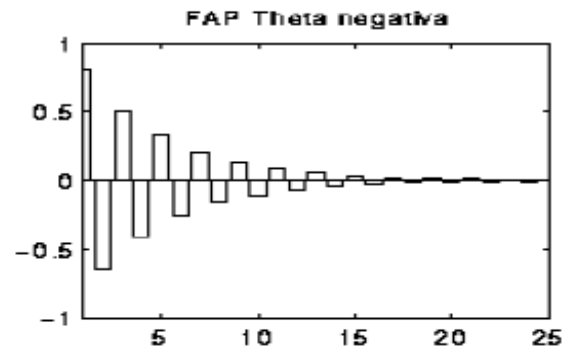
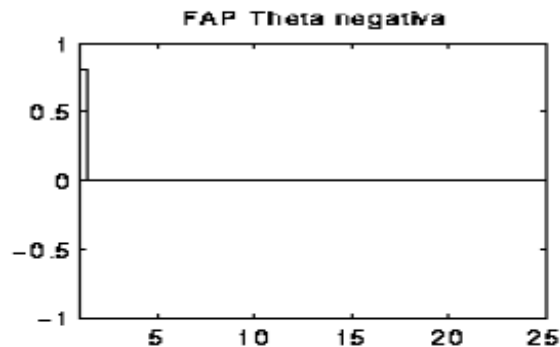
El modelo puede escribirse con la notación de operadores como:

$$X_t = (1 + \theta_1 B) Z_t$$

$$|\theta_1| < 1$$



El proceso es **invertible** y tiene la propiedad de que el efecto de los valores pasados decrece con el tiempo.



### 2.4.2.- El proceso MA(q).

Generalizando la idea podemos escribir procesos cuyo valor actual no sólo dependa de la última innovación, sino de las q últimas innovaciones. Se obtiene entonces el proceso MA(q) cuya expresión general es.

$$X_t = Z_t + \theta_1 Z_{t-1} + \theta_2 Z_{t-2} + \dots + \theta_q Z_{t-q}$$

Introduciendo la notación del operador de retardo:

$$X_t = (1 + \theta_1 B + \theta_2 B^2 + \dots + \theta_q B^q) Z_t \quad \longleftrightarrow \quad X_t = \theta_q (B) Z_t$$

**Este proceso es** la suma de procesos estacionarios y por lo tanto sea cual sea el valor de los parámetros es **estacionario**. **El proceso es invertible y tiene la propiedad de que el efecto de los valores pasados decrece con el tiempo si las raíces del polinomio característico son en módulo mayores que la unidad.**



La función de autocorrelación de un proceso MA(q) tiene la misma “forma” que la función de autocorrelación parcial de un modelo AR(q). **Concluimos que existe una dualidad entre los modelos AR y MA, de manera que la FAS de un MA es como la FAP de un AR y viceversa.**

## 2.5. EL MODELO MIXTO ARMA(p,q).

Una extensión natural de los modelos anteriores es aquella que incluye términos autorregresivos y términos de medias móviles. Se representan por la ecuación:

$$X_t - \phi_1 X_{t-1} - \phi_2 X_{t-2} - \dots - \phi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + \theta_2 Z_{t-2} + \dots + \theta_q Z_{t-q}$$



$$(1 - \phi_1 B - \phi_2 B^2 - \dots - \phi_p B^p) X_t = (1 + \theta_1 B + \theta_2 B^2 + \dots + \theta_q B^q) Z_t$$

Este proceso es estacionario si lo es su parte AR y es invertible si lo es su parte MA.

## 2.6.- PROCESOS INTEGRADOS: EL MODELO ARIMA(p,d,q).

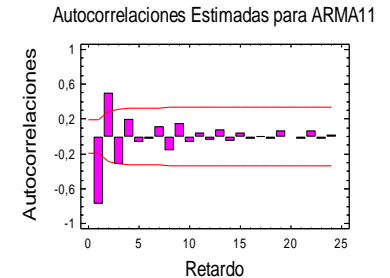
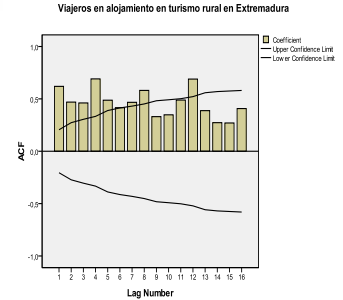
➤ Cuando **el nivel de la serie no es constante en el tiempo**, pudiendo en particular tener tendencia creciente o decreciente, diremos que **la serie es no estacionaria en la media**.

➤ Cuando **la variabilidad o las autocorrelaciones se modifican en el tiempo**, diremos que la serie **es no estacionaria en la varianza o en las autocorrelaciones..**

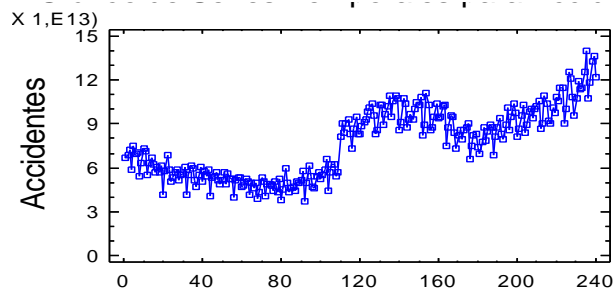
➤ Finalmente si la distribución de la variable varía con el tiempo, diremos que la serie es no estacionaria en distribución.

Los procesos no estacionarios más importantes son **los procesos integrados**, que tienen la propiedad fundamental de que **al diferenciarlos se obtienen procesos estacionarios**.

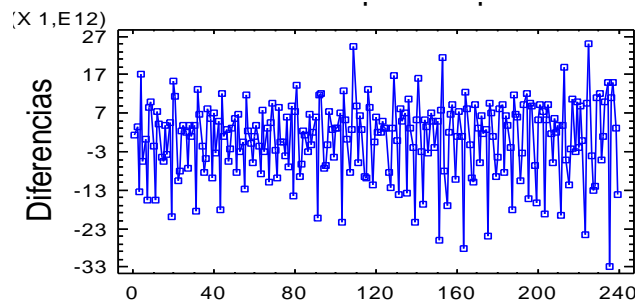
Una propiedad importante que diferencia a los procesos integrados de los estacionarios es la forma en que desaparece la dependencia en el tiempo.



Frecuentemente las series económicas no son estacionarias pero sus diferencias relativas, o las diferencias cuando medimos la variable en logaritmos, son estacionarias. Por ejemplo la serie número de accidentes es no estacionaria en media.



$$\omega_t = X_t - X_{t-1}$$

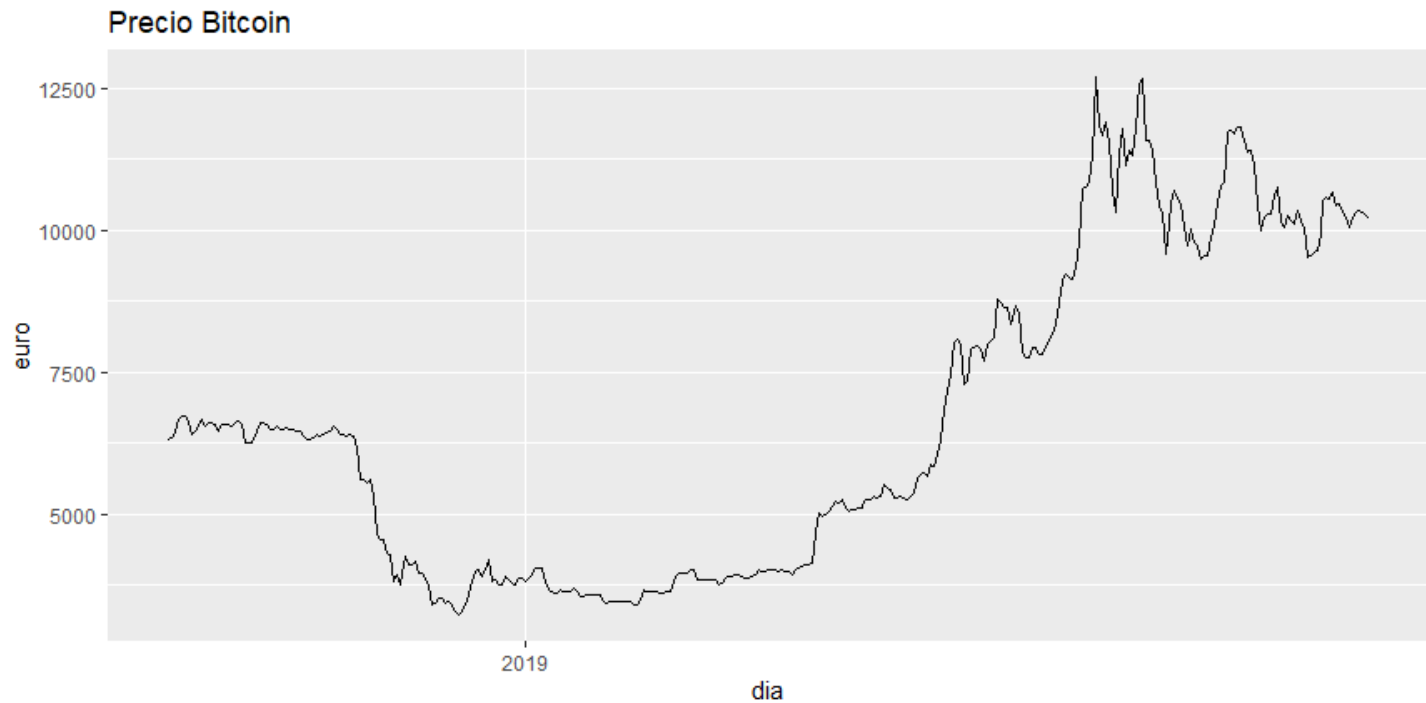


Diremos que un proceso es integrado de orden 1 si  $\omega_t = \nabla X_t = X_t - X_{t-1}$  ya es estacionaria.

En general si necesitamos realizar “d” diferencias diremos que es un proceso integrado de orden d. Así el modelo ARIMA(p,d,q) se expresa:

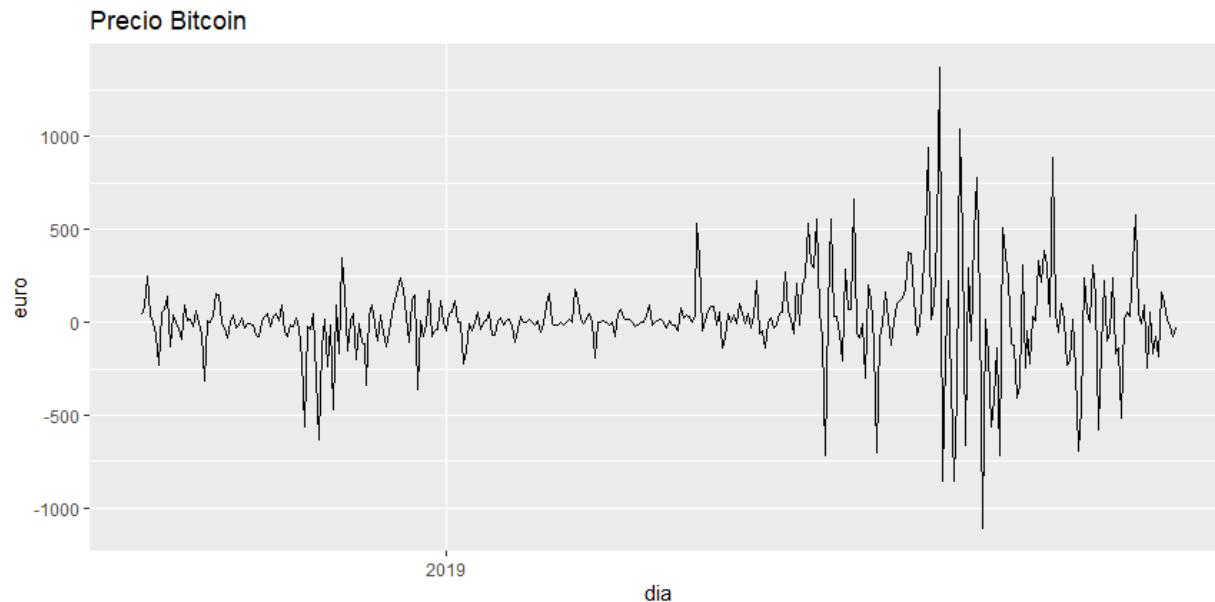
$$\left(1 - \phi_1 B - \phi_2 B^2 - \dots - \phi_p B^p\right) (1 - B)^d X_t = \left(1 + \theta_1 B + \theta_2 B^2 + \dots + \theta_q B^q\right) Z_t$$

```
autoplot(precio)+ ggtitle("Precio Bitcoin") + xlab("dia") +  
ylab("euro")
```



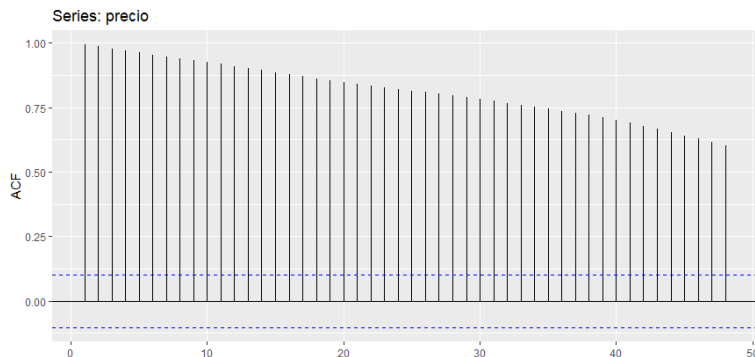
```
#Serie diferenciada
```

```
autoplot(diff(precio))+ ggtitle("Precio Bitcoin") + xlab("dia") +  
ylab("euro")
```



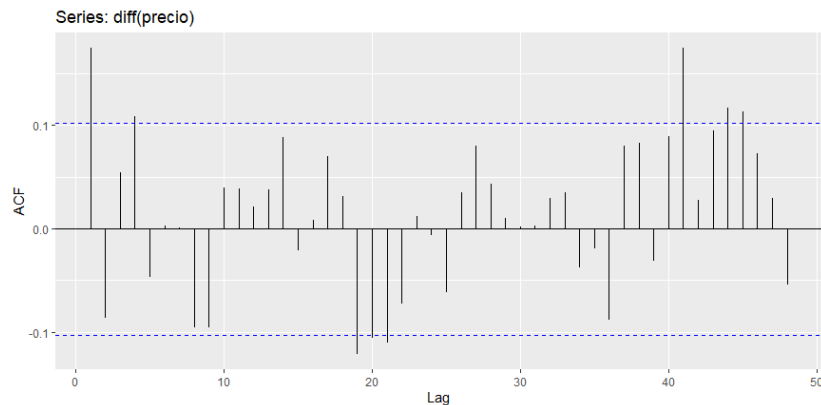
Como vemos, la serie ya tiene media constante aunque podría ser que la varianza no lo fuera.

Si estudiamos el correlograma de la serie no estacionaria vemos que decrece muy lentamente y de forma lineal.



#Calculamos las autocorrelaciones simples de la serie diferenciada hasta el retardo 48

```
ggAcf(diff(precio), lag=48)
```



## 2.7.- EL MODELO ARIMA ESTACIONAL.

En el tema de métodos descriptivos vimos que podíamos eliminar la estacionalidad mediante diferencias con los índices estacionales.

Podemos convertir una serie con estacionalidad en estacionaria mediante las diferencias de orden  $s$ , siendo  $s$  el periodo de la serie.

Definimos el operador diferencia de periodo  $s$  o diferencia estacional de orden 1 como

$$\nabla_s X_t = X_t - X_{t-s} = (1 - B^s) X_t$$

Un modelo estacional general será de la forma

$$ARIMA(p, d, q)(P, D, Q)_s$$

$$\begin{aligned} & (1 - \Phi_1 B^s - \Phi_2 B^{2s} - \dots - \Phi_p B^{ps}) (1 - \phi_1 B - \phi_2 B^2 - \dots - \phi_p B^p) (1 - B^s)^D (1 - B)^d X_t = \\ & (1 + \Theta_1 B + \Theta_2 B^{2s} + \dots + \Theta_q B^{qs}) (1 + \theta_1 B + \theta_2 B^2 + \dots + \theta_q B^q) Z_t \end{aligned}$$

Veamos algunos ejemplos de este modelo:

$$ARIMA(1,0,0)(1,0,0)_{12} \Rightarrow (1 - \Phi_1 B^{12})(1 - \phi_1 B) X_t = Z_t$$

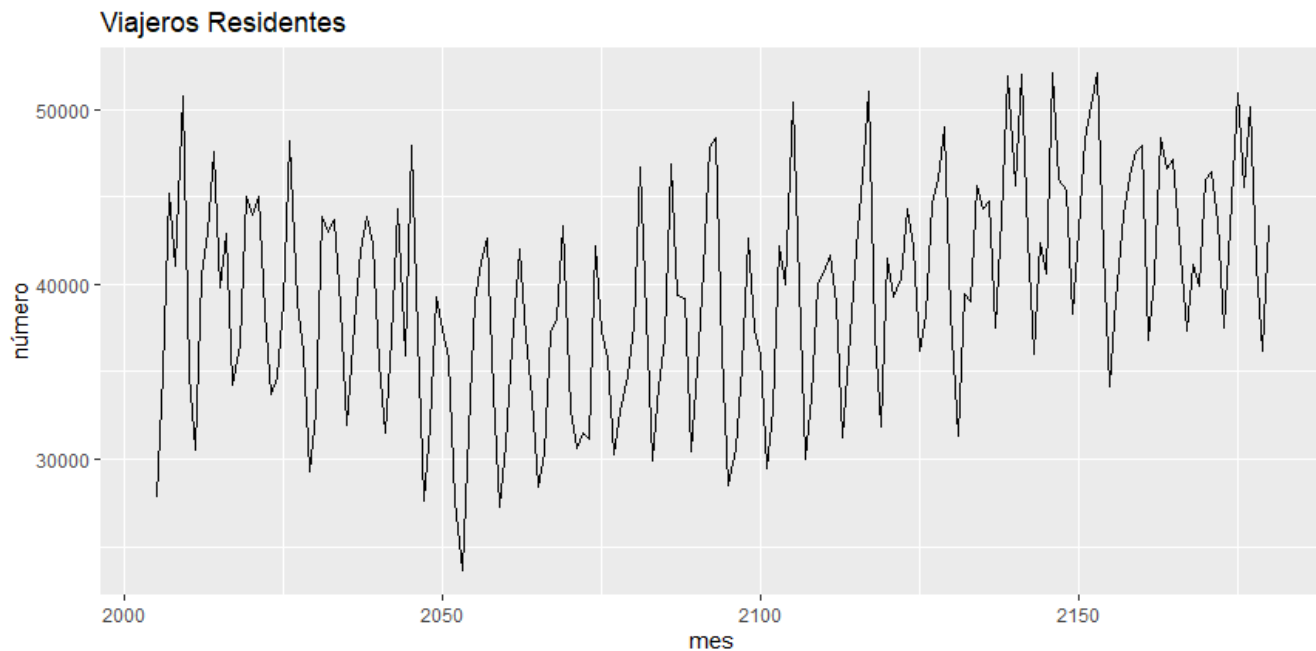
$$ARIMA(1,0,0)(0,0,1)_{12} \Rightarrow (1 - \phi_1 B) X_t = (1 - \Theta_1 B^{12}) Z_t$$

$$ARIMA(2,0,0)(1,0,0)_{12} \Rightarrow (1 - \Phi_1 B^{12})(1 - \phi_1 B - \phi_2 B^2) X_t = Z_t$$



**Ejemplo:** La serie de viajeros en Córdoba es claramente estacional. La siguiente sintaxis nos permite ver la gráfica de la serie y los autocorrelogramas

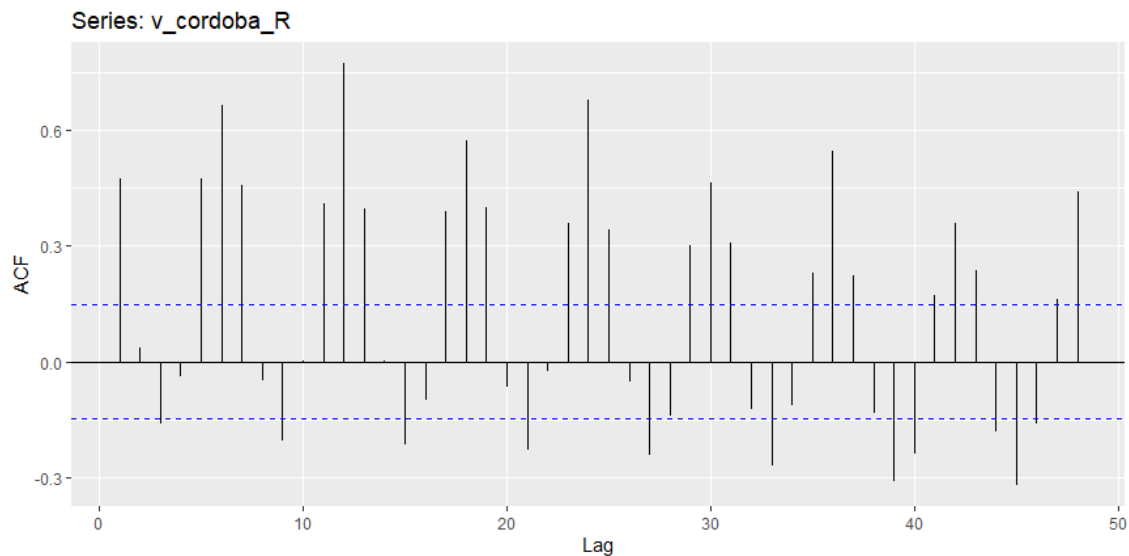
```
autoplot(v_cordoba_R)+ ggtitle("Viajeros Residentes") + xlab("mes") + ylab("número")
```



Si representamos sus correlogramas se observa la estructura estacional y la no estacionariedad en media que se refleja en el gráfico de la serie.

```
#Calculamos las autocorrelaciones simples hasta el retardo 48
```

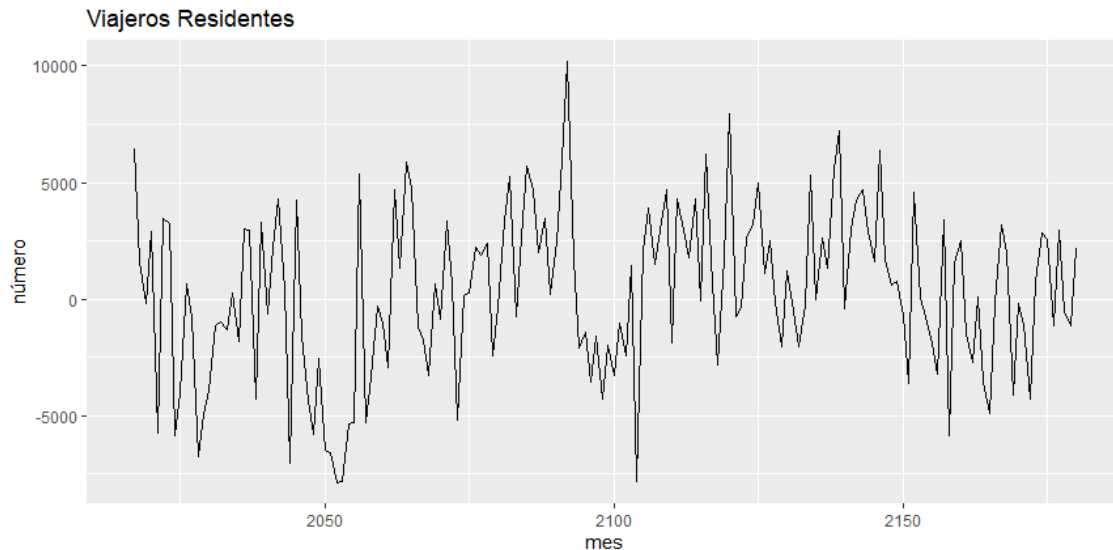
```
ggAcf(v_cordoba_R, lag=48)
```



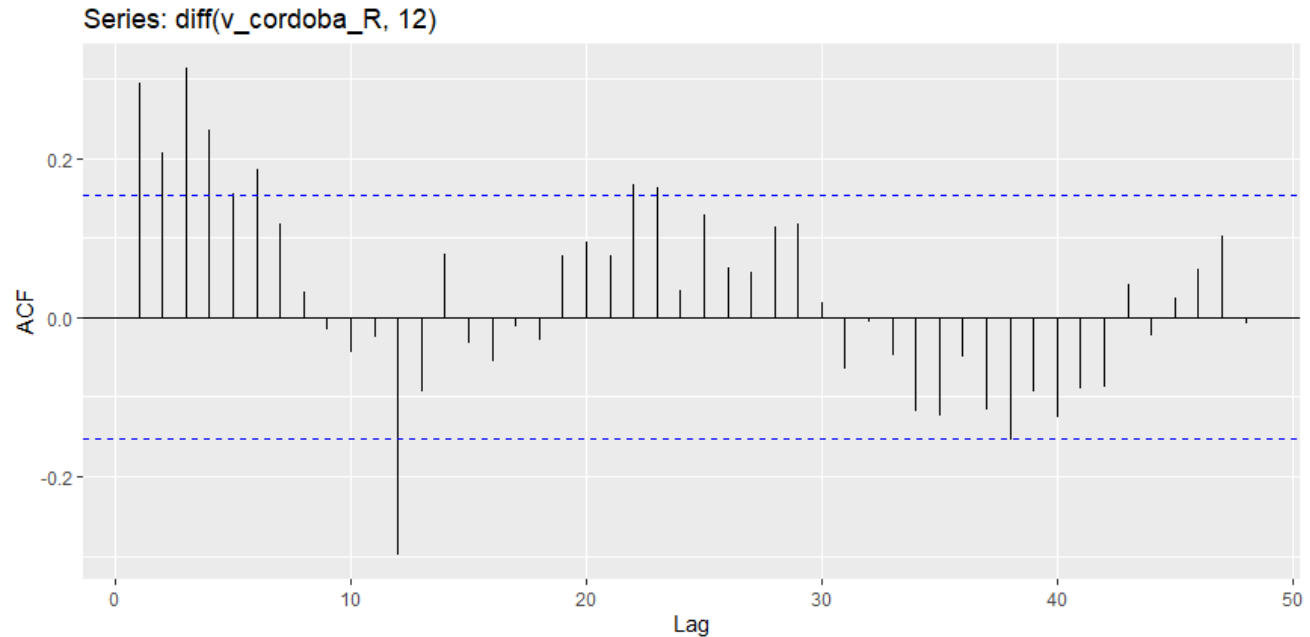
Si diferenciamos la serie, mediante una diferenciación de orden estacional, y calculamos sus funciones de autocorrelación tenemos

```
#Serie diferenciada estacionalmente
```

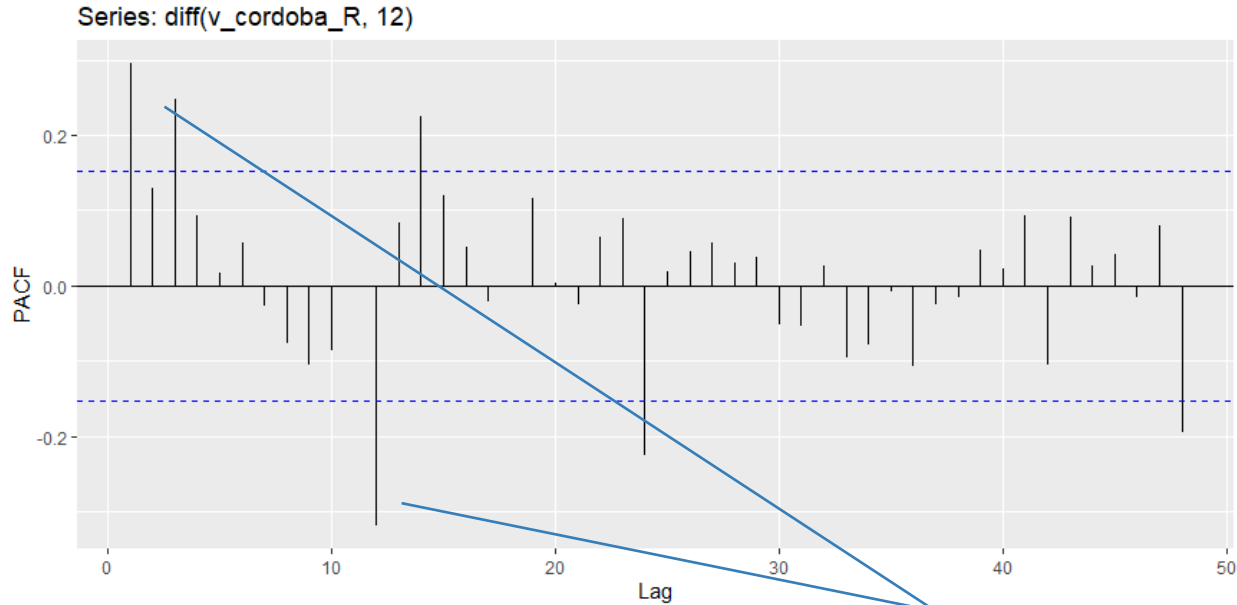
```
autoplot(diff(v_cordoba_R,12))+ ggtitle("Viajeros Residentes") + xlab("mes") + ylab("número")
```



```
#Calculamos las autocorrelaciones simples de la serie diferenciada  
ggAcf(diff(v_cordoba_R,12), lag=48)
```



```
#Calculamos las autocorrelaciones parciales de la serie diferenciada  
ggPacf(diff(v_cordoba_R,12),lag=48)
```



En donde podemos ver que el proceso ya es estacionario  
y teniendo en cuenta las correlaciones distintas de cero  
tanto en las correlaciones simples como en las parciales  
podría ajustarse a un modelo

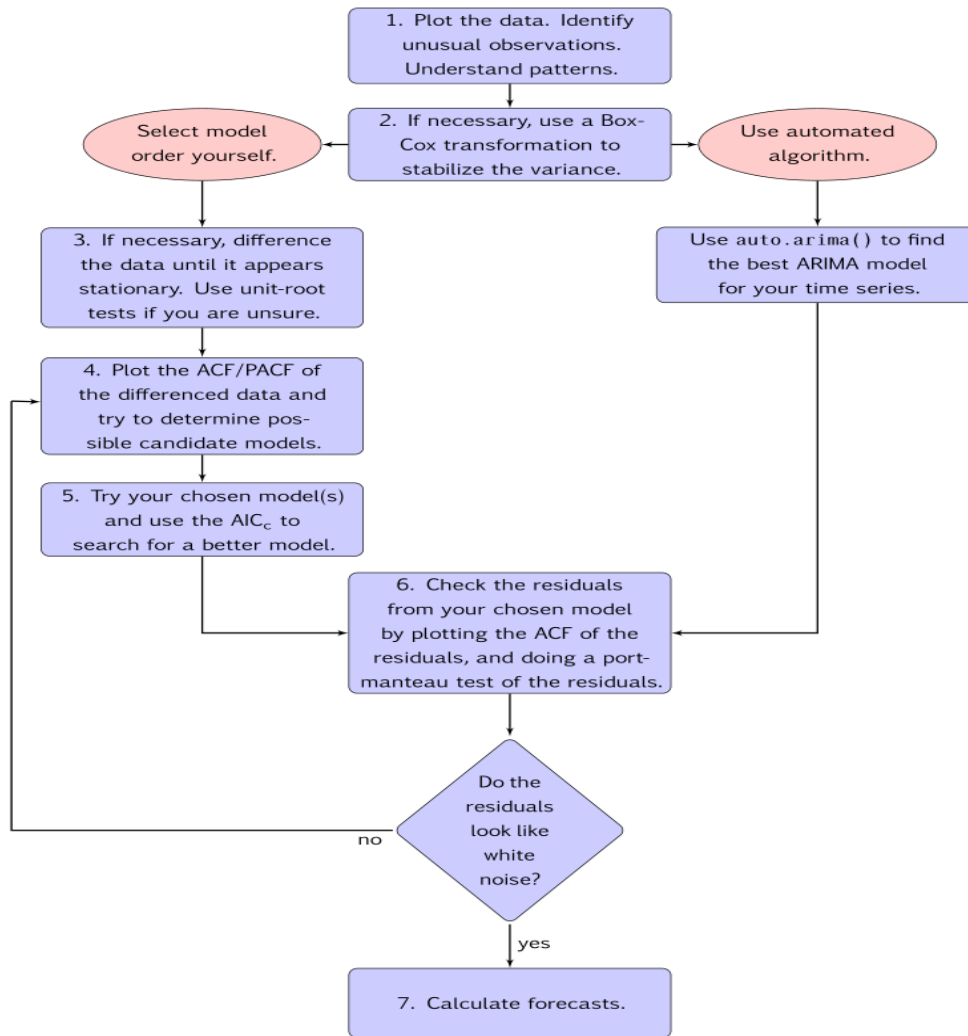
*ARIMA*(3, 0, 0)(0, 1, 1)<sub>12</sub>

### 3.1.- La Metodología Box-Jenkins.

En el tema anterior hemos estudiado las propiedades teóricas de los modelos ARIMA, a continuación vamos a analizar **como ajustar dichos modelos a series reales**. Box y Jenkins propusieron utilizar una metodología que se resume en cuatro etapas.

- **Paso 1. Identificación del modelo:** Utilizamos los datos históricos de la serie para encontrar el modelo apropiado.
- **Paso 2. Estimación:** Estimamos los parámetros del modelo escogido utilizando los datos históricos.
- **Paso 3. Pruebas del modelo:** Realizamos distintos contrastes para decidir si el modelo construido es adecuado. Si no lo es volveríamos al paso 1.
- **Paso 4. Predicción:** Una vez que el modelo ha sido construido y comprobada su adecuación lo utilizamos para hacer predicciones.





### 3.2 TRANSFORMACIONES PARA ESTABILIZAR LA VARIANZA.

Cuando la falta de estacionariedad además viene dada por una varianza que aumenta o disminuye con la media de la serie, podemos utilizar las transformaciones Box-Cox para estabilizarla. Dichas transformaciones tienen la expresión general:

$$\begin{cases} w = \log y \\ w = \frac{y^\lambda - 1}{\lambda} \end{cases}$$

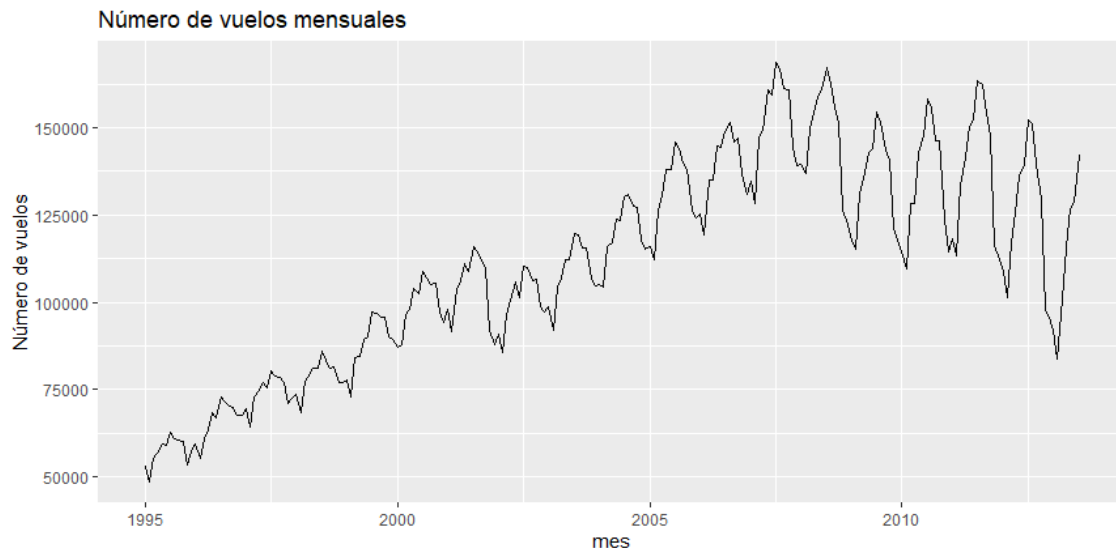
Esta familia se utiliza para valores de  $y$  mayores que cero, pero se puede expresar de forma más general utilizando el valor de  $c$  para que  $y$  sea positivo.

$$\begin{cases} w = \log (y + c) / g \\ w = \frac{(y + c)^\lambda - 1}{\lambda g} \end{cases}$$



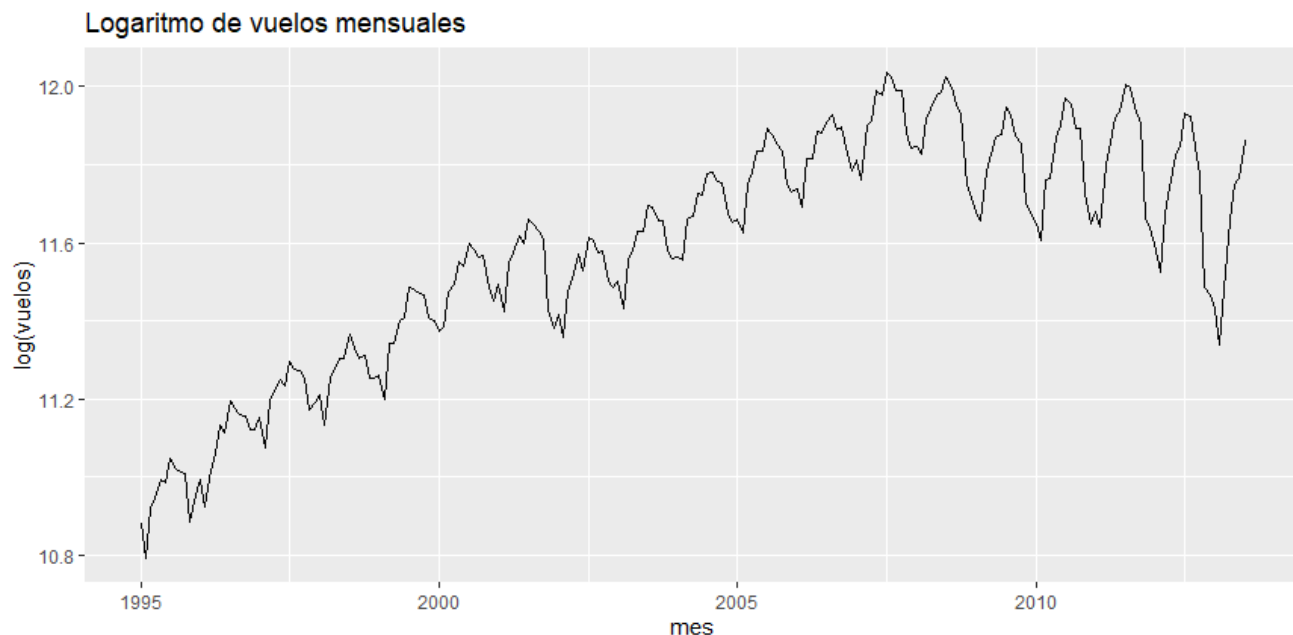
Por ejemplo, la serie número de vuelos España desde Enero de 1995 tiene la siguiente representación gráfica:

```
#Número de vuelos en España  
vuelos <- ts(VUELOS[,-1], start=c(1995,1), frequency=12)  
autoplot(vuelos)+ ggtitle("Número de vuelos mensuales") +  
  xlab("mes") + ylab("Número de vuelos")
```



Para estabilizar la varianza tomamos logaritmos y representamos la serie para observar el efecto de esta transformación.

```
autoplot(log(vuelos))+ ggtitle("Logaritmo de vuelos mensuales") +  
  xlab("mes") + ylab("log(vuelos)")
```

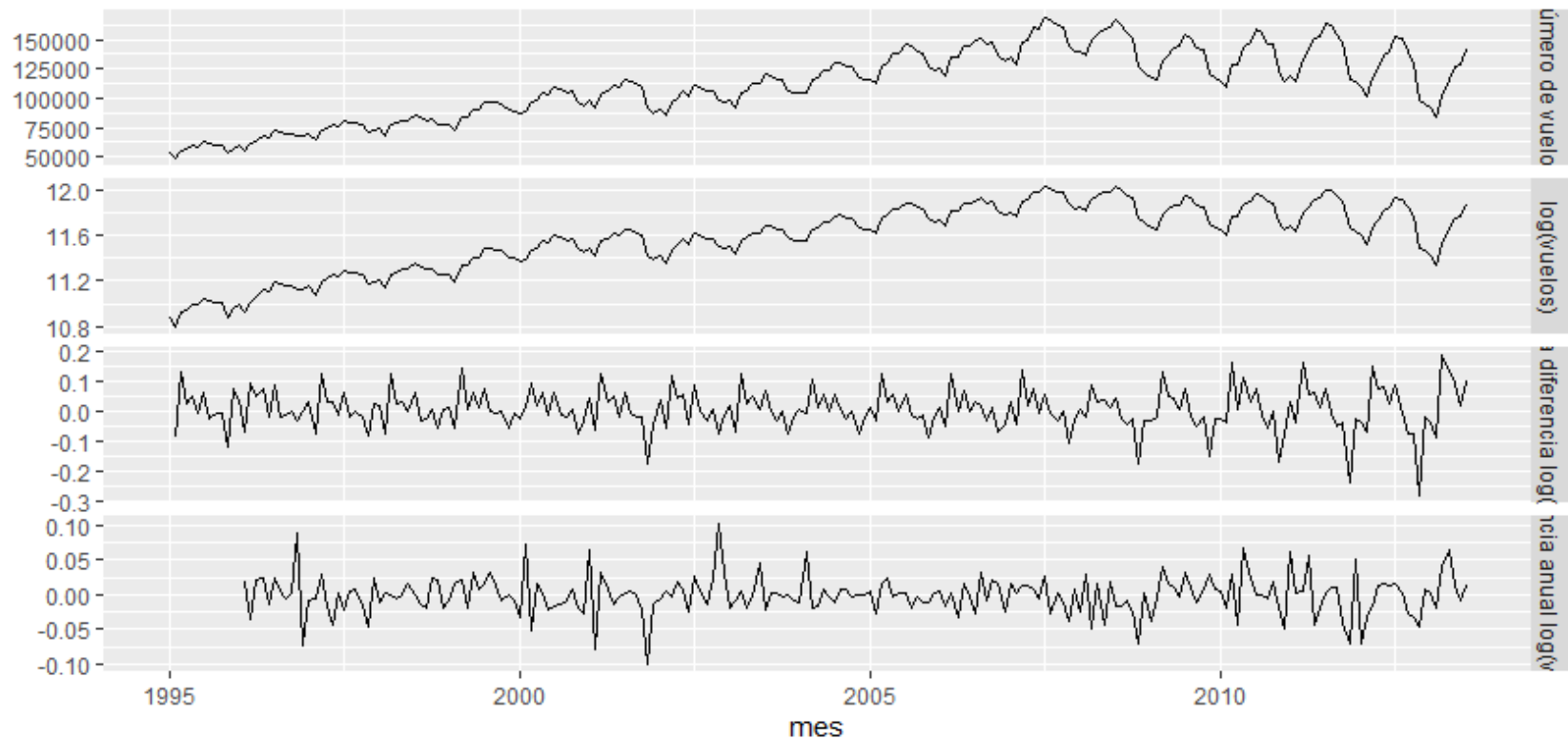


Mediante la **función cbind** creamos un dataframe con la serie original, el logaritmo de la serie, el logaritmo diferenciado una vez y sobre esta la diferenciación estacional. Estas cuatro series son representadas juntas utilizando el **operador pipe** de la **librería dplyr**

```
#Podemos representar a la vez la serie transformada por log y diferenciada  
#la función cbind crea un data frame con las columnas de las series  
#el operador %>% nos facilita la comprensión de funciones anidadas cambiando el orden  
#En este caso la función autoplot se aplica al resultado de cbind
```

```
cbind("número de vuelos" = vuelos,  
      "log(vuelos)" = log(vuelos),  
      "primera diferencia log(vuelos)" = diff(log(vuelos)),  
      "diferencia anual log(vuelos)" = diff(diff(log(vuelos)),12)) %>%  
autoplot(facets=TRUE) + xlab("mes") + ylab("") +  
ggtitle("Número de vuelos")
```

## Número de vuelos



### 3.3. IDENTIFICACIÓN DEL MODELO ARMA

La identificación puede hacerse siguiendo las siguientes reglas:

- 1.- **Decidir cual es el orden máximo de las partes AR y MA a la vista de las funciones de autocorrelación muestrales.**
- 2.- **Evitar la identificación inicial de modelos mixtos** ARMA y comenzar con modelos AR o MA, preferiblemente de orden bajo (*principio de parsimonia*). En la práctica la mayoría de las series reales pueden representarse con modelos ARMA con  $p$  y  $q$  menores que 3.
- 3.- **Intentar identificar la estructura de los valores separados por  $s$  periodos** en el caso de series con componente estacional

Hasta ahora hemos supuesto de los modelos tenían media nula por comodidad para al cálculo teórico de las autocorrelaciones puesto que el valor de la media no influye en estos cálculos. Ahora, para saber si debemos incluir el parámetro media en el modelo, calcularemos la media muestral

$$\bar{X}_n = \frac{\sum_{t=1}^n X_t}{n}$$

Que se distribuye (para n suficientemente grande) como una Normal con varianza

$$n^{-1} \sum_{|k| < \infty} \gamma_k$$

Estos valores de las autocovarianzas se sustituyen por sus estimaciones quedando, por tanto la varianza de la media aproximada por:

$$V(\bar{X}_n) \simeq \frac{\hat{\sigma}_X^2}{n} (1 + 2\hat{\rho}_1 + \dots + 2\hat{\rho}_k)$$

Si  $|\bar{X}_n| > z_{\alpha/2} \sqrt{V(\bar{X}_n)}$

**Rechazaremos que la media es nula**



### 3.4. ESTIMACIÓN DE LOS PARÁMETROS.

Comenzaremos con un modelo sin parte estacional

$$ARMA(p, q)$$

$$Z_t = -\theta_1 Z_{t-1} - \dots - \theta_q Z_{t-q} + (X_t - \mu) - \phi_1 (X_{t-1} - \mu) - \dots - \phi_p (X_{t-p} - \mu)$$

Se trata de estimar

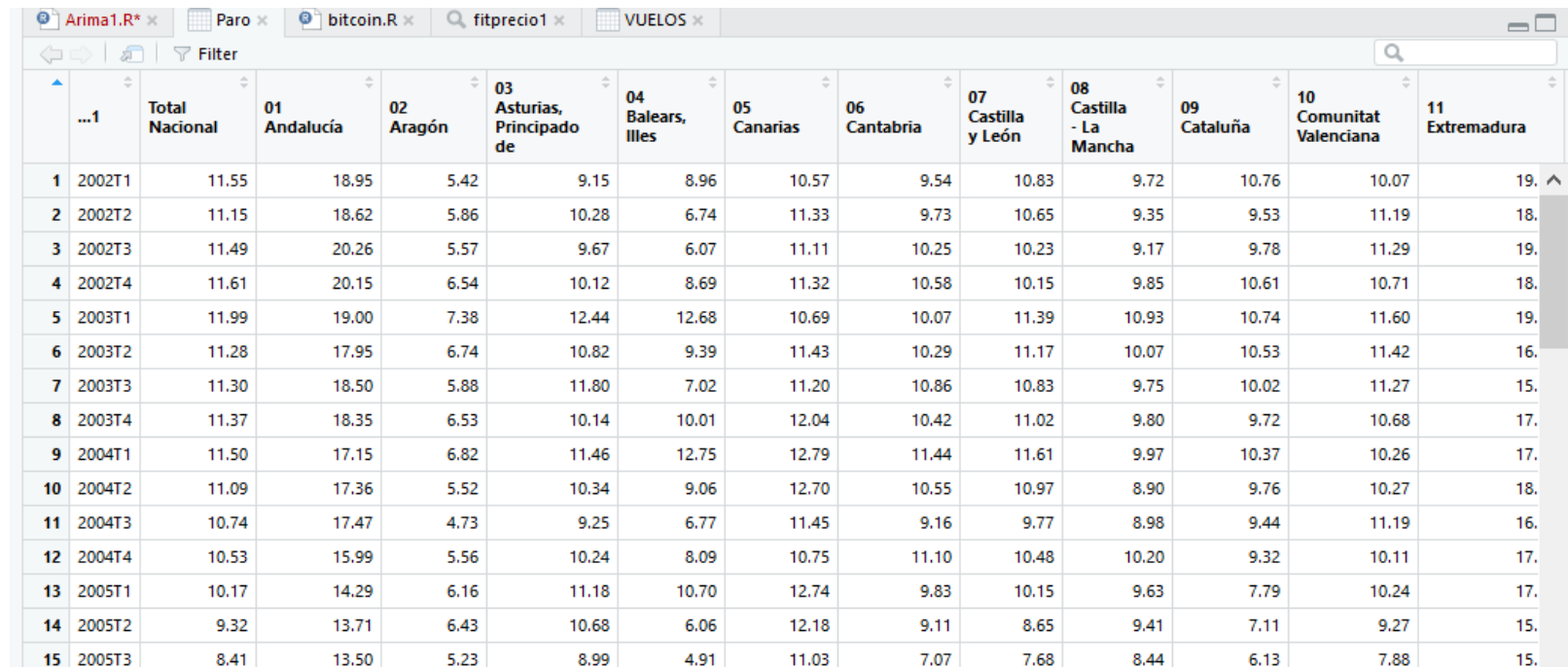
$$\phi_1, \dots, \phi_p, \theta_1, \dots, \theta_q, \sigma^2 \text{ y } \mu$$

Se puede demostrar que si el modelo ARMA es estacionario e invertible, **los procedimientos de estimación máximo verosímil y estimación por mínimos cuadrados condicionales (CLS) o no condicionales conducen a estimadores óptimos.**

**Cuando  $n$  es lo suficientemente grande podemos aproximar la distribución de los estimadores por una Normal multivariante.**

## Ejemplo:

Vamos a estudiar las series de tasa de paro en España, en Cataluña y en la Comunidad de Madrid. Estas series las tenemos en el fichero Paro que tiene por columnas el paro por comunidades autónomas.



The screenshot shows a spreadsheet application with multiple tabs at the top: 'Arima1.R\*', 'Paro', 'bitcoin.R', 'fitprecio1', and 'VUELOS'. The 'Paro' tab is active, displaying a table with 13 columns and 15 rows of data. The columns represent different regions and the total national average. The rows represent time periods from 2002T1 to 2005T3. The table includes a search bar and a filter icon at the top right.

	...1	Total Nacional	01 Andalucía	02 Aragón	03 Asturias, Principado de	04 Balears, Illes	05 Canarias	06 Cantabria	07 Castilla y León	08 Castilla - La Mancha	09 Cataluña	10 Comunitat Valenciana	11 Extremadura
1	2002T1	11.55	18.95	5.42	9.15	8.96	10.57	9.54	10.83	9.72	10.76	10.07	19.
2	2002T2	11.15	18.62	5.86	10.28	6.74	11.33	9.73	10.65	9.35	9.53	11.19	18.
3	2002T3	11.49	20.26	5.57	9.67	6.07	11.11	10.25	10.23	9.17	9.78	11.29	19.
4	2002T4	11.61	20.15	6.54	10.12	8.69	11.32	10.58	10.15	9.85	10.61	10.71	18.
5	2003T1	11.99	19.00	7.38	12.44	12.68	10.69	10.07	11.39	10.93	10.74	11.60	19.
6	2003T2	11.28	17.95	6.74	10.82	9.39	11.43	10.29	11.17	10.07	10.53	11.42	16.
7	2003T3	11.30	18.50	5.88	11.80	7.02	11.20	10.86	10.83	9.75	10.02	11.27	15.
8	2003T4	11.37	18.35	6.53	10.14	10.01	12.04	10.42	11.02	9.80	9.72	10.68	17.
9	2004T1	11.50	17.15	6.82	11.46	12.75	12.79	11.44	11.61	9.97	10.37	10.26	17.
10	2004T2	11.09	17.36	5.52	10.34	9.06	12.70	10.55	10.97	8.90	9.76	10.27	18.
11	2004T3	10.74	17.47	4.73	9.25	6.77	11.45	9.16	9.77	8.98	9.44	11.19	16.
12	2004T4	10.53	15.99	5.56	10.24	8.09	10.75	11.10	10.48	10.20	9.32	10.11	17.
13	2005T1	10.17	14.29	6.16	11.18	10.70	12.74	9.83	10.15	9.63	7.79	10.24	17.
14	2005T2	9.32	13.71	6.43	10.68	6.06	12.18	9.11	8.65	9.41	7.11	9.27	15.
15	2005T3	8.41	13.50	5.23	8.99	4.91	11.03	7.07	7.68	8.44	6.13	7.88	15.



```
#Tasa de paro por comunidades
```

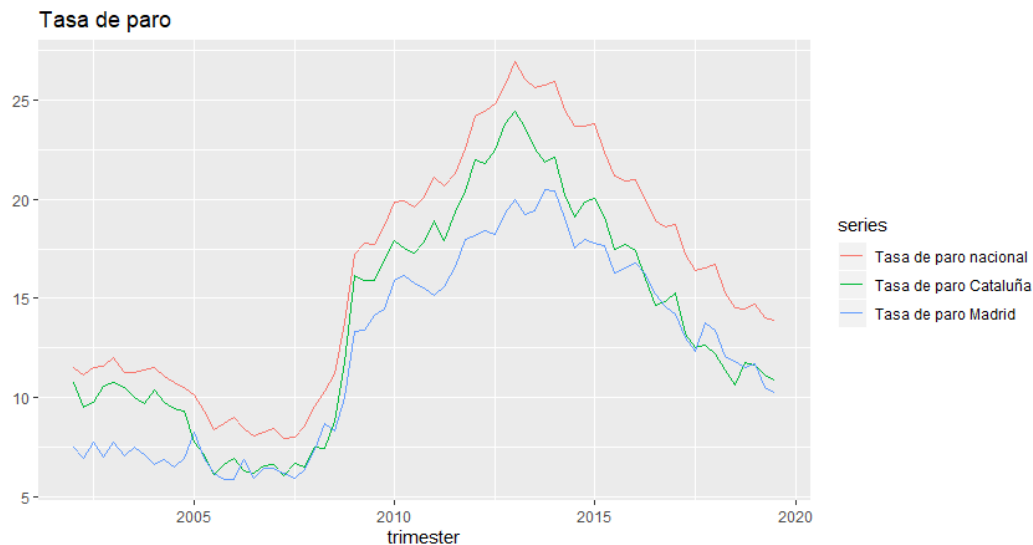
```
Paro_T <- ts(Paro[,-1], start=c(2002,1), frequency=4)
```

```
Paro_N <- ts(Paro[,2], start=c(2002,1), frequency=4)
```

```
Paro_C <- ts(Paro[,11], start=c(2002,1), frequency=4)
```

```
Paro_Ma<- ts(Paro[,15], start=c(2002,1), frequency=4)
```

```
cbind("Tasa de paro nacional" = Paro_N,  
      "Tasa de paro Cataluña" = Paro_C,  
      "Tasa de paro Madrid" = Paro_Ma) %>%  
  autoplot() + xlab("trimester") + ylab("") +  
  ggtitle("Tasa de paro")
```



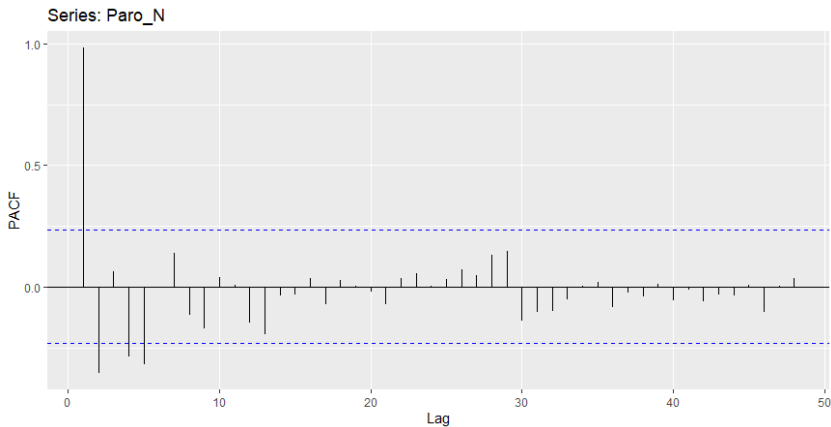
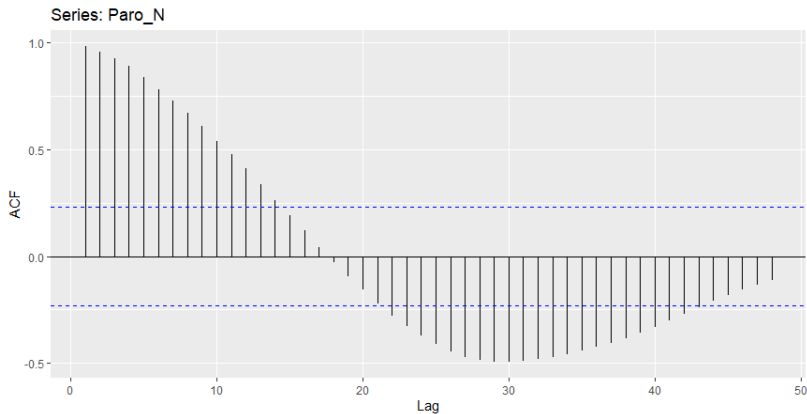
Para la serie Paro nacional comenzamos por calcular los autocorrelogramas

```
#Calculamos las autocorrelaciones simples hasta el retardo 48
```

```
ggAcf(Paro_N, lag=48)
```

```
#Calculamos las autocorrelaciones parciales hasta el retardo 48
```

```
ggPacf(Paro_N, lag=48)
```



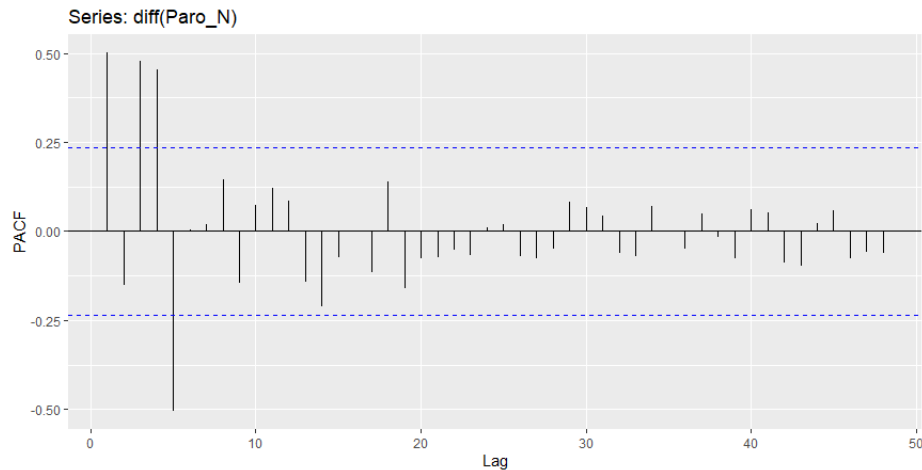
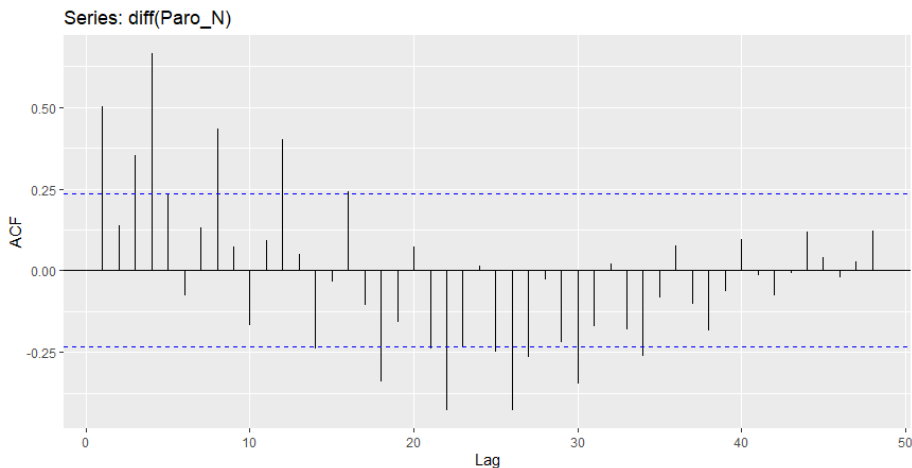
Puesto que en el gráfico de la serie hemos visto que la media no es constante porque la serie tiene tendencia y el ACF decrece de forma lenta es necesario hacer una diferenciación de orden 1.

```
#Calculamos las autocorrelaciones simples hasta el retardo 48
```

```
ggAcf(diff(Paro_N), lag=48)
```

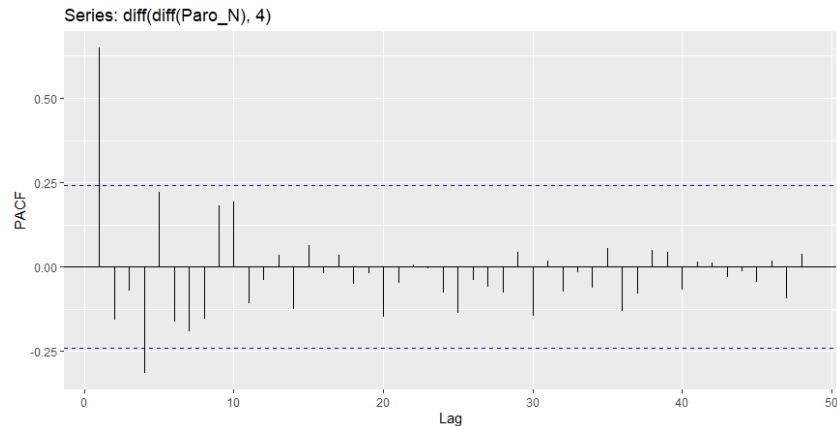
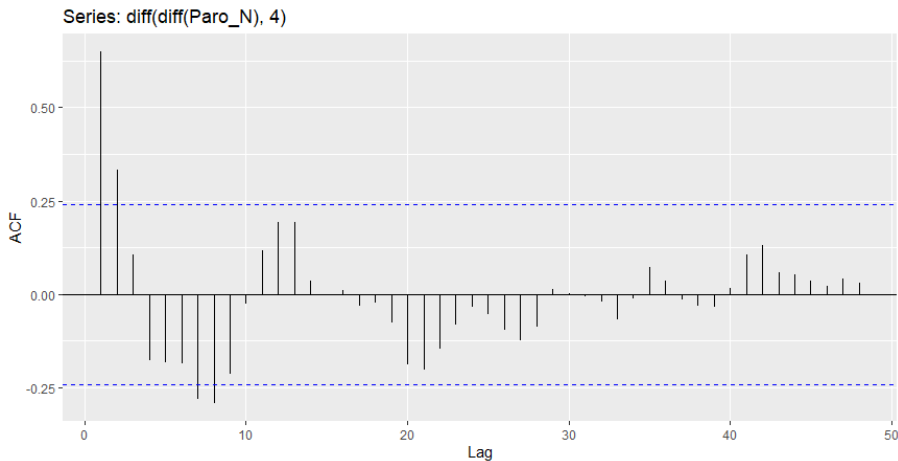
```
#Calculamos las autocorrelaciones parciales hasta el retardo 48
```

```
ggPacf(diff(Paro_N), lag=48)
```



Observamos que las autocorrelaciones ya decrecen de forma más rápida pero los retardos múltiplos de 4 siguen teniendo una correlación muy alta. Por esto, es necesario hacer una diferenciación de orden 4 sobre la de orden 1

```
ggAcf(diff(diff(Paro_N),4), lag=48)  
ggPacf(diff(diff(Paro_N),4), lag=48)
```



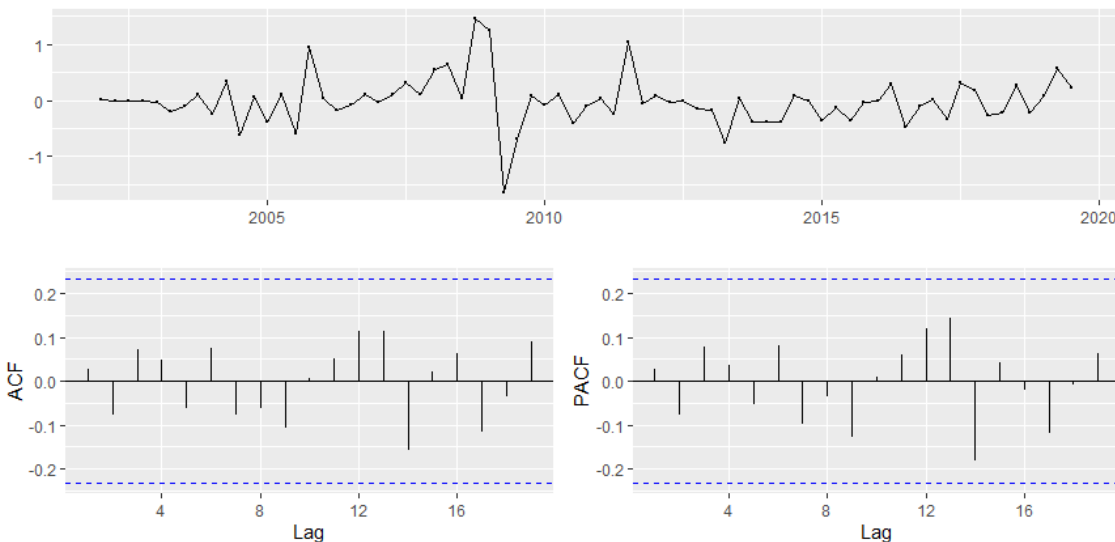
Con la serie doblemente diferenciada vemos, en el PACF, que la autocorrelación de orden 1 sigue siendo significativa y también la de orden 4. Por esto nuestro candidato a ajustar sería:  $ARIMA(1,1,0)(0,1,1)_4$ , o cualquiera de sus variaciones en las posiciones autoregresivas o de medias móviles.

#Ajuste manual

```
Paro_N %>% Arima(order=c(1,1,0), seasonal=c(0,1,1)) %>%
```

```
residuals() %>% ggtsdisplay()
```

Con esta sintaxis ajustamos el modelo ARIMA y directamente dibujamos el análisis de los residuos lo que nos permite comprobar de una forma rápida si el modelo ajustado es correcto. El resultado es el siguiente



Una función muy interesante es la función `auto.arima`, que encuentra el mejor modelo Arima ajustando todos los órdenes hasta que consigue que los residuos estén incorrelados

#Ajuste con la función `auto.arima`

```
fitparon1 <- auto.arima(Paro_N,seasonal=TRUE)
```

```
checkresiduals(fitparon1)
```

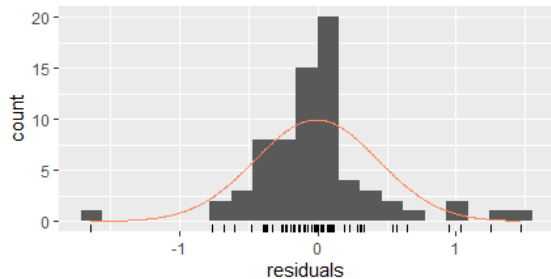
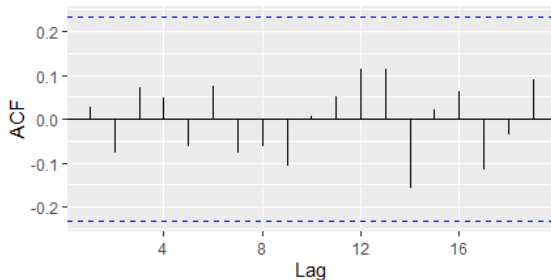
Ljung-Box test

data: Residuals from ARIMA(1,1,0)(0,1,1)

$Q^* = 2.5993$ ,  $df = 6$ ,  $p\text{-value} = 0.8572$

Model df: 2. Total lags used: 8

Residuals from ARIMA(1,1,0)(0,1,1)[4]



Los valores estimados para el modelo ARIMA ajustados los obtenemos con la función print:

```
print(fitparon1)
```

```
Series: Paro_N
```

```
ARIMA(1,1,0)(0,1,1)[4]
```

```
Coefficients:
```

	ar1	sma1
	0.7773	-0.7213
s.e.	0.0813	0.1333

→  $(1 - 0.7773B)(1 - B^4)(1 - B)X_t = (1 - 0.7213B^4)Z_t$

```
sigma^2 estimated as 0.2138: log like
```

```
likelihood=-43.35
```

```
AIC=92.7 AICc=93.08 BIC=99.26
```

$$(1-0.7773B)(1-B^4)(1-B)X_t = (1-0.7213B^4)Z_t$$

$$(1-0.7773B)(1-B^4)(X_t - X_{t-1}) = -0.7213Z_{t-4} + Z_t$$

$$(1-0.7773B)(X_t - X_{t-1} - X_{t-4} + X_{t-5}) = -0.7213Z_{t-4} + Z_t$$

$$X_t - X_{t-1} - X_{t-4} + X_{t-5} - 0.78X_{t-1} + 0.78X_{t-2} + 0.78X_{t-5} - 0.78X_{t-6} = -0.7213Z_{t-4} + Z_t$$

$$X_t - 1.78X_{t-1} + 0.78X_{t-2} - X_{t-4} + 1.78X_{t-5} - 0.78X_{t-6} = -0.7213Z_{t-4} + Z_t$$

$$X_t = 1.78X_{t-1} - 0.78X_{t-2} + X_{t-4} - 1.78X_{t-5} + 0.78X_{t-6} - 0.7213Z_{t-4} + Z_t$$

$$X_t = 1.78X_{t-1} - 0.78X_{t-2} + X_{t-4} - 1.78X_{t-5} + 0.78X_{t-6} - 0.7213Z_{t-4} + Z_t$$





### 3.5.- DIAGNOSIS DEL MODELO

El objetivo perseguido al ajustar un modelo ARIMA es encontrar un modelo adecuado para representar la serie objeto del estudio. Después de que se han estimado los parámetros del modelo, en la etapa de diagnosis tenemos que evaluar la adecuación del modelo, comprobando que se satisfacen las hipótesis del mismo.

#### 3.5.1.- Significación estadística de los parámetros

Para cada uno de los parámetros del modelo especificado debemos realizar el contraste de hipótesis

$$\begin{array}{ll} H_0 : \varphi = 0 & \frac{\hat{\varphi}}{S(\hat{\varphi})} \cong t_{T-k} \\ H_1 : \varphi \neq 0 & \end{array}$$

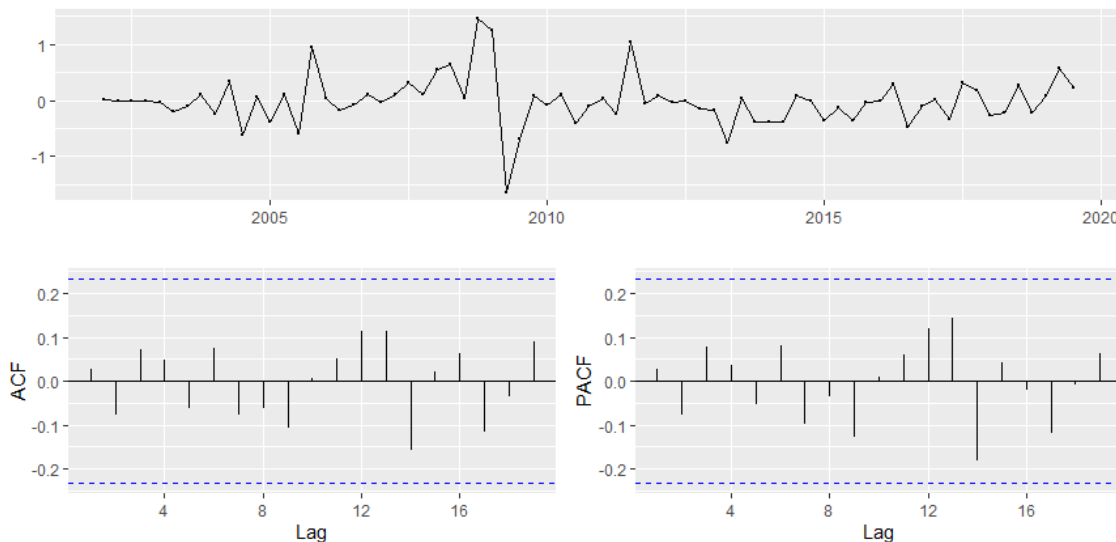
### 3.5.2. Análisis de los residuos.

Llamaremos residuos del modelo a las estimaciones obtenidas para el ruido:

$$\hat{Z}_t = -\hat{\theta}_1 \hat{Z}_{t-1} - \dots - \hat{\theta}_q \hat{Z}_{t-q} + (X_t - \hat{\mu}) - \hat{\phi}_1 (X_{t-1} - \hat{\mu}) - \dots - \hat{\phi}_p (X_{t-p} - \hat{\mu})$$

Estas se obtendrán de forma recursiva utilizando como valores iniciales generalmente 0 para los ruidos y la media de la variable para  $X_t$

```
checkresiduals(fitparon1)
```



De esta forma **los residuos así estimados deben tener media cero, varianza constante y ausencia de correlación para cualquier retardo.** También hay que comprobar que tienen distribución Normal, en este caso la falta de correlación implicaría independencia.

**El procedimiento habitual para comprobar la incorrelación de los residuos es representar su correlograma con las bandas de confianza.**

El primer contraste a realizar es si los residuos están incorrelados. Para ello se calculan su correlograma simple y parcial. Si los residuos son independientes los coeficientes de autocorrelación simples estimados se cumple que

$$\hat{r}_k \cong N\left(0, \sqrt{\frac{T-k}{T(T+2)}}\right)$$

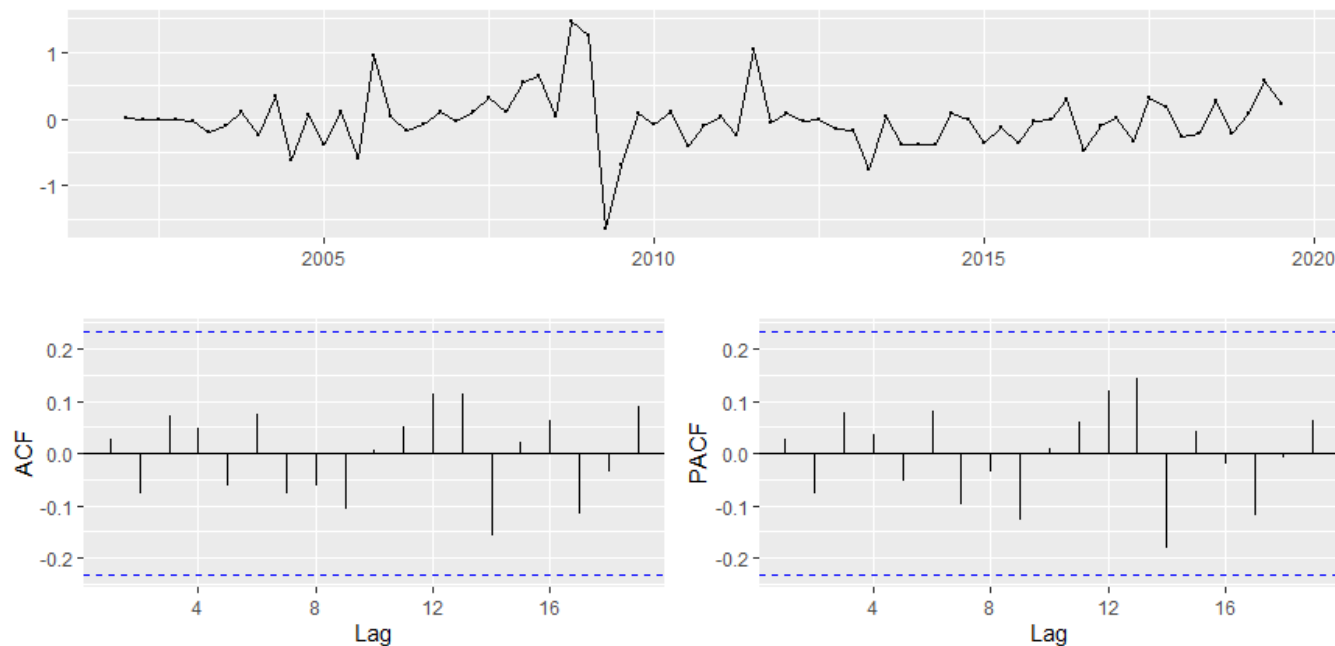
Y para k no muy pequeño se aproxima por

$$\hat{r}_k \cong N\left(0, \frac{1}{\sqrt{T}}\right)$$

Por esto, el procedimiento habitual de verificar la incorrelación de los residuos es dibujar dos líneas paralelas a distancia

$$2/\sqrt{T}$$

y comprobar si todos los coeficientes están dentro de los límites de confianza.



## El contraste Ljung-Box sobre las autocorrelaciones.

Un contraste global de que los primeros  $h$  coeficientes son cero es el contraste de Ljung-Box. Si los residuos son realmente ruido blanco sabemos que

$$\hat{r}_k \cong N\left(0, \sqrt{\frac{T-k}{T(T+2)}}\right)$$

Por tanto si tipificamos, elevamos al cuadrado y sumamos tenemos

$$Q(h) = T(T+2) \sum_{j=1}^h \frac{\hat{r}_j^2}{T-j} \cong \chi_{h-n}^2$$

Donde  $n$  es el número de parámetros estimados. Para modelos no estacionales

$$n = p + q + 1$$

Ljung-Box test

```
data: Residuals from ARIMA(1,1,0)(0,1,1)[4]
```

```
Q* = 2.5993, df = 6, p-value = 0.8572
```

```
Model df: 2. Total lags used: 8
```

### 3.5.3. Medidas de la adecuación del modelo.

Cuando ajustamos un modelo calcularemos las medidas de adecuación de su ajuste basadas en los errores del modelo para las  $T$  observaciones de que disponemos

$$\varepsilon_t = X_t - \hat{X}_t$$

El valor total de estos residuos se resume en diferentes estadísticos que presentamos a continuación y cuya interpretación en general es sencilla: buscamos el menor error total posible medido de diferentes formas.

El Error Absoluto Medio:

$$MAE = \sum_{t=1}^T \frac{|e_t|}{T}$$

La suma de cuadrados de los errores:

$$SSE = \sum_{t=1}^T e_t^2$$

Si dividimos por los grados de libertad de los errores tenemos la Media de los Errores al cuadrado:

$$MSE = \sum_{t=1}^T \frac{e_t^2}{T - k}$$

La desviación estándar del error también llamada raíz de la media de los errores al cuadrado.

$$RMSE = \sqrt{\sum_{t=1}^T \frac{\hat{e}_t^2}{T-k}}$$

En R podemos calcular todas estas medidas con la función `accuracy`.

```
# Accuracy
```

```
round(accuracy(fitparon1),3)
```

	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Training set	-0.009	0.439	0.283	0.153	1.969	0.142	0.028

El criterio de información de Akaike y el criterio bayesiano de Schwarz están basados en el logaritmo de la función de verosimilitud utilizada para calcular los estimadores de la serie bajo el supuesto de Normalidad de los residuos

$$AIC = -2\ln(L) + 2k$$

$$SBC(BIC) = -2\ln(L) + \ln(n)k$$

Donde  $L$  es la función de verosimilitud de la serie,  $k$  el número de parámetros y  $n$  el número de residuos calculados

```
print(fitparon1)
```

```
sigma^2 estimated as 0.2138: log likelihood=-43.35
```

```
AIC=92.7    AICc=93.08    BIC=99.26
```



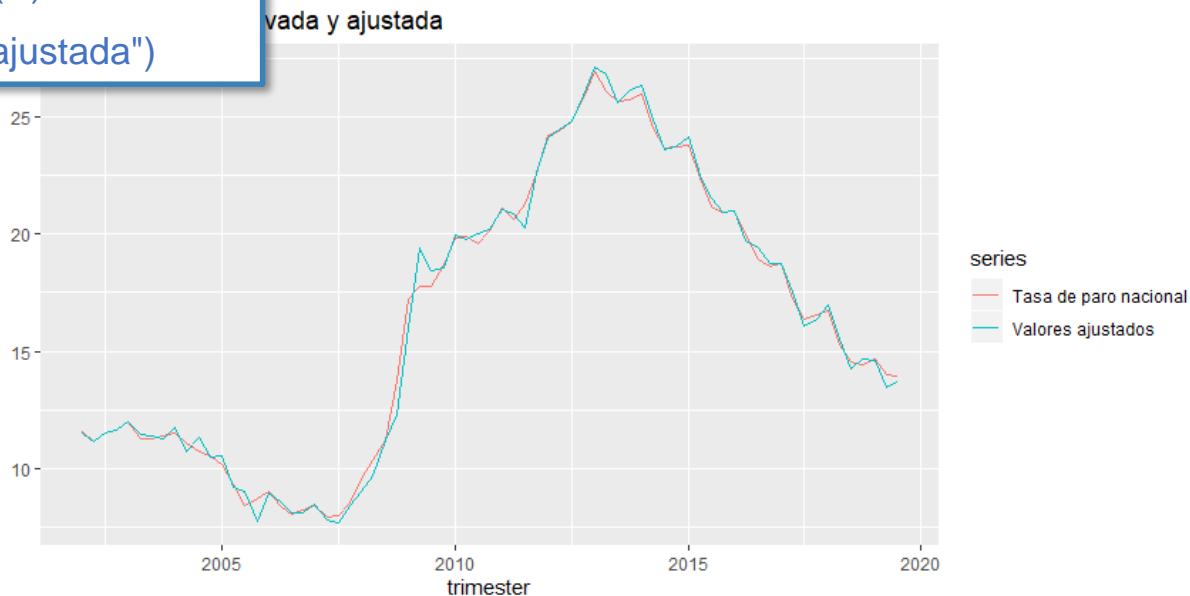
Resumimos a continuación algunas cuestiones sobre la interpretación y utilización de estos criterios:

- **El primer término de la definición del AIC es el que realmente mide el desajuste, su valor aumenta cuando peor es el ajuste;** mientras que el segundo, denominado de penalización, mide la complejidad del modelo a partir del número de parámetros.
- El AIC sigue el principio de parsimonia: Cuando el número de parámetros de un modelo  $k$  aumenta el AIC también, **por tanto escoger el modelo que tiene el mínimo AIC supone elegir el modelo con el menor número de parámetros posible.**
- Si el número de parámetros de un modelo  $k$  aumenta, el modelo gana complejidad y el término de penalización se incrementa, pero a la vez el desajuste disminuye, por tanto el valor final del AIC supone un equilibrio entre reducir la complejidad (principio de parsimonia) y mantener un valor mínimo de desajuste entre el modelo teórico y estimado.

## 2.10. CÁLCULO DE LAS PREDICCIONES

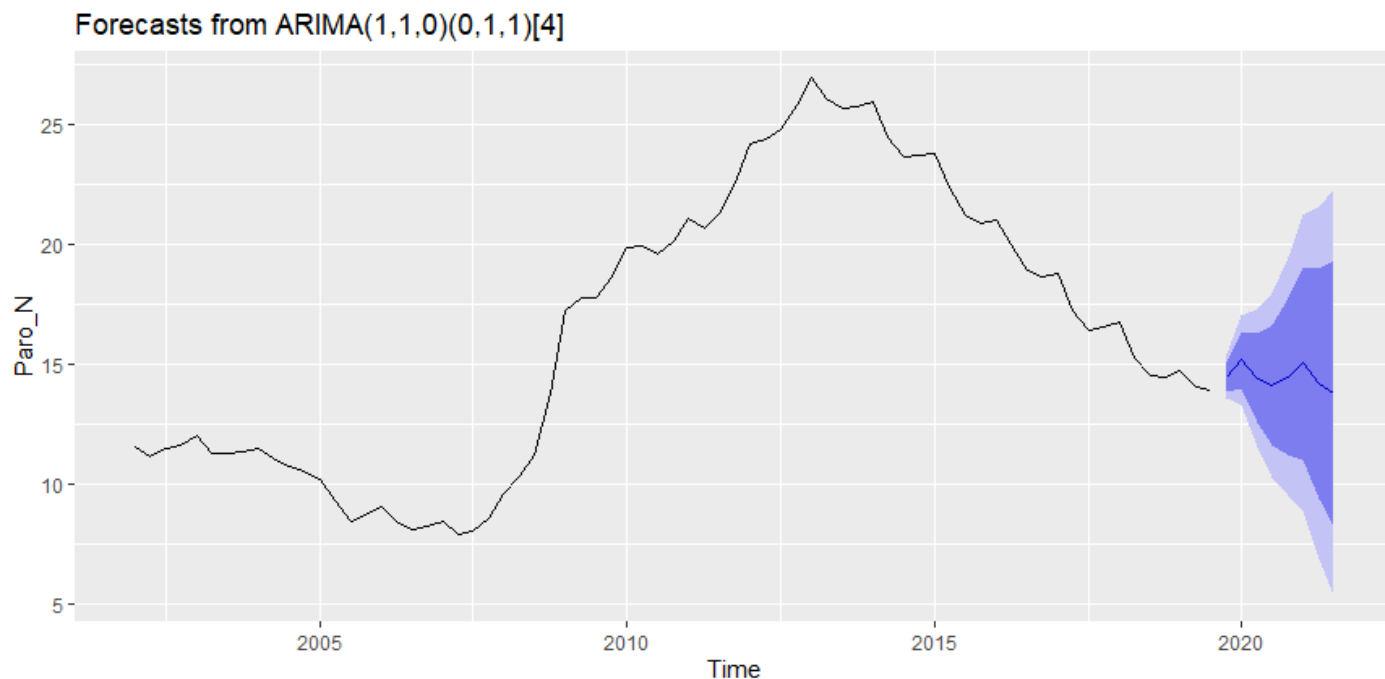
Para calcular las predicciones futuras es necesario previamente calcular los valores estimados mediante el modelo escogido para cada uno de los valores observados.

```
cbind("Tasa de paro nacional" = Paro_N,  
      "Valores ajustados" = fitted(fitparon1)) %>%  
autoplot() + xlab("trimester") + ylab("") +  
ggtitle("Tasa de paro observada y ajustada")
```



A partir de aquí podemos calcular las predicciones para los valores siguientes al último observado. El número de valores a predecir será indicado por  $h=$  , observemos que en el ejemplo  $h=1$  significa un periodo completo de una año.

```
autoplot(forecast(fitparon1),h=1)
```



## Bibliografía:

<https://otexts.com/fpp2/>

Avril Coghlan. A Little Book of R For Time Series.

Librería Forecast de R

