

# Segundo ejercicio de identificación de un modelo ARIMA

## Datos

Cargue la serie de datos simulados [00c296-12.gdt](#)

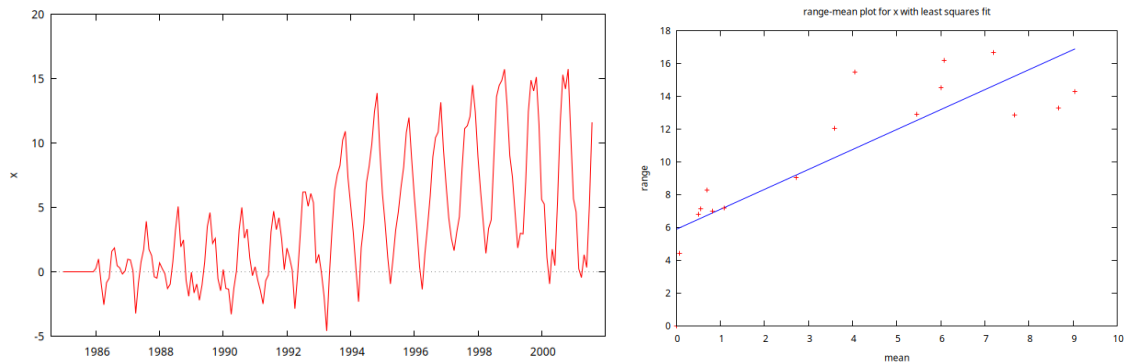
```
open ../datos/IdentificaEstosARIMA/00c296-12.gdt
```

## Tareas a realizar

1. Realice un primer análisis gráfico: haga un gráfico de la serie y un gráfico *rango-media*
2. Determine si es necesario transformar logarítmicamente los datos
3. Determine si es necesario tomar una diferencia estacional de la serie
4. Determine si es necesario tomar una o más diferencias regulares de la serie
5. Encuentre un modelo ARIMA para la serie que sea lo más parsimonioso posible, pero cuyos residuos se puedan considerar *ruido blanco*.
6. Ficheros
  - Versiones: [pdf](#); [html](#).
  - Datos: [00c296-12.gdt](#)
  - Guión de gretl: [P-L07-B-SegundoEjercicioIdentificacionARIMA.inp](#)

## Primer análisis gráfico

```
gnuplot x --time-series --with-lines --output="SerieEnNiveles.png"  
rmpplot x --output="rango-media.png"
```



De estos gráficos se desprende que la serie tiene una acusada pauta estacional y que la volatilidad probablemente depende del nivel de la serie.

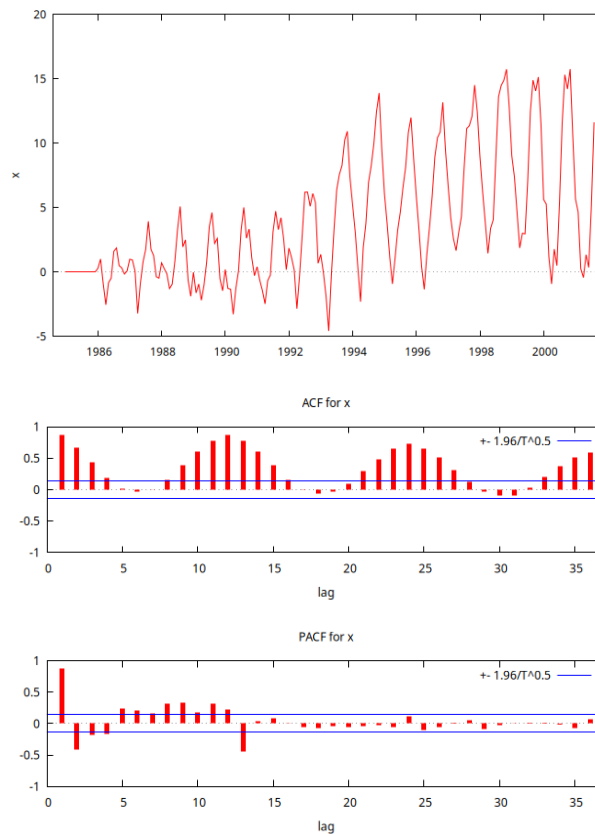
## Estacionariedad en varianza

A la luz de los anteriores gráficos, donde se aprecia que la variabilidad de los datos aumenta con el nivel de la serie, parece necesaria la transformación logarítmica; pero esta serie toma valores negativos, por lo que **no podemos transformar los datos logarítmicamente** (para hacerlo deberíamos sumar previamente un valor constante suficientemente elevado como para que todos los valores fueran positivos). Por el momento, dejemos la serie sin transformarla logarítmicamente.

## Diferencias estacionales

Observemos el gráfico de la serie y su correlograma.

```
corrgm x 36 --plot="x_ACF-PACF.png"
```



En el gráfico de la serie se aprecia una acusada pauta estacional. En la función de autocorrelación simple las correlaciones correspondientes a los retardos estacionales son muy significativas (y con bastantes “satélites”); en la función de autocorrelación parcial los 13 primeros retardos son muy significativos, en particular, el decimotercero (adyacente al 12) es muy importante.

Además, si tratamos de ajustar un AR(1) estacional:

```
ARIMA000X100 <- arima 0 0 0 ; 1 0 0 ; x
```

Function evaluations: 622

Evaluations of gradient: 123

ARIMA000X100:

ARMA, using observations 1985:01-2001:08 (T = 200)

Estimated using AS 197 (exact ML)  
 Dependent variable: x  
 Standard errors based on Hessian

	coefficient	std. error	z	p-value	
const	3.59533	1.28683	2.794	0.0052	***
Phi_1	0.955320	0.0148914	64.15	0.0000	***

Mean dependent var	3.780099	S.D. dependent var	4.838896
Mean of innovations	0.358041	S.D. of innovations	1.535258
R-squared	0.906552	Adjusted R-squared	0.906552
Log-likelihood	-384.1534	Akaike criterion	774.3069
Schwarz criterion	784.2019	Hannan-Quinn	778.3112

	Real	Imaginary	Modulus	Frequency
AR (seasonal)				
Root 1	1.0468	0.0000	1.0468	0.0000

ARIMA000X100 saved

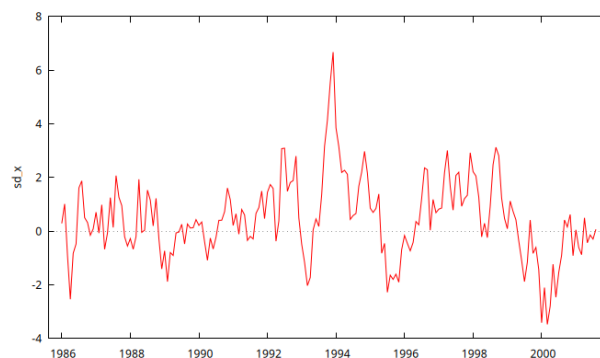
constatamos que la estimación del parámetro  $\Phi_1$  está muy próxima a uno.

**Estas evidencias apuntan a que es necesario tomar una diferencia estacional**

Recuerde que los test ADF y KPSS no sirven para determinar si es necesario tomar diferencias estacionales (solo sirven para las diferencias regulares).

Por tanto, tomamos una diferencia estacional.

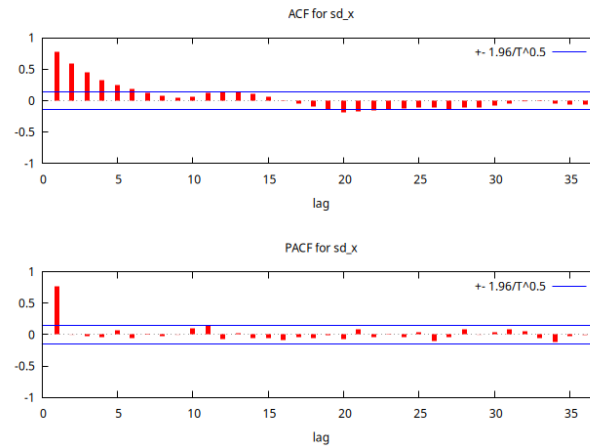
```
sdiff x
gnuplot sd_x --time-series --with-lines --output="SerieEnDiferencias.png"
```



## Repetición del análisis con la serie diferenciada estacionalmente

La serie resultante no muestra signos de estacionalidad. Veamos si se ve algo en el correlograma:

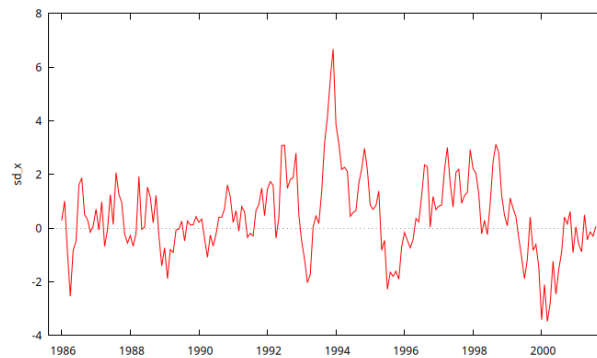
```
corrgram sd_x 36 --plot="sd_x_ACF-PACF.png"
```



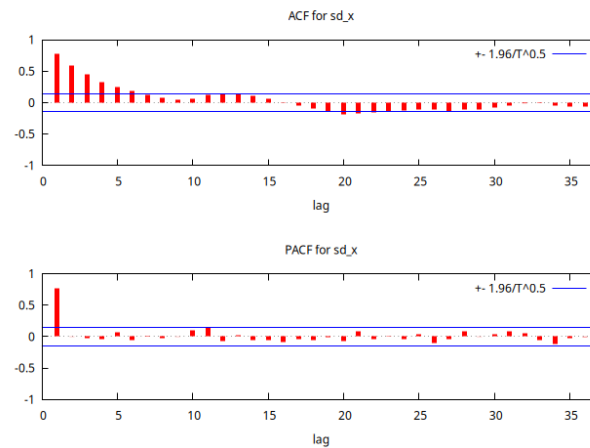
No hay nada que sugiera que persiste algún signo de estacionalidad.

## Estacionariedad en media

El gráfico de la serie diferenciada estacionalmente no muestra tener una clara tendencia o evolución a largo plazo de su nivel.



En el correlograma, la ACF decae rápidamente, indicando que la serie parece ser la realización de un proceso estacionario.



Probemos a ajustar un modelo AR a los datos diferenciados estacionalmente

```
ARIMA110 <- arima 1 1 0 ; x
```

Function evaluations: 20

Evaluations of gradient: 5

ARIMA110: ARIMA, using observations 1985:02-2001:08 (T = 199)

Estimated using AS 197 (exact ML)

Dependent variable: (1-L) x

Standard errors based on Hessian

	coefficient	std. error	z	p-value
const	0.0799826	0.255554	0.3130	0.7543
phi_1	0.405536	0.0659820	6.146	7.94e-10 ***

Mean dependent var	0.058438	S.D. dependent var	2.351529
Mean of innovations	0.000197	S.D. of innovations	2.149835
R-squared	0.824333	Adjusted R-squared	0.824333
Log-likelihood	-434.7714	Akaike criterion	875.5428
Schwarz criterion	885.4227	Hannan-Quinn	879.5415

	Real	Imaginary	Modulus	Frequency
AR				
Root 1	2.4659	0.0000	2.4659	0.0000

ARIMA110 saved

El parámetro  $\phi_1$  está muy lejos de la unidad (consecuentemente, también lo está la raíz autorregresiva). Probemos con los tests formales de raíz unitaria y estacionariedad

## Test ADF

```
adf -1 sd_x --c --glsl --test-down --perron-qu
```

Augmented Dickey-Fuller (GLS) test for sd\_x

testing down from 14 lags, criterion modified AIC, Perron-Qu

sample size 187

unit-root null hypothesis: a = 1

```
test with constant
including 0 lags of (1-L)sd_x
model: (1-L)y = b0 + (a-1)*y(-1) + e
estimated value of (a - 1): -0.225392
test statistic: tau = -4.85952
approximate p-value 0.000
1st-order autocorrelation coeff. for e: 0.002
```

El p-valores es muy bajo, por lo que se rechaza la  $H_0$  de que la serie es  $I(1)$

## Test KPSS

```
kpss -1 sd_x
```

KPSS test for sd\_x

T = 188

Lag truncation parameter = 4

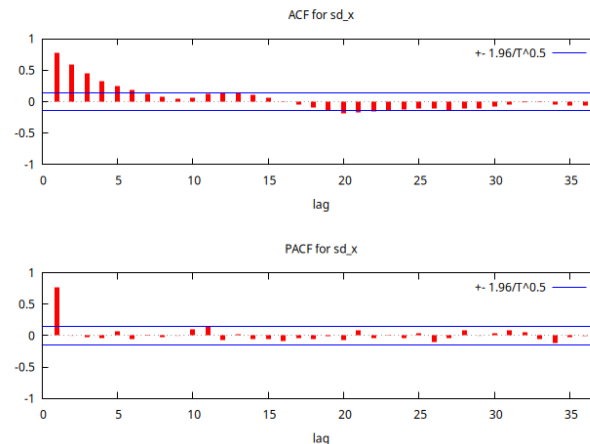
Test statistic = 0.22692

	10%	5%	1%
Critical values:	0.348	0.462	0.739
P-value >	.10		

El p-valor es elevado, por los que NO se rechaza la  $H_0$  de que la serie es  $I(0)$ . Todas estas evidencias indican de manera muy clara que NO es necesario tomar ninguna diferencia ordinaria.

## Primer intento de búsqueda de un modelo ARIMA

Observando al ACF y la PACF se aprecia que la ACF decae a una tasa exponencial, y la PACF se trunca tras el primer retardo, lo cual es compatible con un AR(1).



Por tanto, parece que la serie en logaritmos sigue un modelo ARIMA(1,1,0). Veamos si es así:

```
ARIMA110cte <- arima 1 0 0; 0 1 0 ; x
```

Function evaluations: 28

Evaluations of gradient: 5

ARIMA110cte:

ARIMA, using observations 1986:01-2001:08 (T = 188)

Estimated using AS 197 (exact ML)

Dependent variable: (1-Ls) x

Standard errors based on Hessian

	coefficient	std. error	z	p-value
const	0.440514	0.291954	1.509	0.1313
phi_1	0.768846	0.0458412	16.77	3.92e-63 ***

Mean dependent var	0.449786	S.D. dependent var	1.487227
Mean of innovations	0.000902	S.D. of innovations	0.941527
R-squared	0.962801	Adjusted R-squared	0.962801
Log-likelihood	-255.8802	Akaike criterion	517.7604
Schwarz criterion	527.4697	Hannan-Quinn	521.6942

		Real	Imaginary	Modulus	Frequency
-----					
AR					
Root	1	1.3007	0.0000	1.3007	0.0000
-----					

ARIMA110cte saved

Los parámetros autorregresivos son significativos y el modulo de las raíces es claramente mayor que la unidad en ambos casos. No obstante, la constante no es significativa.

Reestimemos el modelo sin constante:

```
ARIMA110 <- arima(1 0 0; 0 1 0; x --nc
```

Function evaluations: 16

Evaluations of gradient: 3

ARIMA110: ARIMA, using observations 1986:01-2001:08 (T = 188)

Estimated using AS 197 (exact ML)

Dependent variable: (1-Ls) x

Standard errors based on Hessian

	coefficient	std. error	z	p-value
-----				
phi_1	0.787833	0.0440563	17.88	1.62e-71 ***

Mean dependent var	0.449786	S.D. dependent var	1.487227
Mean of innovations	0.095115	S.D. of innovations	0.946555
R-squared	0.962771	Adjusted R-squared	0.962771
Log-likelihood	-256.9190	Akaike criterion	517.8381
Schwarz criterion	524.3110	Hannan-Quinn	520.4606

		Real	Imaginary	Modulus	Frequency
-----					
AR					
Root	1	1.2693	0.0000	1.2693	0.0000
-----					

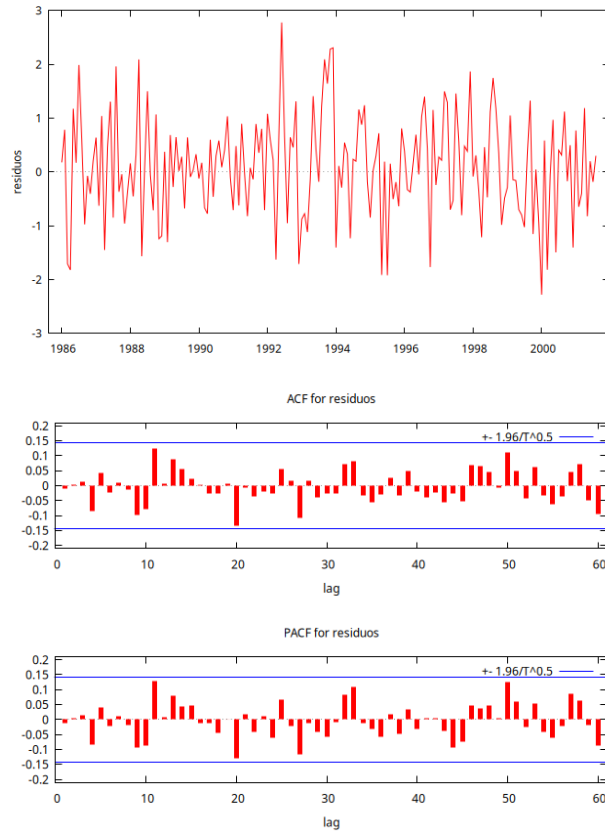
ARIMA110 saved

## Análisis de los residuos

Todo parece OK, pero debemos ver el gráfico de los residuos y su correlograma, así como los estadísticos Q de Ljung-Box para constatar si podemos asumir que son la realización de un proceso de ruido blanco.

```
series residuos = $uhat
```

```
gnuplot residuos --time-series --with-lines --output="Residuos.png"
corrgram residuos 60 --plot="residuosACF-PACF.png"
```



```
corrgm residuos 15
```

Autocorrelation function for residuos

\*\*\*, \*\*, \* indicate significance at the 1%, 5%, 10% levels

using standard error  $1/T^{0.5}$

LAG	ACF	PACF	Q-stat. [p-value]	
1	-0.0117	-0.0117	0.0260	[0.872]
2	0.0043	0.0042	0.0296	[0.985]
3	0.0141	0.0142	0.0677	[0.995]
4	-0.0841	-0.0838	1.4398	[0.837]
5	0.0410	0.0393	1.7673	[0.880]
6	-0.0229	-0.0218	1.8703	[0.931]
7	0.0091	0.0108	1.8867	[0.966]
8	-0.0119	-0.0199	1.9146	[0.984]
9	-0.0982	-0.0921	3.8399	[0.922]
10	-0.0798	-0.0882	5.1169	[0.883]
11	0.1237 *	0.1294 *	8.2065	[0.695]
12	0.0056	0.0074	8.2128	[0.768]
13	0.0892	0.0790	9.8375	[0.707]
14	0.0536	0.0451	10.4283	[0.730]
15	0.0221	0.0466	10.5290	[0.785]

El gráfico de los residuos no presenta ninguna estructura reconocible y ninguna autocorrelación es significativa.



Más importante aún, los correlogramas no muestran ninguna pauta reconocible, se parecen mucho entre sí y los estadísticos Q muestran p-valores muy elevados, por lo que podemos asumir que estos residuos son “ruido blanco”.

También conviene mirar si los residuos tienen distribución gaussiana:

```
normtest residuos --all
```

Test for normality of residuos:

Doornik-Hansen test = 0.0229723, with p-value 0.98858

Shapiro-Wilk W = 0.995253, with p-value 0.819881

Lilliefors test = 0.0417516, with p-value ~ = 0.58

Jarque-Bera test = 0.0946866, with p-value 0.95376

Podemos asumir claramente que tienen distribución normal.

Si en la ventana del modelo estimado pincha en el menú desplegable **Gráficos -->Espectro con respecto al periodograma espectral** verá que el espectro teórico del modelo se ajusta perfectamente al periodograma de la serie.

Por tanto, podemos concluir que la serie 00c296-12.gdt, no requiere la transformación logarítmica (en cualquier caso no se podía tomar sin aumentar previamente su nivel para hacerla positiva), sigue un proceso  $ARIMA(1, 0, 0) \times (0, 1, 0)_S$  con media cero.

## Modelo efectivamente simulado

Veamos si ese es el modelo usado en su simulación. Si miramos la línea 150 del fichero [000-Etiquetas-12.txt](#) que se encuentra en el directorio de donde hemos obtenido los datos encontramos lo siguiente:

```
00c296, , mu = 0.0, ar = '(1 - 0.8B)(1 + 0.8B)', ma = '(1 + 0.55B)', i = '(1 - B12)'
```

Efectivamente, NO requería la transformación logarítmica, la media era 0,0 y era necesaria una diferencia estacional, pero ninguna regular.

No obstante, el modelo simulado tenía un polinomio autorregresivo de de orden dos, AR(2), y un polinomio de media móvil de orden uno, MA(1). Veamos qué pasa si intentamos estimar el verdadero modelo simulado... **¡pues hemos identificado un modelo distinto del simulado!**

## Pruebas con otro modelo ARIMA

Estimemos el verdadero modelo simulado:  $ARIMA(2, 0, 1) \times (0, 1, 0)_S$ :

```
ARIMAsimulado <- arima(2 0 1; 0 1 0 ; x --nc
```

Function evaluations: 36

Evaluations of gradient: 15

ARIMAsimulado:

ARIMA, using observations 1986:01-2001:08 (T = 188)

Estimated using AS 197 (exact ML)

Dependent variable: (1-Ls) x

Standard errors based on Hessian

	coefficient	std. error	z	p-value
phi_1	0.666360	2.21538	0.3008	0.7636
phi_2	0.0987904	1.74518	0.05661	0.9549
theta_1	0.114904	2.21597	0.05185	0.9586
Mean dependent var	0.449786	S.D. dependent var	1.487227	
Mean of innovations	0.094410	S.D. of innovations	0.946522	
R-squared	0.962769	Adjusted R-squared	0.962366	
Log-likelihood	-256.9125	Akaike criterion	521.8250	
Schwarz criterion	534.7708	Hannan-Quinn	527.0701	
	Real	Imaginary	Modulus	Frequency
AR				
Root 1	1.2639	0.0000	1.2639	0.0000
Root 2	-8.0091	0.0000	8.0091	0.5000
MA				
Root 1	-8.7029	0.0000	8.7029	0.5000

ARIMAsimulado saved

El ajuste es parecido (fíjese en los coeficientes de determinación) pero solo el parámetro los parámetros no son significativos (aunque  $\phi_1$  toma un valor relativamente próximo al del modelo anterior). Por tanto...

La estimación del verdadero modelo empleado en la simulación de los datos ¡NO ES MEJOR QUE EL MODELO QUE HEMOS IDENTIFICADO!

La explicación es que en el modelo estimado, el factor  $(1 + 0,09879B)$  del polinomio AR casi se cancela con el polinomio MA  $(1 + 0,1149B)$ . Por eso hemos encontrado un modelo más parsimonioso y que funciona OK.

Ahora escoja al azar nuevas series del [directorio](#) (dispone de centenares de series simuladas con distintos modelos) y practique la identificación hasta que adquiera seguridad.