MODELIZACIÓN DE UNA SERIE TEMPORAL CON DATOS REALES

MODELOS ESTOCÁSTICOS

Irene León
Gabriele Villa
Isabel Villarejo

UNIVERSIDAD DE MÁLAGA

Curso 2023/2024

Índice

1. Introducción

Con este trabajo se pretende analizar la evolución de la producción de energías renovables en Estados Unidos a lo largo de los años.

Producir cada vez más energía renovable y abandonar las fuentes convencionales es una necesidad que comparten todos los países del mundo. Según los datos del último informe de la Agencia Internacional de Energías Renovables (IRENA, por sus siglas en inglés), en 2022 hasta el 83 % de toda la capacidad eléctrica añadida procedía de fuentes renovables. Mientras que en 2021, según un informe publicado por Ember –think tank independiente sobre el clima–, las renovables generaron el 38 % de la electricidad mundial.

Las grandes potencias mundiales como EEUU tienen mayor facilidad y oportunidad para la producción de energías renovables. Por ello, el objetivo de este trabajo es estudiar y analizar la producción de energías renovables a lo largo de los años. Para ello, hemos tomado los datos proporcionados por el EIA (Energy Information Administration) referentes a la producción mensual de energías renovables de los años desde 1973 hasta febrero de 2023.y hemos trabajado utilizando las herramientas de la teoría de Series Temporales mediante modelos ARIMA complementando con el uso de los softwares estadísticos RStudio y Excel.

2. Contexto

La energía es la base en el problema del cambio climático y también algo fundamental para su solución.

Una gran cantidad de los gases de efecto invernadero que cubren la Tierra y atrapan el calor del Sol se generan debido a la producción de energía, mediante la quema de combustibles fósiles con el objetivo de generar electricidad y calor.

Los combustibles fósiles, como el carbón, el petróleo y el gas, son con diferencia los mayores causantes del cambio climático global.

La ciencia lo indica claramente: para evitar los impactos más negativos del cambio climático, es necesario reducir las emisiones a casi la mitad en 2030 y alcanzar el cero neto en el año 2050. Para lograrlo, necesitamos dejar de depender de los combustibles fósiles e invertir en fuentes de energía alternativas que sean limpias, accesibles, asequibles, sostenibles y fiables

Las energías renovables son un tipo de energías derivadas de fuentes naturales que llegan a reponerse más rápido de lo que pueden consumirse. Un ejemplo de estas fuentes son, por ejemplo, la luz solar y el viento; estas fuentes se renuevan continuamente. Las fuentes de energía renovable abundan y las encontramos en cualquier entorno.Además, emiten pocos (o ninguno) contaminantes o gases de efecto invernadero en el aire.

Son la parte más importante de la transición hacia un sistema energético que abandone los combustibles fósiles, contrarrestando así el calentamiento global.

Los combustibles fósiles dan cuenta todavía de más del 80 % de la producción de energía en todo el mundo, aunque las fuentes de energía más limpias cada vez ganan más fuerza. Cerca del 29 % de la electricidad proviene actualmente de fuentes de energía renovables. Algunas ventajas más de aumentar la producción de energías renovables son:

- Reducen la contaminación y mejoran la calidad del aire
- No generan residuos difíciles de tratar
- Es un recurso endógeno, por lo que no es necesario importarla de otros lugares
- Combaten directamente el cambio climático al tener cero emisiones de CO2

3. Modelización de la serie

3.1. Metodología

METODOLOGÍA DE BOX-JENKINS

a) **Identificación:**

En primer lugar, veremos si nuestra serie es estacionaria y si presenta estacionalidad, lo cual necesita ser modelado.

Si nuestra serie es estacional(se observa algún tipo de patrón), buscaríamos en la FAC el periodo (ayudándonos de rejillas si fuese necesario). Una vez encontrado el periodo, propondríamos un modelo para modelizar esta parte de la serie, donde hay que tener en cuenta los residuos multiplos del periodo que se salen de las bandas de confianza de la FAC. Aunque también hay que tener presente la FACP, para la modelización de la parte regular de la serie no será tan importante como el comportamiento de los residuos en la FAC.

Si nuestra serie no es estacionaria (por ejemplo, si presenta una clara tendencia creciente no es estacionaria), trataríamos de diferenciar la serie hasta alcanzarla.

Una vez solucionada la estacionariedad y estacionalidad, procederíamos a proponer un modelo SARIMA(p,d,q)(P,D,Q)[s].

b) **Estimación de los parámetros:**

La salida de *Arima()* incluye los coeficientes ajustados y el error estándar (s.e.) para cada coeficiente. Observando los coeficientes podemos excluir los insignificantes. Podemos usar una función *coeftest()* para este propósito.

c) **Validación del modelo:**

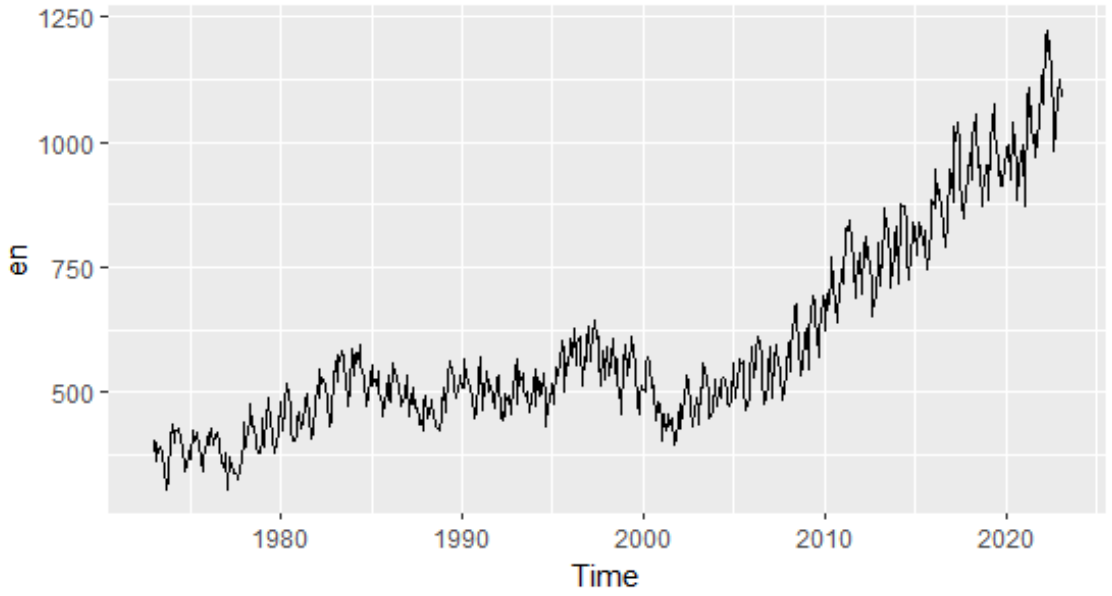
Tenemos que comprobar que los residuos estimados se comportan como un ruido blanco gaussiano. La serie residual debe ser incorrelada, que tengan distribución normal (lo podemos estudiar a través de gráficos Q-Q) y varianza constante (homocedasticidad).

d) **Pronóstico:**

Los parámetros del modelo SARIMA se pueden usar como modelo predictivo para hacer pronósticos de valores futuros de la serie una vez que se selecciona el modelo más adecuado para los datos de la serie temporal.

3.2. Identificación

Cargamos los datos en RStudio y graficamos la serie.



Observamos una tendencia creciente en la gráfica de la serie. La media va cambiando, por tanto no es estacionaria. Aun así realizamos los test de estacionariedad:

```
> adf.test(en)

        Augmented Dickey-Fuller Test

data:  en
Dickey-Fuller = -1.2606, Lag order = 8, p-value = 0.8913
alternative hypothesis: stationary
```

Como el p-valor es 0.8913 >0.05, aceptamos H_0. Esto es, la serie no es estacionaria.

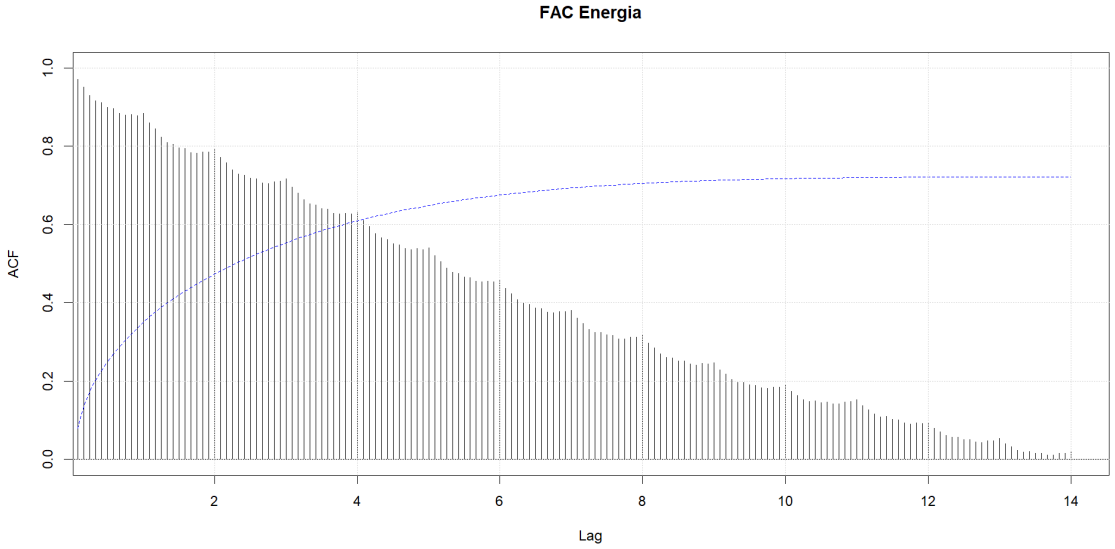
```
> kpss.test(en)

        KPSS Test for Level Stationarity

data:  en
KPSS Level = 6.4511, Truncation lag parameter = 6, p-value = 0.01
```

Como el p-valor es 0.01 <0.05, rechazamos H_0. Esto es, la serie no es estacionaria.

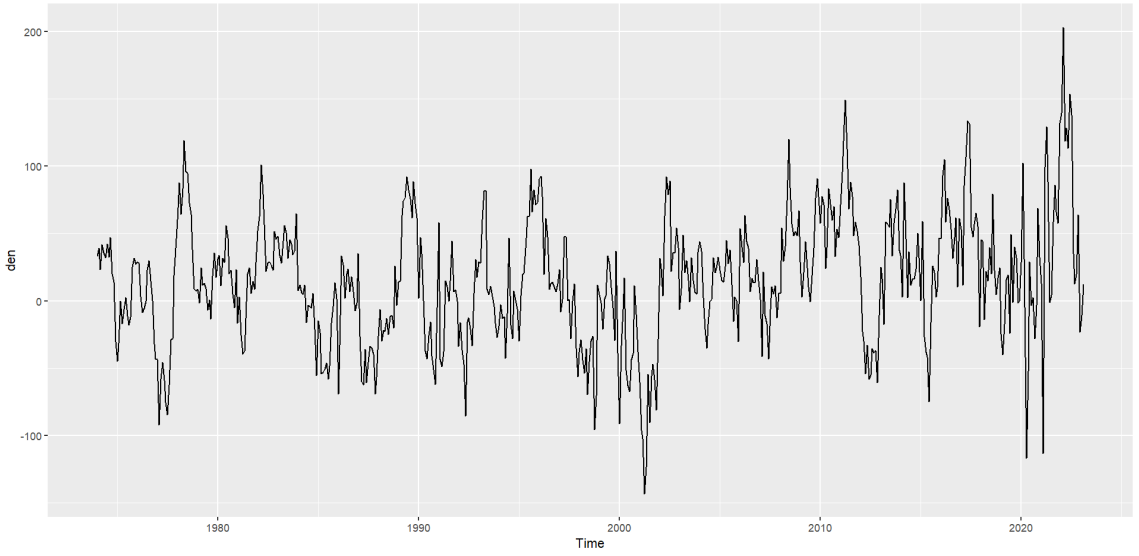
Graficamos a continuación su función de autocorrelación (FAC) y su función de autocorrelación parcial (FACP) y en la FAC se observa estacionalidad cada 12. Esto tiene sentido pues la producción de energía renovable depende en gran medida del clima.
Incluimos en la gráfica una rejilla en la que se observe en el eje x los múltiplos de 12.

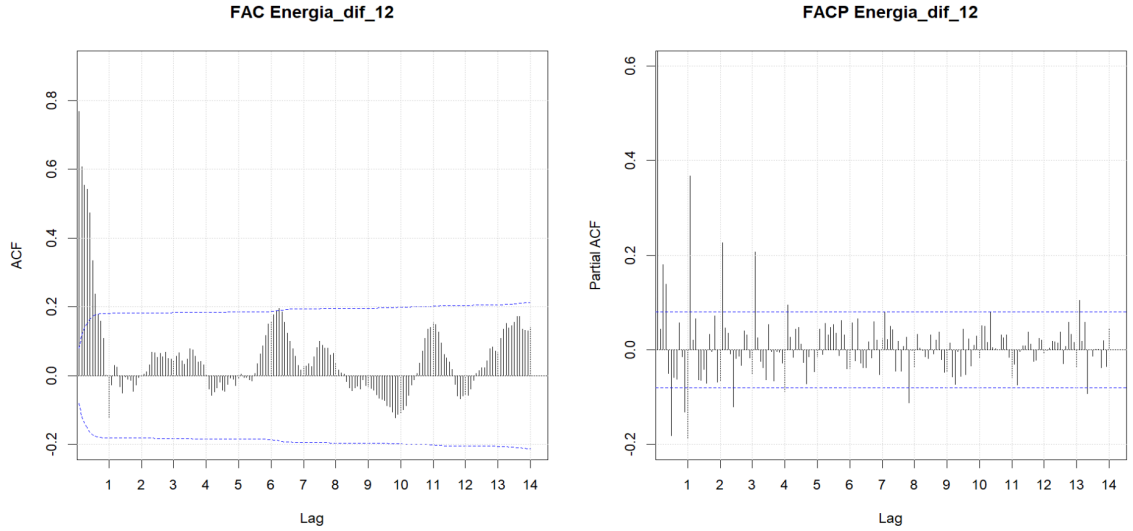


Es claro, por tanto, que hay estacionalidad (periodo de 12 meses).
Así, comencemos modelizando la **parte estacional** de la serie:

MODELIZACIÓN DE LA PARTE ESTACIONAL

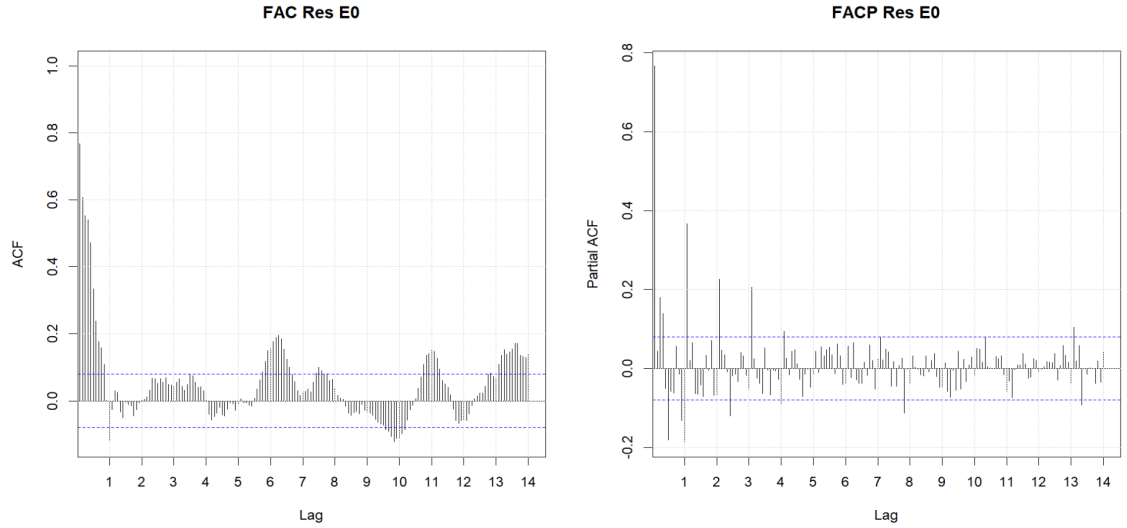
En el gáfico anterior observamos que los 4 primeros retardos múltiplos de 12 en la FAC se salen de las bandas de confianza, por lo que hay persistencia. Esto nos indica también que NO hay estacionariedad.
Ante la falta de estacionariedad, diferenciamos la serie y graficamos la FAC y la FACP de la diferenciación.





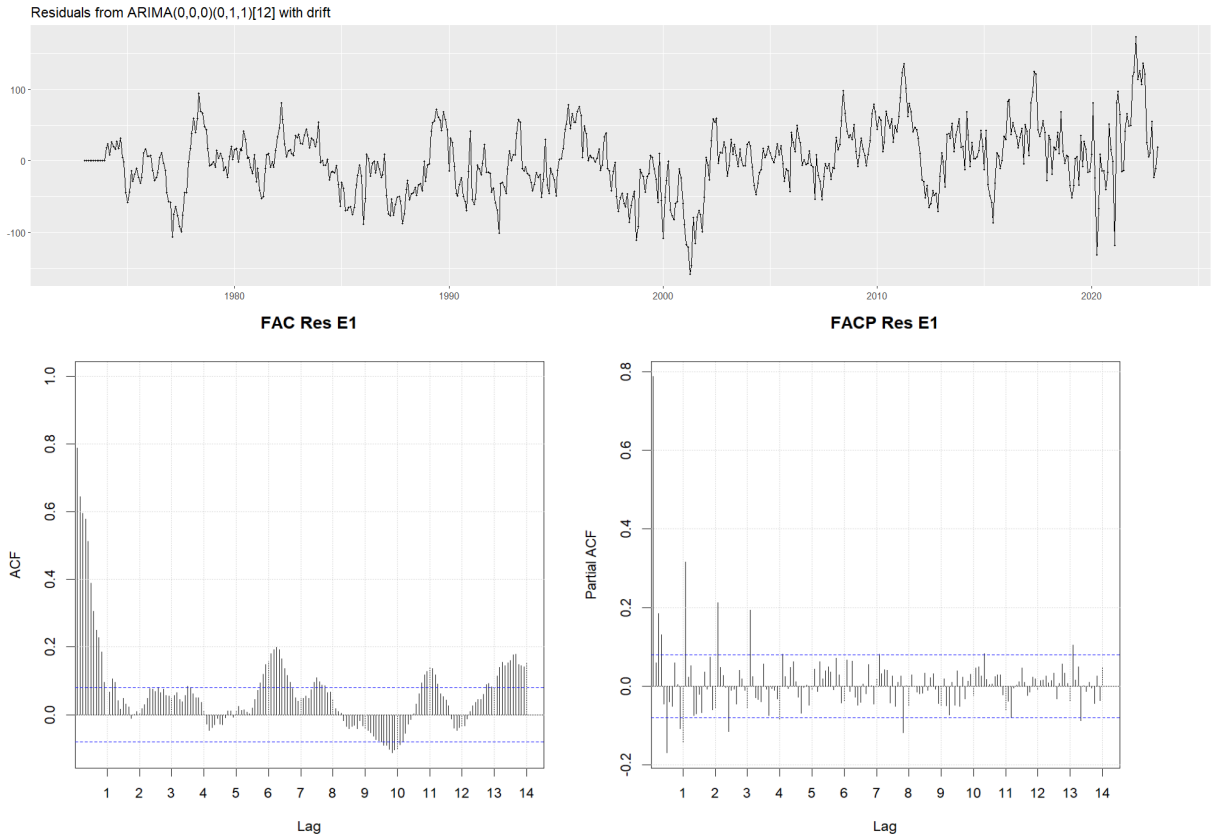
Ha desaparecido la persistencia.
En la FAC todos los residuos parecen estar dentro de las bandas de confianza y en la FACP se sale el primer retardo.

Proponemos como primer modelo: ARIMA(0,0,0)(0,1,0)[12]



Al variar la forma de las bandas de confianza en la FAC, se sale el primer residuo. Al producirse un corte en FAC y decaimiento en FACP, aumentamos el orden de medias móviles a espera de mejorar el modelo propuesto anteriormente.

Por tanto, proponemos modelizarla con medias móviles de orden uno. Esto es, con un modelo ARIMA(0,0,0)(0,1,1)[12]

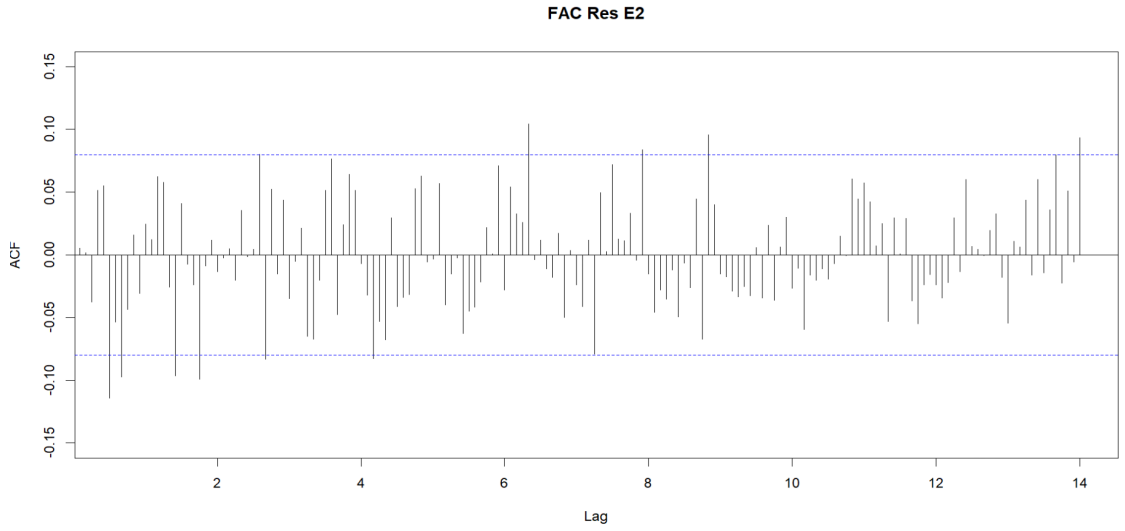


Observamos en la gráfica anterior, que ninguno de los primeros residuos en la FAC se salen de las bandas de confianza. Numéricamente comprobamos que se salen los residuos 6,10,11 y 14. Por lo que podemos aceptar este modelo, en cuanto aún queda por modelizar la parte regular.
Notamos que en la FAC se sale el primer retardo, pero ningún otro modelo consigue corregir esto. Por tanto, como para la modelización de la parte estacional no es realmente importante la FACP de los residuos, aceptaremos esta modelización para la parte estacional de la serie.

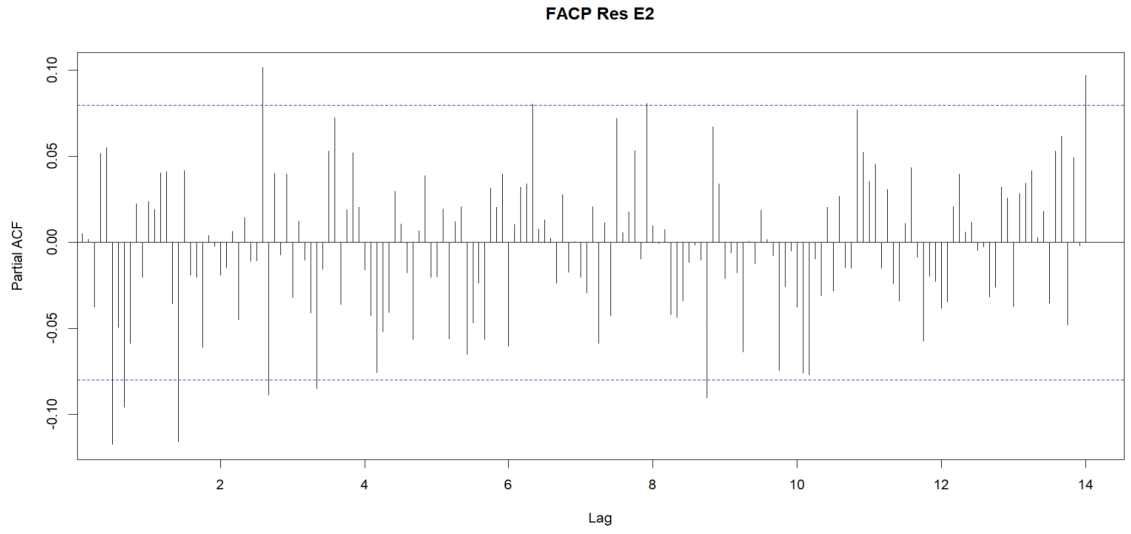
Modelemos ahora la **parte regular**:

MODELIZACIÓN DE LA PARTE REGULAR

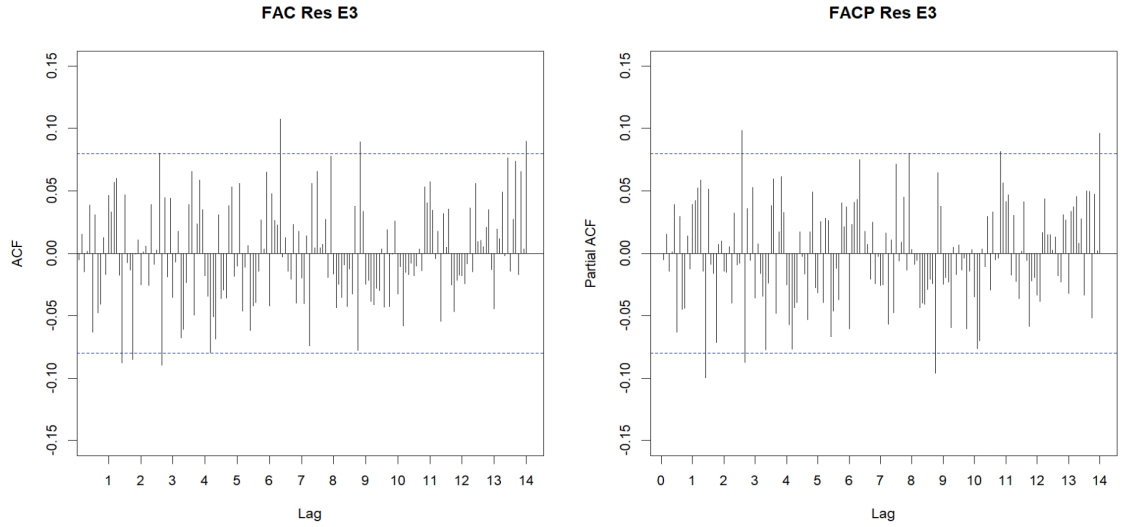
Se percibe que aun podría haber problemas en cuanto a estacionariedad, ya que desde el año 2000 la serie (última serie graficada que aparece en la modelización de la parte estacional) no está centrada en 0 y la varianza va cambiando, por lo que volvemos a diferenciar, esta vez en la parte regular. Además por el corte en la FAC y el decaimiento en la FACP (de la gráfica anterior), propusimos como modelo para la parte regular (0,1,1), pero se salen los primeros residuos en FAC y FACP. Por tanto, decidimos aumentar el orden de medias móviles proponiendo como modelo para la parte regular (0,1,2).



Aunque se salen algunos residuos en la FAC, no se consideraría en principio, un problema para rechazar dicho modelo (el primer retardo que sale es el 6). Al igual que en la FACP.



Sin embargo, continuamos probando modelar la serie de otro modo para conseguir un mejor modelo.
El modelo ARIMA(3,1,3)(0,1,1)[12] tiene mejores FAC y FACP, ya que los residuos que se salen de las bandas de confianza de ambas funciones, son más tardíos.



El primer retardo que sale en la FAC es el 17, por lo que parece que esta es una mejor modelización para nuestra serie.

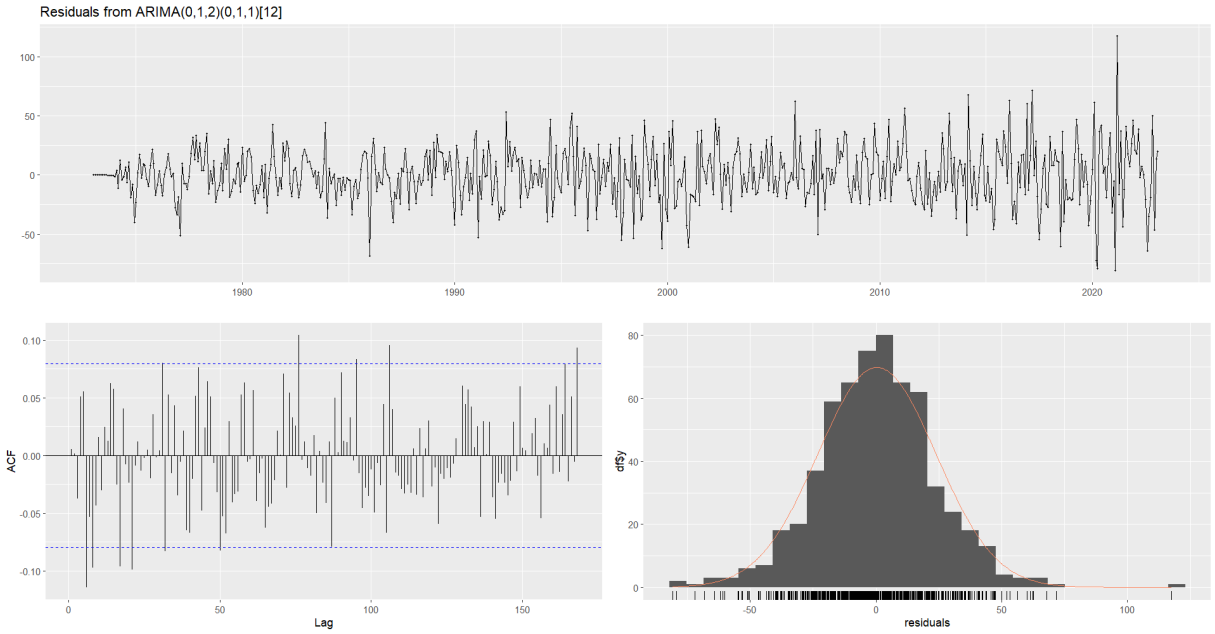
No obstante, al comprobar si hay problemas de invertibilidad en este modelo, vemos que efectivamente los hay:

```
> Phi<-e3$coef[-length(e3$coef)]
> abs(polyroot(c(1,-Phi+qnorm(0.025)*coeftest(e3)[1:3,2]*c(1,1,1))))
[1] 0.6568837 1.0150967 1.0150967 1.0727278 1.0727278 3.7403611
> abs(polyroot(c(1,-Phi+qnorm(0.025)*coeftest(e3)[1:3,2]*c(1,1,-1))))
[1] 0.809383 1.006271 1.006271 1.185037 1.185037 1.283963
> abs(polyroot(c(1,-Phi+qnorm(0.025)*coeftest(e3)[1:3,2]*c(1,-1,1))))
[1] 0.8495067 1.1677647 1.1677647 0.8495067 1.7205838 1.7205838
> abs(polyroot(c(1,-Phi+qnorm(0.025)*coeftest(e3)[1:3,2]*c(1,1,1))))
[1] 0.6568837 1.0150967 1.0150967 1.0727278 1.0727278 3.7403611
```

Por tanto, rechazamos el modelo anterior y volvemos al modelo ARIMA(0,1,2)(0,1,1)[12]. Como ya hemos comentado se salen algunos residuos de las bandas de confianza de las FAC y FACP, pero no se consideran un problema por no ser de las primeros.
Comprobemos que en este caso no hay porblemas de invertibilidad:

```
> Theta<- e2$coef[-length(e2$coef)]
> abs(polyroot(c(1,Theta+qnorm(0.025)*coeftest(e2)[1:2,2]*c(1,1))))
[1] 1.394177 2.605188
> abs(polyroot(c(1,Theta+qnorm(0.025)*coeftest(e2)[1:2,2]*c(1,-1))))
[1] 1.833895 4.719921
> abs(polyroot(c(1,Theta+qnorm(0.025)*coeftest(e2)[1:2,2]*c(-1,1))))
[1] 1.613011 2.251748
> abs(polyroot(c(1,Theta+qnorm(0.025)*coeftest(e2)[1:2,2]*c(-1,-1))))
[1] 2.277830 3.800037
```

No hay problemas de invertibilidad. De este modo proponemos como modelo para pasar a las siguientes fases del estudio del mismo: ARIMA(0,1,2)(0,1,1)[12].



3.3. Estimación de los parámetros

En este apartado utilizaremos los comandos *Arima()* y *coeftest()*. En la parte estacional obtenemos,

```
Series: en
ARIMA(0,0,0)(0,1,1)[12] with drift

Coefficients:
         sma1   drift
      -0.1268  1.2561
s.e.   0.0413  0.1393

sigma^2 = 2152:  log likelihood = -3100.2
AIC=6206.4   AICc=6206.44   BIC=6219.54
```

```
> coeftest(e1)

z test of coefficients:

       Estimate Std. Error z value  Pr(>|z|)
sma1  -0.126769   0.041266 -3.0720  0.002127 **
drift  1.256093   0.139259  9.0199 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

donde los contrates son significativos.

En la parte regular obtenemos,

```
Series: en
ARIMA(0,1,2)(0,1,1)[12]

Coefficients:
          ma1      ma2     sma1
      -0.2546  -0.1954  -0.8263
s.e.   0.0402   0.0408   0.0273

sigma^2 = 566.3:  log likelihood = -2708.11
AIC=5424.22   AICc=5424.29   BIC=5441.74
```

```
> coeftest(e2)

z test of coefficients:

      Estimate Std. Error  z value  Pr(>|z|)
ma1  -0.254639   0.040195  -6.3351 2.371e-10 ***
ma2  -0.195426   0.040765  -4.7940 1.635e-06 ***
sma1 -0.826318   0.027319 -30.2468 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

donde los contrates significativos. Por lo que, hasta ahora el modelo es bueno.

3.4. Validación

Comprobemos que los residuos se comportan como un ruido blanco gaussiano:

- **Normalidad:** Pasamos el test de normalidad.

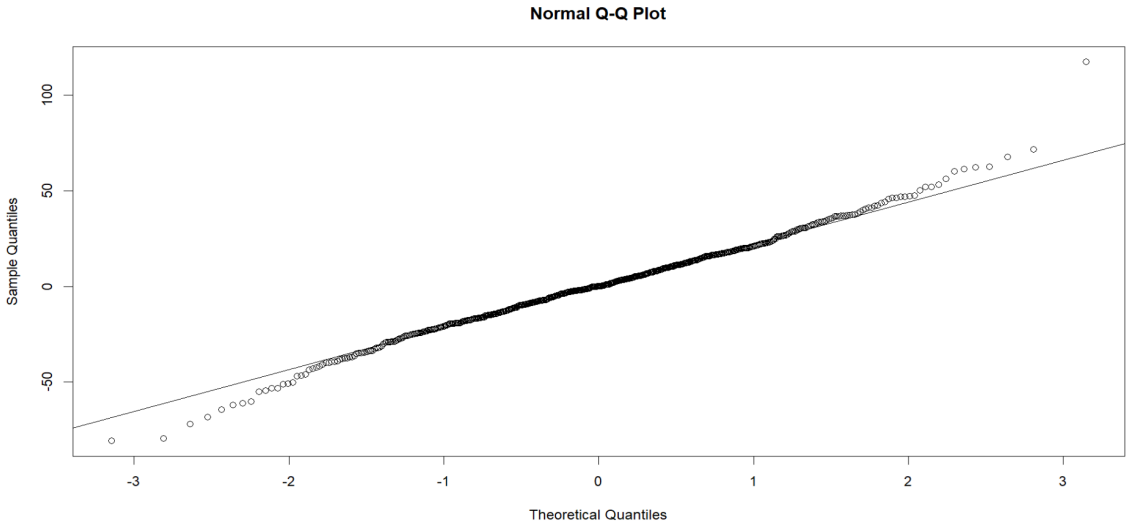
```
> ks.test(e2$residuals,'pnorm', mean(e2$residuals),sd(e2$residuals))

        Asymptotic one-sample Kolmogorov-Smirnov test

data:  e2$residuals
D = 0.035509, p-value = 0.4336
alternative hypothesis: two-sided
```

Como en este test H_0 es que los residuos siguen una distribucion normal, como el p-valor>0.05, aceptamos que esos residuos tengan una distribucion normal.

También observando el gráfico Q-Q, deducimos que hay normalidad en lo residuos.



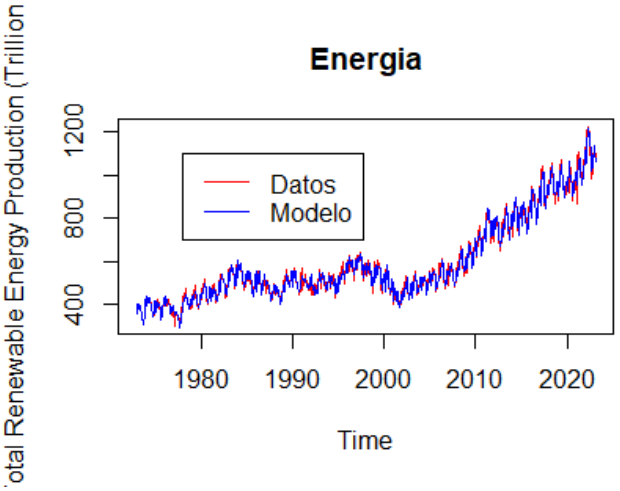
- **Correlación**:

```
> e2$var.coef
                ma1           ma2           sma1
ma1   1.615617e-03 -3.769193e-04 -4.032511e-05
ma2  -3.769193e-04  1.661753e-03  2.581753e-05
sma1 -4.032511e-05  2.581753e-05  7.463356e-04
```

Como todas las covarianzas son cercanas a cero, podemos afirmar que no hay correlación en los residuos.

3.5. Predicción

En este punto, hemos intentado prever cuál será la producción de energía renovable desde marzo hasta diciembre de 2023 (el conjunto de datos terminaba en febrero). Para hacerlo, primero representamos el gráfico de los datos (en rojo) y del modelo (en azul):



Posteriormente, con el comando *forecast(e2,h=10)*, hemos calculado los valores previstos desde marzo hasta diciembre, junto con sus respectivos intervalos de confianza:

```
> forecast(e2,h=10)
         Point Forecast     Lo 80    Hi 80     Lo 95    Hi 95
Mar 2023       1191.898 1161.4009 1222.396 1145.2564 1238.540
Apr 2023       1168.089 1130.0516 1206.126 1109.9159 1226.262
May 2023       1211.166 1169.5956 1252.737 1147.5894 1274.743
Jun 2023       1179.787 1134.9606 1224.613 1111.2310 1248.343
Jul 2023       1144.102 1096.2407 1191.963 1070.9045 1217.299
Aug 2023       1104.551 1053.8362 1155.266 1026.9895 1182.112
Sep 2023       1050.990  997.5741 1104.406  969.2974 1132.683
Oct 2023       1089.918 1033.9312 1145.905 1004.2934 1175.543
Nov 2023       1116.785 1058.3396 1175.230 1027.4006 1206.169
Dec 2023       1153.817 1093.0132 1214.621 1060.8254 1246.809
```

Finalmente, representamos gráficamente el pronóstico junto con su respectivo intervalo de confianza:

```
> ajuste ← fitted(e2)
> prde2←forecast(ajuste,h=10) #prediccion hasta dic 2023
> plot(prde2)
```