Estudio de series temporales:

Consumo bruto interior de energía eléctrica mensual en España.

Arnau Turch Ferreres y Elías Abad Rocamora, 2020

UPC

1. Introducción

En este informe se estudia la evolución del consumo de electricidad en españa desde enero del año 1985 hasta julio del año 2019.

Para llevar a cabo este estudio utilizaremos la metodología Box-Jenkins explicada en clase: En un primer lugar realizaremos una identificación de la serie mediante distintas técnicas como por ejemplo el análisis de su ACF y PACF.

En segundo lugar, una vez tengamos los posibles modelos identificados anteriormente, procederemos a realizar una estimación de dichos modelos utilizando la función “arima” que nos proporciona R.

Seguidamente tendremos que evaluar los modelos estimados; nos centraremos principalmente en su capacidad predictiva y cómo se comportan sus residuos.

Una vez evaluados todos los modelos tendremos que elegir el que mejor se comporte, y con el ya seremos capaces de realizar predicciones a largo plazo.

Por último, aplicaremos un análisis de valores atípicos al modelo elegido y intentaremos interpretar las causas de estos. Además compararemos el comportamiento del modelo calculado a partir de la serie linealizada con el ya calculado previamente.

2. Identificación

2.1. Transformaciones

Para poder generar modelos de predicción para una serie temporal, necesitaremos que la serie sea estacionaria. Esto significa que se tienen que cumplir una serie de propiedades estadísticas:

- Media constante
- Homogeneidad de varianzas (Homocedasticidad)
- Estructura de autocovarianza constante (sólo depende de la distancia entre muestras)

Para conseguir estas propiedades y poder ajustar modelos más precisos, realizaremos las transformaciones que sean necesarias. Primeramente haremos un plot de la serie.

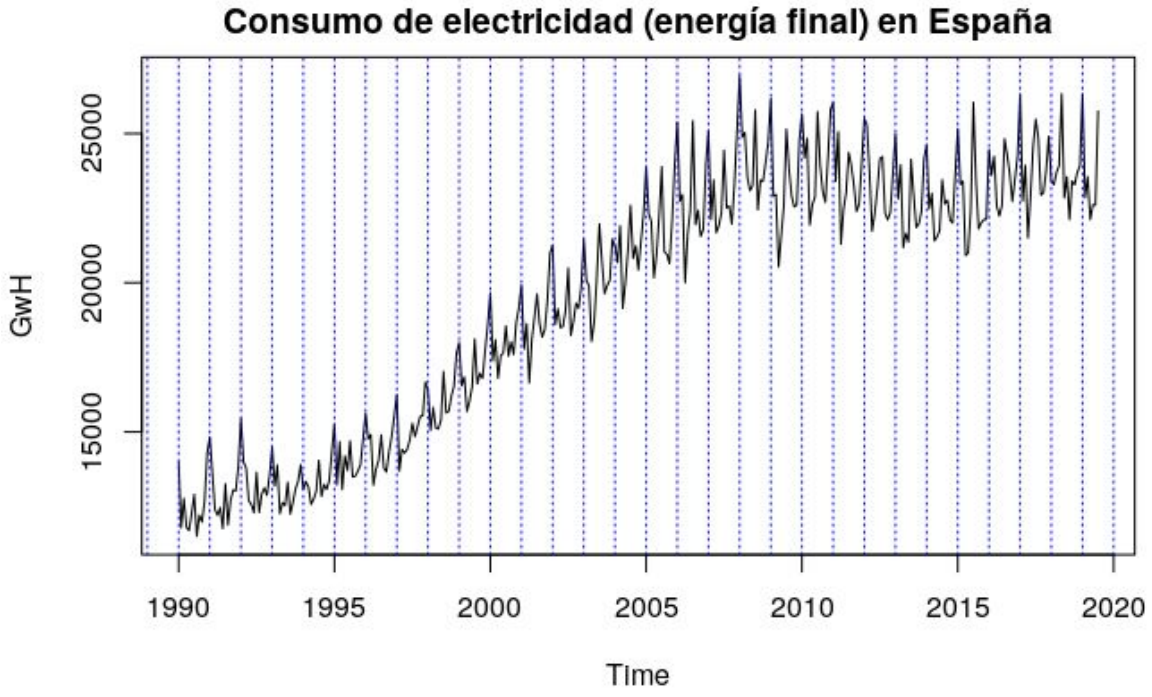

Figura 1. Consumo de la electricidad en España.

Vemos que la serie original es claramente no estacionaria, a simple vista se ve que la media no es constante y la varianza tampoco. Primero intentaremos corregir la diferencia de varianzas:

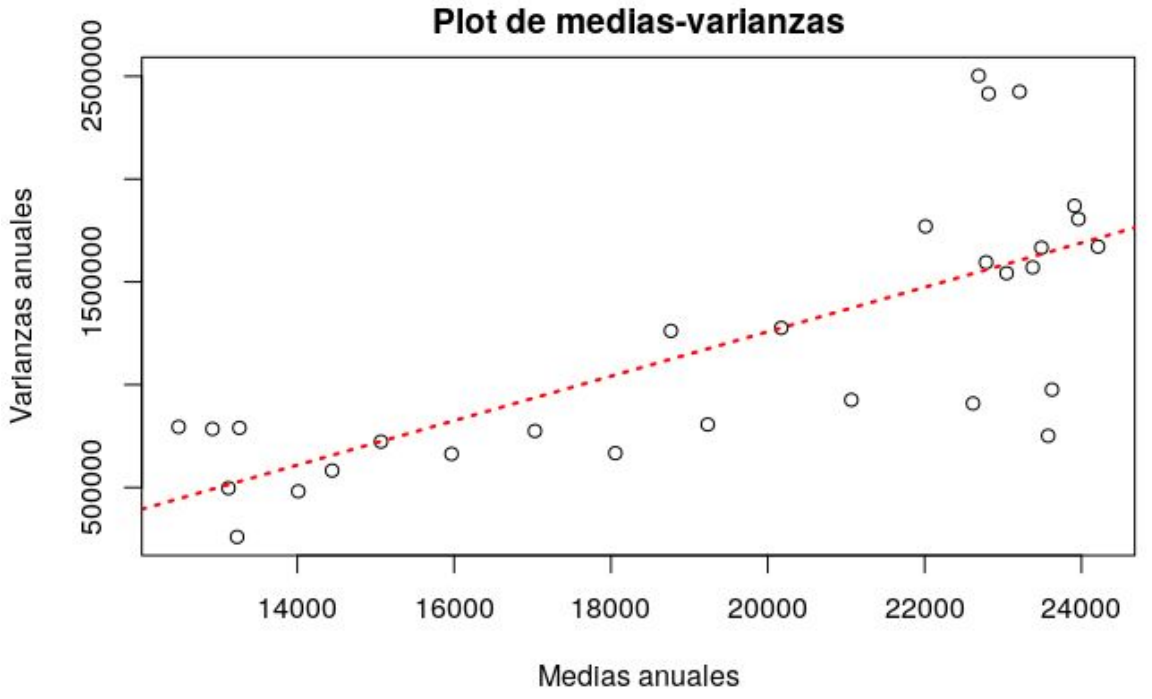

Figura 2. Plot de medias-varianzas de la serie.

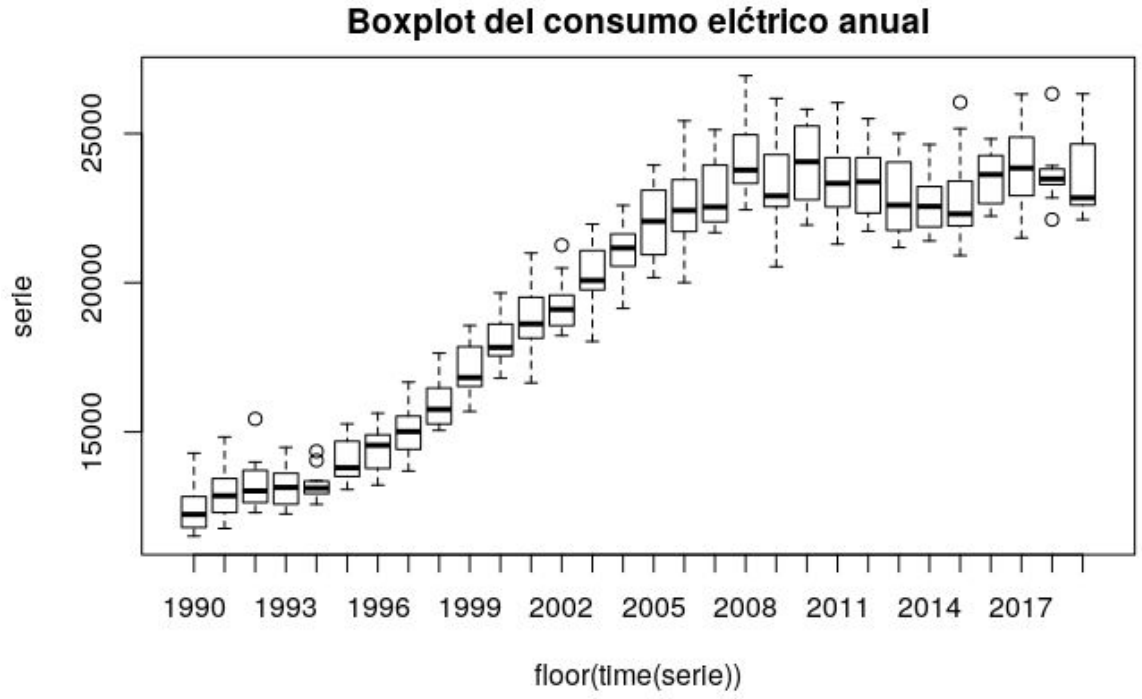

Figura 3. Boxplot por años del consumo eléctrico.

Tanto en el plot de medias varianzas cómo en el boxplot anual de la serie, podemos observar que la varianza aumenta a medida que lo hace la media de la serie. La transformación adecuada para conseguir homocedasticidad sería el logaritmo.

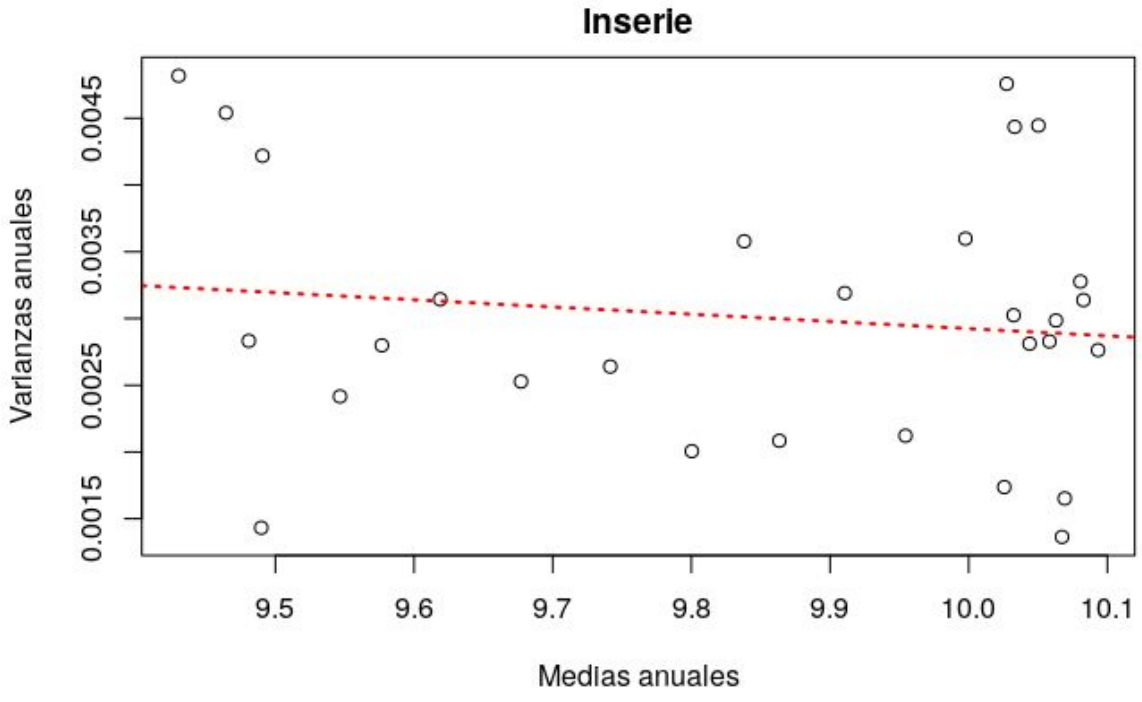

Figura 4. Plot de medias-varianzas para la serie transformada.

Ahora la varianza de la serie es prácticamente independiente de la media y estamos más cerca de que la serie sea estacionaria.

El siguiente paso que debemos dar es comprobar si la serie sigue un cierto patrón estacional, y si es el caso, eliminarlo aplicando una diferenciación estacional.

$$W_t = (1 - B^s)ln(X_t) \quad \textbf{Fórmula 1.}$$

Hacemos un ”monthplot” para observar si sucede dicho fenómeno:

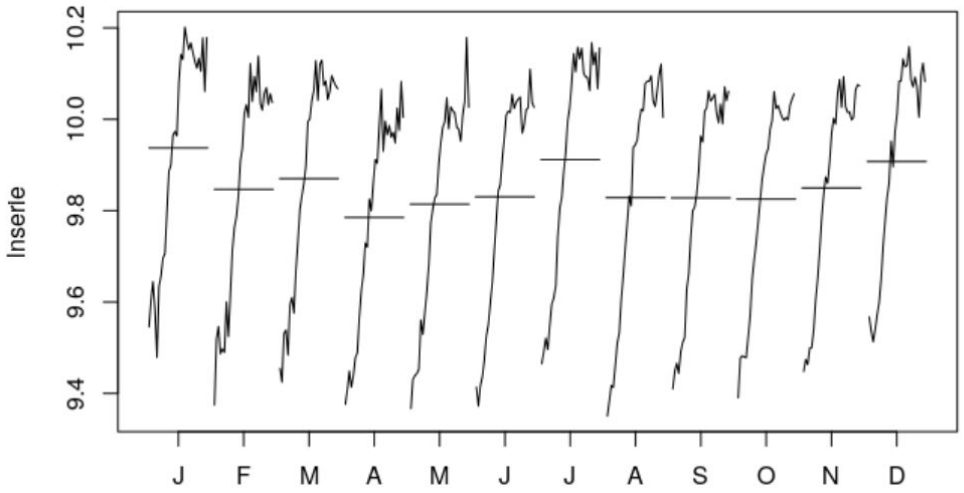

Figura 5. Monthplot

Efectivamente se observa un patrón estacional, la media para los diferentes meses no es constante. Este patrón estacional tendrá que ser eliminado utilizando la fórmula anterior con el parámetro $s = 12$ para obtener una serie estacionaria.

En este plot se muestra la serie $(1 - B^{12})ln(X_t)$, la componente estacional estaría ya eliminada pero la serie aún no tiene una media constante.

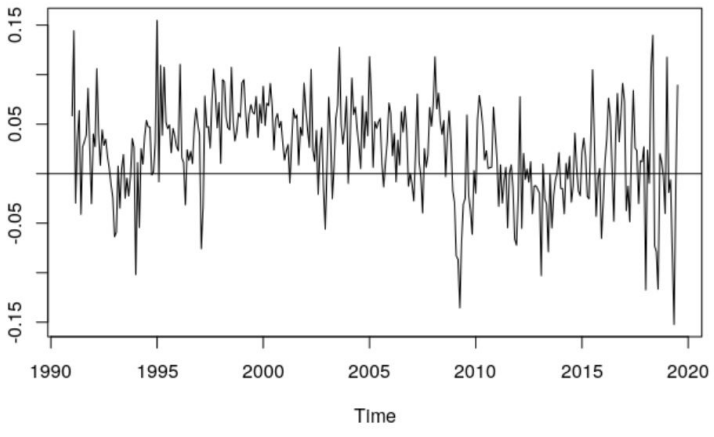

Figura 6. Serie logarítmica diferenciada estacionalmente

Para conseguir la estacionariedad completa, realizaremos las diferenciaciones regulares que sean necesarias. Para realizar estas diferenciaciones volveremos a recurrir a la fórmula 1 utilizando como exponente de B (backshift operator), $s = 1$. Así que la serie tras una diferenciación regular sería:

$$(1 - B)(1 - B^{12})ln(X_t)$$

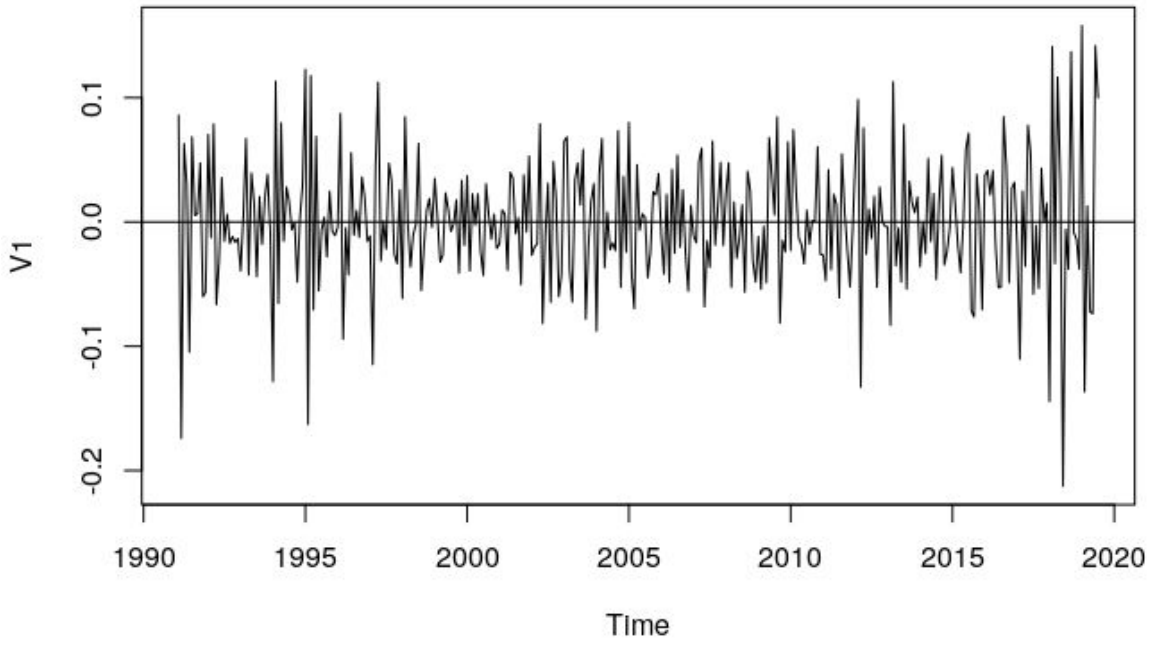

Figura 7. Serie logarítmica diferenciada estacionalmente y regularmente

A simple vista podríamos decir que esta serie ya es estacionaria, pero para ser más rigurosos podemos parar de diferenciar cuándo la varianza de la serie aumente.

```
Varianza d1d12lnserie:
 0.002688037

Varianza d1d1d12lnserie:
 0.007301351
```

Cómo la varianza después de realizar dos diferenciaciones regulares es mayor, trabajaremos de aquí en adelante con la serie diferenciada una vez regularmente:

$$W_t = (1 - B)(1 - B^{12})ln(X_t)\,.$$

2.2. Identificación de posibles modelos

Una vez tenemos la serie estacionaria es el momento de identificar los modelos. Para ello nos ayudaremos de la ACF y de la PACF de la serie.

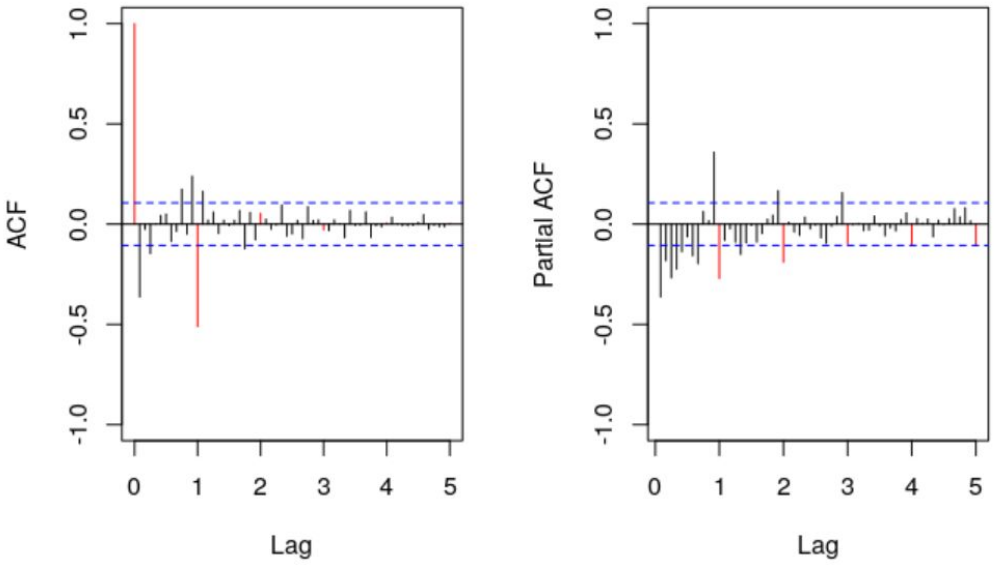

Figura 8. ACF y PACF

En rojo podemos observar los lags múltiplos de s=12, que nos servirán para identificar la parte estacional del modelo.

La parte regular del modelo puede ser asociada a un MA(3), ya que el último lag significativo del ACF se encuentra en la posición 3, o un AR(8), porque el último lag significativo de la PACF está en la posición 8.

Por lo que respecta la parte estacional, podemos considerar que se trata de un MAestacional(1), debido al único lag significativo de la ACF (múltiplo de 12) que se encuentra en la primera posición, o bien un ARestacional(2), ya que el último lag significativo es el segundo pintado de rojo.

Con todo esta información podemos identificar los modelos posibles:

- ARIMA(0,1,3)$(0,1,1)_{12}$
- ARIMA(8,1,0)$(0,1,1)_{12}$
- ARIMA(0,1,3)$(2,1,0)_{12}$
- ARIMA(8,1,0)$(2,1,0)_{12}$

3. Estimación

Para realizar la estimación utilizaremos los modelos que tiene una componente MA(1) estacional:

- ***mod1*** = ARIMA(0,1,3)$(0,1,1)_{12}$
- ***mod2*** = ARIMA(8,1,0)$(0,1,1)_{12}$

Esta decisión se basa en que el patrón MA(1) estacional es el más claro debido a la forma de la ACF y PACF. El último *lag* estacional significativo de la ACF es el primero y en la PACF se observa un claro patrón descendente infinito. Este comportamiento es característico de un MA(1).

```
Call:
arima(x = lnserie, order = c(0, 1, 3), seasonal = list(order = c(0, 1, 1), period = 12))

Coefficients:
          ma1      ma2      ma3     sma1
      -0.5413  -0.1503  -0.0869  -0.8264
s.e.   0.0565   0.0618   0.0602   0.0402

sigma^2 estimated as 0.001064:  log likelihood = 677.89,  aic = -1345.78
```

Figura 9. Estimación de *mod1*

El tercer parámetro del modelo (ma3) no es significativo por lo que utilizaremos la siguiente modificación del modelo.

```
Call:
arima(x = lnserie, order = c(0, 1, 3), seasonal = list(order = c(0, 1, 1), period = 12),
    fixed = c(NA, NA, 0, NA))

Coefficients:
          ma1      ma2  ma3     sma1
      -0.5709  -0.2024    0  -0.8247
s.e.   0.0514   0.0469    0   0.0402

sigma^2 estimated as 0.001071:  log likelihood = 676.87,  aic = -1345.73
```

Figura 10. Estimación de *mod1* con ma3=0

```
Call:
arima(x = lnserie, order = c(8, 1, 0), seasonal = list(order = c(0, 1, 1), period = 12))

Coefficients:
          ar1      ar2      ar3      ar4      ar5      ar6      ar7      ar8     sma1
      -0.5659  -0.4782  -0.5195  -0.4648  -0.3120  -0.1609  -0.1838  -0.1231  -0.8228
s.e.   0.0555   0.0626   0.0670   0.0703   0.0706   0.0669   0.0620   0.0549   0.0413

sigma^2 estimated as 0.001022:  log likelihood = 684.72,  aic = -1349.43
```

Figura 11. Estimación de *mod2*

Se observa que todos los coeficientes son significativos.

4. Validación

4.1. Análisis de residuos

Para poder validar estadísticamente un modelo, necesitamos que los residuos cumplan las siguientes condiciones:

- Homogeneidad de varianzas
- Distribución $N(0, \sigma^2)$
- Independencia

A continuación realizaremos la comprobación de estas hipótesis mediante varios plots y resultados de tests:

4.1.1. Modelo 1

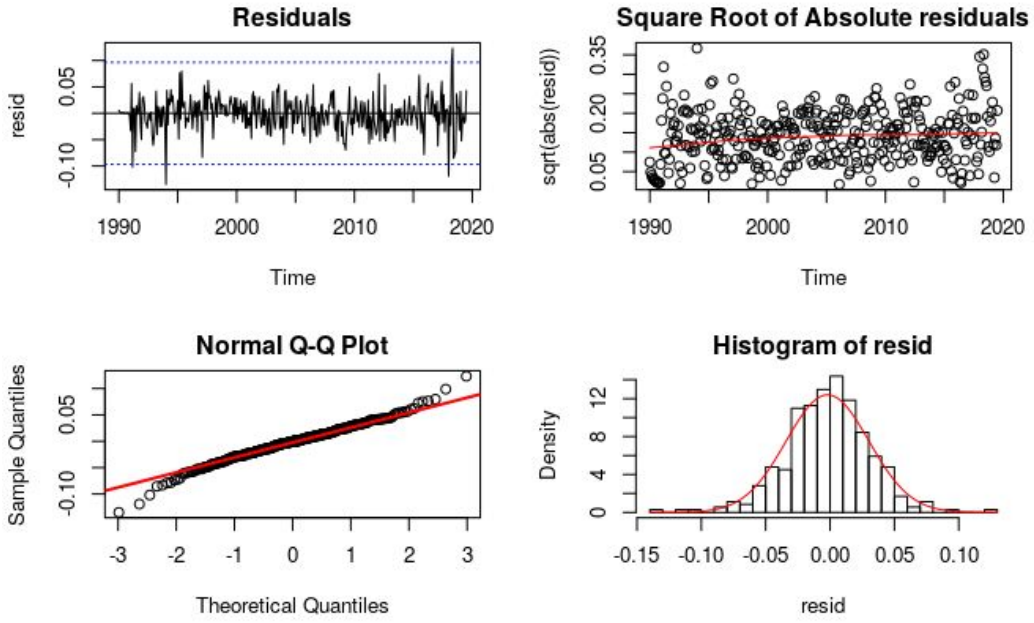

Figura 12. Normalidad y varianza de los residuos del modelo 1.

En el caso del modelo 1, vemos que la hipótesis de homogeneidad de varianzas se cumple. Esto se ve reflejado en el plot de la raíz cuadrada del valor absoluto de los residuos, donde la línea de regresión es prácticamente horizontal.

La hipótesis de normalidad no se cumple para estos residuos. El p-valor del test de Shapiro-Wilk es de 0.000529, con lo cual descartamos que sigan una distribución normal.

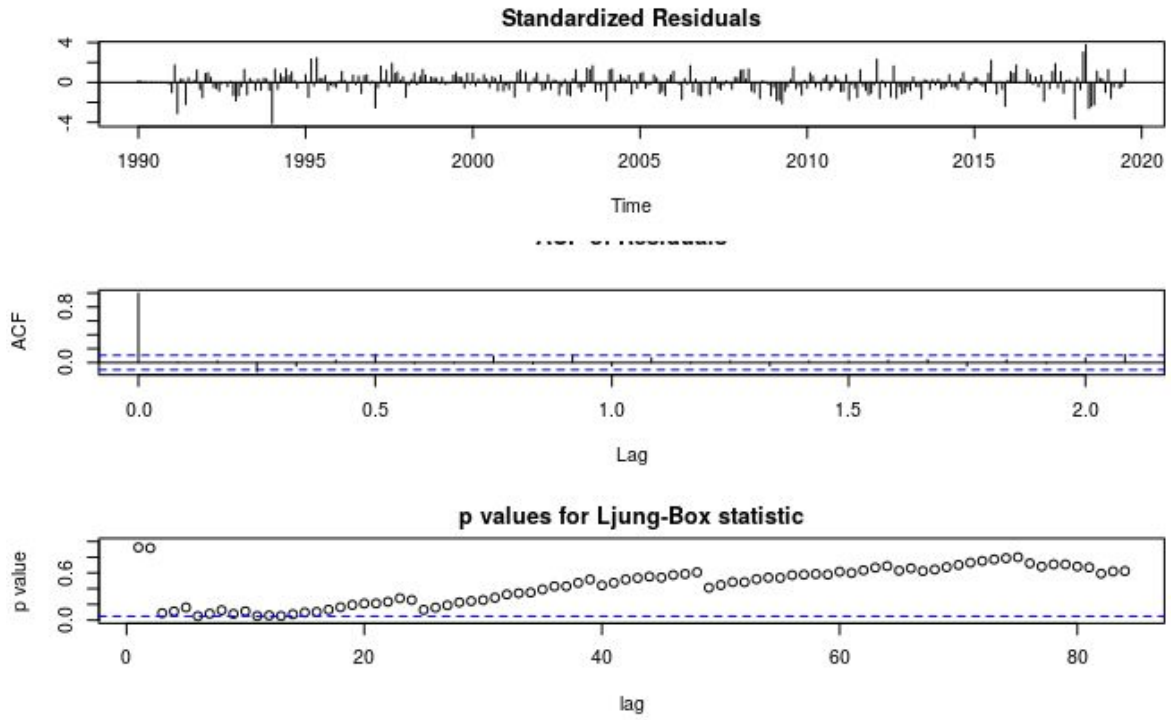

Figura 13. Residuos estandarizados, ACF de los residuos y p-valores del Ljung-Box test del modelo 1.

Podemos observar que aunque no se observa ningún residuo claramente significativo en la ACF, el test de Ljung-Box indica que hay algunos “lags” significativos y por lo tanto rechazamos que se cumpla la hipótesis de independencia.

4.1.1. Modelo 2

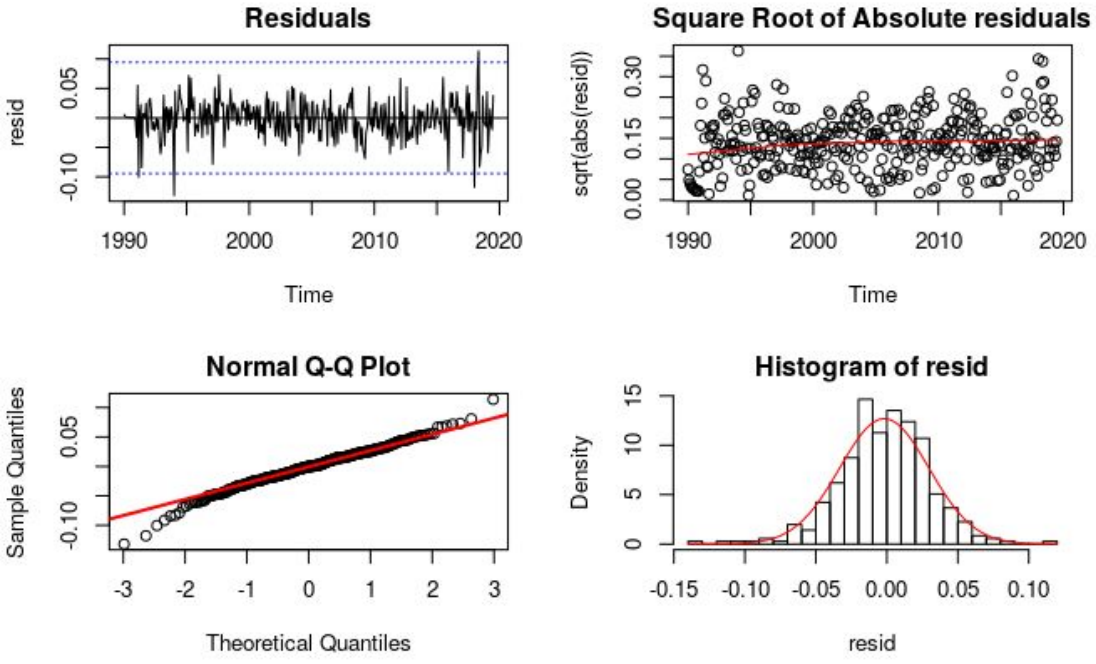

Figura 14. Normalidad y varianza de los residuos del modelo 2.

En cuanto al modelo 2, podríamos decir que ocurre algo parecido a lo del modelo 1. La varianza es homogénea tal y como se observa en el plot de la raíz cuadrada del valor absoluto de los residuos. Y la hipótesis de Normalidad también es rechazada por el test de Shapiro-Wilk.

El hecho de que la hipótesis de normalidad no se cumpla, no es algo que nos preocupe demasiado, ya que esta hipótesis es la más fácil de corregir y seguramente no se rechace cuándo linealicemos la serie eliminando los sucesos atípicos.

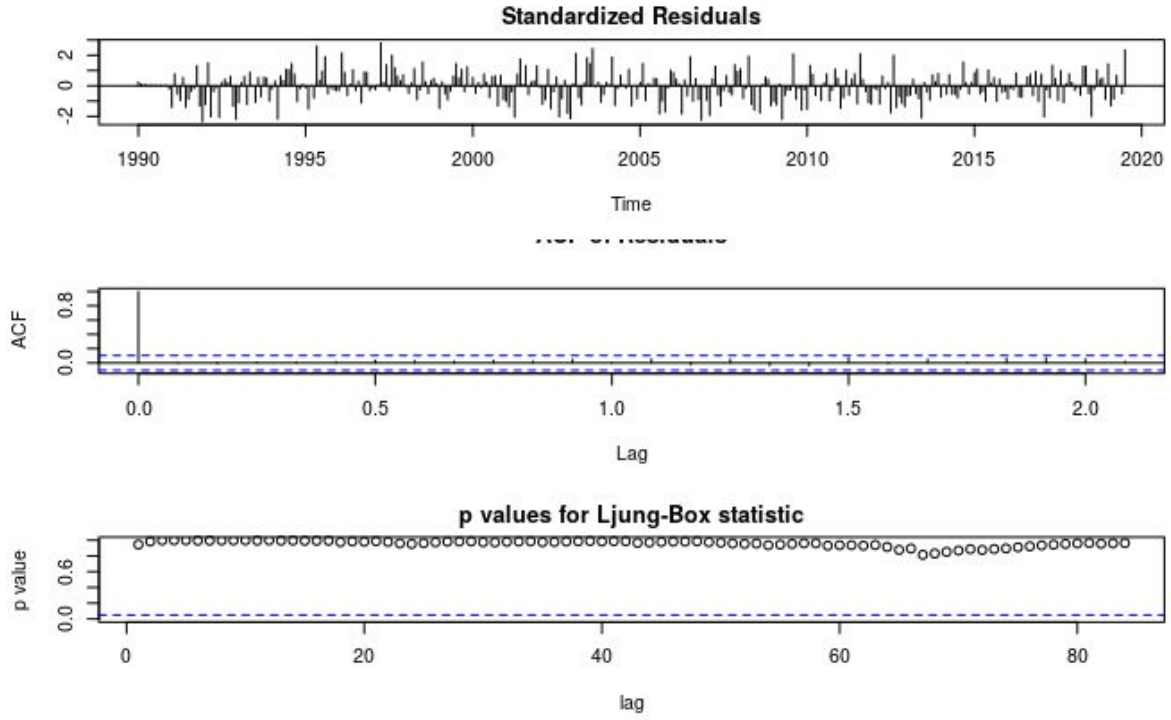

Figura 15. Residuos estandarizados, ACF de los residuos y p-valores del test de Ljung-Box test del modelo 2.

La diferencia clave de este modelo con respecto al anterior, es que la hipótesis de independencia sí que se cumple. En la ACF no se observa ningún residuo significativo y el test de Ljung-Box nos confirma que todos los residuos son claramente no significativos.

El cumplimiento de esta hipótesis es importante porque es más complicado hacer que se cumpla que las demás y muchas veces conlleva rediseñar los parámetros del modelo.

4.2. AR(∞), MA(∞), invertibilidad, estacionalidad, y medidas de adecuación

En primer lugar vamos a calcular las expresiones de ambos modelos como AR(∞) y MA(∞).

```
Psi-weights (MA(inf))

--------------------
     psi 1      psi 2      psi 3      psi 4      psi 5      psi 6      psi 7      psi 8
-0.5708764 -0.2024224  0.0000000  0.0000000  0.0000000  0.0000000  0.0000000  0.0000000
     psi 9     psi 10     psi 11     psi 12     psi 13     psi 14     psi 15     psi 16
 0.0000000  0.0000000  0.0000000 -0.8247367  0.4708227  0.1669452  0.0000000  0.0000000
    psi 17     psi 18     psi 19     psi 20
 0.0000000  0.0000000  0.0000000  0.0000000

Pi-weights (AR(inf))

--------------------
       pi 1        pi 2        pi 3        pi 4        pi 5        pi 6        pi 7        pi 8
-0.57087637 -0.52832227 -0.41716489 -0.34509386 -0.28144946 -0.23052759 -0.18857444 -0.15431665
       pi 9       pi 10       pi 11       pi 12       pi 13       pi 14       pi 15       pi 16
-0.12626742 -0.10332024 -0.08454244 -0.89391431 -0.52742785 -0.48204441 -0.38195099 -0.31562340
      pi 17       pi 18       pi 19       pi 20
-0.25749739 -0.21088843 -0.17251447 -0.14117299
```

Figura 16. 20 primeros coeficientes de la formulación como $AR(\infty)$ y $MA(\infty)$ del modelo 1.

```
Psi-weights (MA(inf))

--------------------
       psi 1        psi 2        psi 3        psi 4        psi 5        psi 6        psi 7
-0.565857331 -0.157956626 -0.159585612 -0.005009679  0.112192797  0.110895581 -0.083031793
       psi 8        psi 9       psi 10       psi 11       psi 12       psi 13       psi 14
-0.005894031  0.059192412 -0.024519103 -0.004864252 -0.848224747  0.448537037  0.149781997
      psi 15       psi 16       psi 17       psi 18       psi 19       psi 20
 0.153146355 -0.001722884 -0.095880904 -0.093664264  0.066110494  0.008086243

Pi-weights (AR(inf))

--------------------
      pi 1       pi 2       pi 3       pi 4       pi 5       pi 6       pi 7       pi 8
-0.5658573 -0.4781511 -0.5195319 -0.4648204 -0.3120336 -0.1609124 -0.1837582 -0.1230884
      pi 9      pi 10      pi 11      pi 12      pi 13      pi 14      pi 15      pi 16
 0.0000000  0.0000000  0.0000000 -0.8227529 -0.4655608 -0.3934002 -0.4274463 -0.3824323
     pi 17      pi 18      pi 19      pi 20
-0.2567265 -0.1323912 -0.1511876 -0.1012713
```

Figura 17. 20 primeros coeficientes de la formulación como $AR(\infty)$ y $MA(\infty)$ del modelo 2

A continuación calcularemos el módulo de las raíces de los polinomios característicos AR y MA de cada modelo.

Si todos los módulos de la parte AR son más grandes que 1 podemos decir que se trata de un modelo causal, y si todos los módulos de la parte MA son más grandes que 1 el modelo será invertible.

```
Modul of AR Characteristic polynomial Roots:

Modul of MA Characteristic polynomial Roots:  1.016187 1.016187 1.016187 1.016187 1.016187 1.016187
1.016187 1.016187 1.016187 1.016187 1.016187 1.016187 1.222108 4.042331
```

Figura 18. Módulos de las raíces de los polinomios característicos de mod1

```
Modul of AR Characteristic polynomial Roots:  1.261952 1.278957 1.278957 1.261952 1.251324 1.411312
1.251324 1.411312

Modul of MA Characteristic polynomial Roots:  1.016391 1.016391 1.016391 1.016391 1.016391 1.016391
1.016391 1.016391 1.016391 1.016391 1.016391 1.016391
```

Figura 19. Módulos de las raíces de los polinomios característicos de mod2

Ambos modelos son invertibles y causales, lo cual nos permitirá truncar los pesos π y ψ cuando estas sean muy pequeñas al expresar los modelos como $AR(\infty)$ y $MA(\infty)$. Esto nos permite obtener predicciones en base a las observaciones anteriores y calcular su varianza fácilmente.

Las medidas de adecuación de los datos, el AIC y BIC, son las siguientes:

```
AIC de mod1:  -1345.731
BIC de mod1:  -1330.392

AIC de mod2:  -1349.432
BIC de mod2:  -1311.084
```

4.3 Estabilidad y capacidad de previsión.

Para comprobar la estabilidad de los modelos, estimaremos dos modelos con la misma configuración que los estimados anteriormente pero esta vez el dataset no incluirá los datos de los últimos 12 meses. Una vez hecho esto observaremos si los parámetros estimados difieren mucho de los parámetros de los modelos que han utilizado la serie completa para ser estimados.

Seguidamente mostramos los resultados de la diferencia porcentual entre los coeficientes del modelo original y el modelo calculado sin la información del último año.

```
diferencia entre los coeficientes:  -0.007220416 0.01133008 NaN 0.03383144

diferencia entre los coeficientes:  0.0005066691 -0.008932891 0.03596283 0.00505759 0.02759786
0.05828711 -0.06415563 -0.1608112 0.02509229
```

Aunque el coeficiente ma8 del modelo 2 se decrementa en alrededor del 16% al eliminar el último año de la base de datos, interpretamos que esta diferencia no es tan grande y podemos decir que ambos modelos son estables.

Para comprobar la capacidad predictora haremos que los modelos estimados sin los datos del último año realizen una predicción de los próximos 12 meses y las compararemos mediante estadísticos de error como el RMSPE y el MAPE. Aparte de esto calcularemos el intervalo de confianza de las predicciones para ver cuál de ellos es más estrecho de media y por lo tanto da una predicción más precisa.

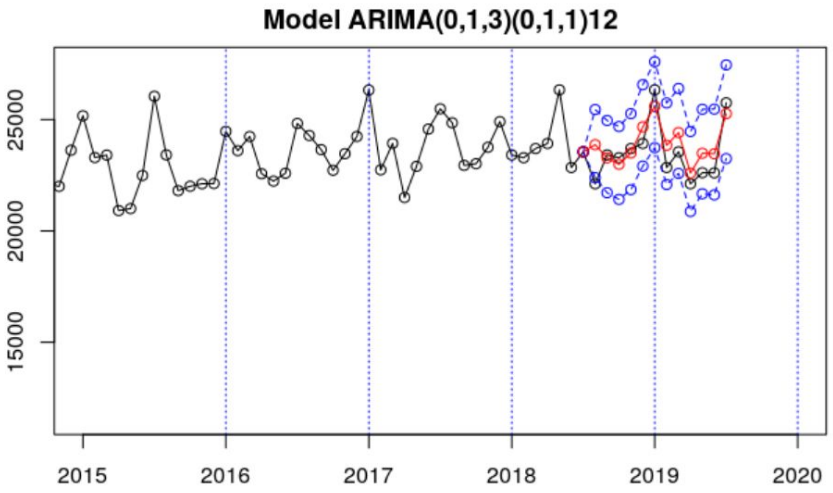

Figura 20. Predicción del modelo 1

```
RMSPE= 0.03575
MAPE= 0.03024639

Media de la amplitud del intervalo de confianza:  3618.13
```

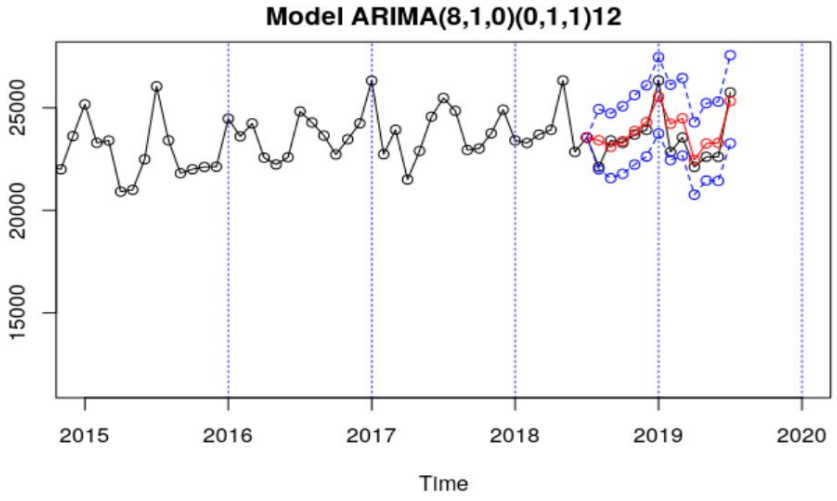

Figura 21. Predicción del modelo 2

```
RMSPE= 0.032045
MAPE= 0.02666732

Media de la amplitud del intervalo de confianza:  3576.473
```

4.4. Selección del modelo

En base a los resultados anteriores, podemos concluir que el modelo 2 es el mejor, pese a tener muchos parámetros (9). La primera diferencia entre los dos modelos, se encuentra en que aunque los dos modelos serían rechazados por no cumplir la hipótesis de normalidad de los residuos, el modelo 2 sí que cumple la hipótesis de independencia. Lo que nos da más esperanzas en que sea aceptable después de realizar el tratamiento de atípicos.

Por otro lado, las medidas de calidad de la predicción (RMSPE, MAPE y amplitud media del intervalo de confianza) son algo mejores en el modelo 2, influyendo positivamente en nuestra decisión de escoger este modelo.

Otro factor a tener en cuenta son las medidas de adecuación a los datos (AIC y BIC). El AIC es algo mejor en el modelo 2 con un valor de -1349.432 frente a -1345.731, sin embargo el BIC es bastante peor, -1311.084 frente a -1330.392. Este incremento en el BIC tiene sentido ya que el BIC penaliza más el aumento en los parámetros del modelo y este modelo tiene 5 parámetros más que el primero.

5. Previsiones

Figura 22. Previsiones para los 12 meses posteriores del modelo 2.

6. Tratamiento de atípicos

6.1. Detección e interpretación de outliers

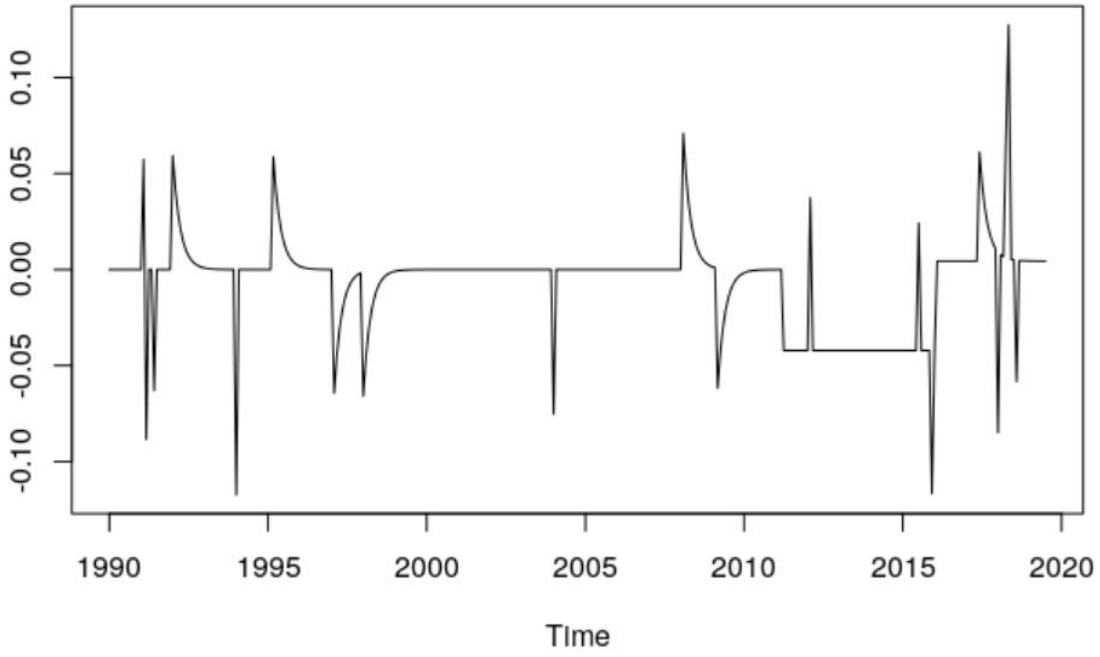

Figura 23. Influencia de los valores atípicos de la serie.

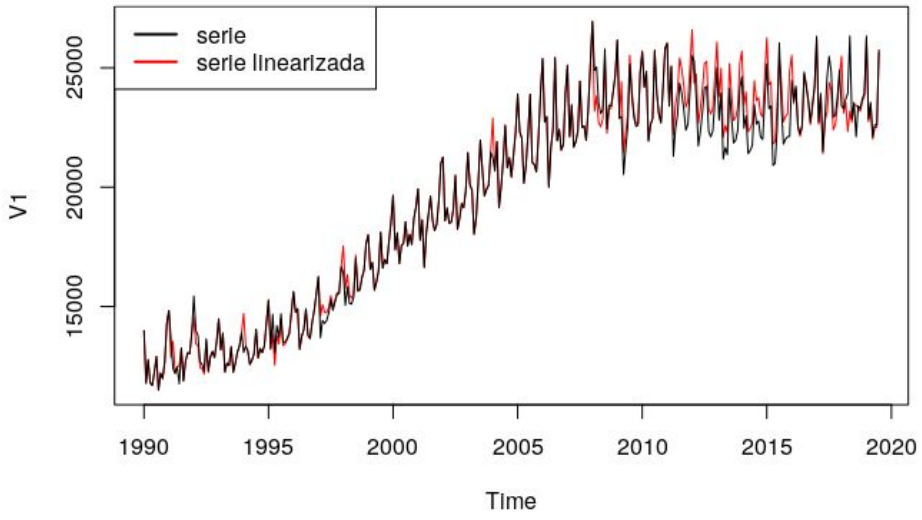

Figura 24. Plot de la serie original y la linealizada.

Buscando en diversas fuentes no hemos sido capaces de relacionar estos outliers con hechos históricos, medidas del gobierno o sucesos similares excepto en el año 2008, año de la crisis económica, donde vemos un transitory change posiblemente relacionado con este acontecimiento.

Si hemos encontrado que el consumo de electricidad va muy ligado con la temperatura. Si hace mucho frío se encienden las calefacciones y aumenta el consumo y si hace mucho calor, se utilizan los aires acondicionados. Así que es probable que los meses en los que haya un AO o un TC, se hayan registrado temperaturas anómalas para ese mes del año.

6.2. Previsiones y comparación

Con la serie linealizada y transformada con las transformaciones anteriores, volveremos a calcular los coeficientes de nuestro modelo 2 (ARIMA(8,1,0)(0,1,1)$_{12}$), comprobaremos si el modelo es ahora válido y compararemos las previsiones futuras con el modelo calculado sin linealizar la serie.

```
arima(x = lnserie2.lin, order = c(8, 1, 0), seasonal = list(order = c(0, 1,
    1), period = 12))

Coefficients:
          ar1      ar2      ar3      ar4      ar5      ar6      ar7      ar8     sma1
      -0.6821  -0.6061  -0.6358  -0.5144  -0.3431  -0.1674  -0.1906  -0.1416  -0.6962
s.e.   0.0557   0.0674   0.0750   0.0805   0.0801   0.0747   0.0669   0.0563   0.0454

sigma^2 estimated as 0.0005425:  log likelihood = 767.67,  aic = -1515.33
```

Figura 25. Estimación de *mod2* con la serie linealizada

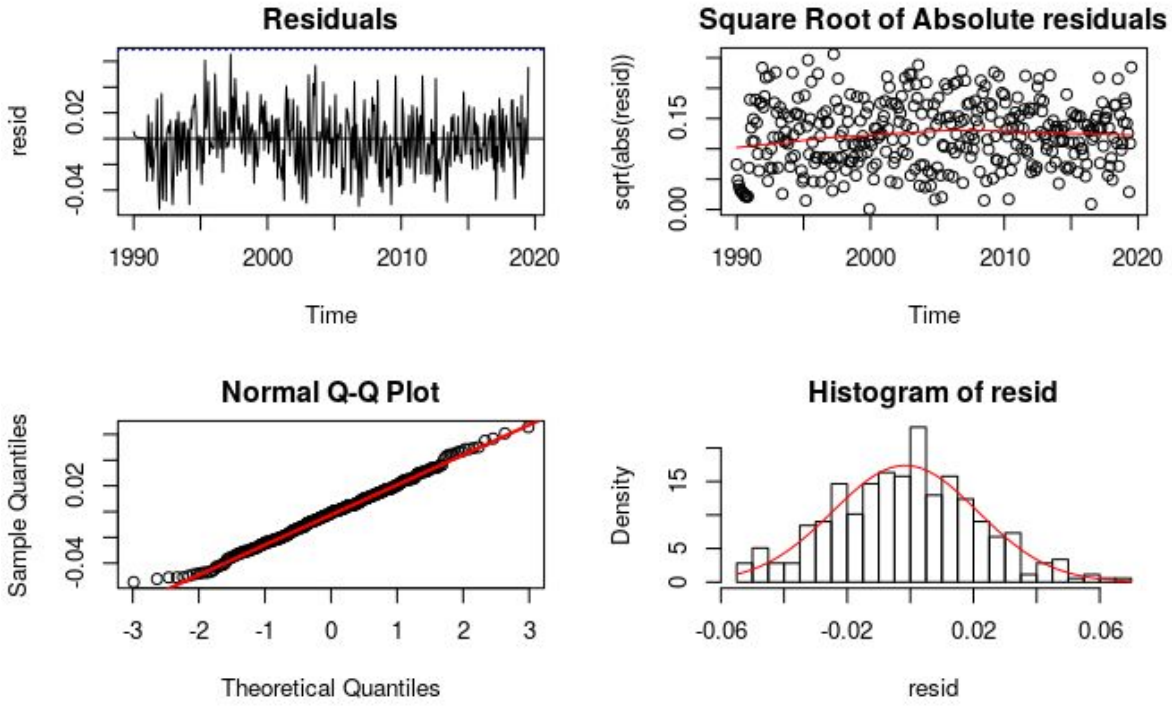

Figura 26. Normalidad y varianza de los residuos de *mod2.lin.*

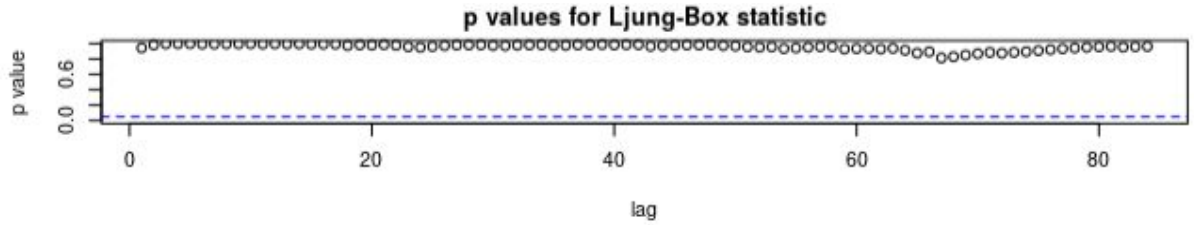

Figura 27. p-valores del test de Ljung-Box.

Tal y como esperábamos, se vuelven a cumplir las hipótesis de homogeneidad de varianzas y de independencia de residuos y además, ahora se cumple la hipótesis de normalidad. El test de Shapiro-Wilk da un p-valor de 0.3417, así que no rechazamos que la distribución de los residuos sea normal.

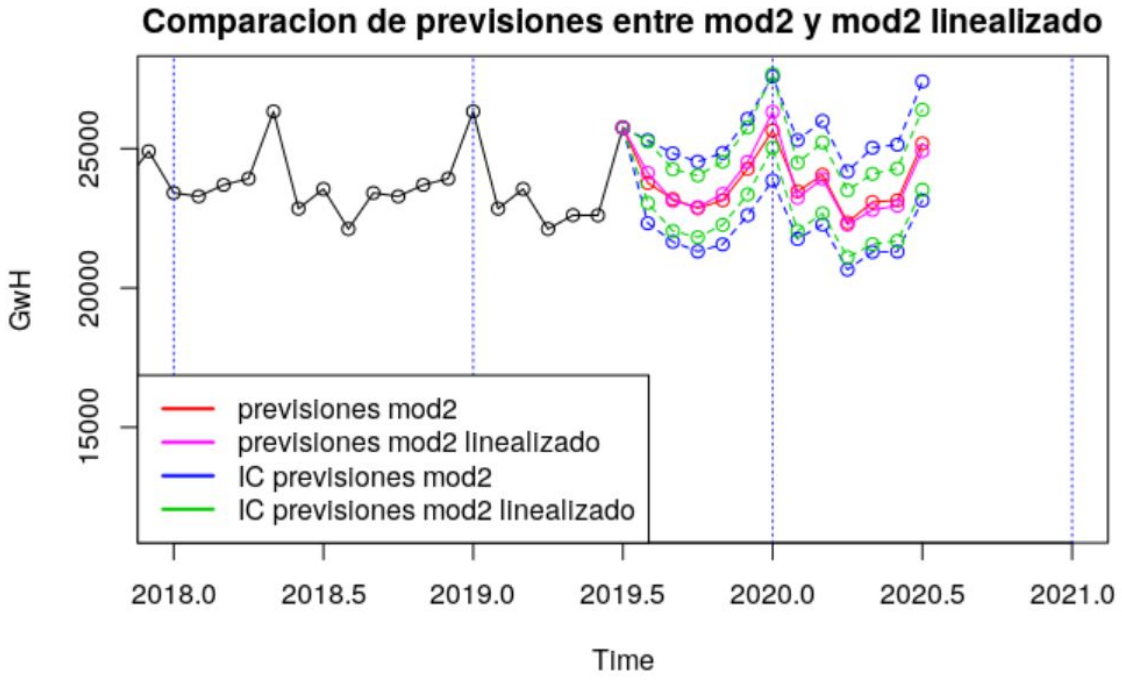

Figura 28. Comparación de las previsiones.

	par	Sigma2Z	AIC	BIC	RMSPE	MAPE	meanLength
ARIMA	9	0.0010224423	-1349.432	-1311.084	0.03204500	0.02666732	3576.473
ARIMA+Atip	30	0.0005442372	-1528.168	-1409.289	0.02186233	0.01779109	2421.796

Figura 29. Comparación entre los dos modelos.

Podemos observar que el modelo utilizando la serie linealizada es bastante mejor que el modelo utilizando la serie original. Este cumple con las tres hipótesis de los residuos y mejora bastante en valores como el AIC y BIC, además tiene una capacidad predictora más elevada como demuestran el RMSP, MAPE y la longitud media de los intervalos de confianza de dichas predicciones.