CONSUMO DE GASOLINA EN ESPAÑA

Hugo Jiménez Muñoz
Jaume Martinez Ara
AD: Series Temporales
GCED-UPC
June 2020

TABLA DE CONTENIDOS

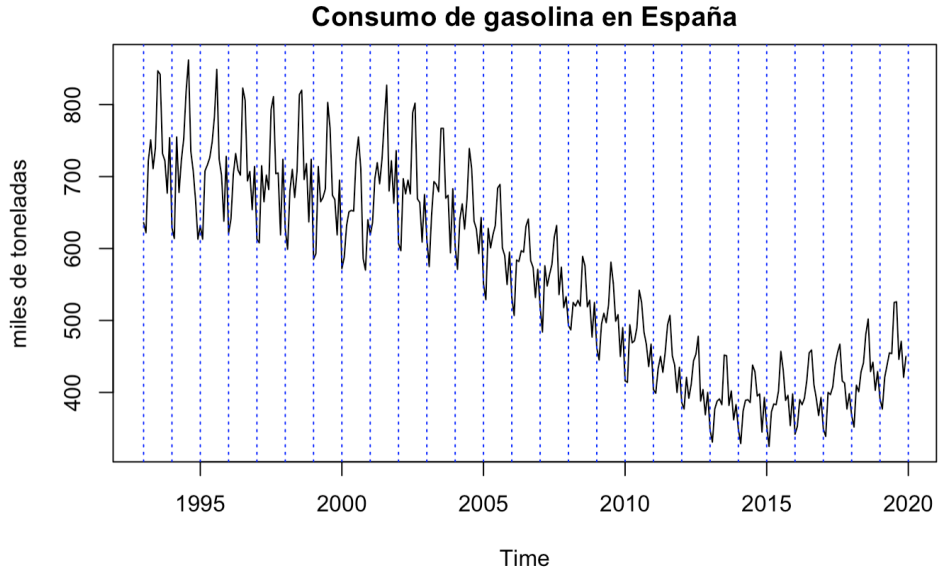
INTRODUCCIÓN

El objetivo principal de este proyecto es predecir futuros valores de una serie temporal, a partir de la última observación registrada. Para realizar esto, utilizaremos la metodología Box-Jenkins mediante modelos ARIMA.
En este caso nuestro nos centraremos en una serie temporal que recoge el consumo de gasolina en España entre los años 1993 y 2019, ambos inclusive.
Para poder llevar a cabo el objetivo propuesto seguiremos las siguientes pautas a la hora de desarrollar la metodología:

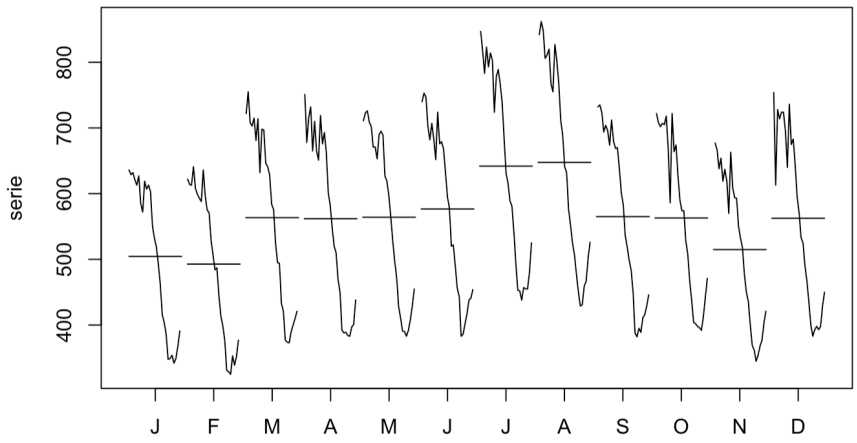
1. Identificaremos 2 modelos una vez hayamos hecho las transformaciones necesarias para obtener una serie estacionaria.
2. Estimaremos los parámetros de los modelos y veremos si son significativos o no.
3. Validaremos los modelos propuestos y nos quedaremos con el que mejor se ajuste a las predicciones.
4. Obtendremos previsiones con el modelo escogido anteriormente de los 12 meses siguientes
5. Trataremos con los atípicos y crearemos otro modelo con la serie linealizada. Una vez obtenido el modelo, lo aplicaremos a la serie original.

INDENTIFICACIÓN

Descripción básica



Como primera toma de contacto con la serie, vemos que durante los primeros 10 años aproximadamente la serie mantiene los mismos valores si los mismos patrones. Llegamos a los primeros años del siglo XXI vemos un ligero y progresivo descenso que tienen como mínimo el año 2015, a partir del qual hay un pequeño repunte.
El primer descenso se debe a la concienciación ciudadana en temas de calentamiento global y de las medidas aplicadas por el gobierno que incluyen ciertas restricciones. Además, también por el surgimiento de propulsiones alternativas en los vehículos como pueden ser coches eléctricos, de gas natural o híbridos.

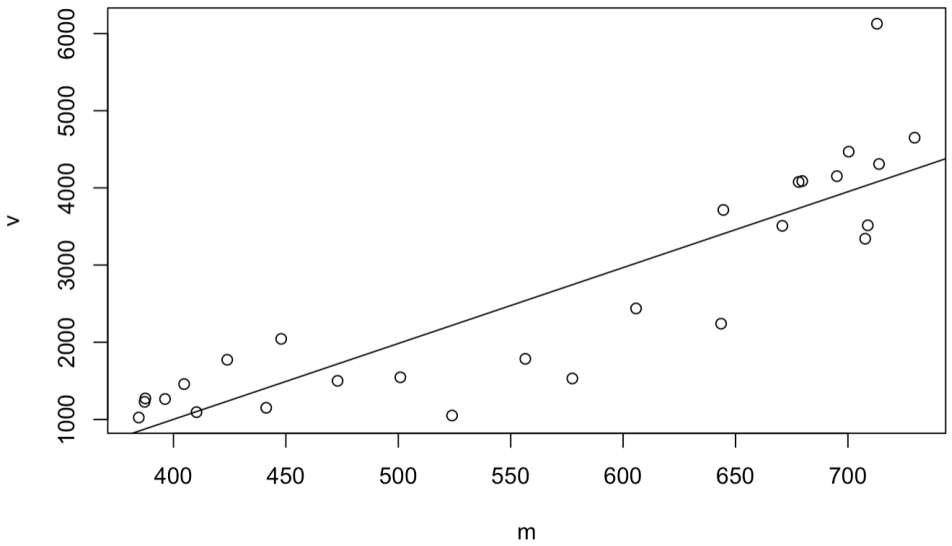


En cuanto a un visión más especifica, la anual, vemos que durante los meses de julio y agosto hay un claro aumento de consumo que viene dado por el aumento de movilizaciones en época de verano. Los meses de invierno vemos un claro descenso del consumo de gasolina debido a que es la época del año con el peor clima. Por lo tanto, la gente tiende más a quedarse en casa en vez de desplazar-se a segundas residencias o realizar actividades al aire libre q impliquen el uso del vehículo. Vemos que en diciembre se estabiliza la decaída de invierno ya que incluye el periodo de vacaciones de Navidad donde la gente suele desplazarse para visitar a familiares o irse de vacaciones.

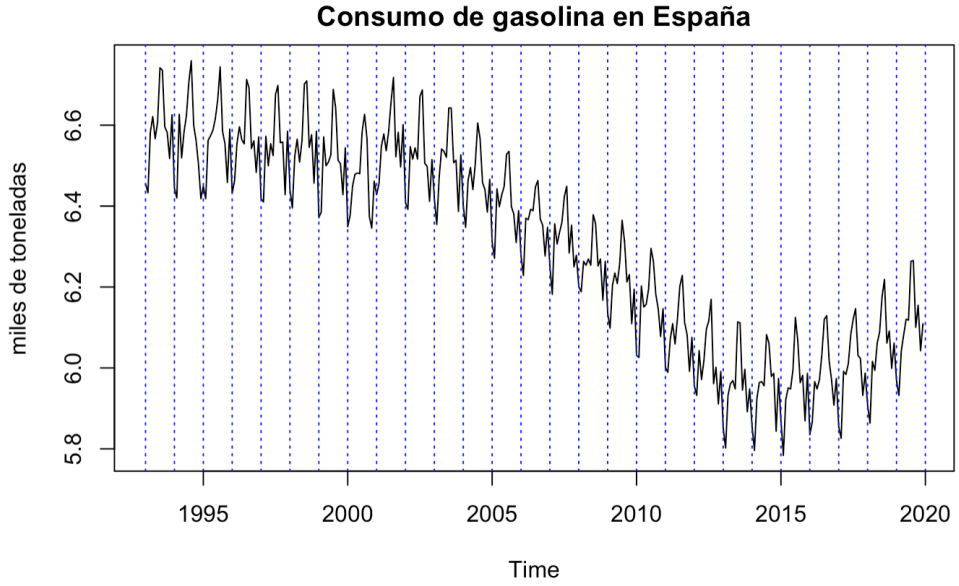
Una vez realizada la descripción básica de la serie, procedemos a transformarla con el fin de hacerla estacionaria y poder identificar el modelo.

Transformaciones

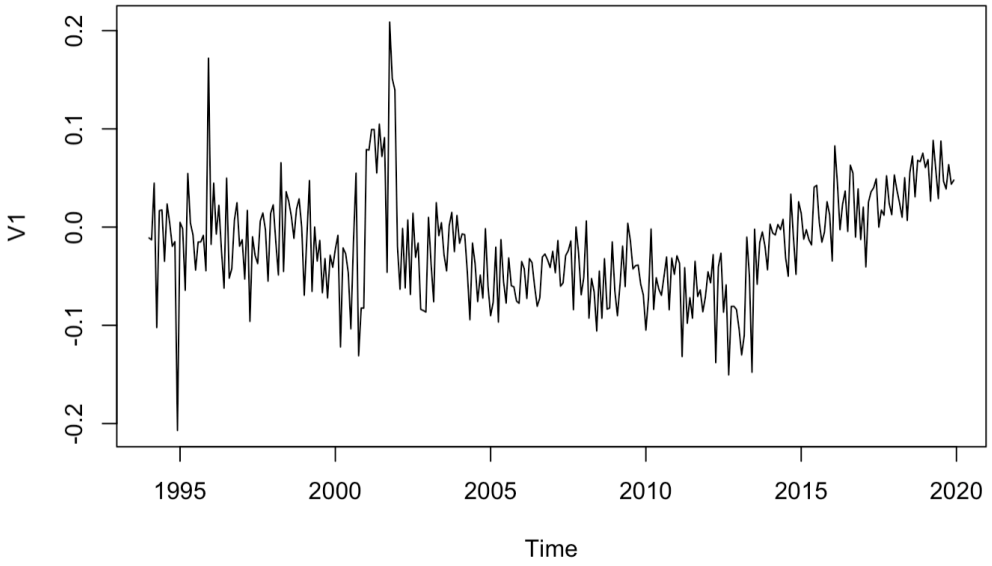
Vemos que la varianza no es constante



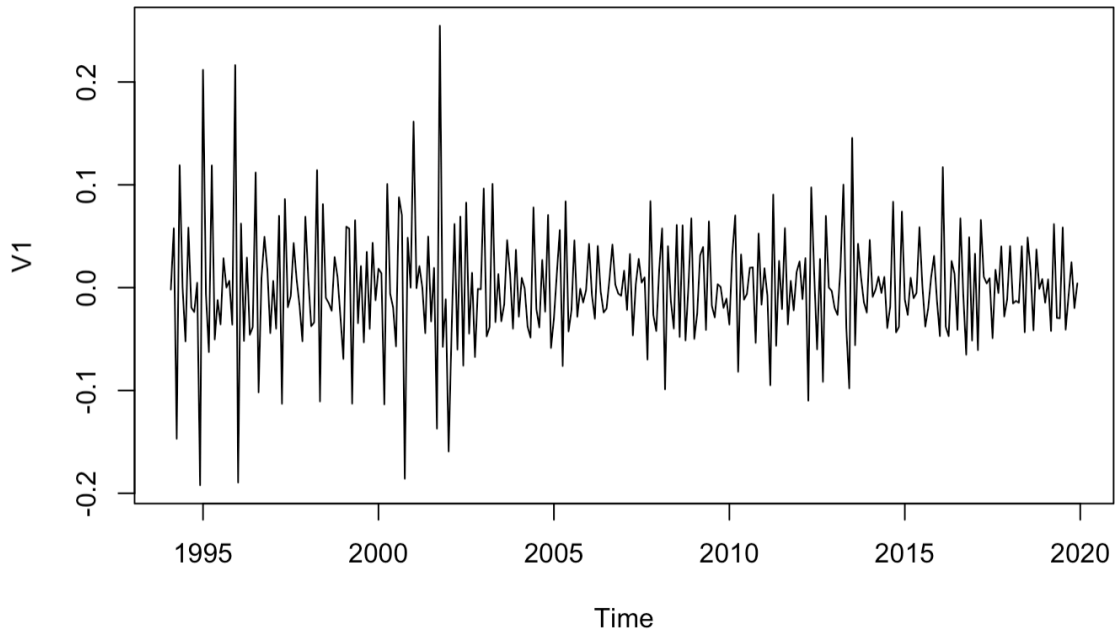
Para homogeneizar la varianza aplicamos el logaritmo y obtenemos lo siguiente:



Como vemos que cada año se repite el mismo patrón, aplicamos una diferenciación estacionaria de orden 12.



Ya que todavía no tenemos media constante aplicamos una diferenciación regular ya que hemos eliminado el patrón estacional.



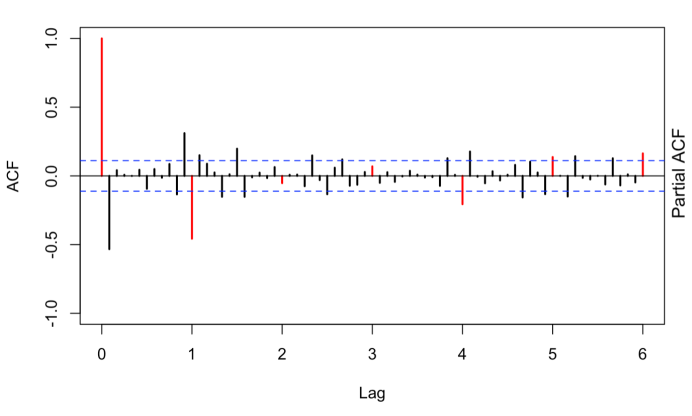
En este caso ya podríamos considerar la media constante aunque vemos que hay algunos outliers aún. Vamos a calcular la varianza con las siguientes diferenciaciones y si aumenta, nos quedamos con estas transformaciones.

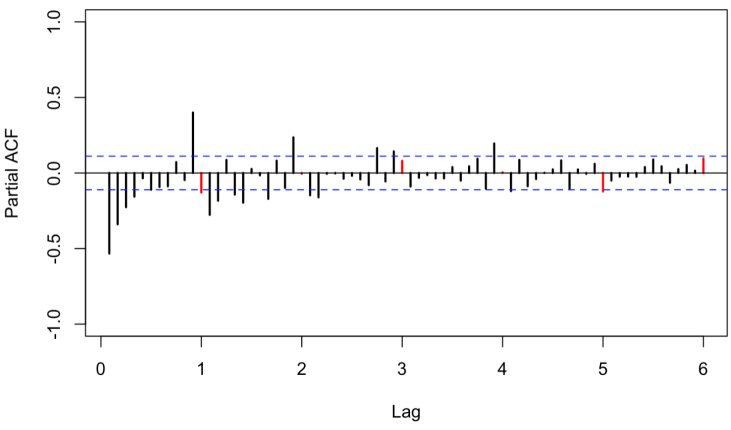
```{r}
var(lnserie)
var(d12lnserie)
var(d1d12lnserie)
var(d1d1d12lnserie)
var(d1d1d1d12lnserie)
var(diff(d1d1d1d12lnserie))
```

```
            V1
V1 0.06263059
            V1
V1 0.002996946
            V1
V1 0.003333053
           V1
V1 0.01025178
           V1
V1 0.03469309
          V1
V1 0.1223437
```

Vemos que a partir de una diferenciación regular y una estacional, las varianzas aumentan, por lo tanto nos quedamos con esas dos transformaciones ya que podemos asumir media y varinza constantes.

Como ya tenemos la serie transformada a continuación analizaremos el ACF y el PACF para determinar que dos modelos son los que mejor encajarían para nuestros datos:





Parte regular: En el caso del ACF vemos que hay retardos infinitos, por lo tanto, vamos a mirar el PACF. En el PACF vemos que los 4 primeros retardos son significativos y los demás son nulos excepto algunos satélites. Por lo tanto tendríamos un ARIMA(4,1,0) con d=1 ya que hemos hecho una diferenciación regular.

Parte estacional: En el ACF vemos lo mismo que anteriormente, infinitos retardos no nulas. Y en el PACF podemos ver como hay el primero significativo, el segundo rezando la significación y el tercero significativo. Por lo tanto vamos a modelizar primero un ARIMA(3,1,0) y como segundo modelo un ARIMA(1,1,0). D = 1 por la diferenciación estacional aplicada anteriormente.

Por lo tanto los modelo que hemos propuesto son los siguientes:

$$ARIMA(4,1,0)(3,1,0)_{12}$$

$$ARIMA(4,1,0)(1,1,0)_{12}$$

ESTIMACIÓN

```
Call:
arima(x = d1d12lnserie, order = c(0, 0, 1), seasonal = list(order = c(0, 0,
    1), period = 12))

Coefficients:
          ma1     sma1  intercept
      -0.6778  -0.8917      1e-04
s.e.   0.0354   0.0476      1e-04

sigma^2 estimated as 0.001079:  log likelihood = 611.26,  aic = -1214.51
```

Primero de todo vemos que no hay que incluir la media en nuestro modelo ya que es 0.
A continuación estimamos el primer modelo y vemos que todos los coeficientes son significativos ya que son mayor que 2 en valor absoluto. Hacemos el modelo con el logaritmo de la serie ya que la varianza aumenta en función de la media como hemos visto en el gráfico anterior.

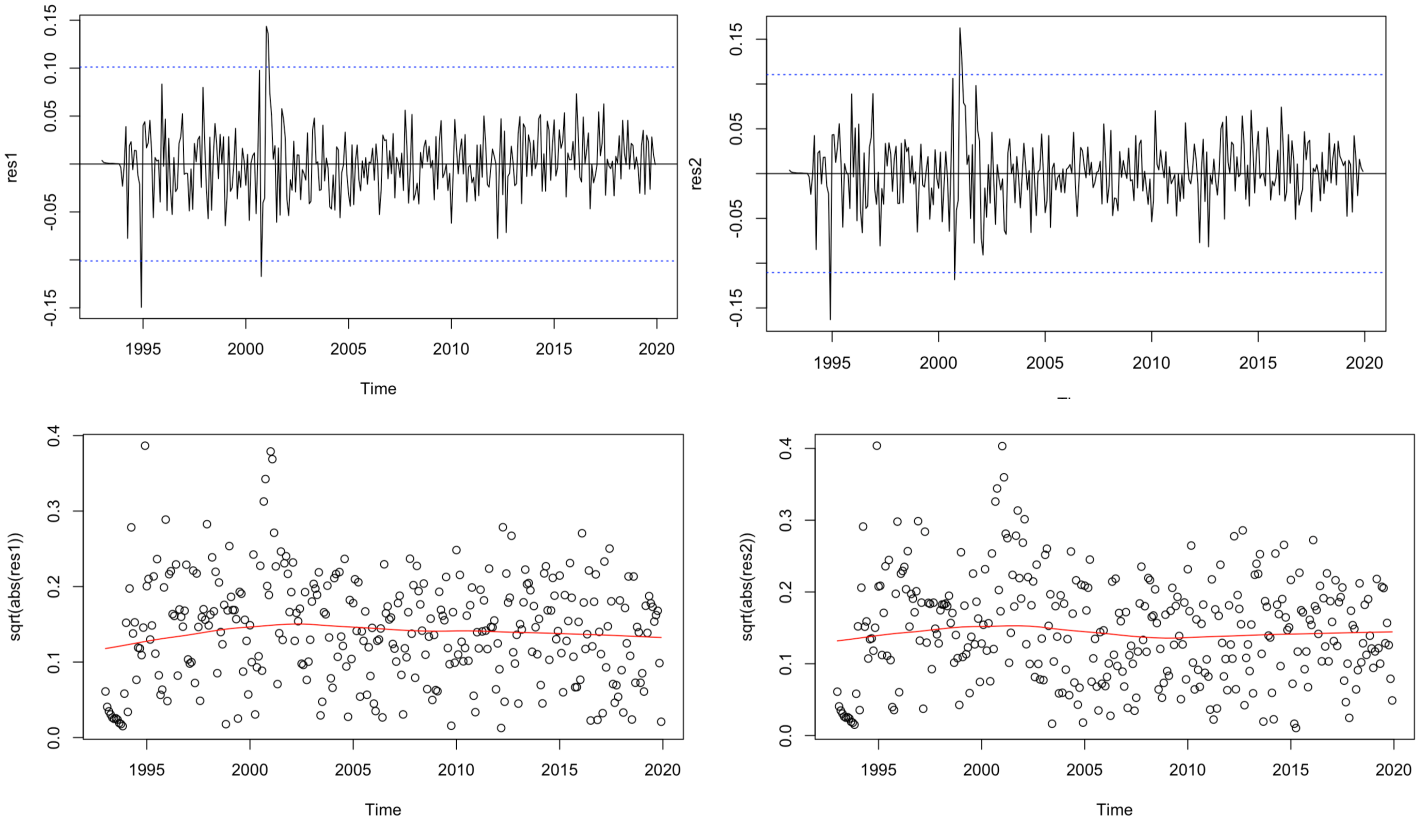
```
Call:
arima(x = lnserie, order = c(4, 1, 0), seasonal = list(order = c(3, 1, 0), period = 12))

Coefficients:
          ar1      ar2      ar3      ar4     sar1     sar2     sar3
      -0.8488  -0.5675  -0.2256  -0.1472  -0.7934  -0.4928  -0.1402
s.e.   0.0561   0.0731   0.0750   0.0566   0.0582   0.0695   0.0599

sigma^2 estimated as 0.001179:  log likelihood = 602.43,  aic = -1188.86
       ar1        ar2        ar3        ar4       sar1       sar2       sar3
-15.124694  -7.758756  -3.009161  -2.602643 -13.621064  -7.086011  -2.341131
```

En el segundo observamos lo mismo que en el primero.

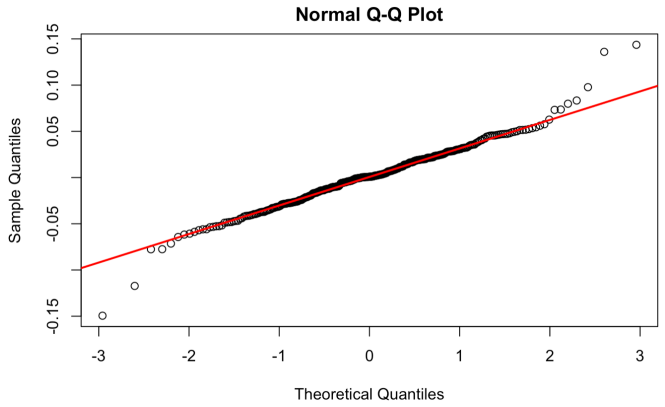
```
Call:
arima(x = lnserie, order = c(4, 1, 0), seasonal = list(order = c(1, 1, 0), period = 12))

Coefficients:
          ar1      ar2      ar3      ar4     sar1
      -0.8447  -0.5934  -0.3059  -0.1956  -0.5424
s.e.   0.0556   0.0713   0.0712   0.0557   0.0493

sigma^2 estimated as 0.001408:  log likelihood = 577.17,  aic = -1142.34
       ar1        ar2        ar3        ar4       sar1
-15.184994  -8.325140  -4.294235  -3.513862 -10.993926
```

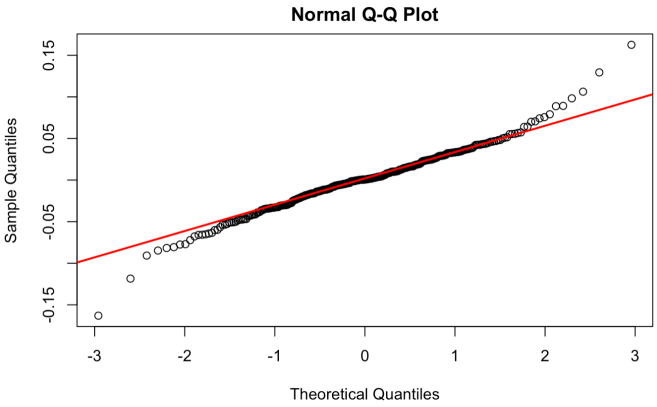
VALIDACIÓN

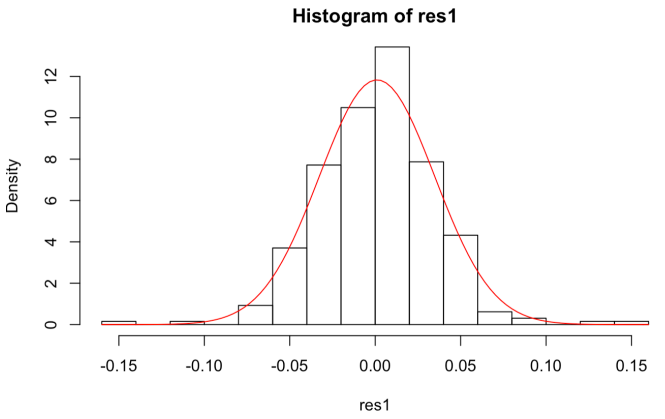
Ahora miraremos los residuos de los modelos y comprobaremos si son constantes, podemos asumir normalidad y si los residuos son independientes entre si.

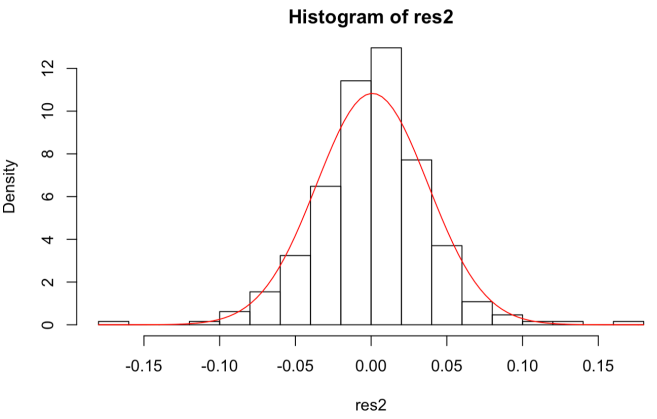


En ambos modelos podemos ver que la varianza acumulada es prácticamente constante ya que sigue una linea recta casi perfecta. Por lo tanto la primera premisa se cumple.

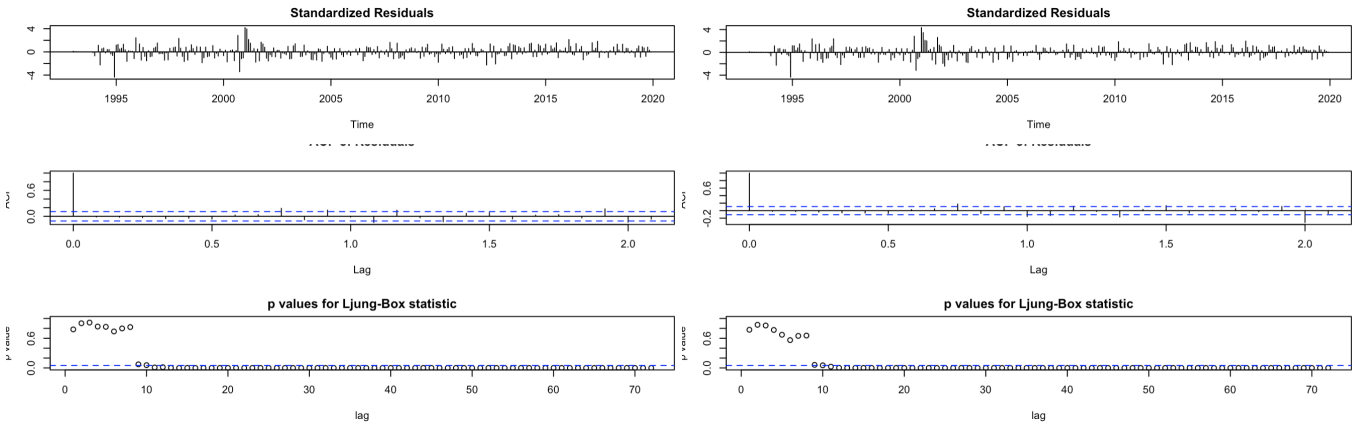








En cuanto a la normalidad, vemos que ambos se ajustan bastante bien a la distribución Gaussiana aunque hay varios outliers que impiden que se ajusten perfectamente. Por lo tanto concluimos que se verifica la premisa de normalidad aunque tratando los atípicos podríamos mejorarlo.



Vemos que hay una dependencia entre los residuos en ambos modelos ya que en el Ljung-Box test vemos que prácticamente la mayoría de p-valores están por debajo del nivel de significación. Esto puede ser debido a la presencia de outliers.

Ahora miraremos las raíces de los polinomios de AR y MA infinitos y veremos si son estacionarios y/o invertibles.

```
 [1] 1.075493 1.046509 1.046509 1.046509 1.075493 1.075493 1.046509 1.046509 1.046509 1.046509 1.046509 1.046509 1.075493 1.046509
[15] 1.046509 1.046509 1.046509 1.046509 1.046509 1.075493 1.046509 1.075493 1.046509 1.046509 1.046509 1.046509 1.075493 1.046509
[29] 1.046509 1.075493 1.075493 1.046509 1.075493 1.075493 1.046509 1.075493 1.445286 1.803314 1.445286 1.803314
numeric(0)
 [1] 1.052295 1.052295 1.052295 1.052295 1.052295 1.052295 1.052295 1.052295 1.052295 1.052295 1.052295 1.052295 1.592431 1.419796
[15] 1.592431 1.419796
numeric(0)
```

Al ser todas las raíces mayores que 1, sabemos que el modelo es causal y como no tiene parte MA, es invertible.

```{r}
AIC(mod); BIC(mod)
AIC(mod1); BIC(mod1)
```

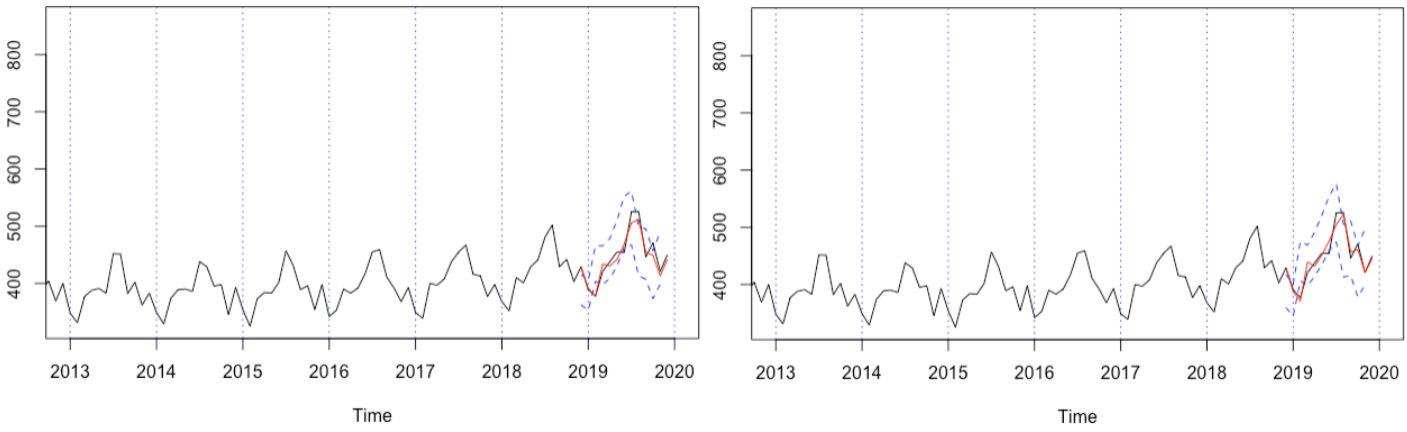
```
[1] -1188.855
[1] -1158.937
[1] -1142.336
[1] -1119.898
```

Para comparar los modelos, utilizamos las medidas de AIC y BIC, que más o menos nos dan resultados muy parecidos por lo tanto, no podemos concluir que un modelo es mejor que otro, pero el segundo modelo es ligeramente mejor.

```{r}
ultim=c(2018,12); serie2=window(serie, end=ultim)
lnserie2 = log(serie2)
(mod1_12 = arima(lnserie2,order=c(4,1,0),seasonal=list(order=c(3,1,0),period=12), include.mean = FALSE))
(mod2_12 = arima(lnserie2,order=c(4,1,0),seasonal=list(order=c(1,1,0),period=12), include.mean = FALSE))
```

```
Call:
arima(x = lnserie2, order = c(4, 1, 0), seasonal = list(order = c(3, 1, 0),
    period = 12), include.mean = FALSE)

Coefficients:
          ar1      ar2      ar3      ar4     sar1     sar2     sar3
      -0.8440  -0.5565  -0.2305  -0.1519  -0.7977  -0.4896  -0.1293
s.e.   0.0572   0.0744   0.0760   0.0576   0.0593   0.0713   0.0617

sigma^2 estimated as 0.001204:  log likelihood = 575.92,  aic = -1135.84

Call:
arima(x = lnserie2, order = c(4, 1, 0), seasonal = list(order = c(1, 1, 0),
    period = 12), include.mean = FALSE)

Coefficients:
          ar1      ar2      ar3      ar4     sar1
      -0.8394  -0.5847  -0.3125  -0.2005  -0.5487
s.e.   0.0567   0.0724   0.0723   0.0568   0.0501

sigma^2 estimated as 0.001437:  log likelihood = 551.6,  aic = -1091.2
```

Se puede observar que los coeficientes de los modelos con las 12 ultimas observaciones y sin ellas dan muy similar. Eso nos indica que el modelo es estable.



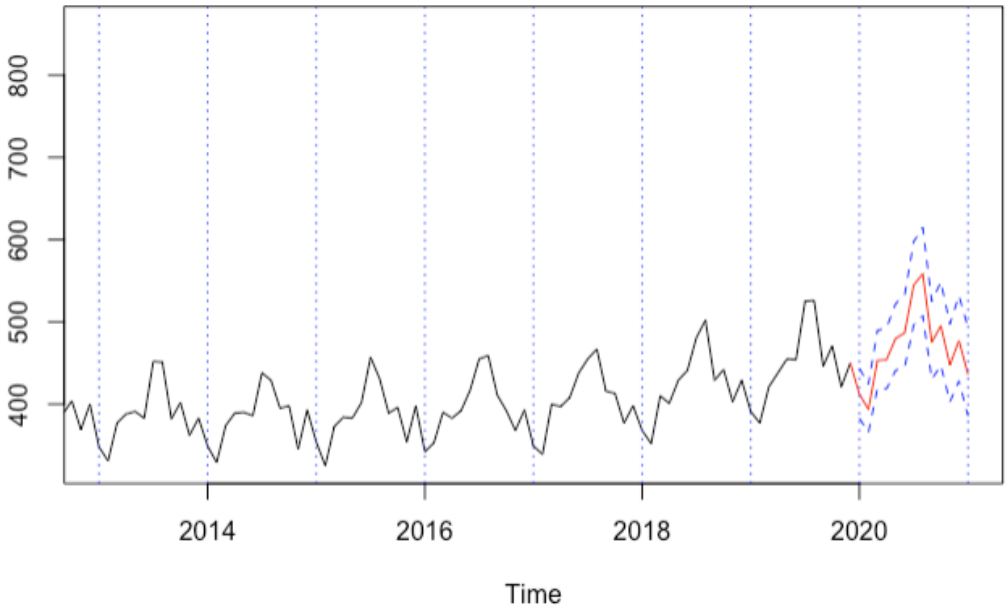
Cuando realizamos la previsión de las 12 últimas observaciones vemos que los dos modelos actúan bastante similar. Vamos a calcular medidas objetivas para determinar cual actúa mejor.

```{r}
primer=c(2018,12);
obs = window(serie, start=primer)
(RMSPE=sqrt(mean(((obs-exp(pr1))/obs)^2)))
(MAPE=mean(abs(obs-exp(pr1))/obs))
mean(tu1-tl1)


obs = window(serie, start=primer)
(RMSPE=sqrt(mean(((obs-exp(pr2))/obs)^2)))
(MAPE=mean(abs(obs-exp(pr2))/obs))
mean(tu2-tl2)
```

```
 [1] 0.02534636
 [1] 0.02178739
 [1] 76.36956
 [1] 0.02395711
 [1] 0.018318
 [1] 82.39678
```

Podemos ver que el error que comete el segundo model es ligeramente menor que el del primer modelo, aunque el intervalo de confianza de las predicciones es mayor. Como el segundo modelo es más simple, por el principio de parsimonia nos quedamos con él para realizar futuras predicciones.

PREVISIONES



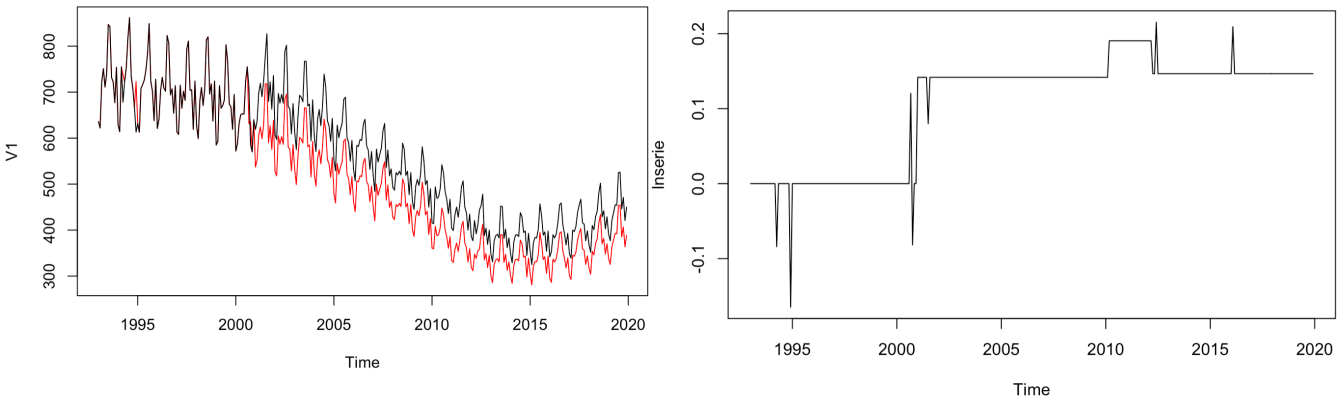
El modelo nos predice el mismo patrón estacional que anteriormente como es lógico, y además que aumentará la media respecto al año pasado por la tendencia alcista que habíamos visto en los últimos años. Este año las predicciones serian incorrectas ya que el consumo de gasolina ha disminuido muy significativamente a causa de las medidas adoptadas por el COVID19 pero este método no tiene en cuenta factores exógenos, por lo tanto, no es capaz de detectarlo.

TRATAMIENTO DE ATÍPICOS

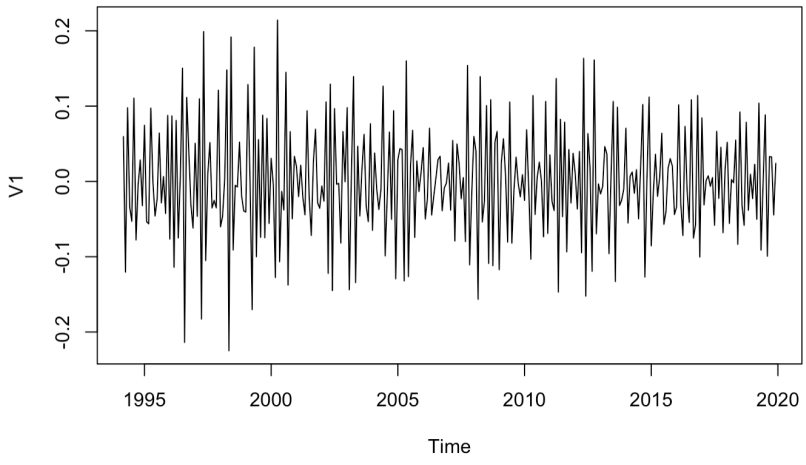
Una vez aplicado el método de identificación de atípicos obtenemos los siguientes:

	Obs <int>	type_detected <fctr>	W_coeff <dbl>	ABS_L_Ratio <dbl>	Fecha <fctr>
4	16	AO	-0.08412361	3.858995	Abr 1994
2	24	AO	-0.16469549	7.075753	Dic 1994
3	93	AO	0.12037153	5.399108	Sep 2000
5	94	AO	-0.08174668	3.834084	Oct 2000
1	97	LS	0.14179822	7.608267	Ene 2001
9	103	AO	-0.06180712	3.078914	Jul 2001
6	207	LS	0.04869955	3.110512	Mar 2010
10	232	LS	-0.04388309	2.973221	Abr 2012
7	234	AO	0.06865209	3.322856	Jun 2012
8	278	AO	0.06258680	3.073120	Feb 2016

1–10 of 10 rows



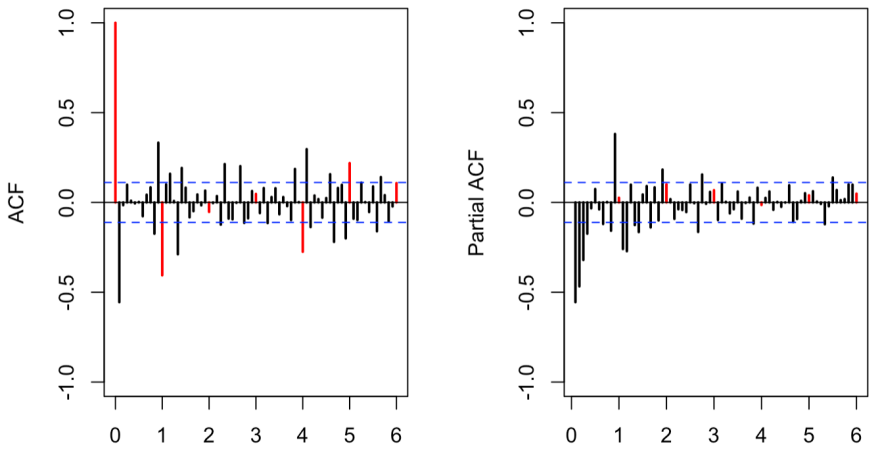
Como hemos hecho anteriormente, como tenemos una serie nueva, volvemos a identificar el modelo que le corresponde una vez hechas las transformaciones. Nos quedamos con la serie logarítmica y una diferenciación estacional y otra regular.



```r
var(lnserie.lin)
var(d12lnserie.lin)
var(d1d12lnserie.lin)
var(d1d1d12lnserie.lin)
var(diff(d1d1d12lnserie.lin))
```

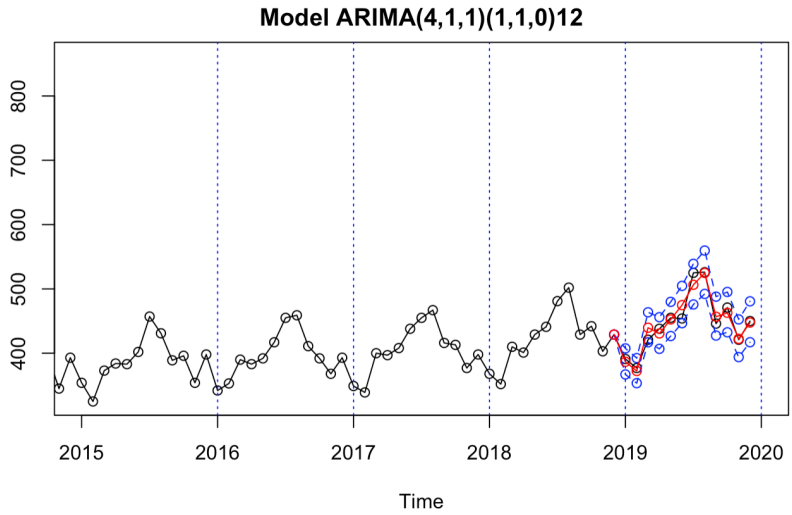
```
            V1
V1 0.09079485
            V1
V1 0.002037728
            V1
V1 0.001873763
            V1
V1 0.005846406
            V1
V1 0.01961803
```

Como en la serie con atípicos, necesitamos una diferenciación estacional de orden 12 y una diferenciación regular ya que la media es constante y si seguimos haciendo diferenciaciones regulares aumenta la varianza.



En el PACF tenemos 4 retardos no nulos y en el ACF retardos no nulos infinitos, en cuanto a la parte regular, por tanto, determinamos un AR(4). En cuanto a la parte estacional, en el ACF tenemos infinitos retardos no nulos, y aunque en el PACF no tenemos ningún retardo significativo, asumimos un AR(1) también, ya que necesitamos parte estacional y el primer retardo por azar no ha salido significativo.

En cuanto a la validación del modelo, se verifican mejor las premisas que en el modelo con atípicos ya que los residuos se ajustan mejor al qq-plot y la varianza es más constante. En cambio, sigue habiendo una dependencia de residuos, con lo que necesitaríamos un modelo para la varianza también.



	par <int>	Sigma2Z <dbl>	AIC <dbl>	BIC <dbl>	RMSPE <dbl>	MAPE <dbl>	meanLength <dbl>
ARIMA(4,1,0)(1,1,0)12	5	0.0014075006	-1142.336	-1119.898	0.02395711	0.01831800	76.36956
ARIMA(4,1,0)(1,1,0)12_NOATIPICS	15	0.0007034234	-1339.191	-1279.354	0.02288504	0.01709872	50.99268

Haciendo una comparación con el modelo con atípicos, vemos que es mejor en todos los aspectos, ya que comete menos error y las predicciones son más fiables ya que el intervalo de confianza es más reducido.