PROJECTE ANÀLISI DE DADES: Entrada Turística Espanya

David Anglada Rotger i Andreu Huguet Segarra

17/5/2019

Contents

1 Introducció **1**

2 Identificació **1**
2.1 Representació gràfica de les dades 1
2.2 Transformació de les dades 3
2.3 ACF/PACF de les dades i proposta de models 7
2.4 Models proposats 9

3 Estimació dels models **9**

4 Validació dels Models **11**
4.1 Estudi dels residus dels models 11
4.2 Estabilitat dels Models 20
4.3 Capacitat de predicció 21
4.4 Elecció de model 24

5 Predicció a llarg termini **24**

6 Tractament de *outliers* **25**
6.1 Identificació i estimació del model per la sèrie linealitzada 27
6.2 Validació del model per la sèrie linealitzada 28
6.3 Estabilitat del model proposat per la sèrie linealitzada 31
6.4 Capacitat de predicció del model proposat per la sèrie linealitzada 32
6.5 Previsions a llarg termini pel model proposat per la sèrie linealitzada 33

7 Comparació dels dos models **33**

8 Comentaris finals **35**

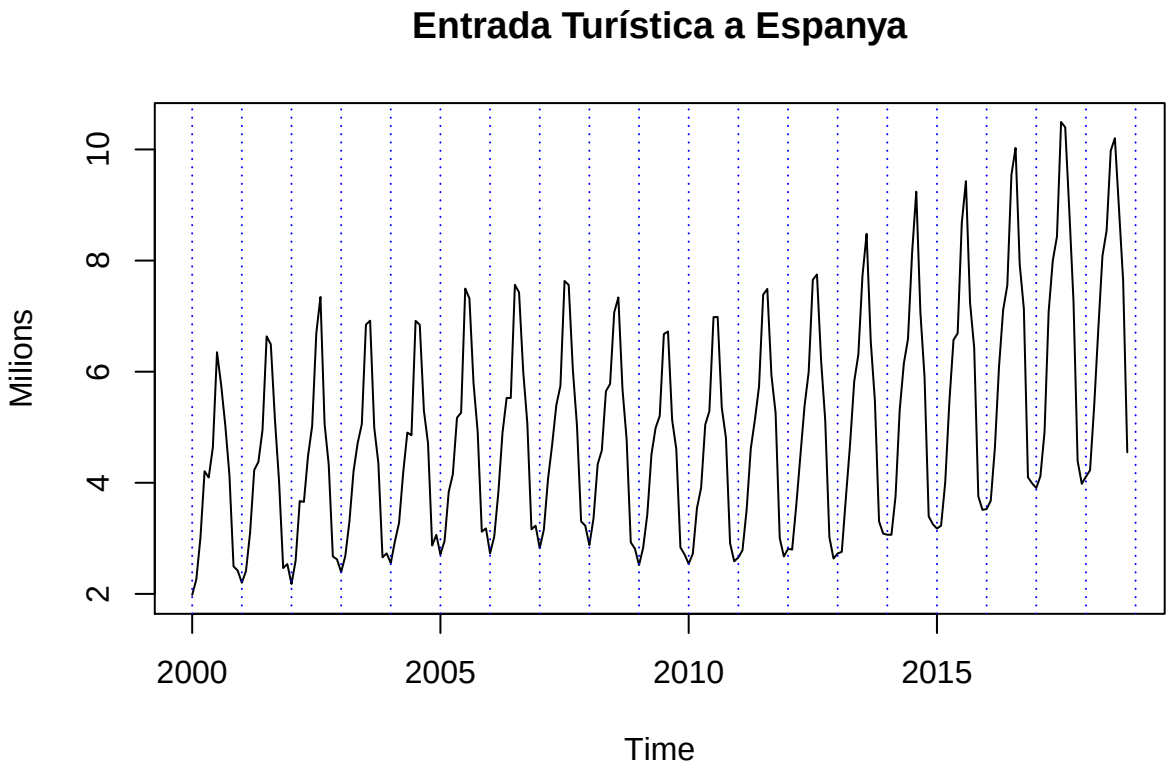
1 Introducció

2 Identificació

2.1 Representació gràfica de les dades

Un cop feta la representació de les dades, s'observa una clara tendència creixent. Tot i així, aquesta tendència no és constant, ja que és menys pronunciada entre els anys 2000 i 2010, fins i tot amb una petita baixada entre els anys 2007i 2010 i sembla que es pronuncia a partir de l'any 2011.

Pel que fa a la variància, s'observa que va augmentant a mesura que augmenta la mitjana dels valors de les dades, és a dir, a mesura que es pronuncia la tendència creixent. És a dir, en els anys 2000-2010, la variància és menor que en els anys 2011-2019, on el creixement augmenta.

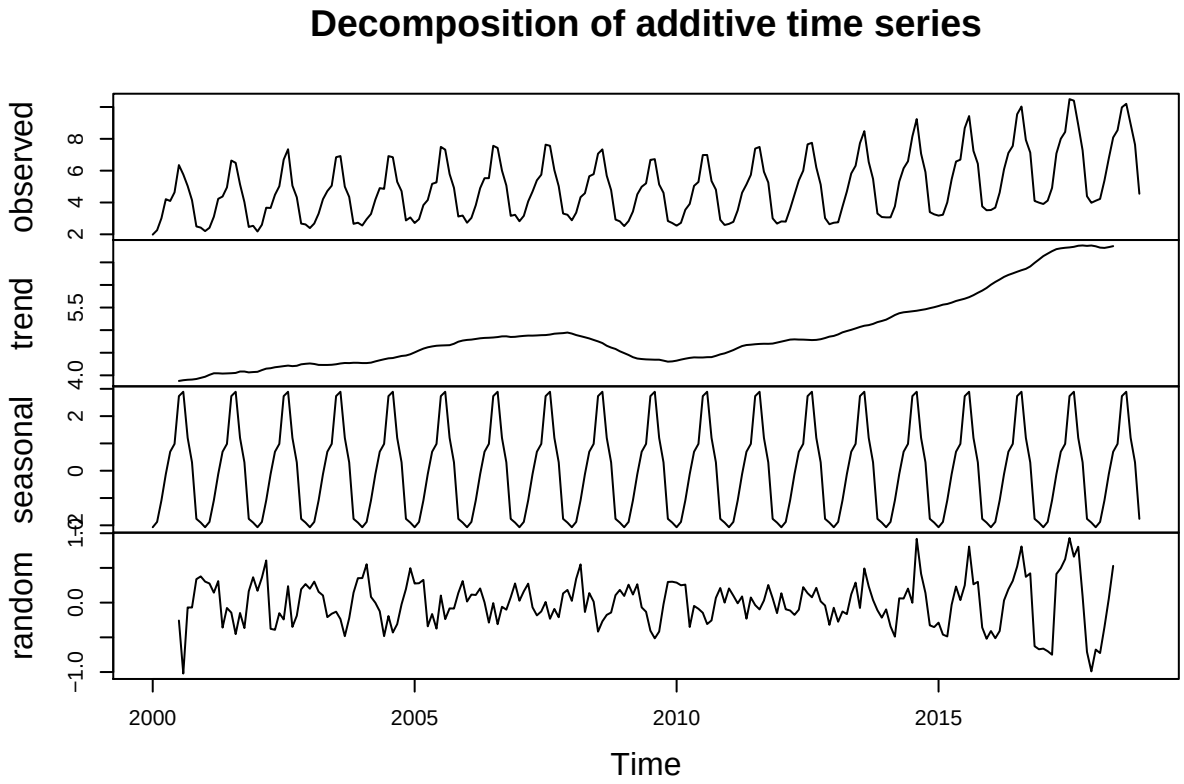


2.1.1 Descomposició en components bàsiques

Per poder analitzar millor les dades, es realitza la seva descomposició en les seves components bàsiques, és a dir, el model aditiu de la serie:

$$X_t = T_t + S_t + C_t + \omega_t$$

on: - T_t és la **tendència** de la sèrie a llarg termini. - S_t és el ***seasonal*** de la sèrie (patró repetit periòdicament amb període constant). - C_t és el **cicle** de la sèrie (patró repetit periòdicament amb període no constant). Aquesta part no surt representada en la descomposició. - ω_t és el soroll aleatòri.



S'observa, tal i com s'havia comentat anteriorment, la clara tendència creixent de la sèrie, amb un creixement menys pronunciat a l'inici, una petita baixada entre els anys 20017 i 2010 i una pujada més pronunciada més cap a l'actualitat. Pel que fa al patró estacional, observem que durant els mesos d'estiu, el número de turistes a Espanya augmenta molt considerablement. Aquest fet que no crida l'atenció, ja que és durant els mesos d'estiu quan més vacanses s'agafa la gent i més aprofiten per venir a les costes espanyoles. Durant els mesos de tardor-hivern, observem que el número de turistes cau en picat.

2.2 Transformació de les dades

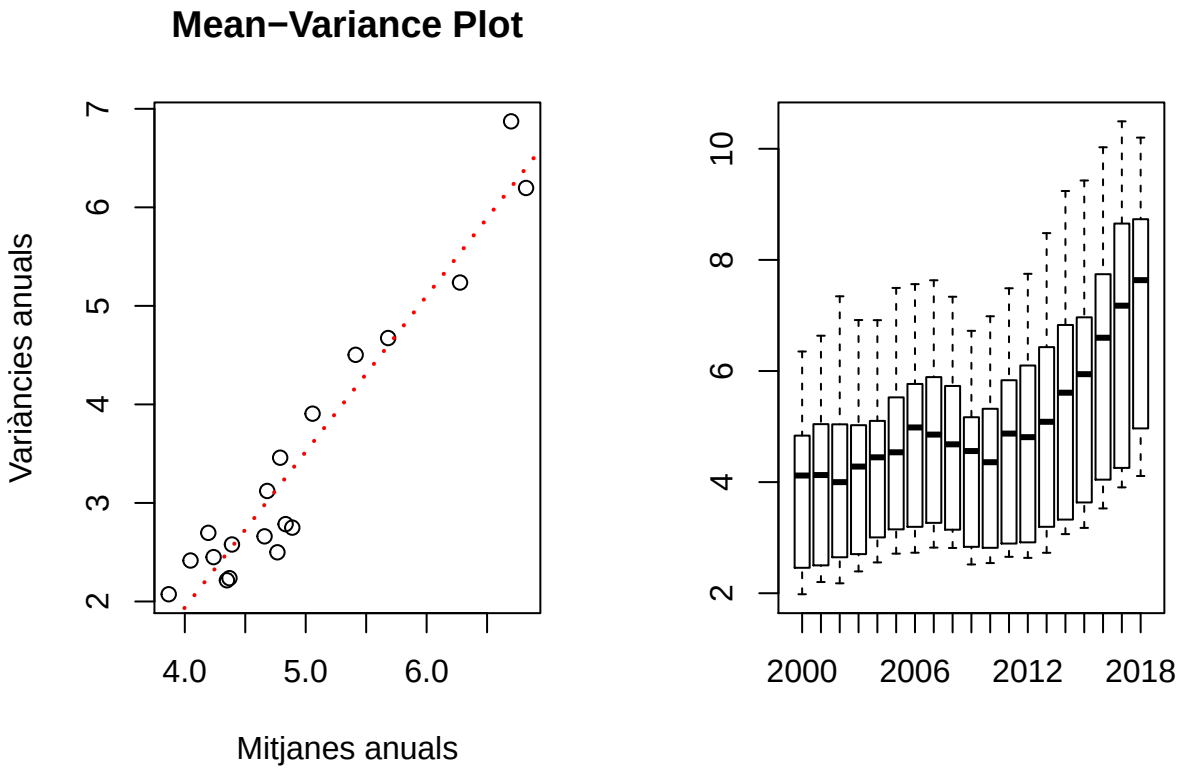
A continuació s'analitzarà la necessitat de realitzar una sèrie de transformacions amb l'objectiu d'aconseguir estacionaritat en la nostra sèrie temporal.

2.2.1 Variància constant

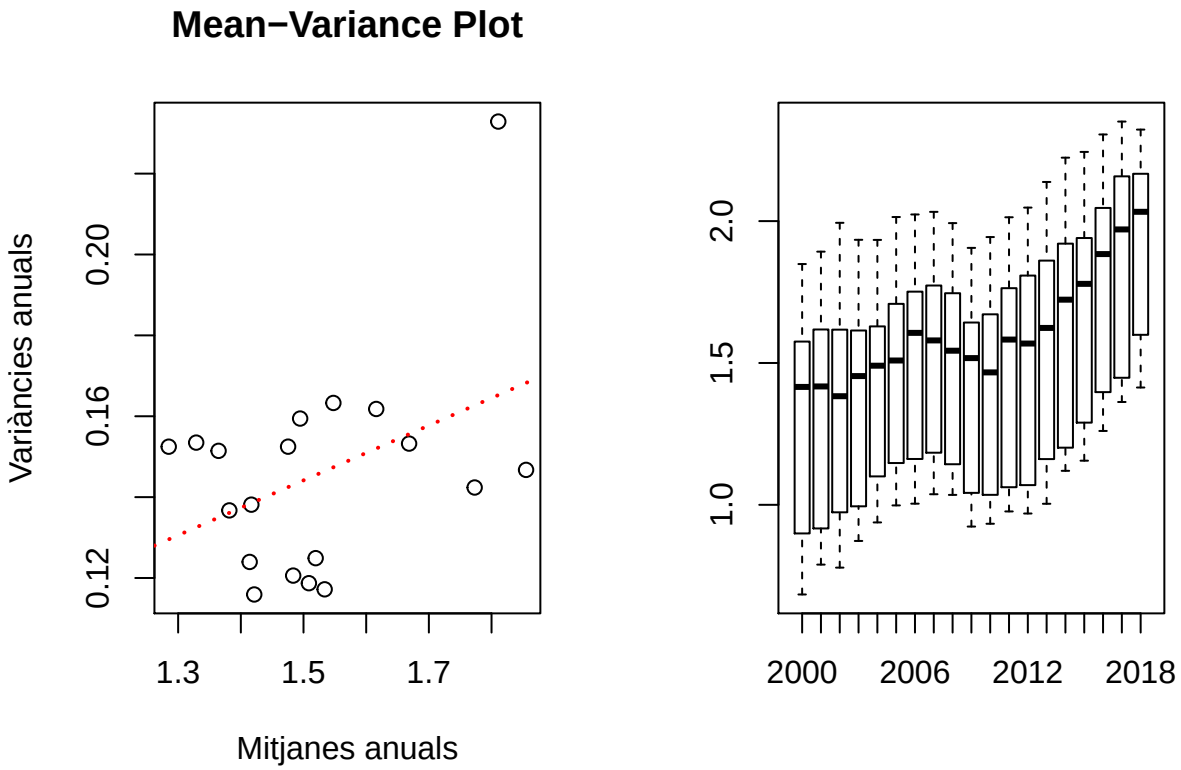
En primer lloc, s'estudiarà si es pot considerar que la variància de les dades sigui constant en el temps. Ja s'ha comentat que a simple vista semblava que no. Tot i així es comprova amb un plot de la variància front la mitjana i un *boxplot* de les dades cada 12 mesos (que és la freqüència de les nostres dades).

```
## Warning in matrix(serie, nr = 12): data length [227] is not a sub-multiple
## or multiple of the number of rows [12]
```

```
## Warning in matrix(serie, nr = 12): data length [227] is not a sub-multiple
## or multiple of the number of rows [12]
```



Tal i com s'havia observat a simple vista, la variància augmenta a mesura que agumenta la mitja. Per tant, no podem assumir variància constant. Amb el *boxplot* es confirma aquesta hipòtesis. Així doncs, es procedeix a realitzar una transformació logarítmica de la sèrie per homogeneïtzar la variància. Els resultats obtinguts són els següents:



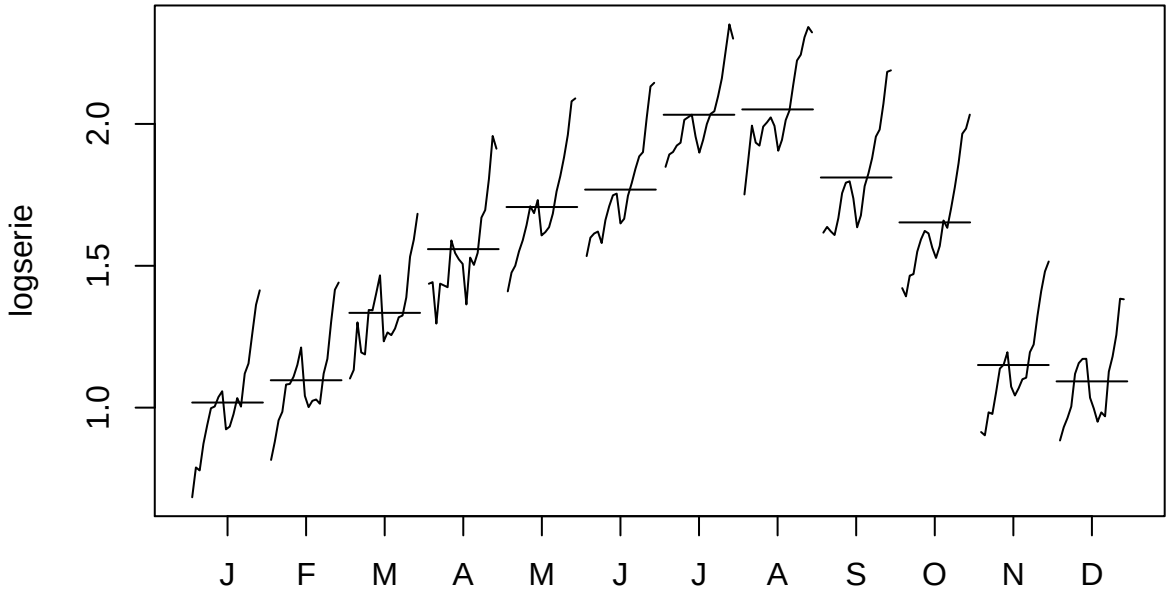
S'observa que la variància s'ha homogeneïtzat, és a dir, ja es pot considerar constant.

2.2.2 Patró estacional

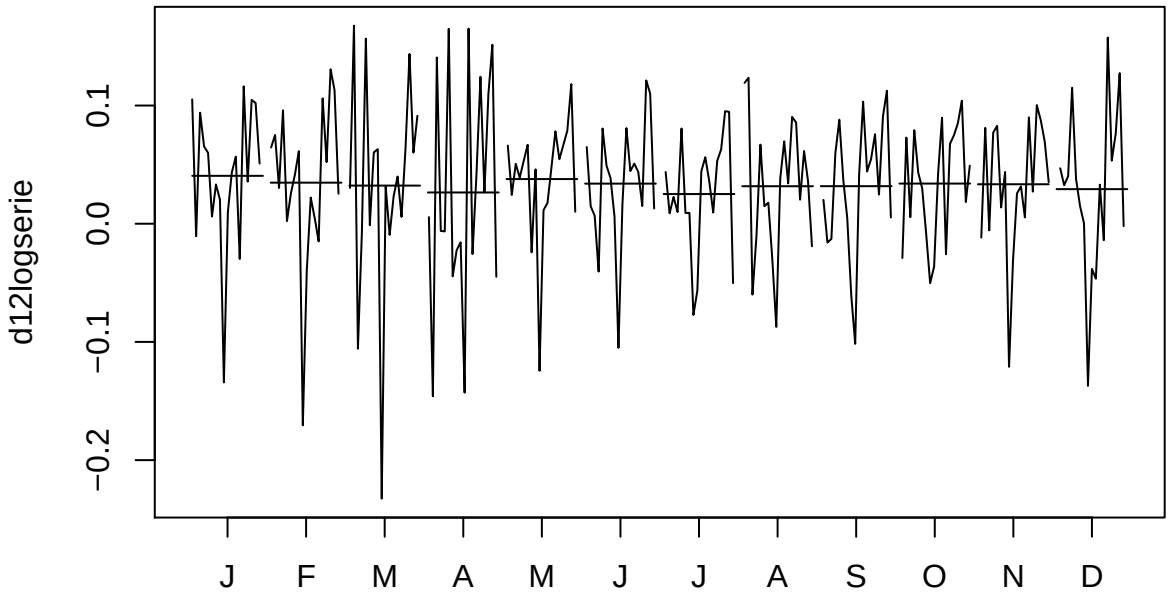
En segon lloc, s'estudiarà l'existència d'un patró estacional en les nostres dades. En cas que hi sigui present, es realitzarà una diferenciació d'ordre 12, és a dir,

$$W_t = X_t - X_{t-12} = (1 - B^{12})X_t$$

on B és el *backshift operator*, per eliminar aquest patró. Es realitza un *monthplot* per comprovar-ne l'existència.



Tal i com s'havia comenta, s'observa una clara pujada de la presència de turistes durant els mesos d'estiu i una baixada en picat en l'entrada de l'hivern/tardor. Així doncs, és necessària una diferenciació d'ordre 12 per eliminar aquest patró.



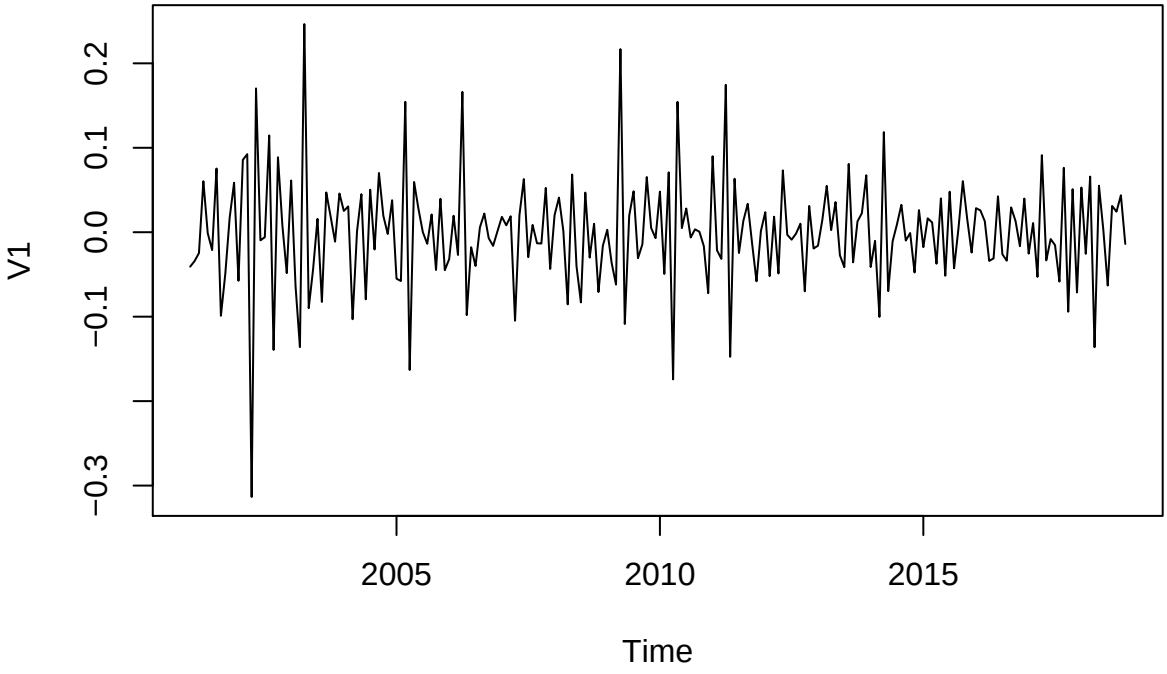
S'observa que amb una diferenciació d'ordre 12 s'ha eliminat el patró estacional. Ara bé, la mitjana de la sèrie encara no és constant.

2.2.3 Mitjana constant

Per últim, es vol aconseguir que la sèrie tingui mitjana constant igual (i si és possible igual a 0) per a poder considerar definitivament la sèrie com un procés estacionari. Per aconseguir-ho, es realitzaran diferenciacions regulars de la sèrie fins que s'obtingui el resultat desitjat

$$W_t = X_t - X_{t-1} = (1 - B)X_t$$

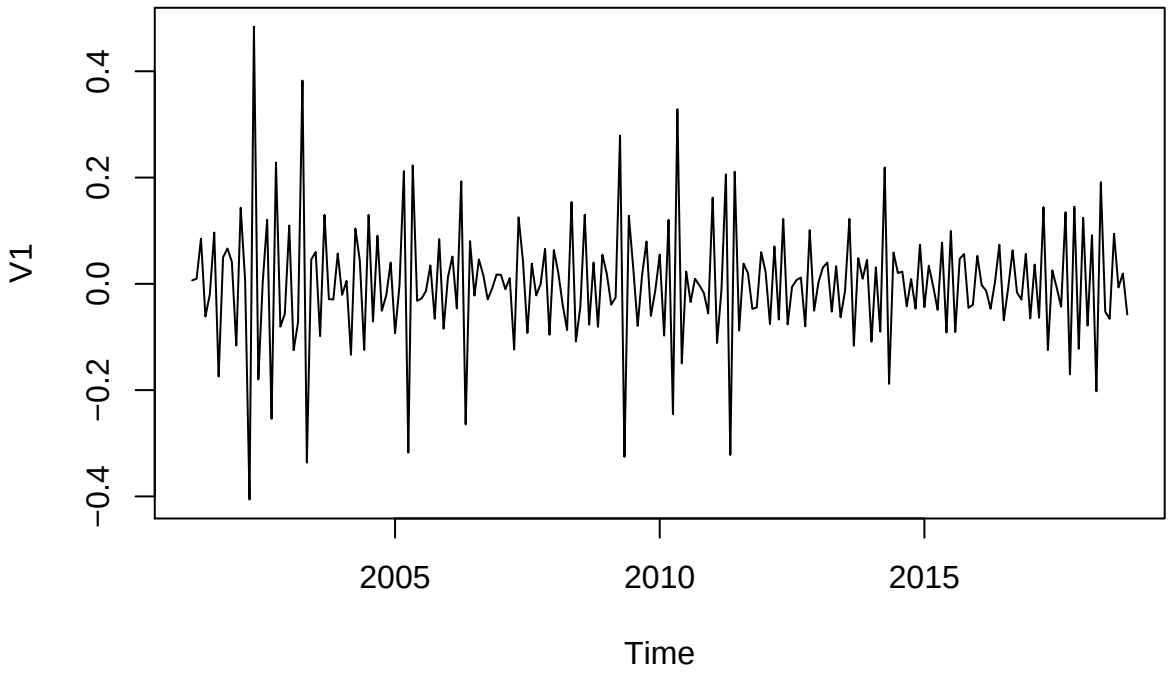
Es realitza la primera diferenciació. Els valors de mitjana i variància aconseguits són els seguents:




```
## [1] -0.0003272785
```

```
##             V1
## V1 0.004266965
```

Com es pot observar, la mitjana del procés diferenciat regularment un cop es pot considerar constant i nula. Ara bé, es mira de diferenciar un segon cop i s'observa que la variància augmenta i, per tant, es té *overdifferentiation*.



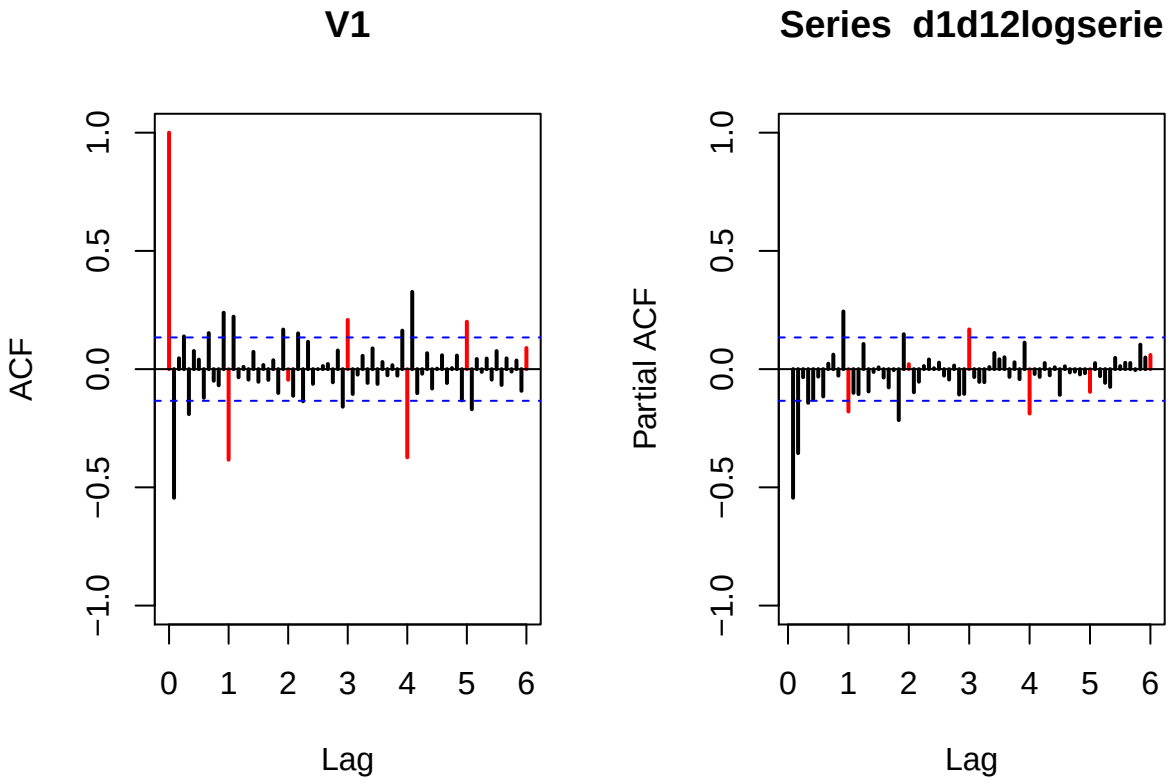

```
## [1] 0.0001260592
```

```
##           V1
## V1 0.0132293
```

En definitiva, la sèrie transformada pel logaritme, diferenciada un cop i amb una diferenciació d'ordre 12 per eliminar el patró estacional (`d1d12logserie`) és un procés estacionari de mitjana 0.

2.3 ACF/PACF de les dades i proposta de models

Tot seguit, es realitza un anàlisi de les funcions *AutoCorrelació* i de *Correlació Parcial* de la sèrie transformada, és a dir, de la sèrie estacionària.



2.3.1 Models proposats per la part regular (p,d,q)

En relació a la part regular de la sèrie, en la funció d'AutoCorrelació (ACF) s'observa un decreixement exponencial alternat en tots els valors. S'observen també valors fora de la banda de confiança en retards llunyans, però poden ser assignats a la aleatorietat del cas 5%. Per tant, en aquest cas, es proposaria $q = 0$. En tot cas, si es volgués mirar d'incloure el primer valor que sobresurt més que la resta, es podria considerar també $q = 1$.

Pel que fa a la funció de Correlació Parcial (PACF) s'observa que els dos primers valors sobresurten més significativament que la resta. La resta de valors es poden considerar nuls, ja que o bé estan dintre de l'interval de confiança, o bé es poden assignar al cas d'aleatorietat del 5%. Per tant, per la part regular, es proposaria $p = 2$.

Donat que s'ha realitzat diferenciació 1 cop, es té que $d = 1$. Per tant, els models proposats per la part regular serien $AR(2)$ o, en tot cas, $ARMA(1, 1)$ sobre la sèrie transformada regular.

2.3.2 Models proposats per la part estacional (P,D,Q)

En relació a la part estacional de la sèrie, en la funció d'AutoCorrelació (ACF) s'observa que el primer valor es força significatiu, però també ho són el tercer, el quart i el cinquè, sobretot el quart. Donat que volem intentar proposar un model simplificat, es proposa $Q = 0$.

Pel que fa a la funció de Correlació Parcial (PACF) s'observa que sobresurt el primer valor una mica i també sobresurten el tercer i el quart valor. Ara bé, no sobresurten de manera tant significativa com en el cas dels valors del ACF i, per tant, podem assignar-ho al cas d'aleatorietat del 5%. Per tant, en aquest cas, es proposaria $P = 1$. En tot cas, es podria proposar $P = 4$ per mirar d'incluir aquests valors que sobresurten de la banda de confiança.

Donat que s'ha realitzat una diferenciació d'ordre 12 per eliminar el patró estacional, es té que $D = 1$. Per tant, el model proposat per la part regular seria un $AR(1)$

2.4 Models proposats

En conclusió, es proposen per la sèrie diferenciada els models estacionals:

-$ARMA(2,0)(1,0)_s$ -$ARMA(2,0)(4,0)_s$ -$ARMA(1,1)(1,0)_s$

I per la sèrie original, tenint en compte les diferenciacions, es proposen:

-$ARIMA(2,1,0)(1,1,0)_{12}$ -$ARIMA(2,1,0)(4,1,0)_{12}$ -$ARIMA(1,1,1)(1,1,0)_{12}$

3 Estimació dels models

A continuació, s'estimen els coeficients dels dos models proposats i es mira que tots siguin significatius. Per mirar-ho, es realitza el test següent (suposant que estem davant d'un model MA:

$$H_0 : \theta_i = 0$$
$$H_1 : \theta_i \neq 0$$

amb l'estadístic

$$\hat{t} = \frac{\hat{\theta}_i}{\text{se}(\hat{\theta}_i)}$$

~

$$t - student_{T-k}$$

, on k és el nombre total de paràmetres i T és el període. Ara bé, a la pràctica es diu que un coeficient és significant si $|\hat{t}| > 2$.

En primer lloc, s'estimen els coeficients dels models proposats, amb intercept i sense.

```
## ######## ARIMA(2,1,0)(1,1,0) ########
```

```
##
## Call:
## arima(x = d1d12logserie, order = c(2, 0, 0), seasonal = list(order = c(1, 0,
##     0), period = 12))
##
## Coefficients:
##           ar1      ar2     sar1  intercept
##       -0.7127  -0.3564  -0.3684    -0.0001
## s.e.   0.0638   0.0635   0.0649     0.0012
##
## sigma^2 estimated as 0.002251:  log likelihood = 347.48,  aic = -684.96
```

```
##
## Call:
## arima(x = logserie, order = pdq.1, seasonal = list(order = PDQ.1, period = 12))
##
## Coefficients:
##           ar1      ar2     sar1
##       -0.7127  -0.3563  -0.3684
## s.e.   0.0638   0.0635   0.0649
##
## sigma^2 estimated as 0.002251:  log likelihood = 347.47,  aic = -686.95
```

```
## ######## ARIMA(2,1,0)(4,1,0) ########
##
## Call:
## arima(x = d1d12logserie, order = c(2, 0, 0), seasonal = list(order = c(4, 0,
##     0), period = 12))
##
## Coefficients:
##           ar1     ar2     sar1     sar2    sar3     sar4  intercept
##       -0.6354  -0.314  -0.3707  -0.1910  0.0143  -0.3299     -1e-04
## s.e.   0.0669   0.066   0.0693   0.0838  0.0833   0.0779      9e-04
##
## sigma^2 estimated as 0.0019:  log likelihood = 361.93,  aic = -707.85
##
## Call:
## arima(x = logserie, order = pdq.1, seasonal = list(order = PDQ.2, period = 12))
##
## Coefficients:
##           ar1     ar2     sar1     sar2    sar3     sar4
##       -0.6354  -0.314  -0.3708  -0.1912  0.0142  -0.3299
## s.e.   0.0669   0.066   0.0693   0.0838  0.0833   0.0778
##
## sigma^2 estimated as 0.0019:  log likelihood = 361.92,  aic = -709.84
## ######## ARIMA(1,1,1)(1,1,0) ########
##
## Call:
## arima(x = d1d12logserie, order = c(1, 0, 1), seasonal = list(order = c(1, 0,
##     0), period = 12))
##
## Coefficients:
##           ar1     ma1     sar1  intercept
##       -0.1469  -0.615  -0.3517     -1e-04
## s.e.   0.0989   0.081   0.0650      8e-04
##
## sigma^2 estimated as 0.002245:  log likelihood = 347.81,  aic = -685.63
##
## Call:
## arima(x = logserie, order = pdq.2, seasonal = list(order = PDQ.1, period = 12))
##
## Coefficients:
##           ar1      ma1     sar1
##       -0.1469  -0.6149  -0.3518
## s.e.   0.0989   0.0811   0.0650
##
## sigma^2 estimated as 0.002245:  log likelihood = 347.8,  aic = -687.61
```

S'observa que, en cap dels casos, l'intercept no és significatiu i, per tant, es descarten els models amb aquest paràmetre. En el cas del model $ARIMA(2,1,0)(1,1,0)_{12}$, els tres coeficients són significatius. Ara bé, en l'altre model proposat, el model $ARIMA(2,1,0)(4,1,0)_{12}$, s'observa que el coeficient `sar3` no és significatiu, però la resta de coeficients sí que ho són. Per últim, el model $ARIMA(1,1,1)(1,1,0)_{12}$ té el coeficient `ar1` no significatiu i els altres dos significatius.

En termes de *loglikelihood* i de *AIC*, el model que sembla el millor és el model $ARIMA(2,1,0)(4,1,0)_{12}$, que

és el millor tant en *AIC* com en *loglikelihood*. Dels altres dos models, donat que el $ARIMA(2,1,0)(1,1,0)_{12}$ té tots els coeficients significatus, es descarta el model $ARIMA(1,1,1)(1,1,0)_{12}$ tot i tenir una mica millor l'*AIC* i la *loglikelihood*. Així doncs, *a-priori*, s'escolliria el primer model $ARIMA(2,1,0)(4,1,0)_{12}$ com a millor model. Tot i així, es realitzarà la validació i la predicció dels dos models escollits en aquest pas.

```
##       ar1       ar2      sar1 intercept
##      TRUE      TRUE      TRUE     FALSE
```

```
##  ar1  ar2 sar1
## TRUE TRUE TRUE
```

```
##       ar1       ar2      sar1      sar2      sar3      sar4 intercept
##      TRUE      TRUE      TRUE      TRUE     FALSE      TRUE     FALSE
```

```
##   ar1   ar2  sar1  sar2  sar3  sar4
##  TRUE  TRUE  TRUE  TRUE FALSE  TRUE
```

```
##       ar1       ma1      sar1 intercept
##     FALSE      TRUE      TRUE     FALSE
```

```
##   ar1   ma1  sar1
## FALSE  TRUE  TRUE
```

4 Validació dels Models

Tot seguit, es realitzarà la validació dels dos models proposat. En el procés de validació es realitzarà un anàlisi dels residus (Z_t) dels models, es comprovarà que aquests siguin estacionaris i invertibles, es verificarà la seva estabilitat i s'evaluarà la seva capacitat de previsió.

4.1 Estudi dels residus dels models

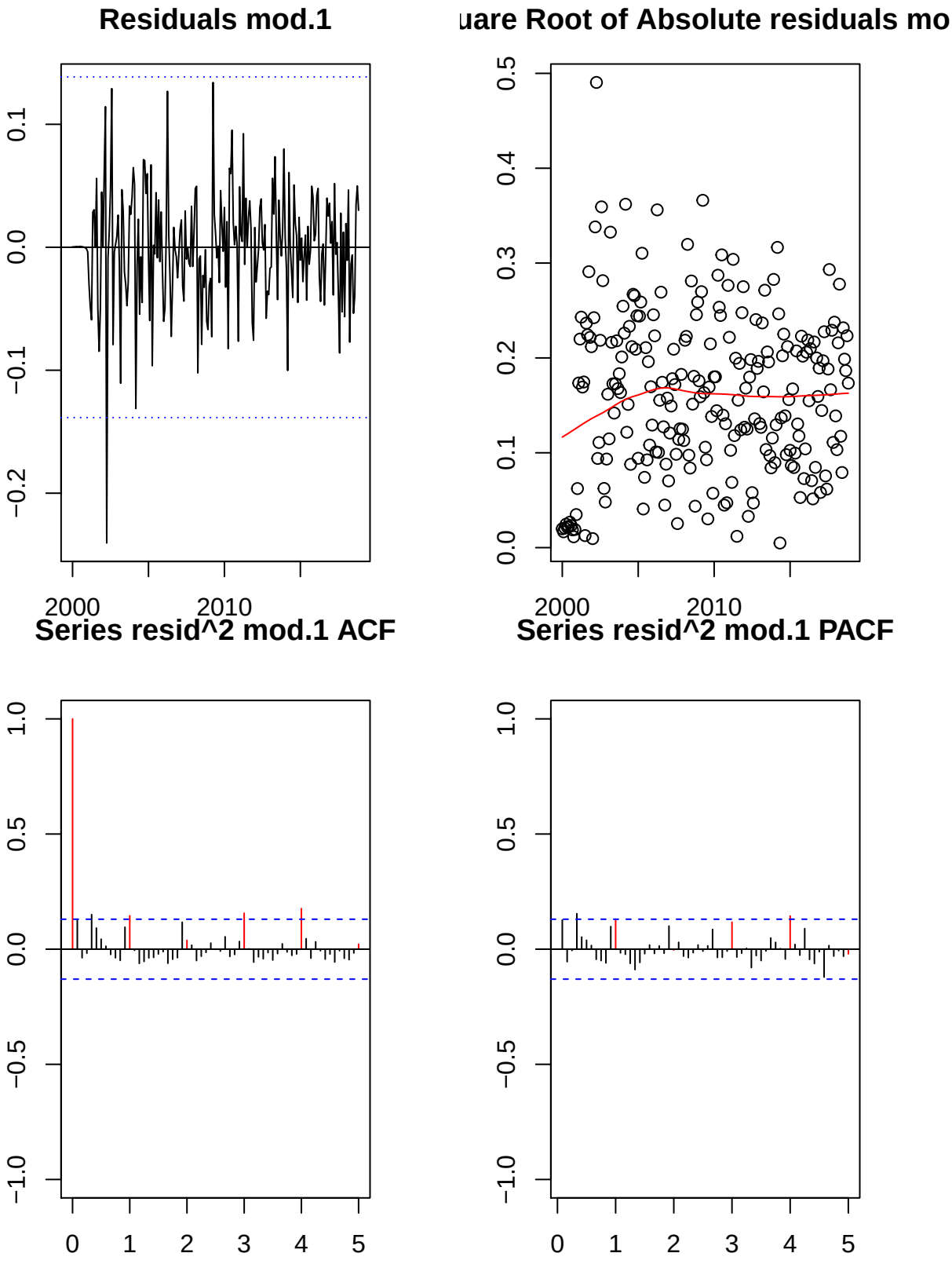
Així doncs, en primer lloc, s'estudiaràn els residus del model i es comprovaran els següents aspectes:

- Homogeneïtat de la variància residual (σ_Z^2 constant).
- Normalitat ($Z_T \sim$ Normal).
- Independència ($\rho(k) = 0 \ \forall k > 0$).

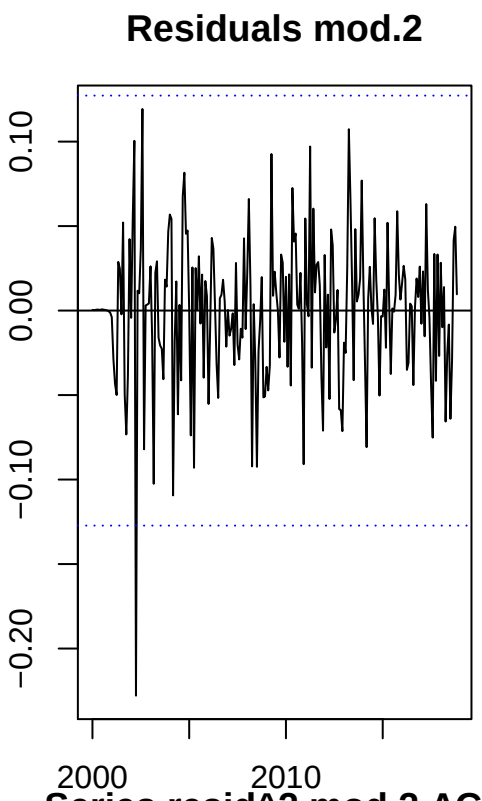
4.1.1 Homogeneïtat de la variància

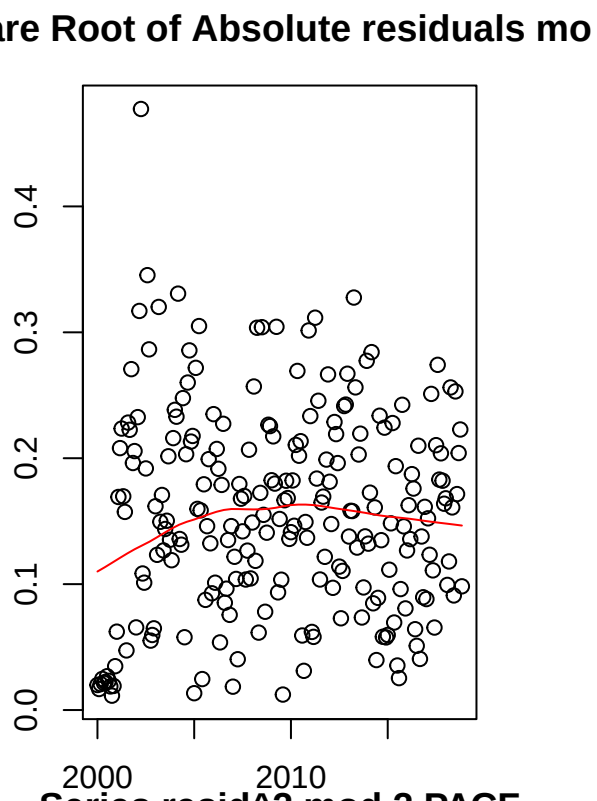
Per comprovar l'homogeneïtat de la variància dels residus, s'analitzen el plot dels mateixos residus, el plot de l'arrel quadrada del seu valor absolut i les funcions ACF i PACF del seu quadrat.

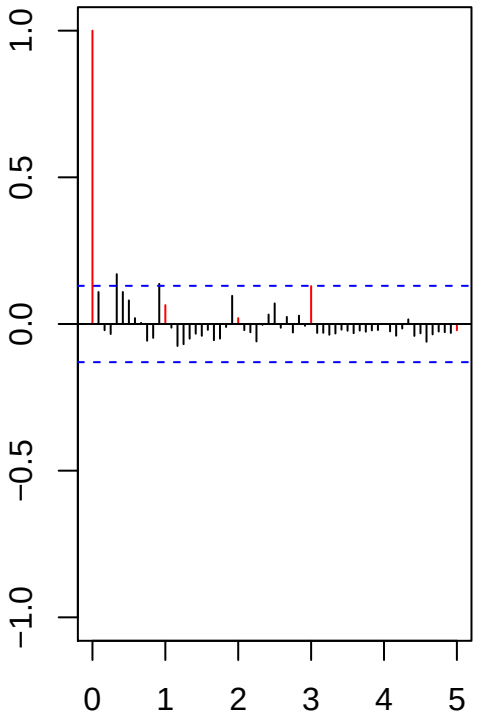
En el cas del primer model (`mod.1`) no s'observa cap tipus de patró (ni creixent ni decreixent) en el plot dels residus o en el plot de l'arrel quadrada del seu valor absolut. A més, en l'ACF i el PACF del quadrat dels residus tots els valors estan dintre de la banda de confiança i, per tant, els podem considerar nuls.

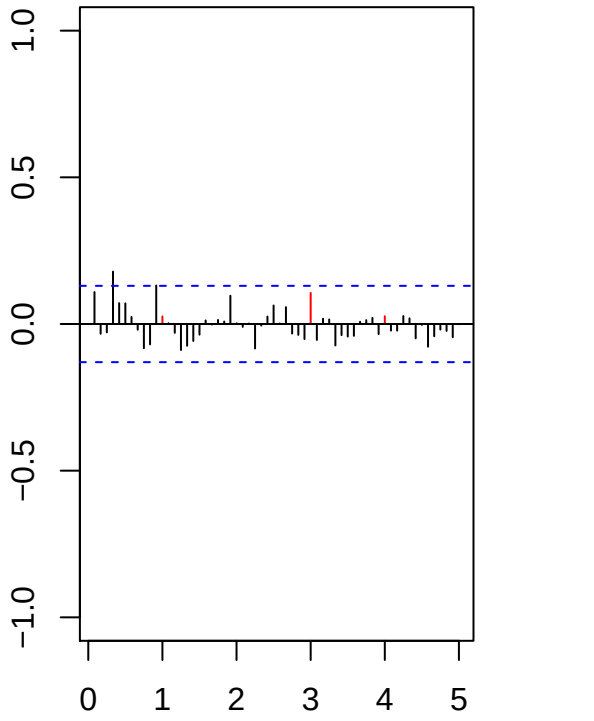


En el cas del segon model (`mod.1`), es poden extreure les mateixes conclusions que en el primer model i, per tant, també es pot assumir homogeneïtat de variància residual.


Residuals mod.2
0.10
0.00
−0.10
−0.20
2000
2010


uare Root of Absolute residuals mo
0.4
0.3
0.2
0.1
0.0
2000
2010

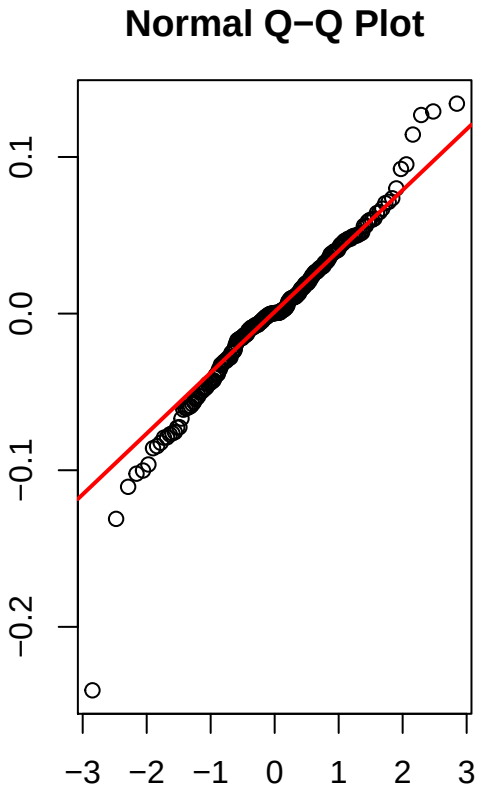

Series resid^2 mod.2 ACF
1.0
0.5
0.0
−0.5
−1.0
0
1
2
3
4
5

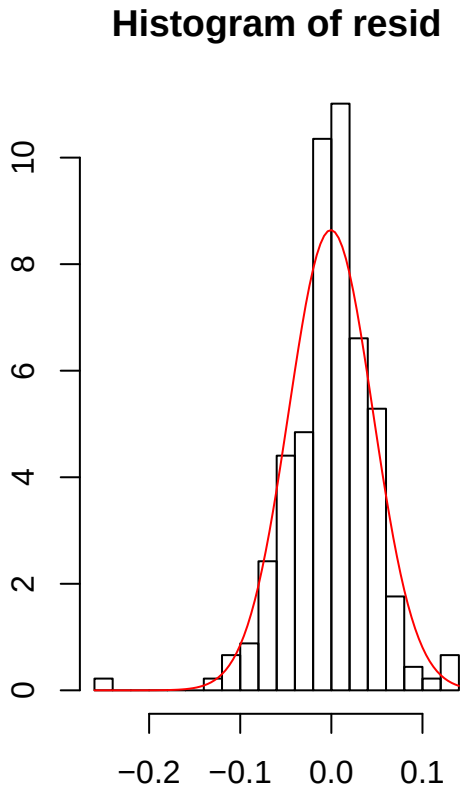

Series resid^2 mod.2 PACF
1.0
0.5
0.0
−0.5
−1.0
0
1
2
3
4
5

4.1.2 Normalitat

Per comprovar la normalitat dels residus dels models proposats s'estudiarà el Q-Q plot, l'histograma dels residus amb la normal que s'hauria de seguir sobreposada i es realitzarà el test de Sharipo-Wilks.

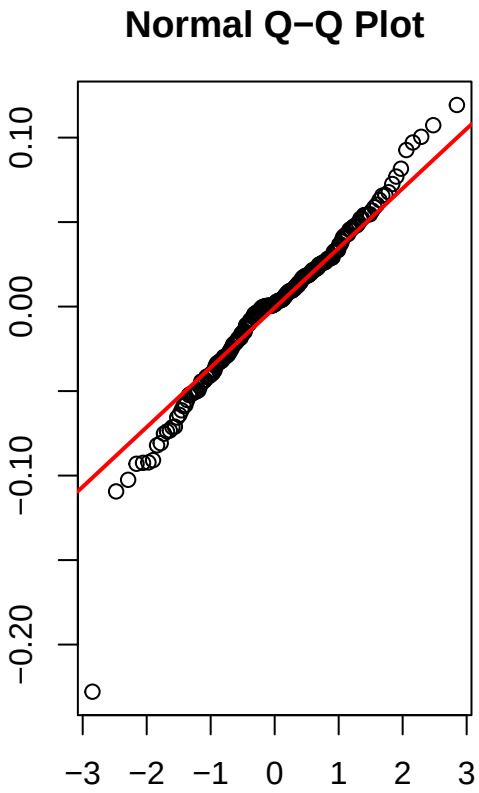
En el cas del model `mod.1`, s'observa en el Q-Q plot que els quartils es situen sobre la línia dels quartils teòrics i que l'histograma s'ajusta a la distribució normal a la que s'hauria d'ajustar (tot i tenir les dues barres més grans una mica per fora de la corba normal). A més, el *p-value* del test de Sharipo-Wilks és 1.713×10^{-05}, menor que 0.05 i, per tant, es pot assumir la hipòtesi de normalitat en els residus.

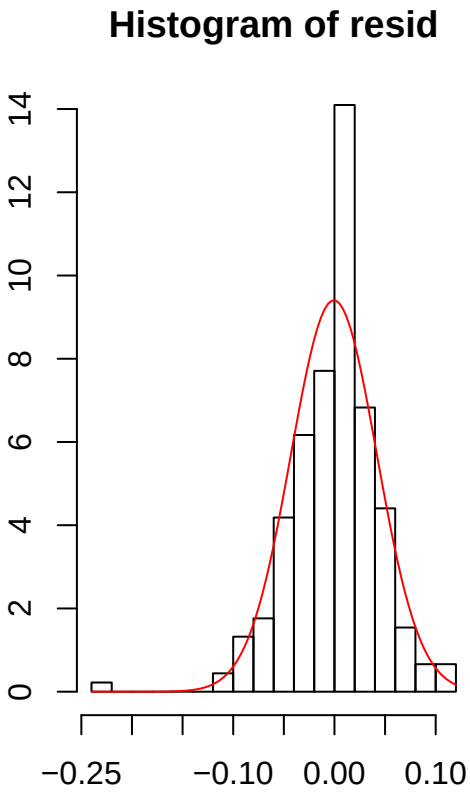





```
##
##  Shapiro-Wilk normality test
##
## data:  resid(model)
## W = 0.96408, p-value = 1.713e-05
```

En el cas del model `mod.2`, les conclusions que s'extreuen són les mateixes. En aquest cas, el *p-value* és de 7.675×10^{-05}. Per tant, també assumim normalitat en aquest cas. En l'histograma, en aquest cas, només hi ha una barra que sobresurt de la corba normal.



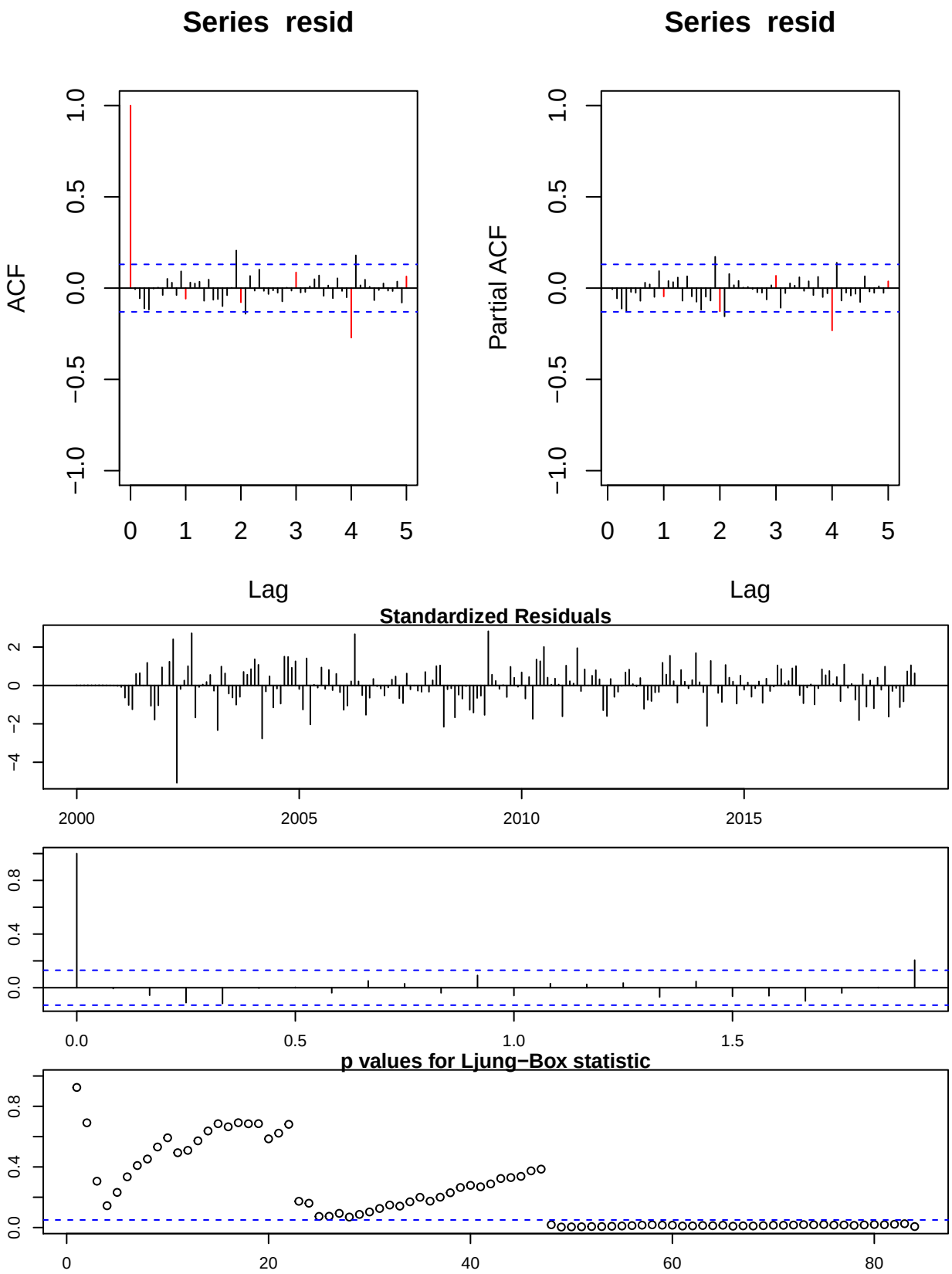



```
##
##  Shapiro-Wilk normality test
##
## data:  resid(model)
## W = 0.96197, p-value = 9.536e-06
```

4.1.3 Independència

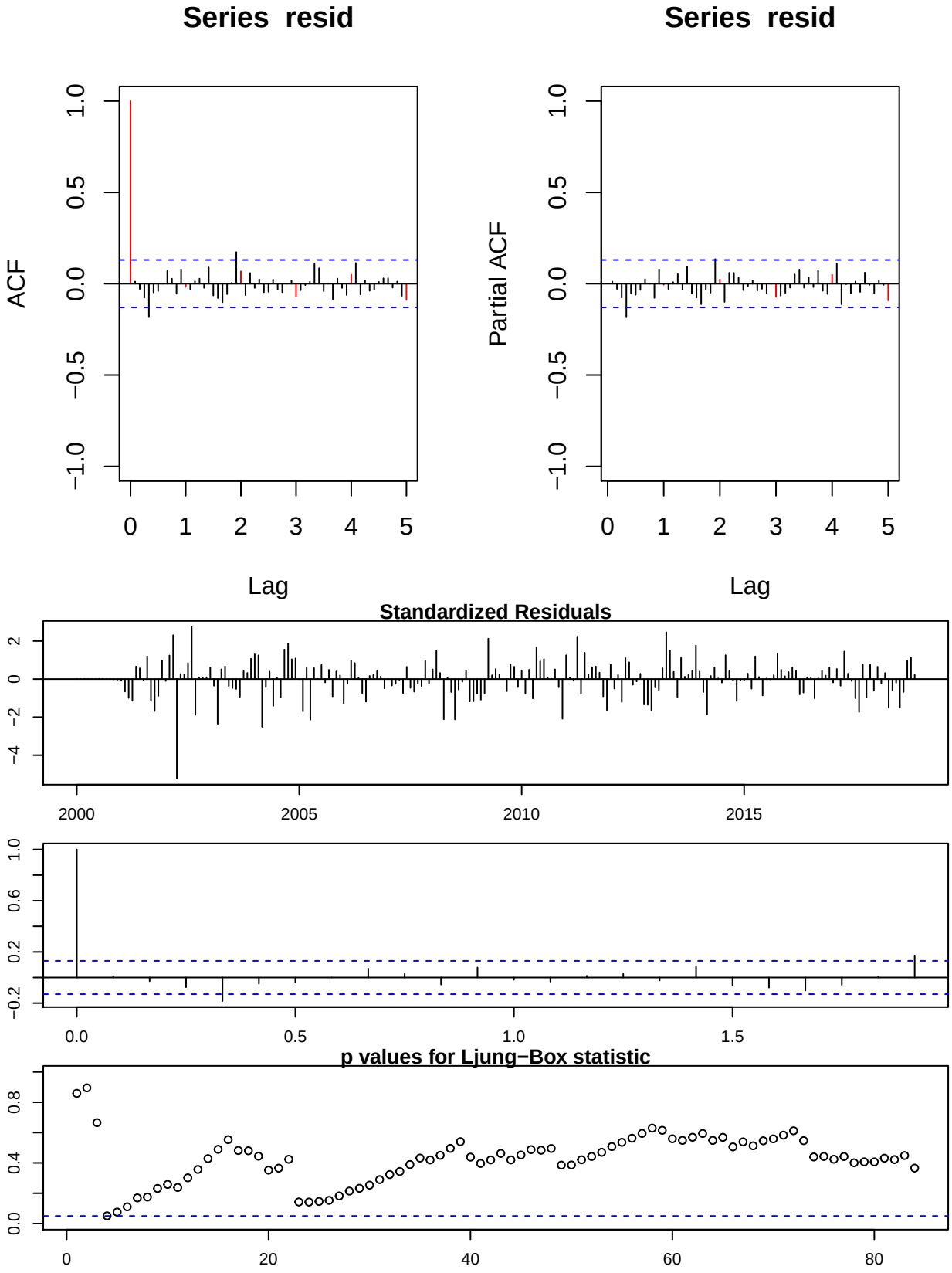
Per comprovar la independència residual, és a dir, que $\rho(k) = 0 \ \forall k > 0$ s'estudiarà el ACF i el PACF dels residus i es realitzarà el test de Ljung-Box.

Pel que fa al primer model, en primer lloc observem que les funcions ACF i PACF prenen valors pràcticament iguals (menys alguns retards llunyans que en un són positius i en l'altre negatius, però estan dins la banda de confiança en els dos casos), cosa que ja fa intuïr que es complirà la independència. Els residus estandaritzats prenen valors dintre de la franja de (-2,2), la gran majoria, que és el comportament esperat. Ara bé, els *p-values* del test Ljung-Box tenen valors superiors a 0.05 en els primers retards però en els retards llunyans no es pot assumir la independència. Tot i així, donat que en els primers retards sí que es té independència, s'assumeix aquesta hipòtesi pel model.



En el segon model, en canvi, observem més diferències entre les gràfiques del ACF i el PACF. Ara bé, en aquest cas els p-values del test de Ljung-Box estan **tots** per sobre de 0.05 i, per tant, podem assumir independència

en tots els residus.



4.1.4 Estacionaritat i invertibilitat dels models

Per analitzar l'estacionaritat i la invertibilitat dels models proposats, s'expresaran els models com a models $AR(\infty)$ i $MA(\infty)$:

$$(1 - \phi_1 B - \cdots - \phi_p B^p) X_t = (1 + \theta_1 B + \cdots + \theta_q B^q) Z_t$$

$$AR(\infty): \quad \frac{1 - \phi_1 B - \cdots - \phi_p B^p}{1 + \theta_1 B + \cdots + \theta_q B^q} X_t = (1 - \pi_1 B - \pi_2 B - \cdots) X_t = Z_t$$

$$MA(\infty): \quad \frac{1 + \theta_1 B + \cdots + \theta_q B^q}{1 - \phi_1 B - \cdots - \phi_p B^p} Z_t = (1 + \psi_1 B + \psi_2 B + \cdots) Z_t = X_t$$

A partir d'aquí el models seran *invertibles* si el mòdul de totes les arrels del polinomi característic $\theta_q(B) = 1 + \theta_1 B + \cdots + \theta_q B^q$ és major que 1, és a dir, si $\sum_{i \geq 0} \pi_i^2 < \infty$. Per per altra banda, seran *estacionaris* si el mòdul de totes les arrels del polinomi característic $\phi_q(B) = 1 - \phi_1 B - \cdots - \phi_q B^q$ és major que 1, és a dir, si $\sum_{i \geq 0} \psi_i^2 < \infty$.

En el cas del primer model, s'observa que es compleixen totes les condicions i, per tant, el `mod.1` és estacionari i invertible.

```
##
## Modul of AR Characteristic polynomial Roots:  1.086772 1.086772 1.086772 1.086772 1.086772 1.086772 1
```

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

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

```
##
## --------------------
```

```
##        psi 1        psi 2        psi 3        psi 4        psi 5
## -0.712679071  0.151578747  0.145923866 -0.158009351  0.060612511
##        psi 6        psi 7        psi 8        psi 9       psi 10
##  0.013106633 -0.030939043  0.017379287 -0.001361261 -0.005222666
##       psi 11       psi 12       psi 13       psi 14       psi 15
##  0.004207147 -0.369553389  0.261873822 -0.054948031 -0.054153898
##       psi 16       psi 17       psi 18       psi 19       psi 20
##  0.058174130 -0.022162680 -0.004934467  0.011413980 -0.006376192
```

```
##
## Pi-weights (AR(inf))
```

```
##
## --------------------
```

```
##       pi 1       pi 2       pi 3       pi 4       pi 5       pi 6
## -0.7126791 -0.3563327  0.0000000  0.0000000  0.0000000  0.0000000
##       pi 7       pi 8       pi 9      pi 10      pi 11      pi 12
##  0.0000000  0.0000000  0.0000000  0.0000000  0.0000000 -0.3684161
##      pi 13      pi 14      pi 15      pi 16      pi 17      pi 18
## -0.2625624 -0.1312787  0.0000000  0.0000000  0.0000000  0.0000000
##      pi 19      pi 20
##  0.0000000  0.0000000
```

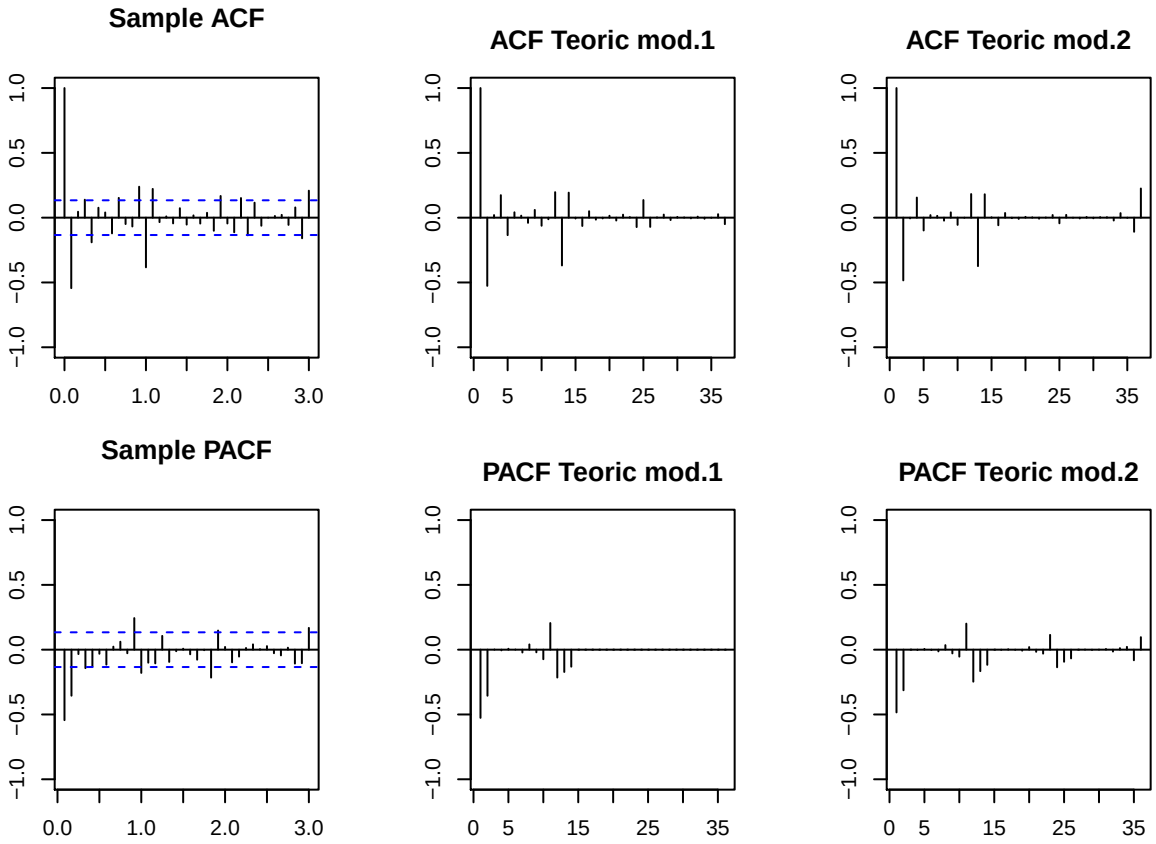
```
##
## Modul of AR Characteristic polynomial Roots:  1.01494 1.01494 1.031881 1.031881 1.031881 1.01494 1.0
```

```
##
## Modul of MA Characteristic polynomial Roots:
##
## Psi-weights (MA(inf))
##
## --------------------
##        psi 1        psi 2        psi 3        psi 4        psi 5
## -0.635389267  0.089751049  0.142465343 -0.118699949  0.030691048
##        psi 6        psi 7        psi 8        psi 9       psi 10
##  0.017767279 -0.020925160  0.007717257  0.001666378 -0.003481774
##       psi 11       psi 12       psi 13       psi 14       psi 15
##  0.001689092 -0.370760186  0.235046721 -0.032939155 -0.052868074
##       psi 16       psi 17       psi 18       psi 19       psi 20
##  0.043933663 -0.011316070 -0.006603676  0.007748794 -0.002850154
##
## Pi-weights (AR(inf))
##
## --------------------
##       pi 1       pi 2       pi 3       pi 4       pi 5       pi 6
## -0.6353893 -0.3139685  0.0000000  0.0000000  0.0000000  0.0000000
##       pi 7       pi 8       pi 9      pi 10      pi 11      pi 12
##  0.0000000  0.0000000  0.0000000  0.0000000  0.0000000 -0.3707801
##      pi 13      pi 14      pi 15      pi 16      pi 17      pi 18
## -0.2355897 -0.1164133  0.0000000  0.0000000  0.0000000  0.0000000
##      pi 19      pi 20
##  0.0000000  0.0000000
```

En el cas del segon model, també es compleix tot i, per tant, també és invertible i estacionari.

4.1.5 Comparació entre els ACF/PACF mostrals i els ACF/PACF teòrics

Per últim, comparem els valors del ACF i el PACF de les dades amb els valors teòric. S'observa que, en el cas del model `mod.2`, els valors teòrics s'aproximen gairebé perfectament als valors mostrals. En el cas del segon model també es podria dir el mateix. Per tant, ambdós models s'aproximen als valors de ACF/PACF de les mostres, potser una mica millor el `mod.2` (ja que té més coeficients per calcular els valors teòrics).



4.2 Estabilitat dels Models

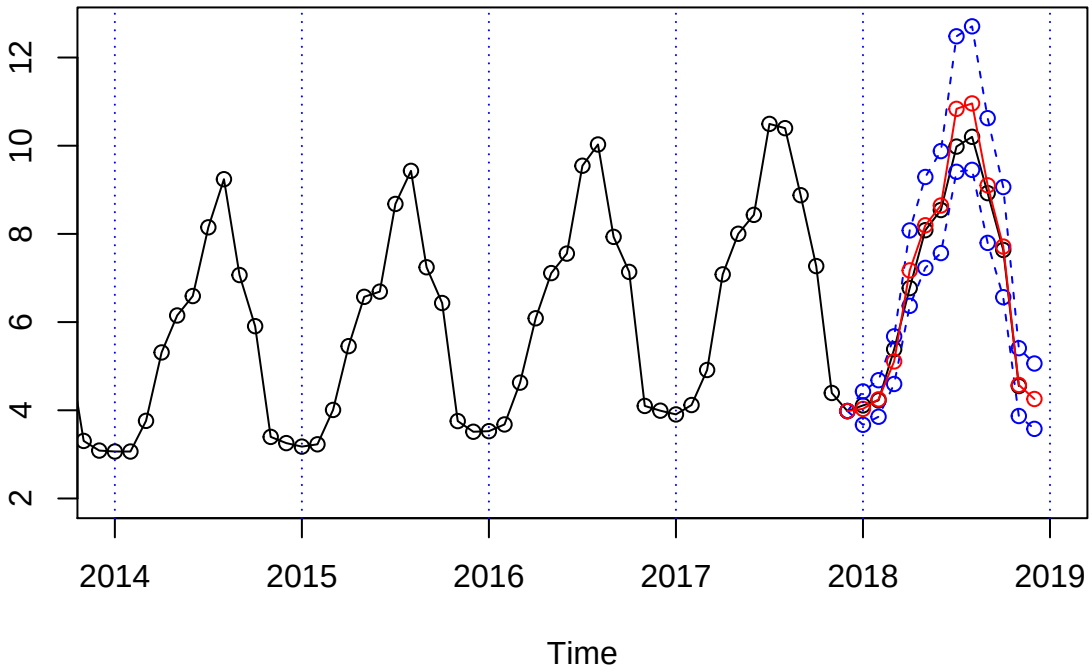
Per comprovar l'estabilitat dels models proposats, calculem els models de la serie ocultant les 12 últimes observacions, és a dir, l'últim període d'observacions. Així doncs, s'observa que el valor dels coeficients varia molt poc, de l'ordre de menys de 0.02 en gairebé tots els casos. Per tant, podem confirmar que els models són estables.

```
## ########## Model ARIMA(2,1,0)(1,1,0)12 amb i sense les 12 últimes observacions ##########

##
## Call:
## arima(x = lnserie1, order = pdq.1, seasonal = list(order = PDQ.1, period = 12))
##
## Coefficients:
##           ar1      ar2     sar1
##       -0.7127  -0.3563  -0.3684
## s.e.   0.0638   0.0635   0.0649
##
## sigma^2 estimated as 0.002251:  log likelihood = 347.47,  aic = -686.95

##
## Call:
## arima(x = lnserie2, order = pdq.1, seasonal = list(order = PDQ.1, period = 12))
##
## Coefficients:
##           ar1      ar2     sar1
##       -0.7173  -0.3619  -0.3597
```

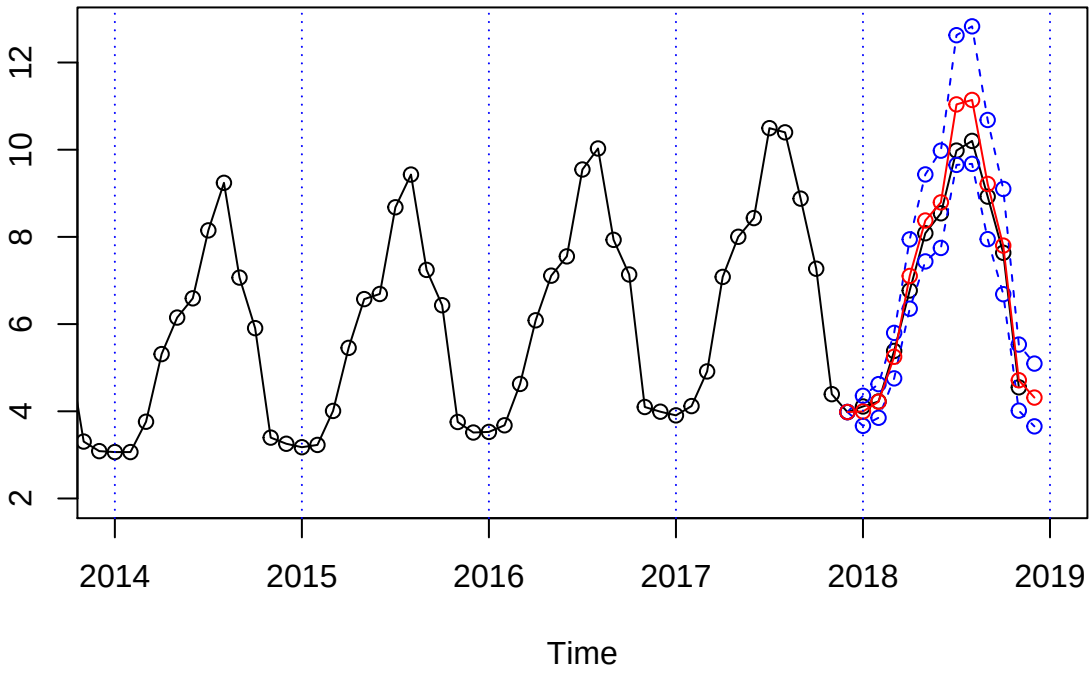
```
## s.e.   0.0654   0.0652   0.0664
##
## sigma^2 estimated as 0.002286:  log likelihood = 328.05,  aic = -648.1

## ########## Model ARIMA(2,1,0)(4,1,0)12 amb i sense les 12 últimes observacions ##########

##
## Call:
## arima(x = lnserie1, order = pdq.1, seasonal = list(order = PDQ.2, period = 12))
##
## Coefficients:
##           ar1     ar2     sar1     sar2    sar3     sar4
##       -0.6354  -0.314  -0.3708  -0.1912  0.0142  -0.3299
## s.e.   0.0669   0.066   0.0693   0.0838  0.0833   0.0778
##
## sigma^2 estimated as 0.0019:  log likelihood = 361.92,  aic = -709.84

##
## Call:
## arima(x = lnserie2, order = pdq.1, seasonal = list(order = PDQ.2, period = 12))
##
## Coefficients:
##           ar1      ar2     sar1     sar2    sar3     sar4
##       -0.6458  -0.3195  -0.3746  -0.2094  0.0023  -0.3236
## s.e.   0.0681   0.0677   0.0715   0.0857  0.0849   0.0799
##
## sigma^2 estimated as 0.001929:  log likelihood = 341.68,  aic = -669.36
```

4.3 Capacitat de predicció

A continuació s'avaluarà la capacitat de predicció dels dos models proposats fent-los predir el valor de les 12 útlimes observacions utilitzant la resta d'observacions conegudes.




```
##                tl        pr        tu     serie  error
## Dec 2017 3.982530  3.982530  3.982530  3.982530  0.000
## Jan 2018 3.670975  4.031601  4.427654  4.110137  0.079
## Feb 2018 3.853540  4.247667  4.682103  4.224826 -0.023
## Mar 2018 4.596207  5.108007  5.676796  5.383687  0.276
## Apr 2018 6.367946  7.172510  8.078727  6.770845 -0.402
## May 2018 7.229586  8.195415  9.290274  8.084173 -0.111
## Jun 2018 7.564745  8.644089  9.877435  8.541181 -0.103
## Jul 2018 9.407498 10.835715 12.480758  9.979779 -0.856
## Aug 2018 9.452025 10.959125 12.706529 10.201456 -0.758
## Sep 2018 7.796220  9.101144 10.624485  8.924326 -0.177
## Oct 2018 6.564132  7.712757  9.062374  7.635569 -0.077
## Nov 2018 3.869744  4.574708  5.408098  4.549899 -0.025
## Dec 2018 3.578129  4.255407  5.060882        NA     NA

## ###### Errors de predicció del model mod.1 ######

## [1] 0.04109617

## [1] 0.02968872
```




```
##                tl        pr        tu     serie  error
## Dec 2017 3.982530  3.982530  3.982530  3.982530  0.000
## Jan 2018 3.666119  3.995709  4.354929  4.110137  0.114
## Feb 2018 3.851341  4.219639  4.623156  4.224826  0.005
## Mar 2018 4.754528  5.250706  5.798665  5.383687  0.133
## Apr 2018 6.353051  7.103871  7.943425  6.770845 -0.333
## May 2018 7.437336  8.375909  9.432927  8.084173 -0.292
## Jun 2018 7.746343  8.791489  9.977646  8.541181 -0.250
## Jul 2018 9.651394 11.039002 12.626110  9.979779 -1.059
## Aug 2018 9.675143 11.141095 12.829164 10.201456 -0.940
## Sep 2018 7.948988  9.214578 10.681667  8.924326 -0.290
## Oct 2018 6.685670  7.799954  9.099954  7.635569 -0.164
## Nov 2018 4.014464  4.712040  5.530831  4.549899 -0.162
## Dec 2018 3.653931  4.314188  5.093751        NA     NA
```

```
## ###### Errors de predicció del model mod.2 ######
```

```
## [1] 0.04874019
```

```
## [1] 0.03802347
```

Com es pot veure, les prediccions (en vermell a la gràfica) de les últimes 12 observacions són semblants i prou bones, ja que en amdós casos s'apropen força a la realitat. A més, el valor real de les observacions (en negre a la gràfica) queda dins l'interval de confiança (en blau a la gràfica) dels valors predits. Per tant, es pot concloure que els models tenen bona capacitat de predicció. A més, els errors de predicció (l'**Error Quadràtic Mitjà** i l'**Error Absolut Mitjà**) dels dos models són semblants i molt petits (> 0.05 en ambdós casos).

4.4 Elecció de model

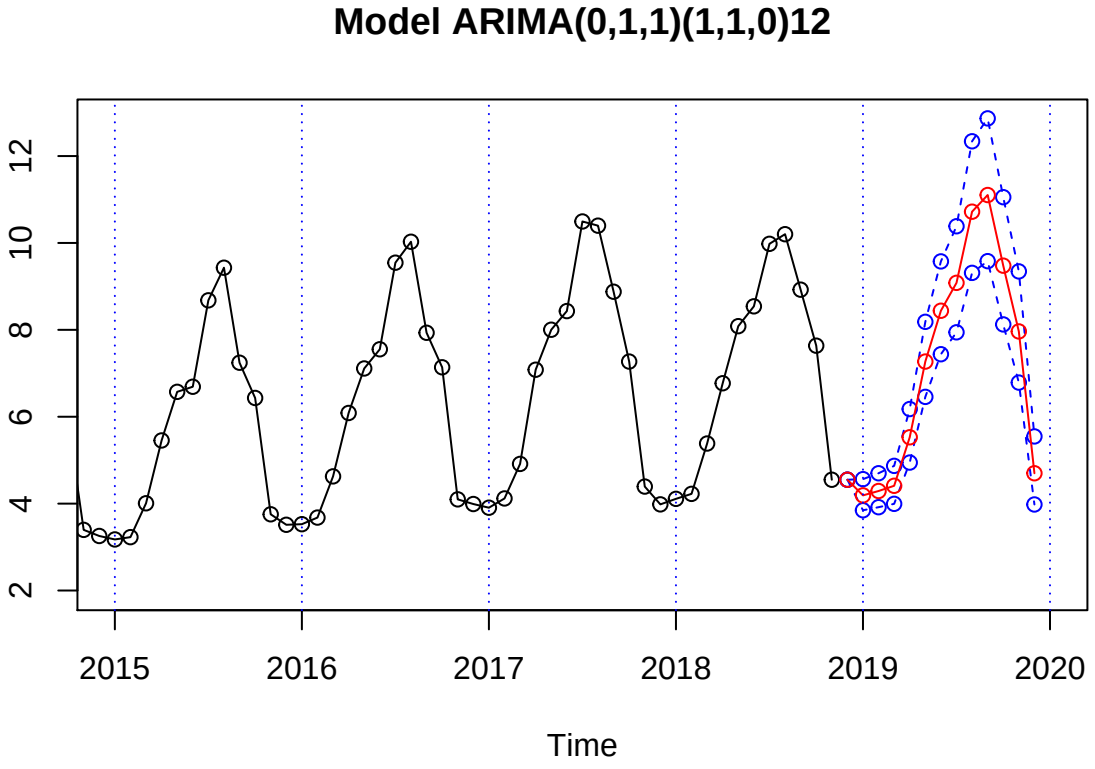
En definitiva, els dos presenten un comportament similar en la predicció de les últimes 12 observacions, però que el test d'independència ha sortit molt millor en el segon model, s'escull el segon model, el model $ARIMA(2,1,1)(4,1,0)_{12}$. A més, com ja s'havia comentat abans, aquest model era el millor en *AIC* i *loglikelihood*.

5 Predicció a llarg termini

A continuació, s'utilitza el model escollit $ARIMA(0,1,1)(1,1,0)_{12}$ per predir el valor de la sèrie els 12 mesos posteriors a l'ultima dada que es té. Com es pot observar, el valor predit (en vermell a la gràfica) sembla prou raonable per 2 motius:

- Segueix amb la tendència general de la sèrie de creixement lleu.
- Segueix amb el patró estacional vist al llarg de tota la sèrie: pujada molt pronunciada del número de turistes durant els mesos de primavera i estiu i baixada en picat els mesos de tardor i hivern.

A més, els intervals de confiança també segueixen aquestes tendències estacionals.




```
##                tl1       pr1       tu1
## Dec 2018 4.549899  4.549899  4.549899
## Jan 2019 3.846837  4.189916  4.563591
## Feb 2019 3.917671  4.290606  4.699042
## Mar 2019 3.995362  4.410566  4.868919
## Apr 2019 4.945079  5.526941  6.177267
## May 2019 6.457154  7.269534  8.184119
## Jun 2019 7.438742  8.439524  9.574947
## Jul 2019 7.942888  9.081850 10.384132
## Aug 2019 9.311565 10.719334 12.339936
```

```
## Sep 2019 9.580515 11.102785 12.866932
## Oct 2019 8.125041  9.476734 11.053296
## Nov 2019 6.786353  7.963653  9.345191
## Dec 2019 3.980613  4.698813  5.546594
```

6 Tractament de *outliers*

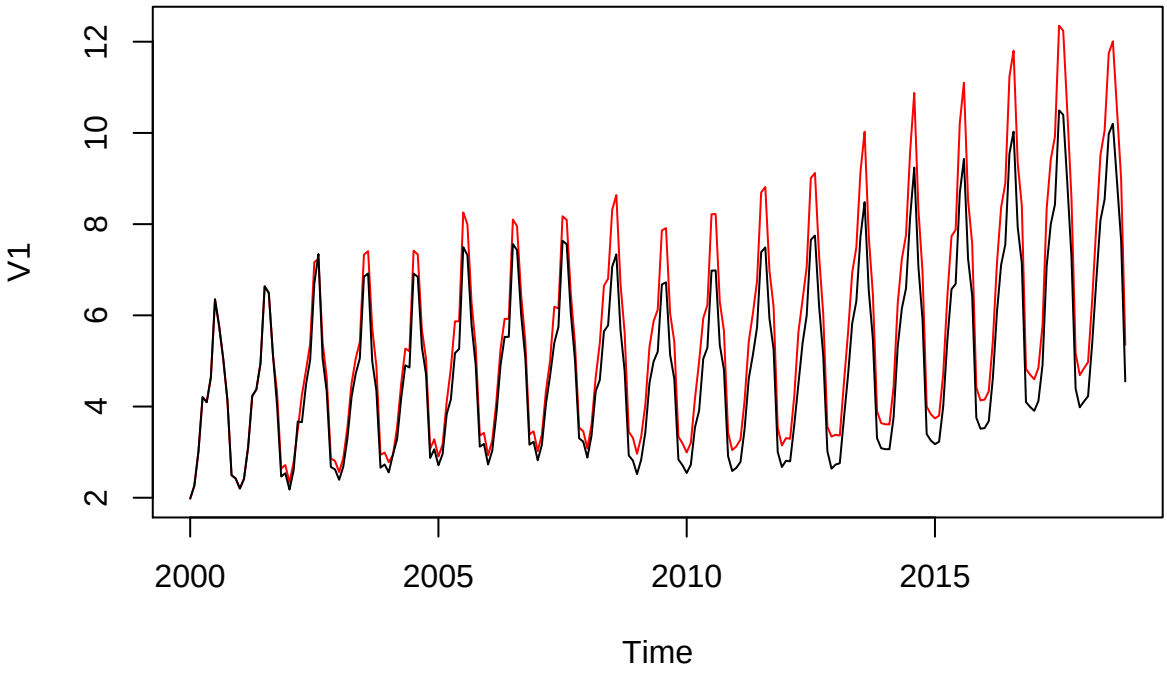
Per acabar l'anàlisi d'aquesta serie, es centrarà l'atenció en la detecció i la correcció de possibles *outliers* en la serie. Aquests valors atípics poden ser de tres tipus diferents:

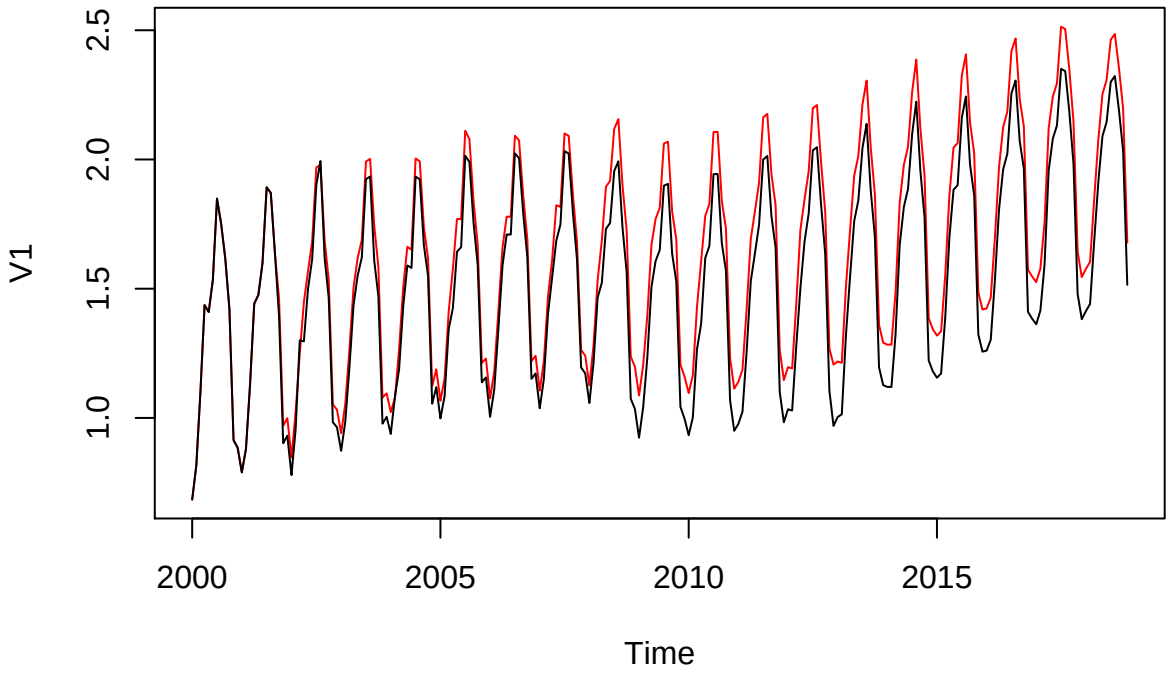
- **Outlier Aditiu (AO)**: Afecta només a un període ($X_t = 1_{t=TO}(t)$).
- **Canvi Transitori (TC)**: Afecta a un període i el seu efecte decreix exponencialment en els següents períodes ($X_t = \delta^{(t-T0)}1_{t\geq TO}(t)$).
- **Level Shift (LS)**: Afecta a un període i el seu efecte es manté durant els següents períodes ($X_t = 1_{t\geq TO}(t)$).

A continuació es mostren els *outliers* detectats, així com la seva influència sobre les dades i la data en què tenen aquest efecte. S'observa que tenim 10 valors atípics en total, dels quals la majoria són puntuals (6 són AO), 2 són TC i tenim 2 LS. L'*outlier* que més efecte ha tingut és del març del 2002, que ha tingut a més un efecte positiu, és a dir, ha fet créixer la sèrie, tot i només afectar a aquell període, ja que era era de tipus AO. En segon lloc, tenim un del tipus LS a l'abril del 2008. Aquest últim es podria associar perfectament a l'entrada en crisi econòmica del pais.

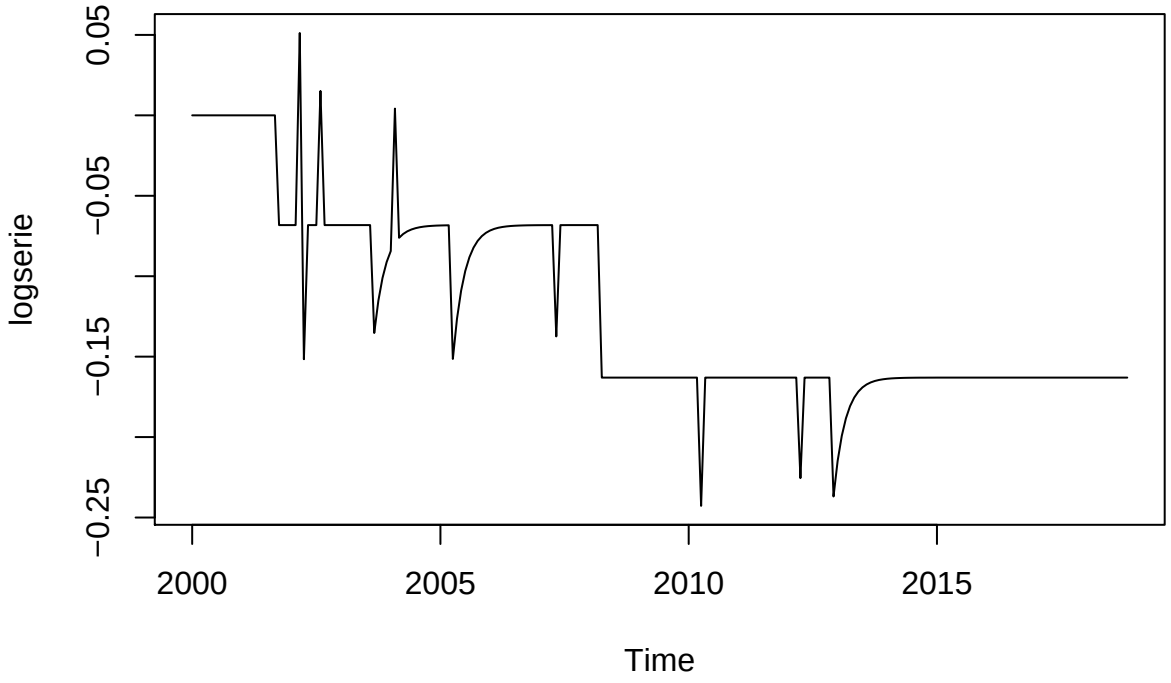
```
## [1] 0.0009590089
```

En el següent gràfic, es poden observar la gràfica amb els *outliers* (en negre) i la gràfica linealitzada, és a dir, sense *outliers* (en vermell). S'observa com hi ha outliers que han provocat que la sèrie tingui valors més baixos i outliers que han provocat que la sèrie tingui valors més alts. El fet que crida més l'atenció és que, a causa dels outliers, la sèrie des del 2005 (sobretot a partir del 2008) té valors més baixos del que hauria de tenir (s'observen els pics negres per sota dels vermells).





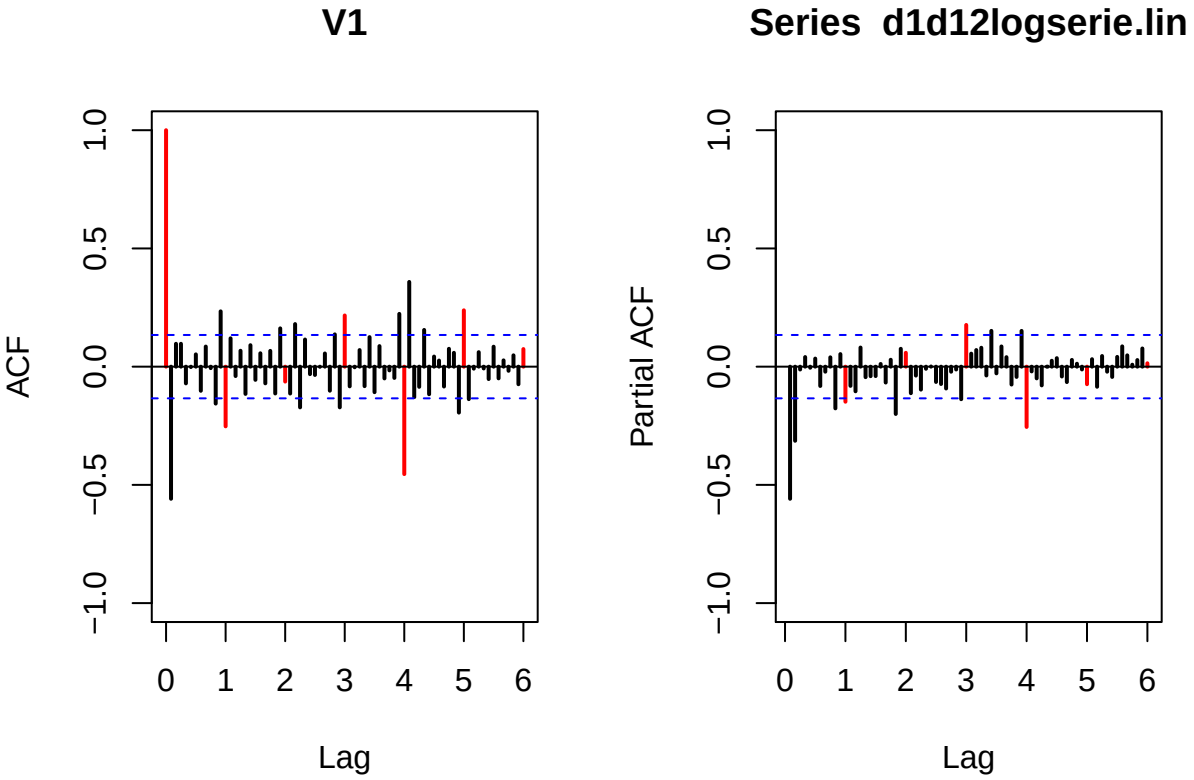
Per veure més clar l'efecte dels outliers, s'exposa la gràfica dels valors de la sèrie menys els valos de la sèrie linealitzada. Com es pot veure, l'abril del 2008 hi ha un *outlier* de tipus LS que fa que la sèrie agafi valors més petits des d'aquesta data en endavant. Crida l'atenció també els pics amunt i avall del principi de la sèrie, a les dates de març del 2002 i abril de 2002, que en només un més de diferència es va tenir una gran entrada turística el març i una baixa entrada turísica l'abril. Per últim, cal remarcar un *outlier* de tipus LS a l'octubre de 2001, que també fa que des de llavors la sèrie prengui valors inferiors.



6.1 Identificació i estimació del model per la sèrie linealitzada

Un cop eliminats els *outliers* de la sèrie, es calculen un altre cop les funcions ACF i PACF de la sèrie linealitzada. S'observa clarament que, pel que fa a la part regular, al ACF es té decreixement exponencial alternat durant tots els valors, tinguent valors infinits fora de la banda de confiança als valors inicials. Pel que fa al PACF, observem que els dos primers valos són encara més significants que en el cas del PACF de la sèrie sense linealitzar. També hi ha altres valors fora de la banda que poden associar-se al cas del 5%. Per tant, pel que fa a la part regular, igual que en el cas de la sèrie amb valors atípics, es confirma la hipòtesi que el model adequat és un AR(2).

Pel que fa a la part estacional, l'anàlisi és molt semblant al de la sèrie sense linealitzar: un ACF amb força valors fora de la banda de confiança, sobretot el quart valor (molt més que el primer) i un PACF on es podrien considerar significatius el primer, el tercer i el quart valor. Per tant, igual que en el cas de la sèrie sense linealitzar, proposem un AR(4) per la part estacional.



Ara bé, a diferència del que s'ha vist anteriorment, en aquest cas surten no significatius tant el coeficient de `sar2` com el de `sar3`

```
##
## Call:
## arima(x = logserie.lin, order = pdq.1, seasonal = list(order = PDQ.2, period = 12))
##
## Coefficients:
##           ar1      ar2     sar1     sar2    sar3     sar4
##       -0.6611  -0.3138  -0.1435  -0.1237  0.1143  -0.4575
## s.e.   0.0663   0.0662   0.0652   0.0726  0.0703   0.0684
##
## sigma^2 estimated as 0.0009259:  log likelihood = 436.23,  aic = -858.46
```

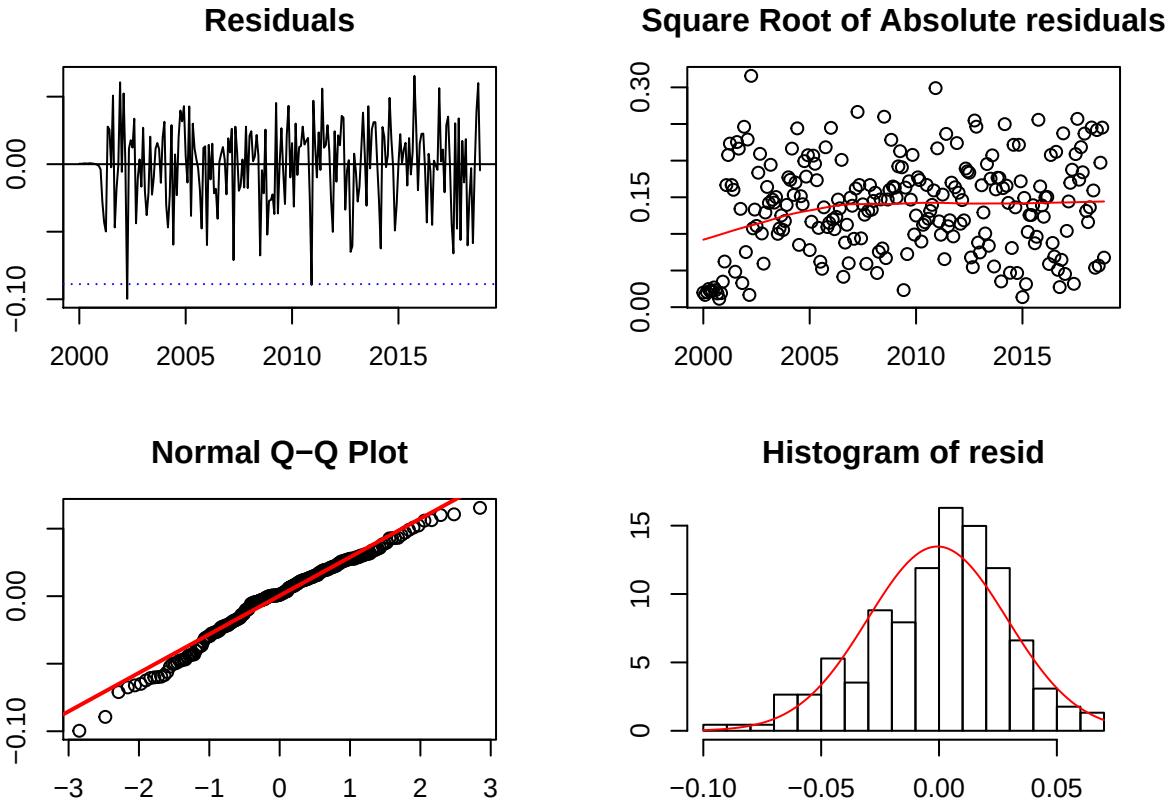
```
##   ar1   ar2  sar1  sar2  sar3  sar4
##  TRUE  TRUE  TRUE FALSE FALSE  TRUE
```

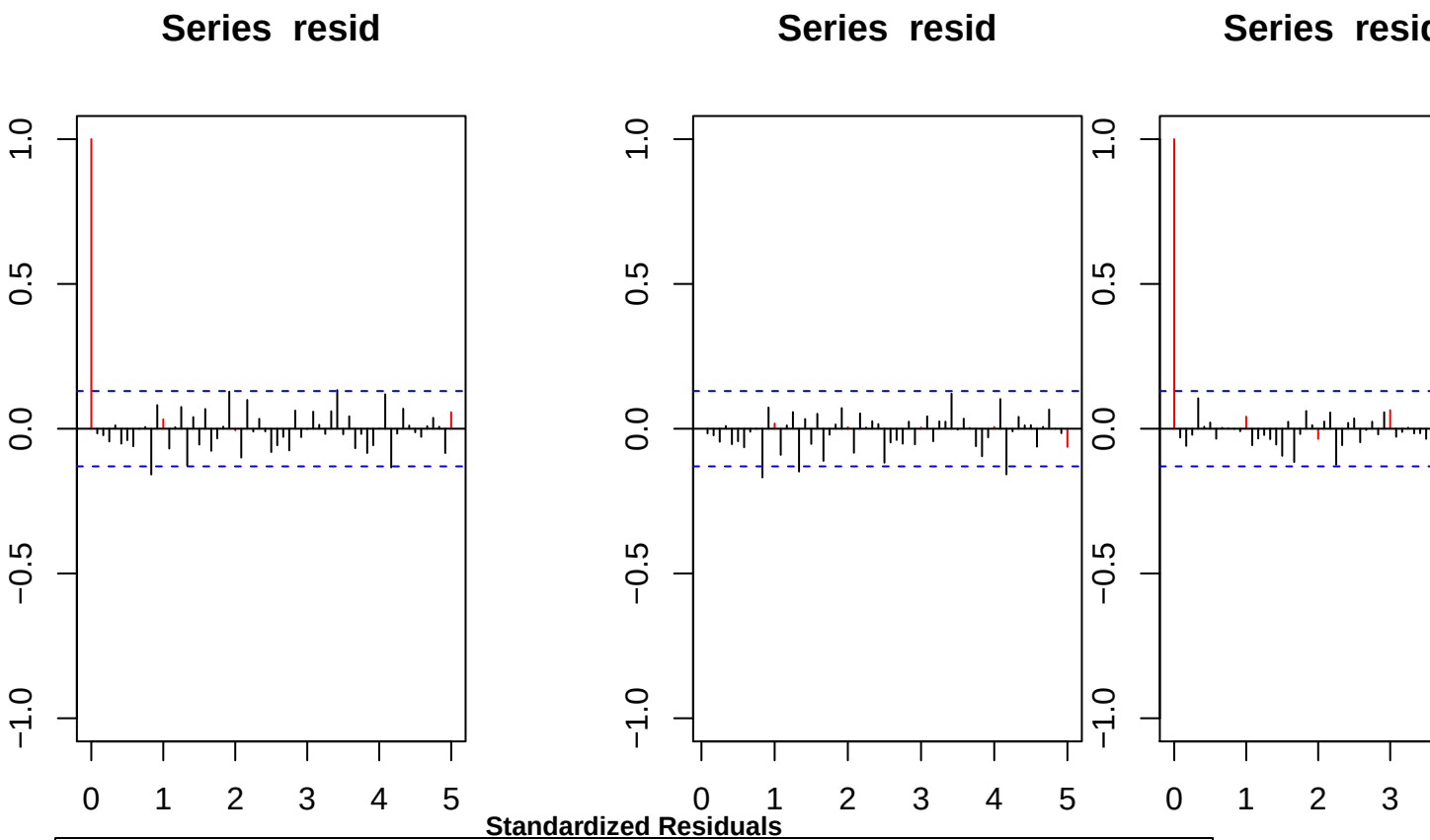
6.2 Validació del model per la sèrie linealitzada

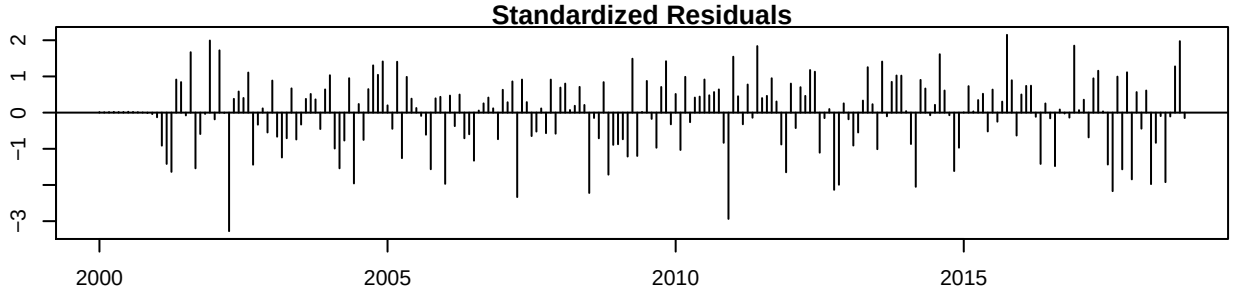
Pel que fa a la validació del model per la sèrie linealitzada, s'observa en les gràfiques els mateixos anàlisis realitzats anteriorment, és a dir:

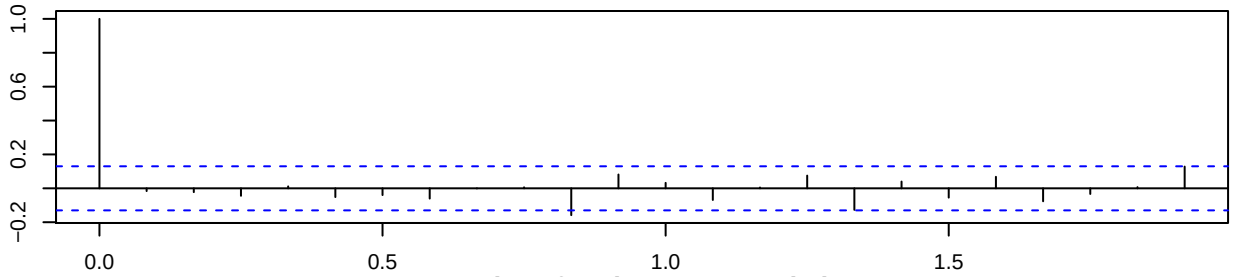
- Es conclueix que es poden assumir les hipòtesis d'homogeneïtat en la variància dels residus (no patrons en les gràfiques de la seva variància i ACF i PACF dels residus al quadrat nuls), de normalitat dels residus (Q-Q plot amb relació lineal, histograma s'ajusta a la corba normal) i d'independència dels residus (*p-values* de Ljung-Box per sobre de 0.05 i ACF i PACF dels residus molt iguals).
- Es pot dir que el model és causal i invertible, ja que totes les arrels dels polinomis característics tenen mòdul major que un.
- El ACF i el PACF teòrics són molt semblants al ACF i PACF mostrals.

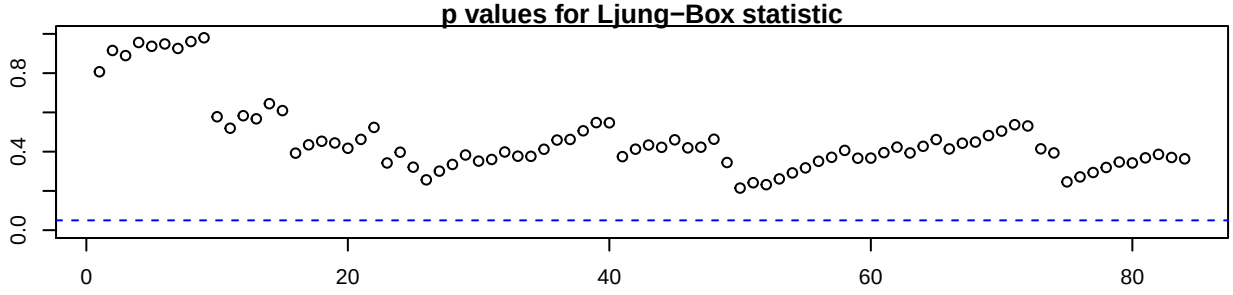
Per tant, es conclou que el model per la sèrie linealitzada és un model vàlid.




Series resid
Series resid
Series resid

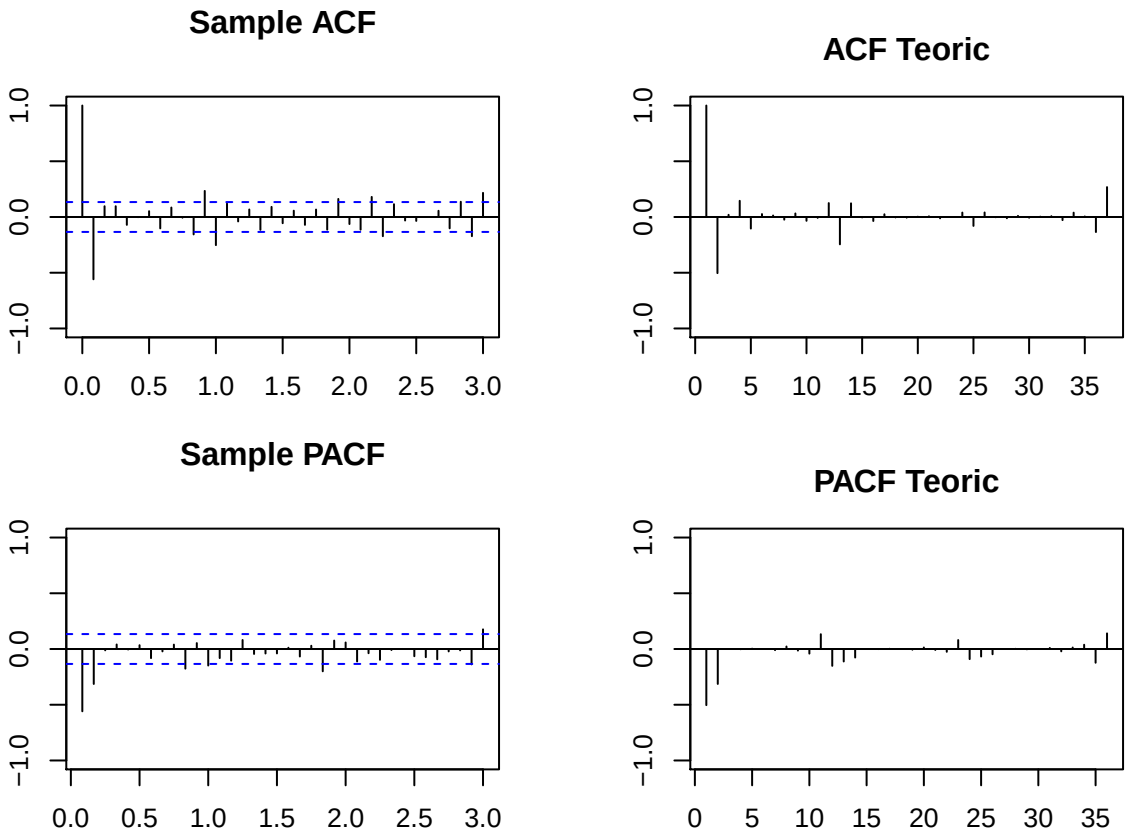

Standardized Residuals




p values for Ljung−Box statistic

##

```
## ------------------------------------------------------------------
##
## Call:
## arima(x = logserie.lin, order = pdq.1, seasonal = list(order = PDQ.2, period = 12))
##
## Coefficients:
##           ar1      ar2     sar1     sar2    sar3     sar4
##       -0.6611  -0.3138  -0.1435  -0.1237  0.1143  -0.4575
## s.e.   0.0663   0.0662   0.0652   0.0726  0.0703   0.0684
##
## sigma^2 estimated as 0.0009259:  log likelihood = 436.23,  aic = -858.46
##
## Modul of AR Characteristic polynomial Roots:  1.010509 1.022373 1.022373 1.022373 1.010509 1.010509 1
##
## Modul of MA Characteristic polynomial Roots:
##
## Psi-weights (MA(inf))
##
## --------------------
##         psi 1         psi 2         psi 3         psi 4         psi 5
## -0.6611483410  0.1232717842  0.1259973933 -0.1219912432  0.0411106127
##         psi 6         psi 7         psi 8         psi 9        psi 10
##  0.0111061704 -0.0202452005  0.0098994609 -0.0001911502 -0.0029805211
##        psi 11        psi 12        psi 13        psi 14        psi 15
##  0.0020305582 -0.1439245742  0.0945182123 -0.0173205016 -0.0182126800
##        psi 16        psi 17        psi 18        psi 19        psi 20
##  0.0174772420 -0.0058390847 -0.0016246499  0.0029067041 -0.0014118738
##
## Pi-weights (AR(inf))
##
## --------------------
##        pi 1        pi 2        pi 3        pi 4        pi 5        pi 6
## -0.66114834 -0.31384534  0.00000000  0.00000000  0.00000000  0.00000000
##        pi 7        pi 8        pi 9       pi 10       pi 11       pi 12
##  0.00000000  0.00000000  0.00000000  0.00000000  0.00000000 -0.14351750
##       pi 13       pi 14       pi 15       pi 16       pi 17       pi 18
## -0.09488635 -0.04504230  0.00000000  0.00000000  0.00000000  0.00000000
##       pi 19       pi 20
##  0.00000000  0.00000000
##
##  Shapiro-Wilk normality test
##
## data:  resid(model)
## W = 0.97956, p-value = 0.002271
```



6.3 Estabilitat del model proposat per la sèrie linealitzada

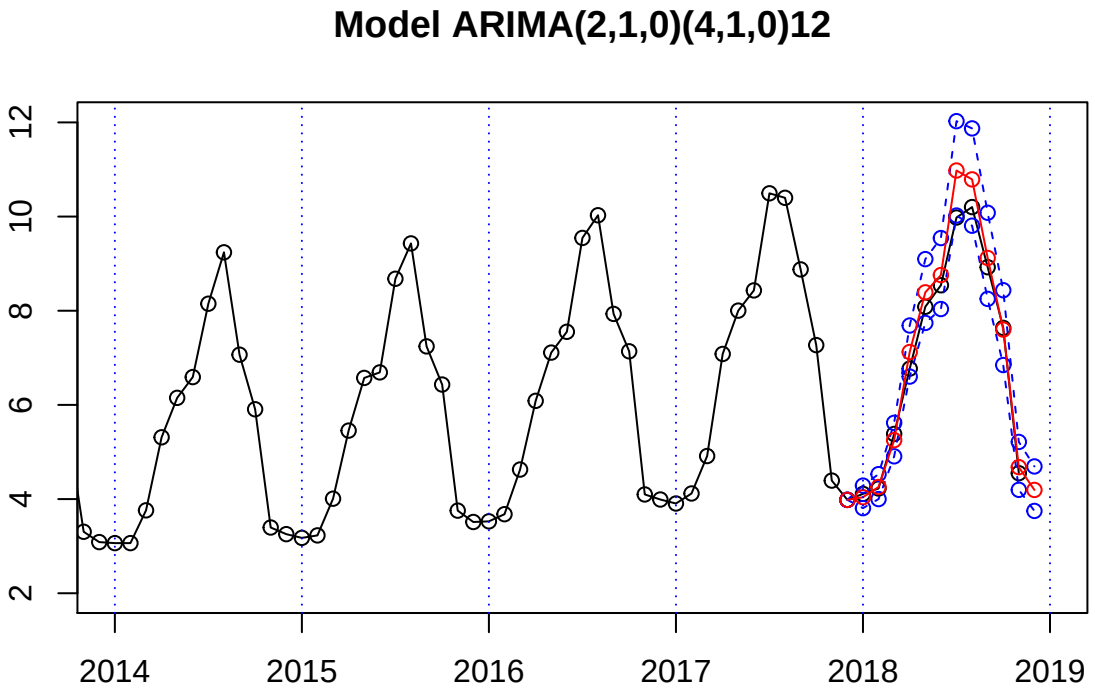
En relació a l'estabilitat del model per la sèrie linealitzada, s'observa que el valor dels coeficients varia molt poc, de l'ordre de menys de 0.02 en gairebé tots els casos. Per tant, podem confirmar que és estable.

```
##
## Call:
## arima(x = logserie1.lin, order = pdq.1, seasonal = list(order = PDQ.2, period = 12))
##
## Coefficients:
##           ar1      ar2     sar1     sar2    sar3     sar4
##       -0.6611  -0.3138  -0.1435  -0.1237  0.1143  -0.4575
## s.e.   0.0663   0.0662   0.0652   0.0726  0.0703   0.0684
##
## sigma^2 estimated as 0.0009259:  log likelihood = 436.23,  aic = -858.46
```

```
##
## Call:
## arima(x = logserie2.lin, order = pdq.1, seasonal = list(order = PDQ.2, period = 12))
##
## Coefficients:
##           ar1      ar2     sar1     sar2    sar3     sar4
##       -0.6778  -0.3276  -0.1295  -0.1432  0.1111  -0.4494
## s.e.   0.0676   0.0682   0.0671   0.0742  0.0713   0.0705
##
## sigma^2 estimated as 0.000911:  log likelihood = 415.46,  aic = -816.91
```

6.4 Capacitat de predicció del model proposat per la sèrie linealitzada

Pel que fa a la capacitat de predicció del model per la sèrie linealitzada, es pot observar que és millor que el model per la sèrie sense linealitzar, ja que s'ajusta molt més (de fet, tenim tant el EQM com el EAM més baix en aquest cas). De fet, en les zones de pujada i baixada els intervals de confiança estan gairebé a sobre del valor real de la sèrie. En els valors més alts és on es té més error, fet que no extranya, ja que és en els mesos de l'any on més ha anat variant el valor de la sèrie al llarg dels anys

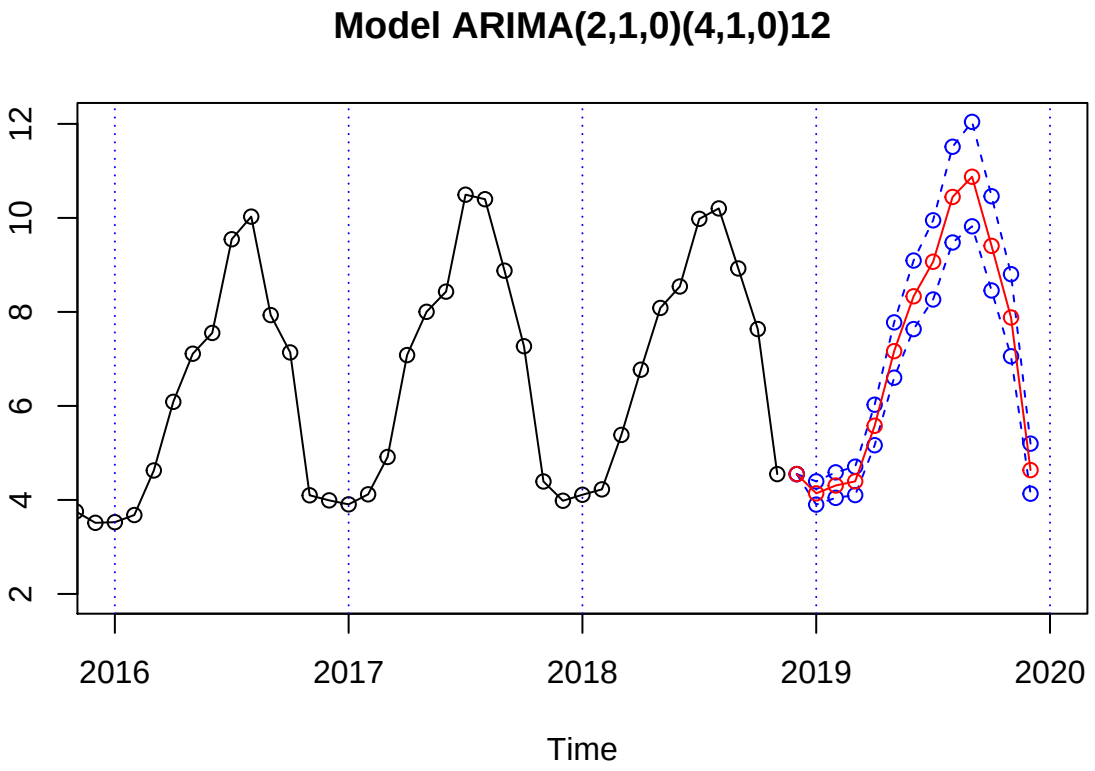



```
##                tl        pr        tu     serie  error
## Dec 2017  3.982530  3.982530  3.982530  3.982530  0.000
## Jan 2018  3.807303  4.039328  4.285492  4.110137  0.071
## Feb 2018  3.999645  4.256121  4.529043  4.224826 -0.031
## Mar 2018  4.907499  5.251276  5.619135  5.383687  0.132
## Apr 2018  6.601429  7.123301  7.686429  6.770845 -0.352
## May 2018  7.740691  8.391078  9.096113  8.084173 -0.307
## Jun 2018  8.036484  8.757653  9.543537  8.541181 -0.216
## Jul 2018 10.022139 10.978290 12.025662  9.979779 -0.999
## Aug 2018  9.805936 10.789681 11.872117 10.201456 -0.588
## Sep 2018  8.251762  9.120408 10.080496  8.924326 -0.196
## Oct 2018  6.846393  7.599599  8.435669  7.635569  0.036
## Nov 2018  4.196912  4.677580  5.213298  4.549899 -0.128
## Dec 2018  3.748080  4.193886  4.692716        NA     NA
```

```
## [1] 0.04120862
```

```
## [1] 0.03142106
```

6.5 Previsions a llarg termini pel model proposat per la sèrie linealitzada



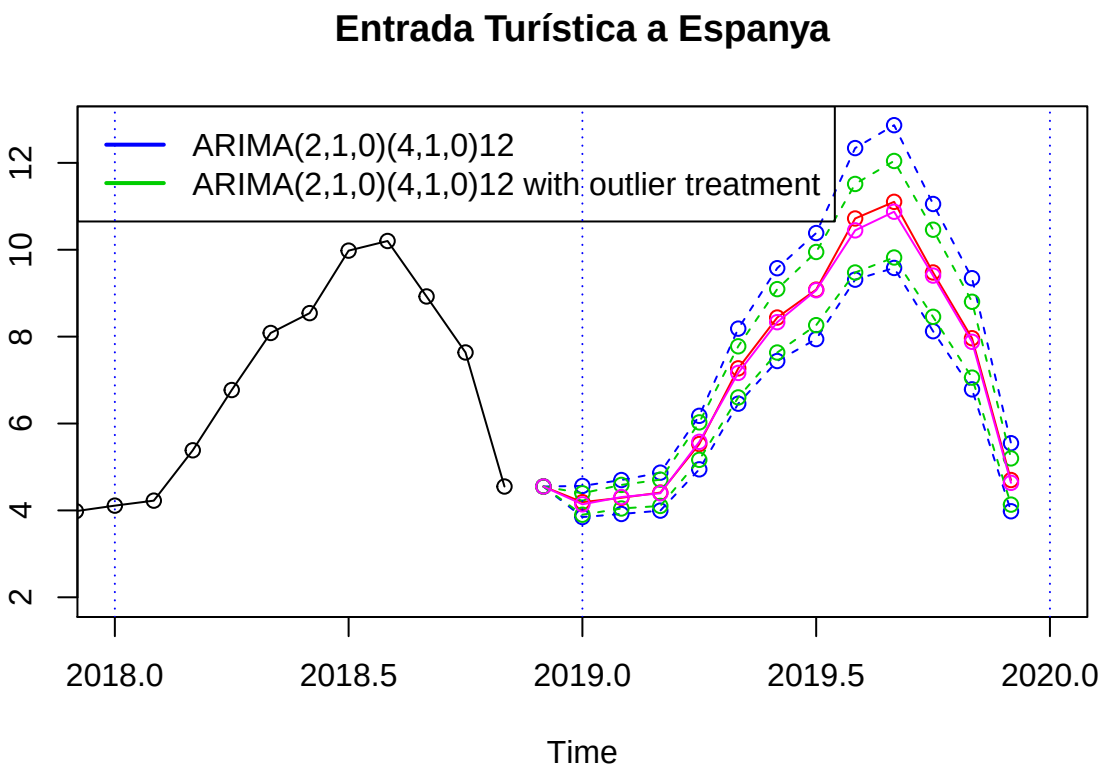
7 Comparació dels dos models

Com a última part d'aquest anàlisi, es realitza una comparació entre els valors predits pel model per la sèrie sense linealitzar i els valors predits per la sèrie linealitzada (és a dir, sense *outliers*). Tot i que els valors són força semblants, el que crida més l'atenció és que l'interval de confiança d'aquests valors és més estret en el cas del model per la sèrie linealitzada. Aquest fet es deu a que la variància dels valors sense *outliers* és molt més petita i, per tant, hi ha més marge d'acotació a l'hora de calcular tant els valors com els intervals. Per tant, podem dir que el model sense *outliers* és millor

```
##                tl2       pr2       tu2
## Dec 2018 4.549899  4.549899  4.549899
## Jan 2019 3.900358  4.140052  4.394477
## Feb 2019 4.044126  4.306979  4.586916
## Mar 2019 4.102884  4.394829  4.707548
## Apr 2019 5.164192  5.578496  6.026037
## May 2019 6.600659  7.164904  7.777382
## Jun 2019 7.634515  8.332016  9.093241
## Jul 2019 8.264364  9.067381  9.948425
## Aug 2019 9.476601 10.445786 11.514091
## Sep 2019 9.821204 10.875595 12.043183
## Oct 2019 8.456203  9.405362 10.461059
## Nov 2019 7.056392  7.881295  8.802630
## Dec 2019 4.132904  4.634817  5.197685
```

```
##          previs1.tl1 previs1.pr1 previs1.tu1 previs2.tl2 previs2.pr2
## Dec 2018    4.549899    4.549899    4.549899    4.549899    4.549899
```

```
## Jan 2019    3.846837    4.189916    4.563591    3.900358    4.140052
## Feb 2019    3.917671    4.290606    4.699042    4.044126    4.306979
## Mar 2019    3.995362    4.410566    4.868919    4.102884    4.394829
## Apr 2019    4.945079    5.526941    6.177267    5.164192    5.578496
## May 2019    6.457154    7.269534    8.184119    6.600659    7.164904
## Jun 2019    7.438742    8.439524    9.574947    7.634515    8.332016
## Jul 2019    7.942888    9.081850   10.384132    8.264364    9.067381
## Aug 2019    9.311565   10.719334   12.339936    9.476601   10.445786
## Sep 2019    9.580515   11.102785   12.866932    9.821204   10.875595
## Oct 2019    8.125041    9.476734   11.053296    8.456203    9.405362
## Nov 2019    6.786353    7.963653    9.345191    7.056392    7.881295
## Dec 2019    3.980613    4.698813    5.546594    4.132904    4.634817
##          previs2.tu2
## Dec 2018    4.549899
## Jan 2019    4.394477
## Feb 2019    4.586916
## Mar 2019    4.707548
## Apr 2019    6.026037
## May 2019    7.777382
## Jun 2019    9.093241
## Jul 2019    9.948425
## Aug 2019   11.514091
## Sep 2019   12.043183
## Oct 2019   10.461059
## Nov 2019    8.802630
## Dec 2019    5.197685
```



Per acabar de confirmar-ho, es mostren tota una sèrie de mesures de bondat d'ajust dels models. Sobretot ens crida l'atenció el *AIC* i el *BIC*, on el model sense *outliers* és clarament millor. Aquest model també té millor RMSPE i MAPE.

8 Comentaris finals

Així doncs, un com realitzat aquest anàlisi, es conclou que la presència de valors atípics en una sèrie pot influir (i molt) en les previsions que pugui fer un model basant-se en ella. Per tant, es confirma la importància de la seva detecció i correcció. Pel que fa a les previsions de la sèrie, sembla que la tendència general de creixement es mantindrà, així com també el patró estacional (és a dir, que Espanya tenint un munt de turistes a l'estiu).

(En l'apartat de models proposats, s'ha volgut proposar també un $ARIMA(1,1,1)(4,1,0)_{12}$, però per problemes del R no s'ha pogut calcular.)