UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat de Matemàtiques i Estadística

Ocupación de viajeros en alojamientos de turismo rural en España (2009 - 2019)

Análisi de Series Temporales

Sara Montañés González

Raquel Ortiz Martín

El objetivo de este informe es ver el comportamiento de la demanda en alojamientos de turismo rural durante los años 2009 - 2019 y hacer una predicción para el período 2020 si no se tuvieran en cuenta los efectos causados por la pandemia del SARS-CoV-2.

Se observa que el mejor modelo para explicar los datos y hacer una predicción a largo plazo es un $ARMA(1,1)SMA(1)_{12}$.

En cuanto a la predicción de un tiempo extrapolable a los datos, se puede ver que la demanda sigue aumentando con la misma tendencia y continuando con el mismo patrón estacional.

Contents

1 Introducción

1.1 Descripción del entorno y el problema - motivación del trabajo

Desde el estallido de la pandemia del SARS-CoV-2 las preferencias de la población respecto a sus destinos vacacionales parecen haber cambiado. Actualmente, los viajeros están más interesados en lugares naturales, al aire libre, debido a que en la época del confinamiento mucha parte de la población se sintió agobiada dentro de sus hogares. Adicionalmente, los españoles se han centrado en un turismo nacional debido a la situación de crisis sanitaria que estamos viviendo. Así pues, la demanda en alojamientos nacionales de turismo rural parece haber aumentado considerablemente respecto a años anteriores.

El objetivo principal de este informe es ver cuál era el comportamiento de la demanda en alojamientos de turismo rural en los años anteriores a la pandemia. De esta manera se podrá ver si este aumento ha sido causado únicamente a las circunstancias mencionadas o si la demanda ya presentaba una tendencia creciente en el periodo 2009 - 2019.

Para la elección de los años analizados, se ha evitado estudiar aquellos momentos temporales en los que había un factor condicionante. Por tanto, los datos tomados dejan fuera la crisis económica del 2008 y la pandemia del 2020. Teniendo esto en cuenta, solo se han seleccionado los datos mensuales pertenecientes al espacio temporal comprendido entre 2009 y 2019, ambos períodos incluidos.

1.2 Definición precisa de la variable

La variable a estudiar es el número de viajeros en alojamientos de turismo rural durante el periodo ya comentado anteriormente. Cabe destacar que esta variable únicamente recoge el total nacional, es decir, aquellos viajeros procedentes de las comunidades y ciudades autónomas españolas.

1.3 Descripción de la fuente de información

Los datos han sido extraídos de la página web del Instituto Nacional de Estadística (INE) de España. Este organismo público realiza mensualmente la Encuesta de ocupación en alojamientos de turismo rural, proporcionando información sobre la oferta y la demanda de los servicios de este tipo de alojamiento. Los datos se encuentran en el siguiente enlace: *https://ine.es/jaxiT3/Tabla.htm?t=2941*

1.4 Resumen de la estructura del trabajo

El estudio sigue el siguiente esquema:

- **Metodología empleada**

  En este apartado se transforma la serie en estacionaria, se grafica el ACF y el PACF con la finalidad de identificar modelos pausibles y se estiman los modelos propuestos. Finalmente, se hace una validación de estos para asegurarnos que los residuos cumplen las hipótesis de normalidad, homocedasticidad e independencia.

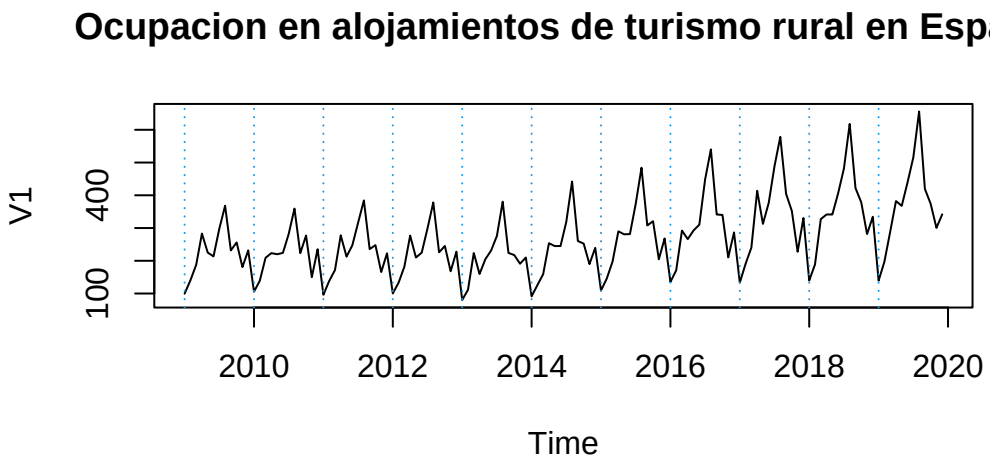
- **Resultados**

  A partir del análisis realizado en el apartado anterior se decide qué modelo ajusta mejor los datos. Para saberlo, se verifica la estabilidad de éstos y se realiza una predicción *out-of-sample* para así poder calcular el RMSE y el MAPE. En funcón de los resultados anteriores, se selecciona el mejor modelo, con el que se llevará a cabo el objetivo principal del estudio, hacer una predicción a largo plazo del número de viajeros en alojamientos rurales en los próximos años si no existiera ningún factor condicionante que actuara como brecha, en el caso expuesto, si no existiera la pandemia del Covid-19.

2 Metodología empleada

2.1 Análisis descriptivo de la serie

El análisis descriptivo de la serie consistirá en un gráfico que permitirá observar la distribución de los datos a lo largo de los 11 años estudiados. El gráfico temporal tiene la siguiente forma:



Como ya se explicó con detalle en la práctica 1, la serie estudiada presenta una tendencia creciente y una estacionalidad multiplicativa de orden s = 12.

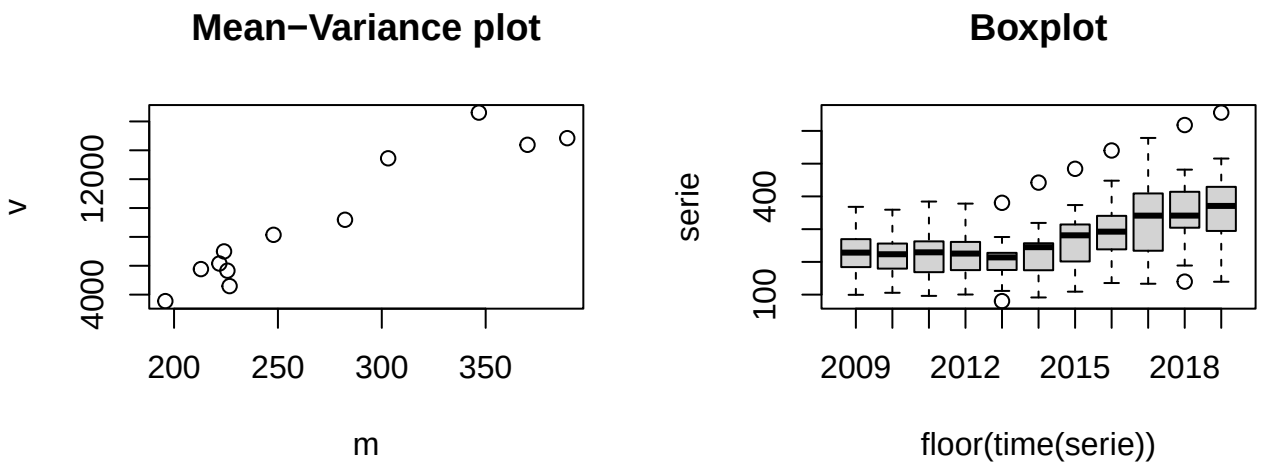
2.2 Identificación del modelo

2.2.1 Transformación de la serie temporal en una serie estacionaria

Para poder trabajar con los datos de una forma más estable es necesario que la serie estudiada sea estacionaria. ¿Qué requisitos tiene que cumplir para conseguirlo? Varianza constante, que no exista la componente estacional y que presente media constante. Se procede a estudiar dichos aspectos:

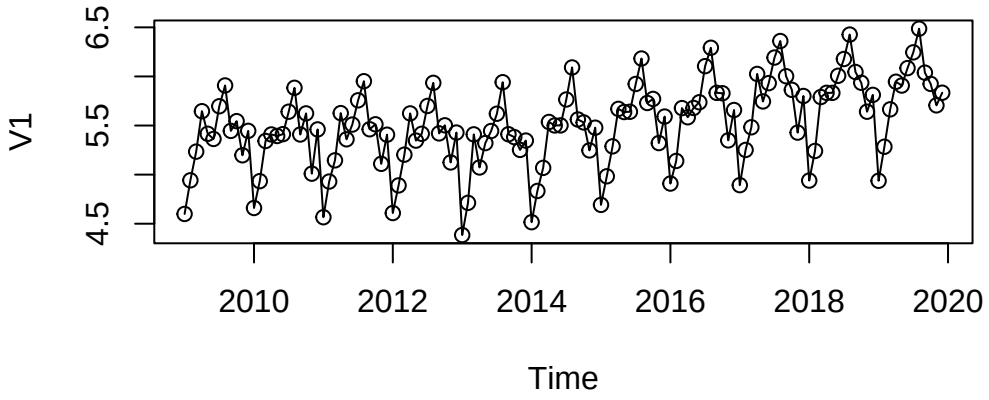
¿Parece ser la varianza constante?

Se dispone de dos soportes gráficos que permiten ver si la varianza de la serie es constante o no: *Gráfico medias vs varianzas* y *Boxplot por periodos*.



En el caso del primer gráfico, se observa que un augmento de la media (eje x) implica un aumento de la varianza (eje y). Para el boxplot se observa que el rango intercuartílico (amplitud de las cajas) aumenta conforme aumenta el nivel de la serie.

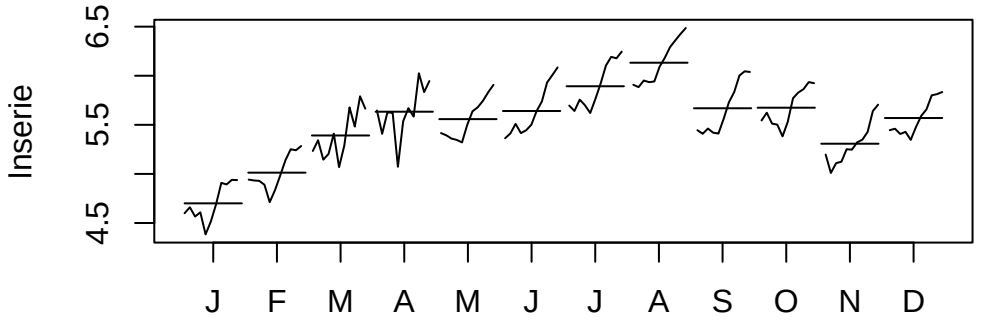
Por tanto, gracias a los resultados obtenidos en ambas representaciones, se conluye que **la varianza de la serie no es constante**. Para conseguir que si lo sea, es necesario aplicar una transformación, en este caso, se ha optado por una **transformación logarítmica**. Para asegurarnos que se ha realizado la transformación correcta, se vuelven a graficar los datos, pero ahora, en escala logarítmica.



La varianza de la serie se ha estabilizado al aplicar la transformación logaritmica.

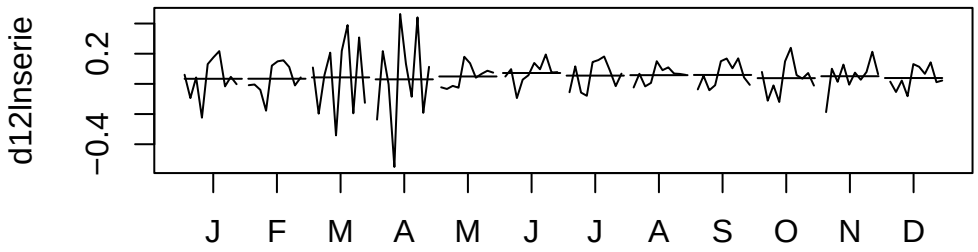
¿Presenta la serie un patrón estacional?

Para dar respuesta a la pregunta expuesta, se hace uso de la función `montplot`, la cual calcula la media en cada orden estacional, en el presente caso, para cada mes.



Se observa que la media no es la misma para cada mes, ya que las líneas horizontales se encuentran en alturas distintas. Así pues la serie **presenta un patrón estacional de orden S = 12**. Para poder eliminarlo, se realiza una diferenciación estacional (D = 1), es decir, se restan las observaciones del patrón estacional anterior. Por tanto, se eliminarán S = 12 observaciones, las 12 primeras.

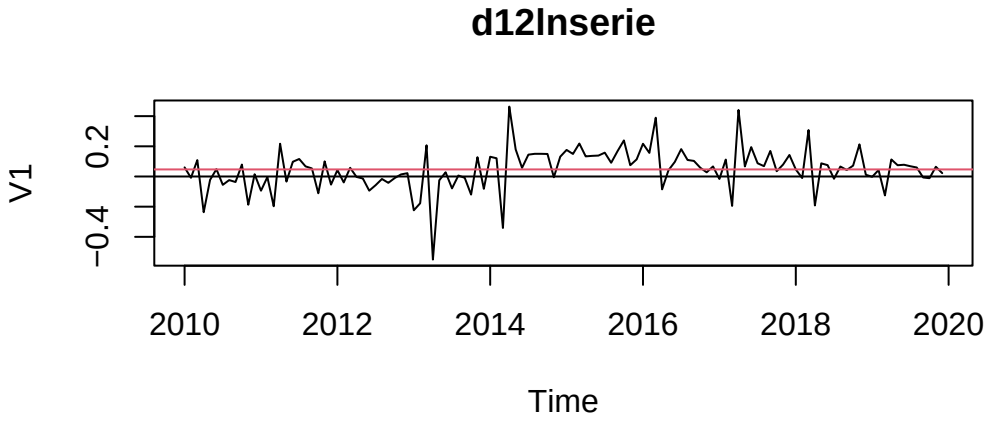
Se vuelve a utilizar la función `monthplot` sobre la serie diferenciada estacionalmente para verificar que el patrón estacional se ha eliminado.



Efectivamente, diferenciando la serie estacionalmente una vez se consigue eliminar el patrón estacional, ya que ahora las medias de cada mes son muy parecidas entre ellas.

¿Es la media de la serie constante?

Se obtiene el gráfico temporal de la serie en escala logarítmica y con una diferenciación estacional.



La media de la serie si parece constante pero se observa que la media no toma el valor 0. Por acuerdo académico no se trabajará con medias distintas a 0. Así pues, para cumplir el acuerdo, se hace una diferenciación regular (d = 1). Para verificar que la media de la serie toma el valor 0 se vuelve a graficar la serie temporal una vez aplicada dicha diferenciación.

La media de la serie ahora si toma el valor 0. Para assegurar que la serie no presenta sobrediferenciación se comparan las varianzas. La mejor serie será aquella que tenga una menor varianza.

Modelo	lnserie	d12lnserie	d1d12lnserie
Varianza	0.17921	0.0186969	0.0403553

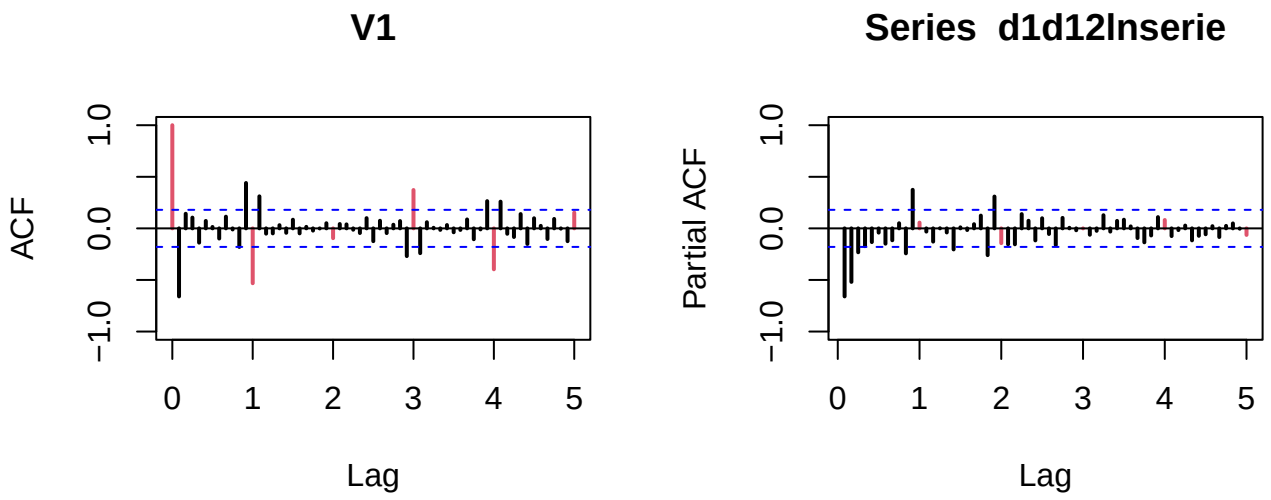
Se observa que la serie con menor varianza es aquella que presenta escala logarítmica y una diferenciación estacional. Por tanto, significa que la diferenciación regular no es necesaria. Aún así, ésta se mantendrá por el motivo expuesto anteriormente, para conseguir que la media tome el valor 0.

2.2.2 Identificación de modelos pausibles

Para identificar algunos modelos se grafican el ACF y el PACF de la serie estacionaria.

Cabe destacar que primeramente se seleccionarán los modelos más parsimoniosos. Si en el apartado de *validación* no se valida ninguno de ellos, se volverá al presente apartado para proponer otros modelos más complejos.

Para una mejor identificación, en el siguiente gráfico se muestra en color rojo los "s" rezagos estacionales de la función (P)ACF. En negro, todos los rezagos regulares.



Antes de la proposición de modelos, cabe especificar que se confirma que se ha llegado a la serie estacionaria ya que los valores de la ACF decaen de manera rápida hacia el 0.

Observando las graficas se concluye que:

- *Opciones para la parte regular*: MA(1), AR(3), ARMA(1, 1)
- *Opciones para la parte estacional*: MA(1)

Todas ellas con las diferenciaciones aplicadas, una regular (d = 1) y otra estacional (D = 1 con S = 12).

Así pues, se obtienen 3 posibles combinaciones, 3 modelos propuestos.

2.2.3 Forma compacta y usando el operador de retardo B de los modelos identificados

$W_t = (1-B)^d(1-B^S)^D ln(W_t)$

Modelo 1: $MA(1)SMA(1)_{12}$ para W_t

Forma compacta de modelo ARIMA para $ln(X_t)$

$$(1-B)(1-B^{12})ln(X_t) = \theta_1(B)\Theta_1(B^{12})Z_t, d = 1,\ y\ D = 1\ con\ S = 12$$

Sustituyendo cada polinomio característico se obtiene:

$$(1-B)(1-B^{12})ln(X_t) = (1+\theta_1 B)(1+\Theta_1 B^{12})Z_t$$

Modelo 2: $AR(3)SMA(1)_{12}$ para W_t

Forma compacta de modelo ARIMA para $ln(X_t)$

$$\phi_3(B)(1-B)(1-B^{12})ln(X_t) = \Theta_1(B^{12})Z_t, d = 1,\ y\ D = 1\ con\ S = 12$$

Sustituyendo cada polinomio característico se obtiene:

$$(1-\phi_1 B-\phi_2 B^2-\phi_3 B^3)(1-B)(1-B^{12})ln(X_t) = (1+\Theta_1 B^{12})Z_t$$

Modelo 3: $ARMA(1,1)SMA(1)_{12}$ para W_t

Forma compacta de modelo ARIMA para $ln(X_t)$

$$\phi_1(B)(1-B)(1-B^{12})ln(X_t) = \theta_1(B)\Theta_1(B^{12})Z_t, d = 1,\ y\ D = 1\ con\ S = 12$$

Sustituyendo cada polinomio característico se obtiene:

$$(1-\phi_1 B)(1-B)(1-B^{12})ln(X_t) = (1+\theta_1 B)(1+\Theta_1 B^{12})Z_t$$

2.3 Estimación de los modelos

Una vez se han expusto los modelos posibles, se procede a realizar su estimación a partir de la funció `arima`. Cabe destacar que dicha estimación se hará para la serie estacionaria y para la serie no estacionaria. Así pues, en total se verán seis modelos. Además, se calcularan los `T-ratios` para cada uno de los coeficientes. Éstos, deberán tomar un valor superior a 2 para que sean necesarios en la explicación de los datos. Por contra, si son inferiores a 2, no sern necesarios. Por esta misma razón, en los modelos de la serie estacionaria, se quiere que el valor del `intercept` salga no significativo, ya que se ha realizado una diferenciación regular extra para conseguir este mismo propòsito, que la media, el `intercept`, sea igual a 0.

2.3.1 Modelo 1A, MA(1)SMA(1)$_{12}$

Se impone a la serie estacional `d1d12lnserie` el modelo identificado: ARIMA(0, 0, 1)(0, 0, 1)$_{12}$.

Se calculan los T-ratios para cada coeficiente para ver cuáles son estadísticamente significativos:

```
## Modelo 1a
##  T-ratios: -18.6 -8.04 1.69
```

El T-ratio del `intercept` es menor a 2, así que se plantea el modelo sin este coeficiente.

2.3.2 Modelo 1B, usando la serie no-estacionaria $ln(X_t)$ (lnserie)

Se elimina el coeficiente `intercept`, estimando el modelo con la serie original en logaritmos (sin aplicar ninguna diferenciación).

Se calculan los T-ratios para cada coeficiente para ver cuáles son estadísticamente significativos:

```
## Modelo 1b
## T-ratios: -17.93 -8.46
```

Todos los coeficientes son estadísticamente significativos, no es necesario eliminar nada más. Se calcula el AIC de los modelos con y sin `intercept` para saber cuál de ellos tiene un valor menor de esta medida.

Modelo	AIC
Con `intercept`	-191.8086452
Sin `intercept`	-191.4156782

Es mejor el modelo con `intercept`, aunque la diferencia entre ambos modelos es muy pequeña.

2.3.3 Modelo 2A, AR(3)SMA(1)$_{12}$

Se impone a la serie estacional `d1d12lnserie` el modelo identificado: ARIMA(3, 0, 0)(0, 0, 1)$_{12}$

Se calculan los T-ratios para cada coeficiente para ver cuáles son estadísticamente significativos:

```
## Modelo 2a
## T-ratios: -10.22 -4.81 -1.4 -8.37 0.74
```

El T-ratio del `intercept` y el `ar3` es menor a 2. Se empieza eliminando el `intercept`.

2.3.4 Modelo 2B, usando la serie no-estacionaria $ln(X_t)$ (lnserie)

Se elimina el coeficiente `intercept`, estimando el modelo con la serie original en logaritmos (sin ninguna diferenciación).

Se calculan los T-ratios para cada coeficiente para ver cuáles son estadísticamente significativos:

```
## Modelo 2b
## T-ratios: -10.16 -4.77 -1.37 -8.36
```

El T-ratio del coeficiente `ar3` es inferior a 2 en valor absoluto, así que se elimina dicho coeficiente usando la función `fixed` para comprobar si mejora el modelo. Se calcula el AIC para el modelo con y sin `intercept` y con el coeficiente `ar3` y sin éste.

Modelo	AIC
Con `intercept` y `ar3`	-193.0518395
Sin `intercept` y con `ar3`	-194.5101628
Sin `intercept` ni `ar3`	-194.6513328

El AIC disminuye al eliminar el `intercept` y el coeficiente de `ar3`, así que es mejor quedarse con el modelo más parsimonioso.

2.3.5 Modelo 3A, ARMA(1, 1)SMA(1)$_{12}$

Se impone a la serie estacional `d1d12lnserie`el modelo identificado: ARIMA(1, 0, 1)(0, 0, 1)$_{12}$

Se calculan los T-ratios para cada coeficiente para ver cuáles son estadísticamente significativos:

```
## Modelo 3a
## T-ratios: -2.95 -10.15 -8.42 1.34
```

El T-ratio del `intercept` es menor a 2, así que se plantea el modelo sin este coeficiente.

2.3.6 Modelo 3B, usando la serie no-estacionaria $ln(X_t)$ (lnserie)

Se calculan los T-ratios para cada coeficiente para ver cuáles son estadísticamente significativos:

```
## Modelo 3b
## T-ratios: -3.1 -9.66 -8.34
```

No es necesario eliminar nada más. Se calcula el AIC de los modelos con y sin `intercept`, para saber cuál de ellos tiene un menor valor de esta medida.

Modelo	AIC
Con `intercept`	-197.625761
Sin `intercept`	-198.0288343

En este caso, es mejor el modelo sin `intercept`, pues su AIC es menor.

2.3.7 Tabla resumen

Se realiza una tabla resumen para los tres modelos definitivos con el objetivo de poder ver de forma más clara la comparación entre ellos.

```
##
## Results
## ============================================================
##                              Dependent variable:
##                    ---------------------------------------
##                     d1d12lnserie           lnserie
##                          1A            2B2           3B
##                         (1)            (2)          (3)
## ------------------------------------------------------------
## ma1                 t = -18.598                  t = -9.657
##                        -0.820                     -0.707
##
## ar1                                 t = -10.974   t = -3.101
##                                       -0.901        -0.317
##
## ar2                                 t = -5.570
##                                       -0.457
##
## ar3                                  t = 0.000
##                                        0.000
##
## sma1                 t = -8.038    t = -9.661    t = -8.343
```

```
##                         -0.827        -0.794        -0.732
##
## intercept            t = 1.688
##                         0.001
##
## ---------------------------------------------------------
## Observations             119           119           119
## Log Likelihood         99.904        101.326       103.014
## sigma2                  0.010         0.010         0.009
## Akaike Inf. Crit.    -191.809      -194.651      -198.029
## =========================================================
## Note:                t = T-statistic value = coeff/SE(coeff)
```

En la tabla anterior se observan las estimaciones y los T-ratios de los coeficientes de los 3 modelos estimados. Antes de proceder a la explicación, se recuerda a que hace referencia cada uno de ellos:

- Modelo 1: $MA(1)SMA(1)_{12}$
- Modelo 2: $AR(3)SMA(1)_{12}$
- Modelo 3: $ARMA(1,1)SMA(1)_{12}$
- Modelos A: estimación con la serie estacionaria
- Modelos B: estimación con la serie NO estacionaria

Como se deseaba, todos los T-ratios son mayores de 2, excepto los que hacen refererencia al `intercept` de la serie estacionaria. Se concluye que todos los coeficientes son necesarios para explicar los datos.

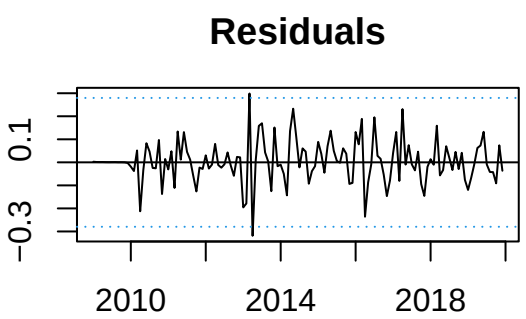
Observando el críterio del AIC, se puede decir que los mejores modelos son aquellos que tienen un valor menor en esta medida de calidad relativa. Así pues, a partir de ahora, solo se seguirán estudiando los modelos 2B2 y 3B, ya que son estos los que tienen un valor más negativo.

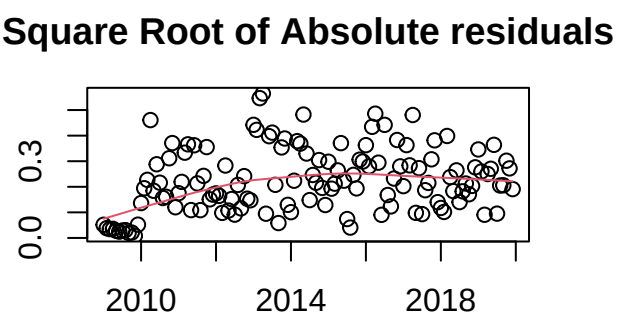
3 Validación de los modelos ajustados

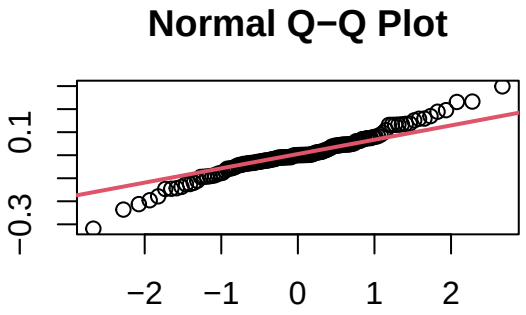
Como se ha comentado anteriormente, solo se va a comprobar la validación de 2 de los 6 modelos propuestos. En caso que esta no se cumpla, se seleccionarán otros modelos y se repetirá el proceso que se verá a continuación. Pero, *¿Qué necesita un modelo para ser validado?*

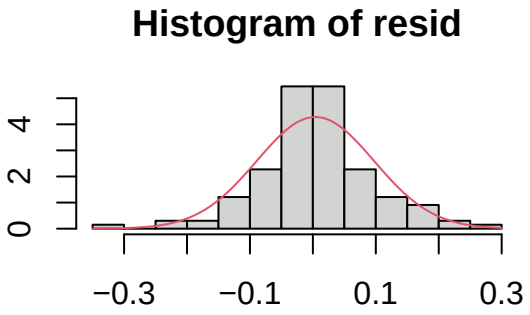
- Los residuos tienen que tener varianza constante y ser normales e independientes
- En el ACF y PACF de los residuos se tiene que observar que estos se distribuyen como un `Ruido Blanco`.
- El modelo debe de ser causal e invertible

3.1 Validación modelo 2B2









Residuals: Los residuos parecen aleatorios y la mayoría se encuentran dentro de las bandas de confianza, centrados en 0.

Square Root of Absolute residuals: La línea roja no es nada constante, de manera que la varianza tampoco lo es.

Normal QQ-Plot: Hay algunos valores atípicos en las colas.

Histogram of resid: Hay algún valor outlier y los datos tienen kurtosis, de manera que no se ajustan bien del todo a una distribución Normal. La kurtosis puede ser debida a la presencia de outliers.

Conclusión: Parece que la varianza no es constante y hay presencia de atípicos en los residuos.

A continuación se realiza el test de Breusch-Pagan y se contrastan las siguientes hipótesis:

- H_0: Los residuos son homocedasticos
- H_1: Los residuos no son homocedasticos

Se obtiene un p-valor superior a 0.05, así que se afirma que los residuos son homocedasticos.

A continuación se realizan los tests de Shapiro-Wilk, Anderson-Darling y Jarque Bera y contrastan las siguientes hipótesis:

- H_0: Los residuos son normales
- H_1: Los residuos no son normales

Se obtienen p-valores inferiores a 0.05, así que se afirma que los residuos no son normales. Hay que tener en cuenta que los tests de Normalidad son muy sensibles a valores atípicos, de manera que se prestará más atención al gráfico QQ-plot, donde se ha visto que los residuos centrales si cumplen la normalidad pero los pertenecientes a las colas no.

A continuación se contrastan las siguientes hipótesis:

Durbin-Watson test

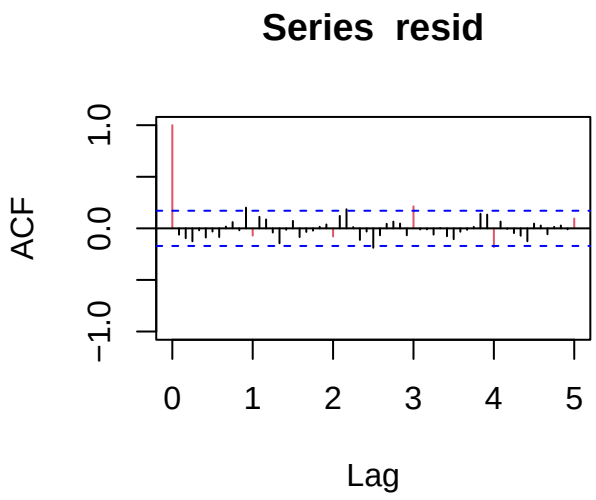
- H_0: No hay autocorrelación en los residuos en el retardo k
- H_1: Hay autocorrelación en los residuos en el retardo k

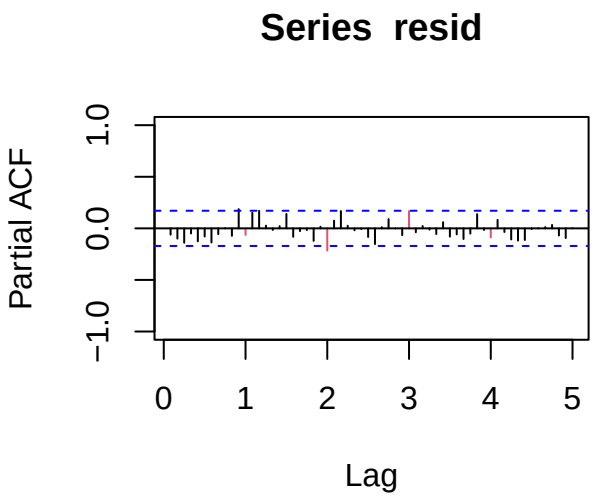
Se obtiene un p-valor superior a 0.05, así que se afirma que no hay autocorrelación en los residuos.

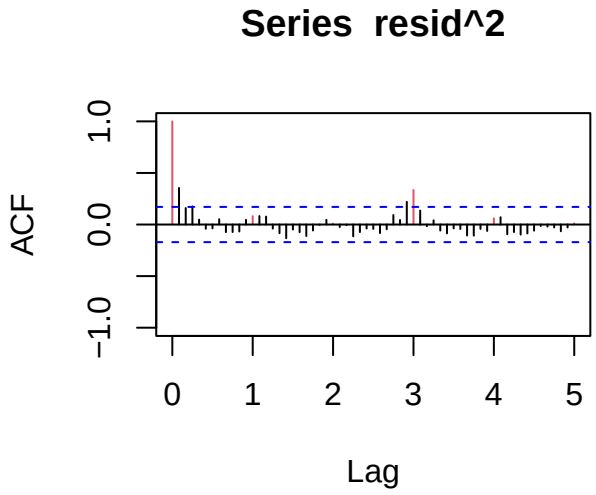
Ljung-Box test

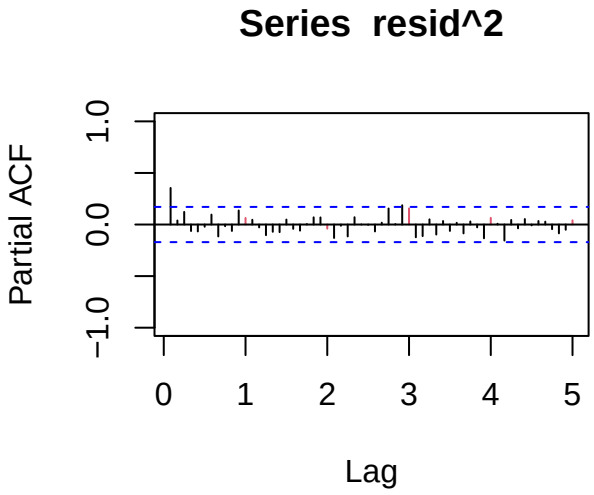
- H_0: No hay autocorrelación hasta el retardo k
- H_1: Hay autocorrelación hasta el retardo k

Se obtiene un p-valor inferior a 0.05 en el retardo (lag) 36. Tal y como está programada la función significa que entre el retardo 24 y el 36 se obtiene una estructura de autocorrelación, la cual se arrastra en todos los retardos posteriores a este hecho. Cabe destacar que la pérdida de independencia se ha dado en un retardo lejano al origen, por tanto, ésta no es preocupante en la validación de los residuos.





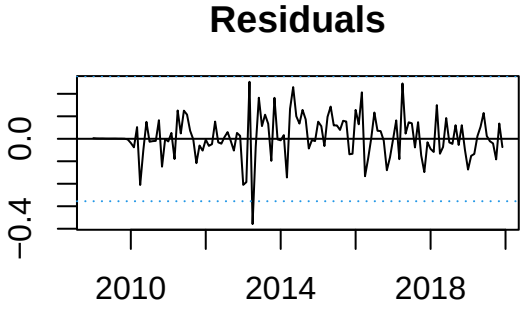


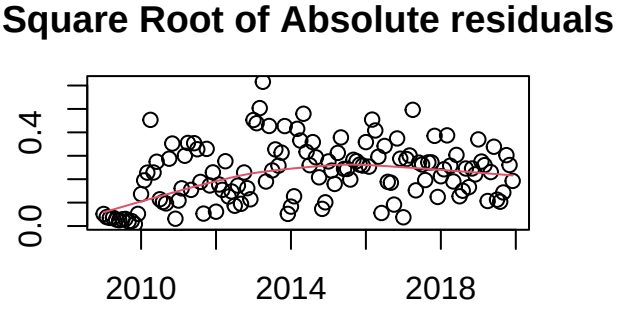


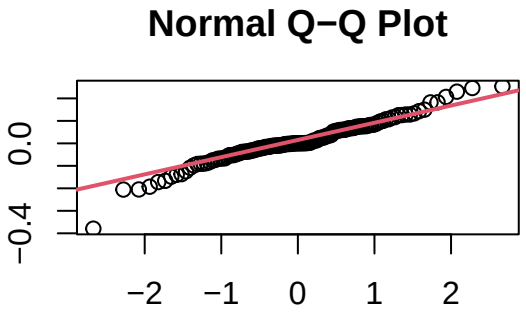
Con los gráficos ACF y PACF de los residuos se reafirma el resultado del test de Durbin-Watson: No hay autocorrelación en los residuos, ya que todos los retardos se encuentran dentro de las bandas de confianza.

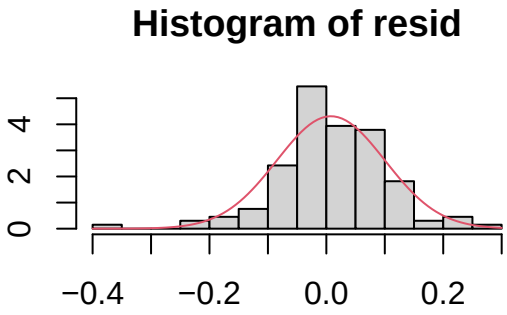
En los gráficos ACF y PACF de los residuos al cuadrado si se observa autocorrelación. Esta puede ser debida a la presencia de atípicos o a la volatilidad. Como se ha comentado anteriormente, en los datos hay presencia de atípicos, por tanto, los retardos que se encuentran fuera pueden ser fruto del azar u outliers.

3.2 Validación modelo 3B









Residuals: Los residuos parecen aleatorios y la mayoría se encuentran dentro de las bandas de confianza, centrados en 0.

Square Root of Absolute residuals: La línea roja no es nada constante, de manera que la varianza tampoco lo es.

Normal QQ-Plot: Hay algunos valores atípicos en las colas, sobretodo en la cola izquierda.

Histogram of resid: Hay algún valor outlier, sobretodo en la cola izquierda. La parte central tampoco ajusta bien una distribución Normal, pues los datos están desplazados hacia la izquierda.

Conclusión: Parece que la varianza no es constante y hay presencia de atípicos en los residuos.

A continuación se realiza el test de Breusch-Pagan y se contrastan las siguientes hipótesis:

- H_0: Los residuos son homocedasticos
- H_1: Los residuos no son homocedasticos

Se obtiene un p-valor superior a 0.05, así que se afirma que los residuos son homocedasticos.

A continuación se realizan los tests de Shapiro-Wilk, Anderson-Darlin y Jarque Bera y se contrastan las siguientes hipótesis:

- H_0: Los residuos son normales
- H_1: Los residuos no son normales

Se obtienen p-valores inferiores a 0.05, así que se afirma que los residuos no son normales. Hay que tener en cuenta que los tests de Normalidad son muy sensibles a valores atípicos, de manera que se prestará más atención al gráfico QQ-plot, donde se ha visto que los residuos centrales si cumplen la normalidad pero los pertenecientes a las colas no.

A continuación se contrastan las siguientes hipótesis:

Durbin-Watson test

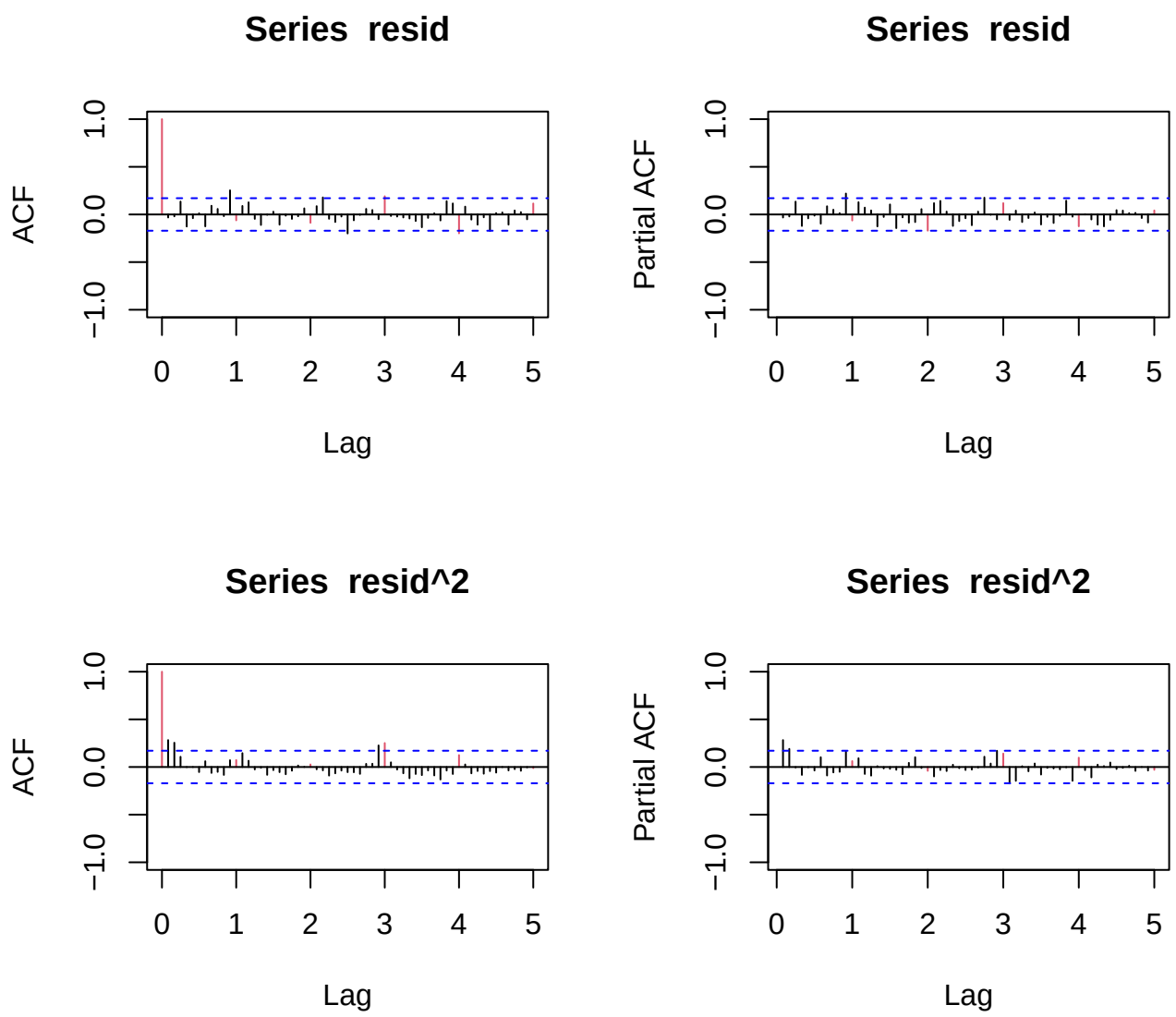
- H_0: No hay autocorrelación en los residuos en el retardo k
- H_1: Hay autocorrelación en los residuos en el retardo k

Se obtiene un p-valor superior a 0.05, así que se afirma que no hay autocorrelación en los residuos.

Ljung-Box test

- H_0: No hay autocorrelación hasta el retardo k
- H_1: Hay autocorrelación hasta el retardo k

Se obtiene un p-valor inferior a 0.05 en el retardo (lag) 36. Tal y como está programada la función significa que entre el retardo 24 y el 36 se obtiene una estructura de autocorrelación, la cual se arrastra en todos los retardos posteriores a este hecho. Cabe destacar que la pérdida de independencia se ha dado en un retardo lejano al origen, por tanto, ésta no es preocupante en la validación de los residuos.



Con los gráficos ACF y PACF de los residuos se reafirma el resultado del test de Durbin-Watson: No hay autocorrelación en los residuos, ya que casi todos los retardos se encuentran dentro de las bandas de confianza.

En los gráficos ACF y PACF de los residuos al cuadrado tampoco se observa autocorrelación. Así pues, se puede afirmar que no hay presencia de outliers ni hay volatilidad.

4 Predicción de modelos ARIMA ajustados y validados

4.1 Modelo 2B2

4.1.1 Verificación estabilidad del modelo

Se deja fuera el último año y se crea de nuevo el modelo. Si el modelo es estable, los coeficientes estimados obtenidos con la serie completa y con la serie incompleta son parecidos. Parecidos significa: coeficientes de la misma magnitud, con mismo signo y misma significancia.

##

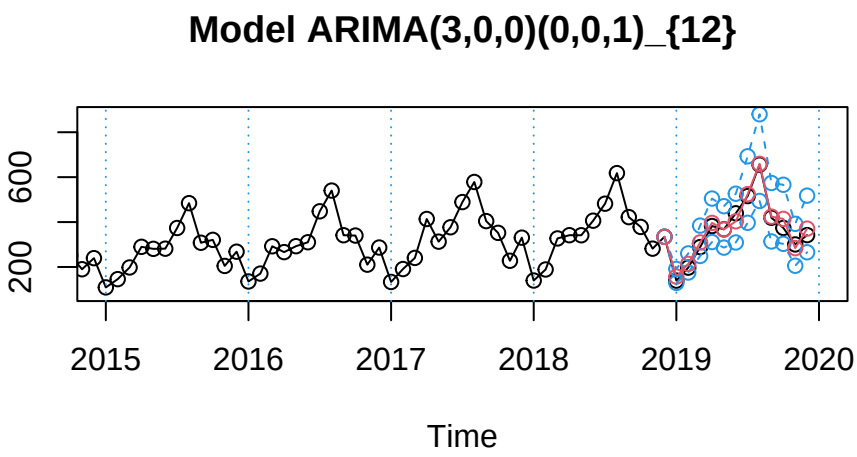
```
## Call:
## arima(x = lnserie1, order = c(3, 1, 0), seasonal = list(order = c(0, 1, 1),
##     period = 12), fixed = c(NA, NA, 0, NA))
##
## Coefficients:
##           ar1     ar2  ar3     sma1
##       -0.9014  -0.457    0  -0.7935
## s.e.   0.0821   0.082    0   0.0930
##
## sigma^2 estimated as 0.009576:  log likelihood = 101.33,  aic = -194.65
```

```
##
## Call:
## arima(x = lnserie2, order = c(3, 1, 0), seasonal = list(order = c(0, 1, 1),
##     period = 12), fixed = c(NA, NA, 0, NA))
##
## Coefficients:
##           ar1      ar2  ar3     sma1
##       -0.9162  -0.4695    0  -0.7905
## s.e.   0.0866   0.0857    0   0.1082
##
## sigma^2 estimated as 0.01005:  log likelihood = 87.93,  aic = -167.86
```

El modelo es estable pues se cumplen las 3 condiciones establecidas.

4.1.2 Predicción out-of-sample y cálculo RMSPE/MAPE

A continuación se grafica la serie completa (sin dejar el último año fuera) y se hace una predicción del último año para ver si el modelo estima correctamente los datos.



El modelo propuesto estima de manera bastante precisa los datos (estimación en color rojo). Además, los datos reales se encuentran dentro de las bandas de confianza (estimaciones en color azul).

Cálculo de medidas RMSPE/MAPE y mean Length

Medida	Valor
EQM	0.0700276
EAM	0.0570966

4.2 Modelo 3B

4.2.1 Verificación estabilidad del modelo

Para comprobar si el modelo es estable se usa el mismo criterio explicado en el modelo 2B2.

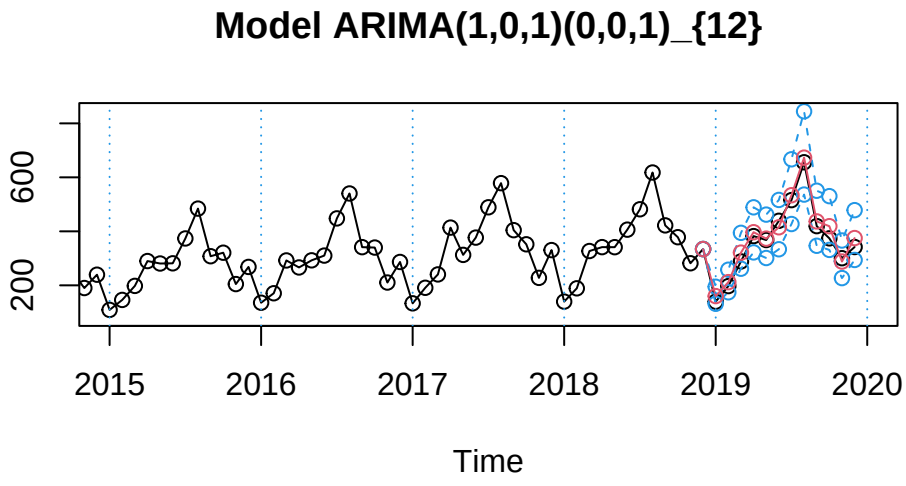
```
##
## Call:
## arima(x = lnserie1, order = c(1, 1, 1), seasonal = list(order = c(0, 1, 1),
##     period = 12))
##
## Coefficients:
##           ar1      ma1     sma1
##       -0.3167  -0.7069  -0.7320
## s.e.   0.1021   0.0732   0.0877
##
## sigma^2 estimated as 0.009497:  log likelihood = 103.01,  aic = -198.03
```

```
##
## Call:
## arima(x = lnserie2, order = c(1, 1, 1), seasonal = list(order = c(0, 1, 1),
##     period = 12))
##
## Coefficients:
##           ar1      ma1     sma1
##       -0.3385  -0.6992  -0.7348
## s.e.   0.1061   0.0762   0.0997
##
## sigma^2 estimated as 0.009984:  log likelihood = 89.37,  aic = -170.74
```

El modelo es estable pues se cumplen las 3 condiciones establecidas.

4.2.2 Predicción out-of-sample y cálculo RMSPE/MAPE

A continuación se grafica la serie completa (sin dejar el último año fuera) y se hace una predicción del último año para ver si el modelo estima correctamente los datos.



Igual que en el modelo 2B2, la estimación es muy cercana a los valores reales.

Cálculo de medidas RMSPE/MAPE y mean Length

Medida	Valor
EQM	0.0787035
EAM	0.0670278

5 Selección del mejor modelo

Una vez vistas las validaciones de los modelos y el análisis de sus capacidades predictivas, se debe elegir un solo modelo con el que se llevará a cabo el objetivo principal del presente estudio, la realización de la predicción a largo plazo.

Así pues, en este punto del estudio, es necesario comparar las medidas de adecuación a los datos (AIC) y sus capacidades de predicción, dando más importancia a estas últimas, con el fin de seleccionar un solo modelo, el mejor entre los dos propuestos.

Para ello, se realiza una tabla con los valores más importantes a tener en cuenta.

```
##
## ================================
##                  mod2B2   mod3B
## --------------------------------
## Log Likelihood 101.326  103.014
## AIC            -194.651 -198.029
## RMSPE           0.070    0.079
## MAPE            0.057    0.067
## Mean Length    210.172  172.116
## --------------------------------
```

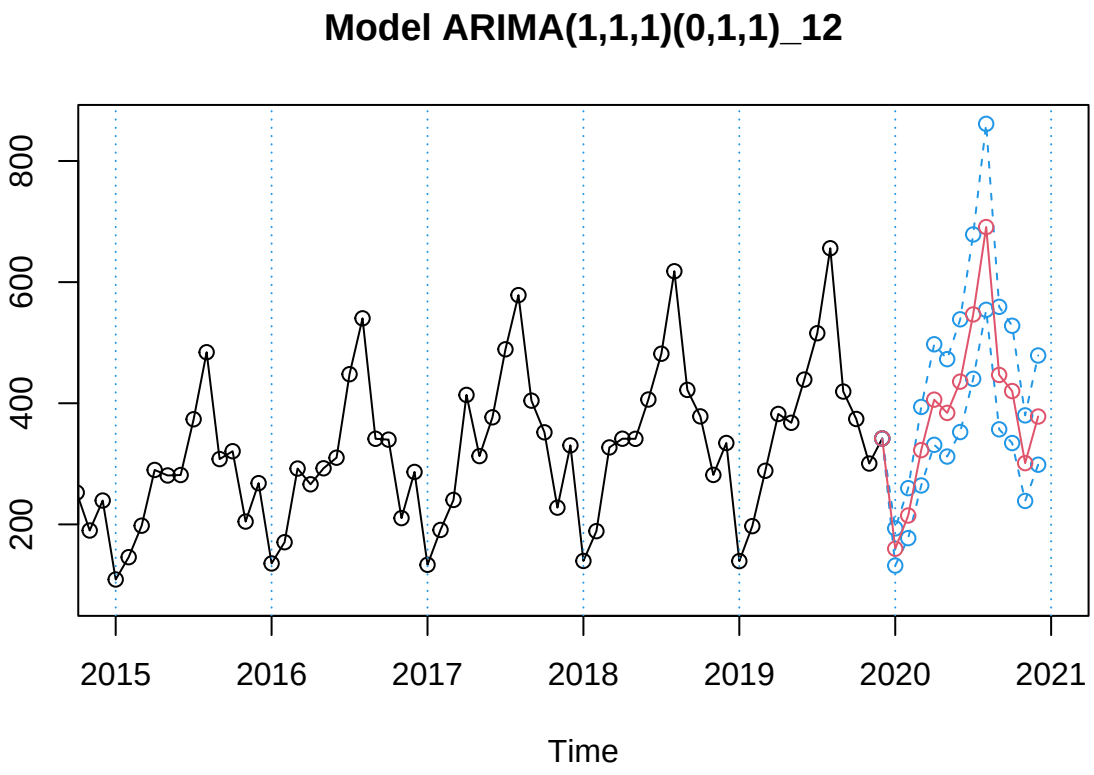
En cuanto a la medida de adecuación a los datos, se observa que el modelo 3B es preferible al 2B2, ya que éste tiene un AIC menor. Por contra, la capacidad predictiva es ligeramente mejor en el modelo 2B2, ya que aquí el error es de 7% en el caso del EQM y un 5.7% en EAM, mientras que el modelo 3B tiene un error del 8% en el EQM y uno del 6.7% en el caso del EAM. Aún así cabe destacar que ambos modelos tienen un error en la capacidad predictiva inferior al 10%, por lo que se concluye que los dos son buenos modelos para predecir los datos estudiados. Además, existe otra medida comparable entre modelos: la amplitud de la predicción. Este valor, interesa que sea pequeño, pues querrá decir que el intervalo de confianza de los valores predichos estará más acotado, será más estrecho. Observando la tabla anterior, se puede decir que el modelo 3B tiene una amplitud de predicción bastante menor que la del modelo 2B2.

Así pues, después de toda esta explicación, se concluye que el mejor modelo para realizar la predicción a largo plazo es el modelo 3B, ya que es el que tiene menor AIC y menor amplitud de predicción. Además, el error no presenta casi diferencias entre ambos modelos.

Antes de pasar a realizar la predicción, es necesario recordar la expresión del modelo elegido. Esta era la siguiente: $ARMA(1,1)SMA(1)_{12}$.

5.0.1 Realización predicción a largo plazo

Una vez seleccionado el mejor modelo, se procede a realizar la predicción. Para llevarla a cabo, se cogen los valores de la serie completa y se estima el número de viajeros en turismo rural que hubiera habido en el año 2020 si no hubiera existido la pandemia del Covid-19.



Se observa que la predicción en un tiempo extrapolable al de los datos sigue la misma tendencia lineal creciente y el mismo patrón que lo visto hasta el momento.

Por tanto, gracias a la predicción observada, se puede decir que ya se ha cumplido el objetivo principal del estudio.

6 Conclusiones

Después de realizar las evaluaciones correspondientes y utilizar los métodos ya conocidos y mencionados con anterioridad, se concluye que la serie temporal correspondiente al número de viajeros españoles en alojamientos de turismo rural se puede explicar como un modelo $ARMA(1,1)SMA(1)_{12}$.

Para llegar a esta conclusión ha sido necesario transformar la serie original con el fin de convertirla en una serie estacionaria. Para ello, se ha debido de aplicar una transformación logarítmica y hacer una diferenciación regular y otra estacional. Una vez conseguida la serie en el formato deseado, se han identificado 3 posibles modelos que podrían encajar con los datos estudiados, aunque uno de ellos ha sido descartado por tener una medida de adecuación a los datos bastante inferior en comparación al resto de opciones. Así pues, a los dos modelos resultantes, se les ha hecho una validación, se ha verificado su estabilidad y se ha observado su capacidad predictiva. A partir de los resultados obtenidos, se ha seleccionado un solo modelo, el mejor de ellos, el cual ha resultado ser un $ARMA(1,1)SMA(1)_{12}$. Este modelo tenía un AIC igual a -198, un error de predicción de 8% y 6.7% en los casos de EQM y EAM, respectivamente y una amplitud de predicción de 172.

Finalmente y a partir del modelo seleccionado, se ha llevado a cabo el objetivo principal del estudio: hacer una predicción para el número de viajeros en turismo rural que hubiera habido en el año 2020 si no hubiera existido la pandemia del Covid-19. En este supuesto, se hubiera esperado la misma tendencia lineal creciente y el mismo patrón mensual que lo visto hasta el momento.

7 Anexo

```
serie = window(ts(read.table("datos.txt", header = F)/1000, start = 2009, freq = 12), star
plot(serie, main = "Ocupacion en alojamientos de turismo rural en España")
abline(v = 2009:2019, col = 4, lty = 3)
```

```
par(mfrow=c(1,2))

m = apply(matrix(serie,ncol=12),2,mean)
v = apply(matrix(serie,ncol=12),2,var)
plot(v~m,main="Mean-Variance plot")

boxplot(serie~floor(time(serie)), main = "Boxplot")
```

```
lnserie=log(serie)
plot(lnserie,type="o")
```

```
monthplot(lnserie)
```

```
d12lnserie<-diff(lnserie,lag=12)
```

```
monthplot(d12lnserie)
```

```
plot(d12lnserie,main="d12lnserie")
abline(h=0)
abline(h=mean(d12lnserie), col=2)
```

```
d1d12lnserie <- diff(d12lnserie)
```

```
plot(d1d12lnserie,main="d1d12lnserie")
abline(h=0)
abline(h=mean(d1d12lnserie), col=2)
```

```
v1 <- var(lnserie)
v2 <- var(d12lnserie)
v3 <- var(d1d12lnserie)
```

```
par(mfrow=c(1,2))

acf(d1d12lnserie,ylim=c(-1,1),lag.max=60,col=c(2,rep(1,11)),lwd=2)
pacf(d1d12lnserie,ylim=c(-1,1),lag.max=60,col=c(rep(1,11), 2),lwd=2)
```

```
mod1A = arima(d1d12lnserie, order = c(0,0,1), seasonal = list(order = c(0,0,1), period=12)
```

```
cat("Modelo 1a \n T-ratios:",round(mod1A$coef/sqrt(diag(mod1A$var.coef)),2))
```

```
mod1B = arima(lnserie, order = c(0,1,1), seasonal = list(order = c(0,1,1), period=12))
```

```
cat("Modelo 1b \nT-ratios:",round(mod1B$coef/sqrt(diag(mod1B$var.coef)),2))
```

```
mod2A = arima(d1d12lnserie, order = c(3,0,0), seasonal = list(order = c(0,0,1), period=12)
```

```
cat("Modelo 2a \nT-ratios:",round(mod2A$coef/sqrt(diag(mod2A$var.coef)),2))
```

```
mod2B = arima(lnserie, order = c(3,1,0), seasonal = list(order = c(0,1,1), period=12))
```

```r
cat("Modelo 2b \nT-ratios:",round(mod2B$coef/sqrt(diag(mod2B$var.coef)),2))

mod1B2 = arima(lnserie,  order = c(3, 1, 0), seasonal = list(order = c(0, 1, 1), period =

mod3A = arima(d1d12lnserie, order = c(1,0,1), seasonal = list(order = c(0,0,1), period=12)

cat("Modelo 3a \nT-ratios:",round(mod3A$coef/sqrt(diag(mod3A$var.coef)),2))

mod3B = arima(lnserie, order = c(1,1,1), seasonal = list(order = c(0,1,1), period=12))

cat("Modelo 3b \nT-ratios:",round(mod3B$coef/sqrt(diag(mod3B$var.coef)),2))

#install.packages("stargazer")
library("stargazer")
stargazer(mod1A, mod1B2, mod3B, title="Results", type="text", notes.append = FALSE, report
notes = c("t = T-statistic value = coeff/SE(coeff)"), digits = 3,  column.labels = c("1A",

validation_grafic <- function(model, dades){

s = frequency(get(model$series))
resid = model$residuals
par(mfrow = c(2, 2), mar = c(3, 3, 3, 3))

#Residuals plot
plot(resid, main = "Residuals")
abline(h = 0)
abline(h = c(-3*sd(resid, na.rm = T), 3*sd(resid, na.rm = T)), lty = 3, col = 4)

#Square Root of absolute values of residuals (Homocedasticity)
scatter.smooth(sqrt(abs(resid)), main = "Square Root of Absolute residuals",
lpars = list(col = 2))

#Normal plot of residuals
qqnorm(resid)
qqline(resid, col = 2, lwd = 2)

#Histogram of residuals with normal curve
hist(resid, breaks = 20, freq = F)
curve(dnorm(x, mean = mean(resid, na.rm = T), sd = sd(resid, na.rm = T)), col = 2, add = T
}

validation_test_homoce <- function(model, dades){

suppressMessages(require(lmtest, quietly = TRUE, warn.conflicts = FALSE))

##Breusch-Pagan test (vs. order)
obs = get(model$series)
print(bptest(resid(model)~I(1:length(resid(model)))))

##Breusch-Pagan test (vs. predictions)
obs = get(model$series)
print(bptest(resid(model)~I(obs-resid(model))))
}
```

```r
validation_test_normal <- function(model, dades){

##Shapiro-Wilks Normality test
print(shapiro.test(resid(model)))
suppressMessages(require(nortest, quietly = TRUE, warn.conflicts = FALSE))

##Anderson-Darling test
print(ad.test(resid(model)))
suppressMessages(require(tseries, quietly = TRUE, warn.conflicts = FALSE))

##Jarque-Bera test
print(jarque.bera.test(na.omit(c(resid(model)))))
}
```

```r
validation_test_indepen <- function(model, dades){
s = frequency(get(model$series))
resid = model$residuals

##Durbin-Watson test
print(dwtest(resid(model)~I(1:length(resid(model)))))

##Ljung-Box test
cat("\nLjung-Box test\n")
print(t(apply(matrix(c(1:4, (1:4)*s)), 1, function(el) {
te = Box.test(resid(model), type = "Ljung-Box", lag = el)
c(lag = (te$parameter), statistic = te$statistic[[1]], p.value = te$p.value)})))
}
```

```r
validation_acf_pacf <- function(model, dades){
  s = frequency(get(model$series))
  resid = model$residuals
  #ACF & PACF of residuals
  par(mfrow=c(1,2))
  acf(resid,ylim=c(-1,1),lag.max=60,col=c(2,rep(1,s-1)),lwd=1)
  pacf(resid,ylim=c(-1,1),lag.max=60,col=c(rep(1,s-1),2),lwd=1)
  par(mfrow=c(1,1))

  #ACF & PACF of square residuals
  par(mfrow=c(1,2))
  acf(resid^2,ylim=c(-1,1),lag.max=60,col=c(2,rep(1,s-1)),lwd=1)
  pacf(resid^2,ylim=c(-1,1),lag.max=60,col=c(rep(1,s-1),2),lwd=1)
  par(mfrow=c(1,1))
}
```

```r
model = mod1B2
validation_grafic(model)
```

```r
validation_test_homoce(model)
```

```r
validation_test_normal(model)
```

```r
validation_test_indepen(model)
```

```
validation_acf_pacf(model)
```

```
model = mod3B
validation_grafic(model)
```

```
validation_test_homoce(model)
```

```
validation_test_normal(model)
```

```
validation_test_indepen(model)
```

```
validation_acf_pacf(model)
```

```
ultim = c(2018,12)                          #Dic 2018

serie1 = window(serie, end = ultim + c(1,0))  #complete series: 2009-2019
lnserie1 = log(serie1)                      #log transformed
serie2 = window(serie, end = ultim)           #series without last year obsrvations: 2009-2
lnserie2 = log(serie2)                      #log transformed
```

```
# Fit the model to the complete series: lnserie1
(mod1B2 = arima(lnserie1,  order = c(3, 1, 0), seasonal = list(order = c(0, 1, 1), period

#Fit the model to the subset series (without 2019 data): lnserie2
(mod1B22 = arima(lnserie2,  order = c(3, 1, 0), seasonal = list(order = c(0, 1, 1), period
```

```
pred = predict(mod1B22, n.ahead=12)                                  #outputs point predictio
pr <- ts(c(tail(lnserie2,1),pred$pred),start = ultim, freq=12)  #point predictions

se <- ts(c(0,pred$se), start = ultim, freq=12)                       #Standard errors for poi

#Prediction Intervals (back transformed to original scale using exp-function)
tl <- ts(exp(pr-1.96*se), start = ultim, freq = 12)
tu <- ts(exp(pr+1.96*se), start = ultim, freq = 12)
pr <- ts(exp(pr), start = ultim, freq = 12)              #predictions in original scale

#Plot of the original airbcn series (thousands) and out-of-sample predictions: only time
ts.plot(serie,tl,tu,pr,lty=c(1,2,2,1),col=c(1,4,4,2),xlim=ultim[1]+c(-3,+2),type="o",main=
abline(v=(ultim[1]-3):(ultim[1]+2),lty=3,col=4)
```

```
obs=window(serie,start=ultim)

mod.EQM1=sqrt(sum(((obs-pr)/obs)^2)/12)    # Error = obs - pred
mod.EAM1=sum(abs(obs-pr)/obs)/12
mod.ML1=sum(tu-tl)/12
```

```
# Fit the model to the complete series: lnserie1
(mod3B = arima(lnserie1, order = c(1,1,1), seasonal = list(order = c(0,1,1), period=12)))

#Fit the model to the subset series (without 2019 data): lnserie2
(mod3B2 = arima(lnserie2, order = c(1,1,1), seasonal = list(order = c(0,1,1), period=12)))
```

```
pred = predict(mod3B2, n.ahead=12)                                  #outputs point prediction
pr <- ts(c(tail(lnserie2,1),pred$pred),start = ultim, freq=12)  #point predictions
```

```
se <- ts(c(0,pred$se), start = ultim, freq=12)                    #Standard errors for poi

#Prediction Intervals (back transformed to original scale using exp-function)
tl <- ts(exp(pr-1.96*se), start = ultim, freq = 12)
tu <- ts(exp(pr+1.96*se), start = ultim, freq = 12)
pr <- ts(exp(pr), start = ultim, freq = 12)              #predictions in original scale

#Plot of the original airbcn series (thousands) and out-of-sample predictions: only time
ts.plot(serie,tl,tu,pr,lty=c(1,2,2,1),col=c(1,4,4,2),xlim=ultim[1]+c(-3,+2),type="o",main=
abline(v=(ultim[1]-3):(ultim[1]+2),lty=3,col=4)
```

```
obs=window(serie,start=ultim)
mod.EQM2=sqrt(sum(((obs-pr)/obs)^2)/12)    # Error = obs - pred
mod.EAM2=sum(abs(obs-pr)/obs)/12
mod.ML2=sum(tu-tl)/12
```

```
selection=function(model){
 s=frequency(get(model$series))
  resid=model$residuals
  par(mfrow=c(2,2),mar=c(3,3,3,3))

resumen<-data.frame(Pruebas=1:5)
colnames(resumen)<-paste0("mod1B")
rownames(resumen)<-c("Log Likelihood","AIC","RMSPE", "MAPE","Mean Length")

  resumen[1,1]=model$loglik
  resumen[2,1]=model$aic
  resumen[3,1]=NA
  resumen[4,1]=NA
  resumen[5,1]=NA
  return(resumen)
}
```

```
model = mod1B2
resumen1 <- selection(model)
colnames(resumen1) <- c("mod2B2")
resumen1[3,1] = mod.EQM1
resumen1[4,1] = mod.EAM1
resumen1[5,1] = mod.ML1

model = mod3B
resumen2 <- selection(model)
colnames(resumen2) <- c("mod3B")
resumen2[3,1] = mod.EQM2
resumen2[4,1] = mod.EAM2
resumen2[5,1] = mod.ML2


tablef<-cbind.data.frame(resumen1,resumen2)
stargazer(tablef, summary=FALSE, type="text")
```

```
##### Previsions a llarg termini amb el model complet ######
```

```
pred=predict(mod3B,n.ahead=12)
pr<-ts(c(tail(lnserie,1),pred$pred),start=ultim+c(1,0),freq=12) #starts Dec 2019!
se<-ts(c(0,pred$se),start=ultim+c(1,0),freq=12)

#Intervals
tl1<-ts(exp(pr-1.96*se),start=ultim+c(1,0),freq=12)
tu1<-ts(exp(pr+1.96*se),start=ultim+c(1,0),freq=12)
pr1<-ts(exp(pr),start=ultim+c(1,0),freq=12)

ts.plot(serie,tl1,tu1,pr1,lty=c(1,2,2,1),col=c(1,4,4,2),xlim=c(ultim[1]-3,ultim[1]+3),type
abline(v=(ultim[1]-3):(ultim[1]+3),lty=3,col=4)
```