Miguel Ángel Castaño Ibáñez

ANÁLISIS Y PREDICCIÓN DE SERIES TEMPORALES

1. **Introducción: Presentación de la serie a analizar. Descripción de la misma y objetivo que se persigue con su estudio. (0.75 puntos)**

   El turismo rural ha experimentado un gran auge, culpa de ello la tienen los Millennials principales precursores de este turismo en los últimos años. También una tendencia de viaje mas al norte ha favorecido a ello. El tipo de alojamiento preferido más utilizado son los albergues, casas rurales o campings.

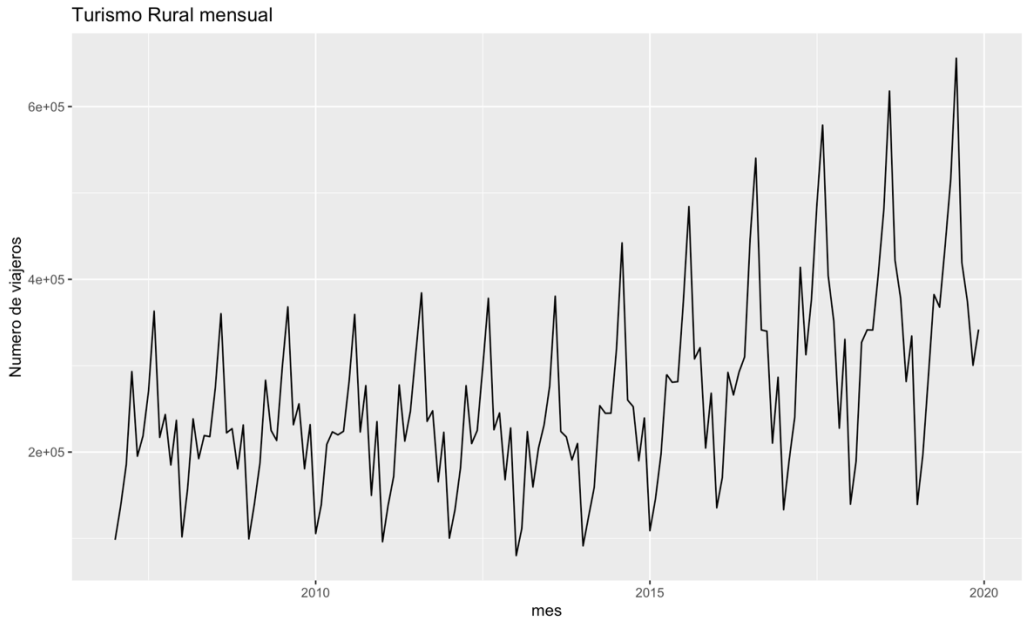
   Por todo ello he decidido analizar el turismo rural en España en base al número de viajeros, tanto nacionales como internacionales, que deciden hospedarse en casas rurales. Además, he escogido los denominados viajeros (más de una noche) en lugar de aquellos que deciden solo pernoctar.

   Los datos han sido extraídos del INE (Instituto Nacional de Estadística), pueden obtenerse aquí. Realizamos la siguiente consulta:

   El objetivo de este estudio de la serie temporal pretende predecir de los datos futuros de la serie temporal, analizando la evolución del numero de viajeros en los últimos 13 años.

2. **Representación gráfica y descomposición de la misma. (1.5 puntos)**

   En la gráfica que se muestra a continuación podemos observar la evolución de la serie temporal para el turismo rural en los años 2007-2019 (líneas 33-37).

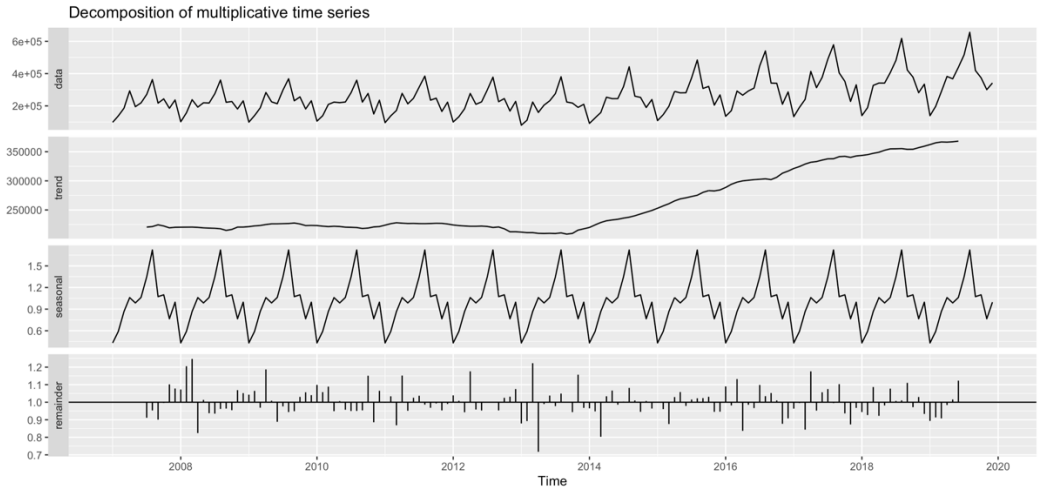


Observamos que es una serie estacional con una tendencia creciente y con una variabilidad mayor. Por estos motivos descomponemos la serie temporal conforme al esquema multiplicativo.

Con la descomposición conseguiremos separar sus componentes: tendencia, componente estacional, componente cíclica y componente aleatoria o irregular. La componente cíclica se incluirá en la tendencia ya que involucrar grandes periodos de tiempo y es difícil detectar con claridad (línea 40).

$$X_t = T_t \cdot S_t \cdot C_t \cdot Z_t \quad \longrightarrow \quad X_t = T_t * S_t * Z_t$$

La serie temporal descompuesta la podemos ver en la siguiente gráfica (línea 43), donde mostramos los datos originales, la tendencia, la parte secuencial, y las irregularidades.

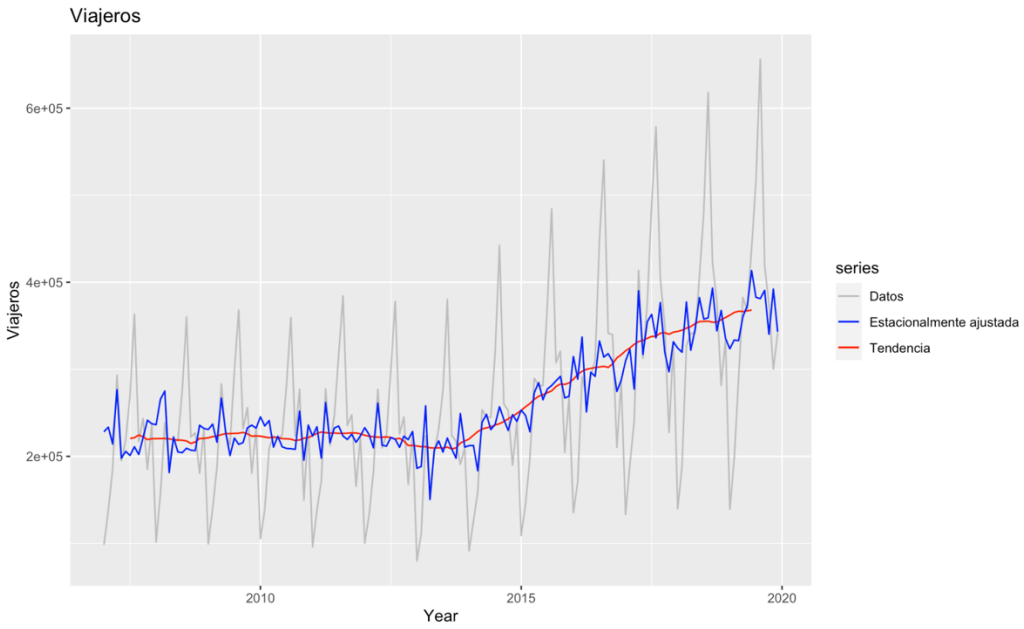


Analizando la gráfica podemos ver como tendencia presenta al principio un comportamiento estable e incluso decreciente, pero a partir de el último trimestre de 2013 crece de forma lineal y finalmente en el ultimo tramo presenta forma logarítmica. Por otro lado, observamos como la irregularidad no toma valores muy grandes, esto quiere decir que nuestra serie no presenta una gran irregularidad cada año.

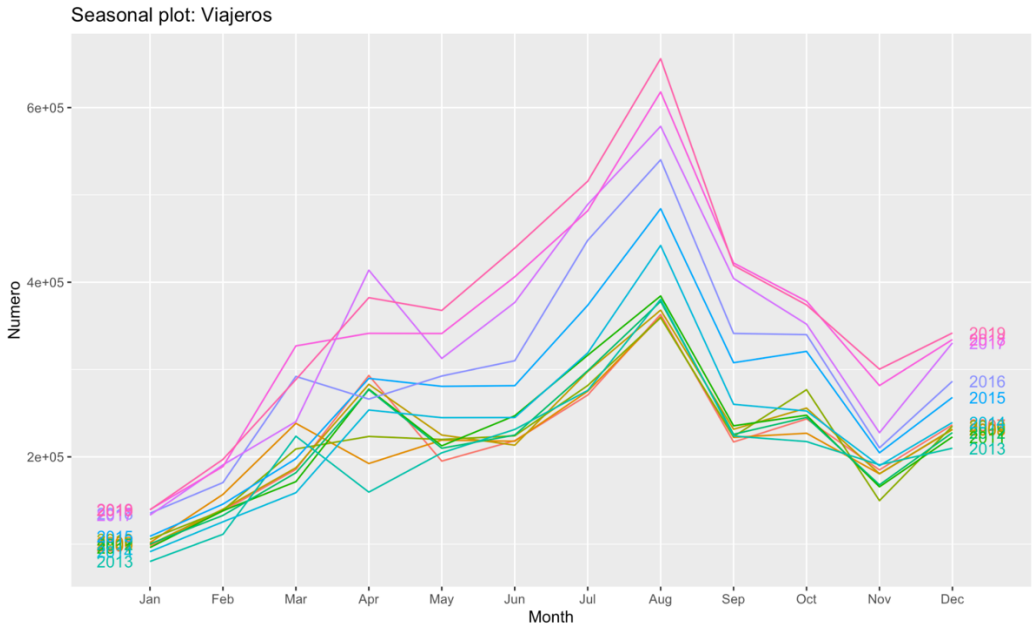
En la línea 46 obtenemos los coeficientes de estacional, este dato nos indica el porcentaje en el que varia en numero de viajeros cada mes, respecto a la media estacional.

```
          Jan       Feb       Mar       Apr       May       Jun       Jul       Aug       Sep       Oct       Nov       Dec
2007 0.4306267 0.5908783 0.8667219 1.0604054 0.9863176 1.0621812 1.3469432 1.7206267 1.0734222 1.0990444 0.7661028 0.9967295
```

A continuación, superponemos a los datos originales el ajuste estacional y la tendencia. Podemos observar como realmente no se produce un gran ajuste, ya que como hemos comprobado anteriormente no presenta una gran irregularidad.



Finalmente toca ver el comportamiento de la serie estacional cada año, superponiendo todos ellos. Aquí, una vez mas comprobamos como los primeros años están muy juntos en las líneas inferiores tras 2014 están mas separadas, exceptuando los dos últimos años que presenta el crecimiento logarítmico en la tendencia. Podemos destacar el año 2013 como el que menos valor tienes ya que la tendencia hasta ese año es decreciente muy lenta y 2019 el año con mayores viajeros. A demás, podemos ver que, aunque a veces se crucen entre si, como casi todas las líneas presentan la misma prácticamente la misma traza sin muchas irregularidades (líneas 61- 65). La gráfica a continuación:



3. **Para comprobar la eficacia de los métodos de predicción que vamos a hacer en los siguientes apartados reservamos los últimos datos observados (un periodo en las series estacionales o aproximadamente 10 observaciones) para comparar con las predicciones realizadas por cada uno de los métodos. Luego ajustamos los modelos sobre la serie sin esos últimos datos en los siguientes apartados.**

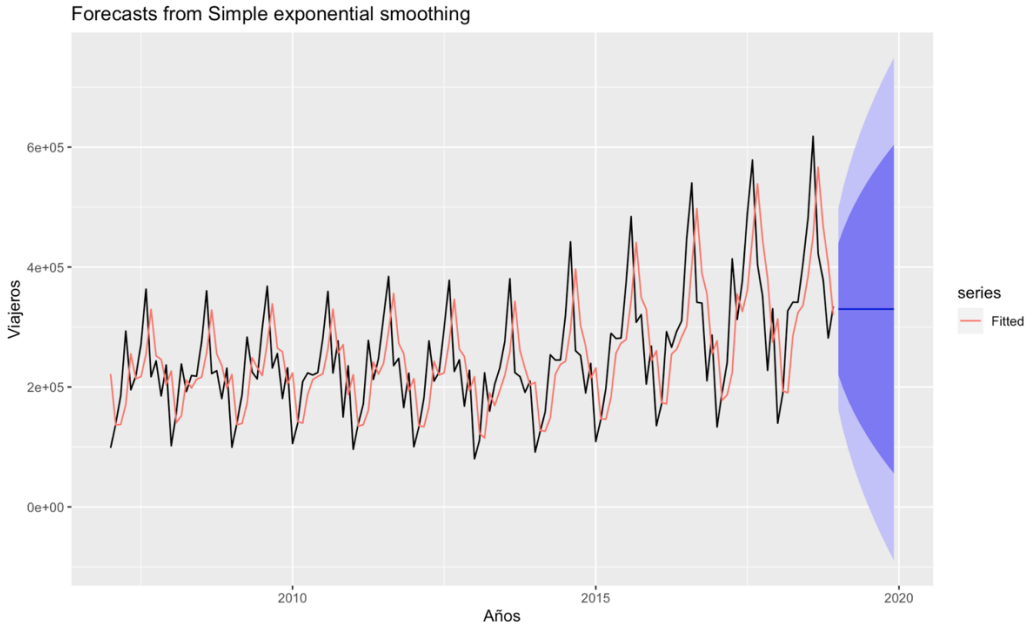
   Eliminamos el ultimo periodo, es decir las 12 observaciones correspondientes a los meses del año 2019, para ajustar los modelos de la serie sin estos datos y finalmente comprobar nuestra predicción. Esto lo haremos en la línea 72.

```
71  # eliminar del fichero las observaciones del ultimo periodo (12 meses)
72  viajeros_TR<-window(viajeros,start=c(2007,1), end=c(2018,12))
```

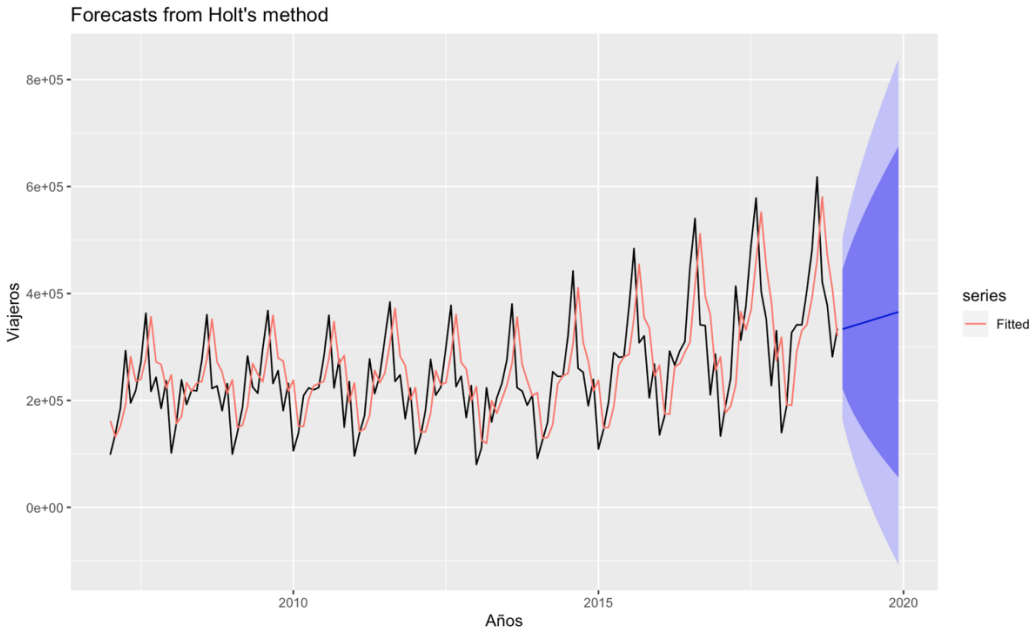
4. **Encontrar el modelo de suavizado exponencial más adecuado. Para dicho modelo, representar gráficamente la serie observada y la serie suavizada con las predicciones para un periodo que se considere adecuado. Escribir la expresión del modelo obtenido. Explicar detalladamente los pasos a realizados y sus justificaciones. (2 puntos)**

   En los modelos de suavizado exponencial podemos destacar 3:

   Método alisado simple. Este método es idóneo para series que no presentan tendencia y tampoco estacionalidad, debido que nos proporcionara solo el próximo valor que podría tomar la serie, en lugar de predecir todos los valores del periodo siguiente. Esto se puede ver en la siguiente gráfica como se representa en línea recta para los 12 valores siguientes (líneas 83 - 98).



Método de alisado doble de Holt. Este método esta recomendado para series que tendencia, pero no estacionalidad. Este método nos revelara una serie de valores con una tendencia creciente o decreciente. Este modelo creo que tampoco es adecuado para nuestra serie estacional ya que nos proporcionara 12 valores con una tendencia creciente, como se puede apreciar en la gráfica siguiente, pero sin tener en cuenta la estacionalidad que se presenta en los periodos (líneas 101 - 116).



Método Holt-Winters. Este método si está recomendado para las series que presentan tendencia y estacionalidad, por lo tanto, nos proporcionará el modelo de suavizado correcto. Teine en cuenta la estacionalidad y la tendencia que presenta nuestra serie (líneas 119-134).

En la línea 124 con la función summary podemos observar los parámetros AIC para ver la bondad del ajuste, los intervalos de confianza en cada pronostico y los paramentos alfa (α) ,beta (β) y gamma (γ).

```
Smoothing parameters:
  alpha = 0.1509
  beta  = 0.0147
  gamma = 1e-04
```

Una vez obtenidos los parámetros podemos sustituirlos en las funciones para el modelo multiplicativo que son:

$$L_t = \alpha \frac{x_t}{S_{t-s}} + (1-\alpha)(L_{t-1} + b_{t-1})$$

$$\boldsymbol{L_t = 0.1509 \frac{x_t}{S_{t-1}} (1 - 0.1509)(L_{t-1} + b_{t-1})}$$

$$b_t = \beta(L_t - L_{t-1}) + (1-\beta) b_{t-1}$$

$$\boldsymbol{b_t = 0.0147(L_t + L_{t-1}) + (1 - 0.0147) b_{t-1}}$$

$$S_t = \gamma \frac{x_t}{L_t} + (1-\gamma) S_{t-s}$$

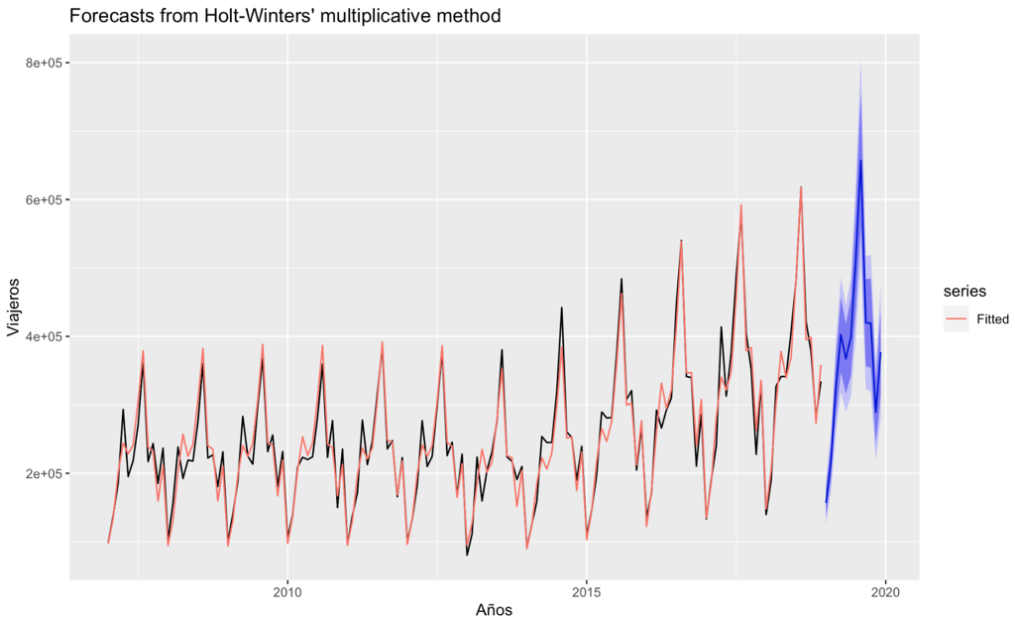
$$\boldsymbol{S_t = 0.0001 \frac{x_t}{L_t} (1 - 0.0001)\, S_{t-1}}$$

Finalmente, para predecir el valor de cada observación aplicamos esta formula, donde en el ejemplo mostramos como la usaríamos para calcular el valor de enero de 2019, y así sucesivamente para los meses posteriores. Tenemos 156 observaciones a las que le hemos sustraído las últimas 12 (el último año), por lo tanto, nos quedan 144 observaciones, es decir t = 144.

$$\hat{x}_{t+1} = (L_t + b_t) S_{t+1-s}$$

$$\boldsymbol{\hat{x}_{145} = (L_{144} + b_{144}) b_{144+1-12}}$$

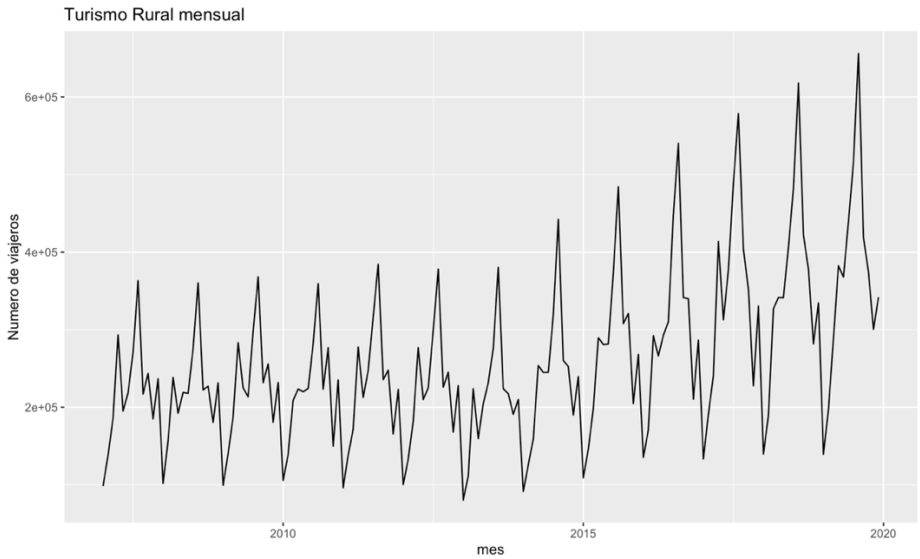
R ya calcula automáticamente todos estos puntos y como resultado obtenemos la siguiente gráfica con el modelo ajustado y la partes azules los intervalos de confianza.



5. **Representar la serie y las funciones de autocorrelación y autocorrelación parcial. Decidir que modelo puede ser ajustado. Ajustar el modelo adecuado comprobando su idoneidad. (Sintaxis, tablas de los parámetros estimados y gráficos) Explicar detalladamente los pasos a realizados y su justificación. (3 puntos)**

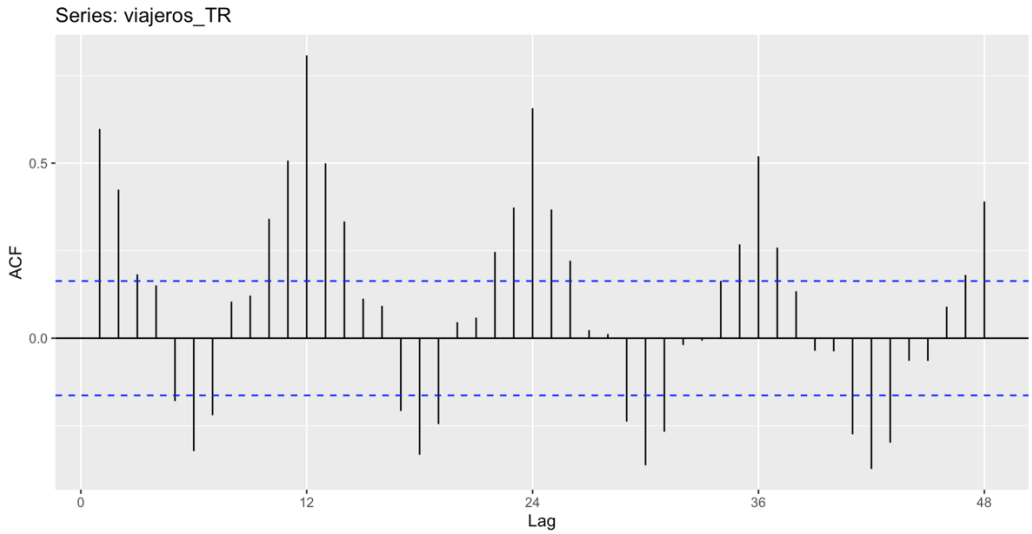
   Antes de trabajar para sacar las funciones de autocorrelación simple y parcial, podemos considerar hacer una transformación logarítmica de nuestra serie, ya que esta va ganando variabilidad con respecto al tiempo. Yo por mi parte no he considerado hacer necesaria esta transformación ya que en el esquema multiplicativo se tendrá en cuenta el crecimiento de la variabilidad respecto al tiempo. Aun así, este modelo tendrá que ser comprobado y validado o si por el contrario tendríamos que tomar una transformación logarítmica o de cualquier otro tipo.

   Representamos de nuevo la serie temporal (líneas 129-134):

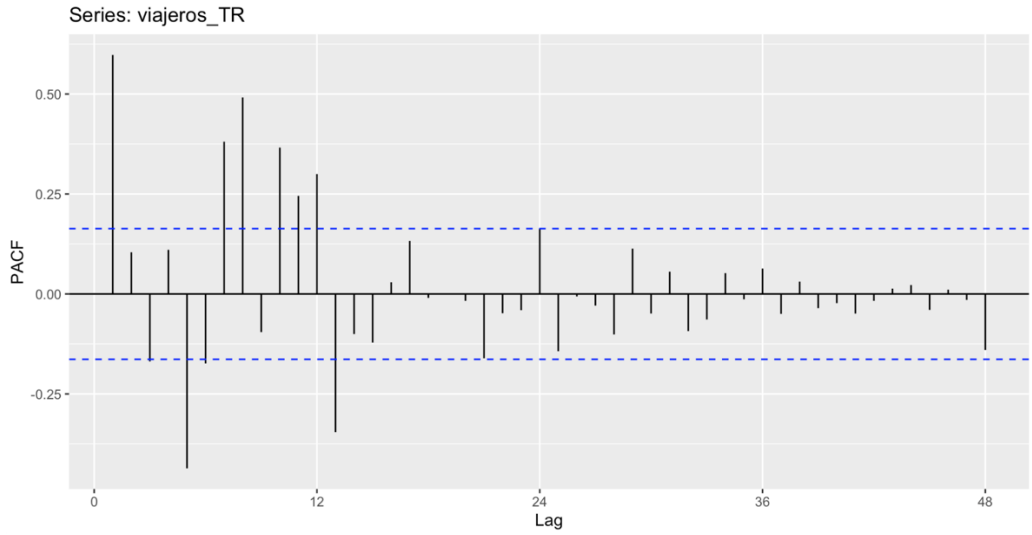


Obtenemos las funciones de autocorrelación simple y parcial de nuestra serie temporal con hasta 48 retardos (48 meses), es decir, veremos el efecto que tiene en 4 años.

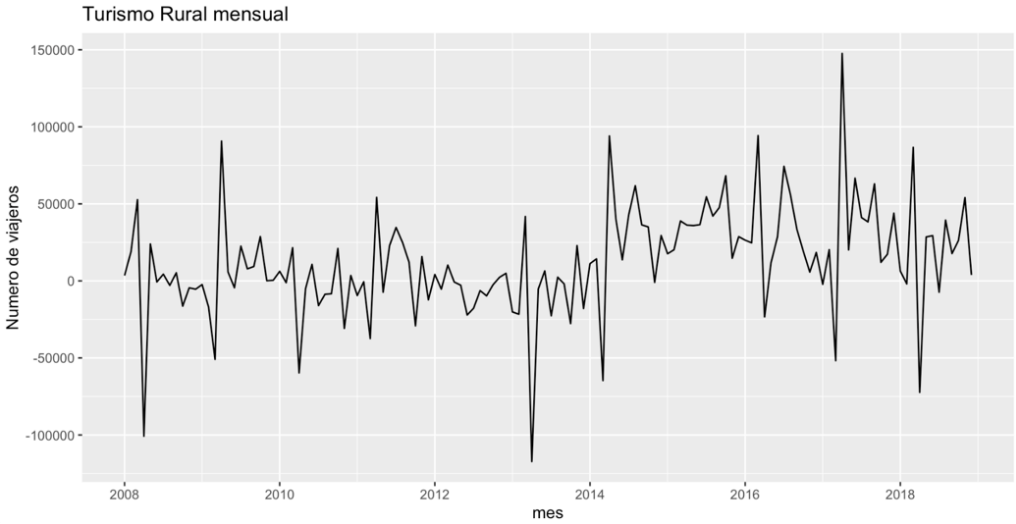
Función de autocorrelación simple (ACF): vemos en la gráfica a continuación, como en los retardos que corresponden con la estacionalidad (múltiplos de 12) los retardos son significativos con comportamientos similares, un claro indicio de la estacionalidad de la serie. A demás, vemos como va decreciendo lentamente en el valor del retardo en los retardos múltiplos a la estacionalidad (12), por ello tenemos que tomar una diferenciación de orden estacional para desestacionalizar esta serie (línea 155).



Función de autocorrelación parcial (PACF): no le vamos a prestar especial atención a la gráfica resultante que se muestra a continuación, ya que en este momento como hemos dicho anteriormente tenemos que aplicar a nuestra serie una diferenciación de orden estacional y volver a calcular ambas funciones de autocorrelación (línea 157).

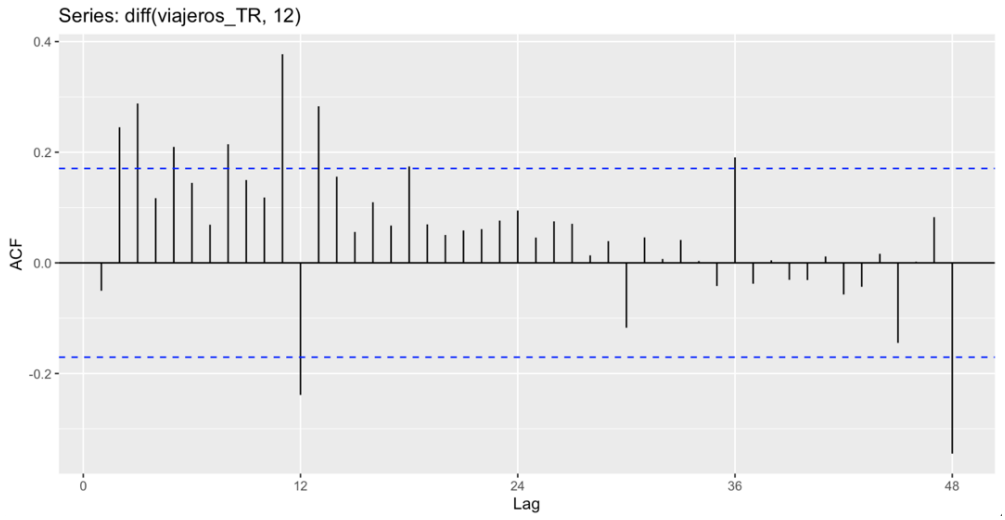


Representamos nuevamente la serie una vez diferenciada en la siguiente gráfica y vemos como la serie ya no presenta estacionalidad de orden 12, pero si que tendremos que volver ha hacer una diferenciación de orden uno ya que la variabilidad es creciente en la serie original. A demás, podemos apreciar en la siguiente gráfica como la media no es constante.

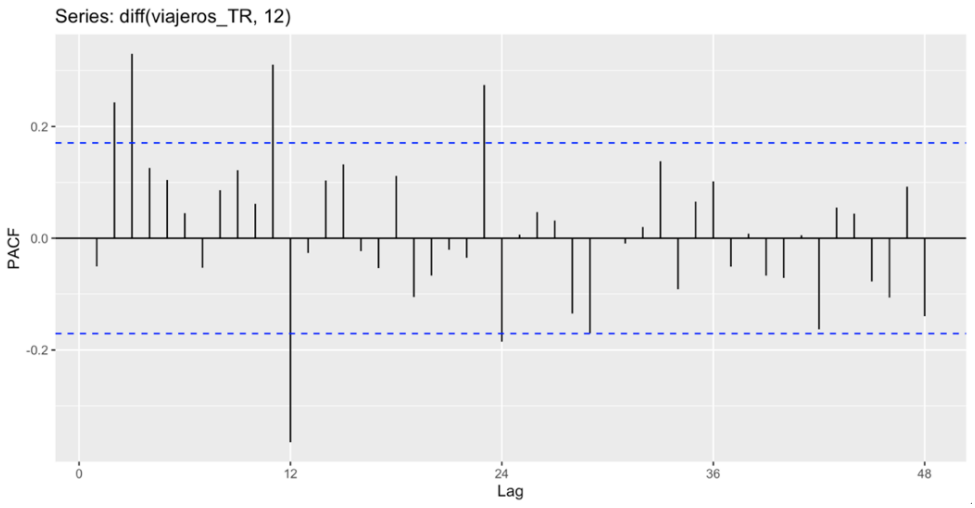


Vamos a volver a representar los correlogramas para comprobar si se ha perdido la estacionalidad (líneas 159-165).

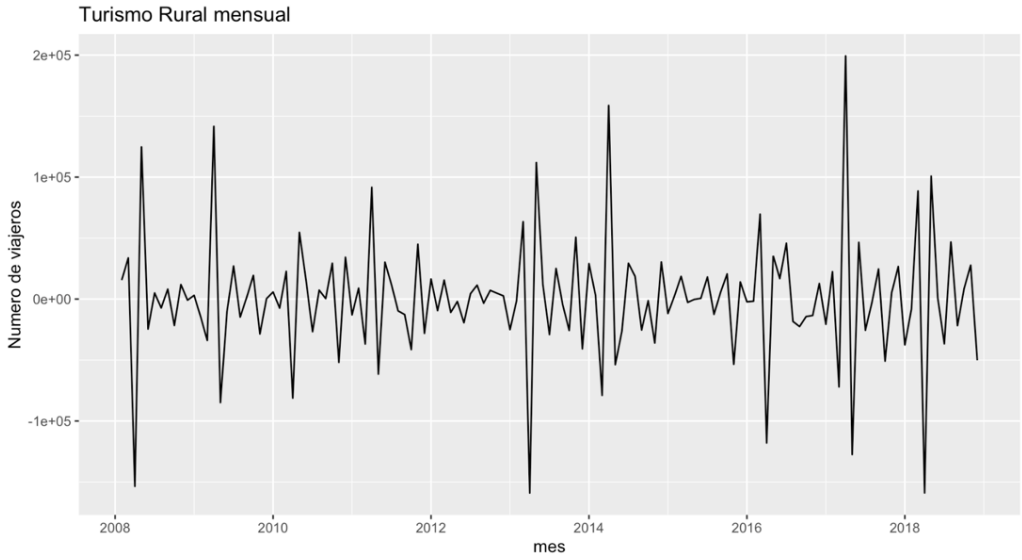
Función de autocorrelación simple (ACF): como podemos ver en la gráfica de a continuación, ya hemos perdido la estacionalidad en los retardos múltiplos a 12, sin embargo, podríamos considerar que decrece lentamente en los primeros retardos por ello será necesaria otra diferenciación regular de orden 1 (línea 170).



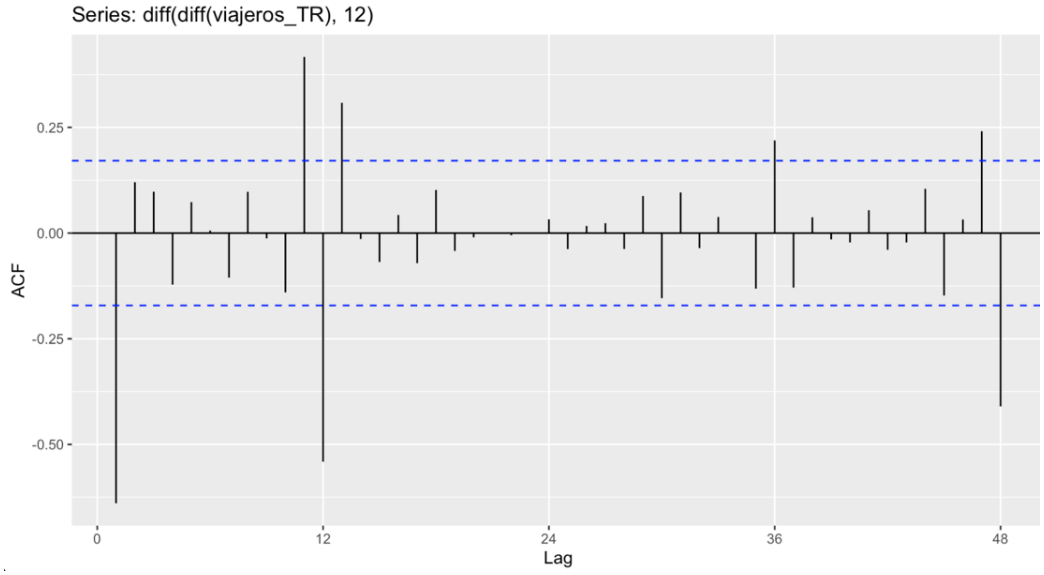
Función de autocorrelación parcial (PACF): que tampoco prestamos especial atención, ya que vamos a diferenciar nuevamente nuestra serie temporal (línea 172).



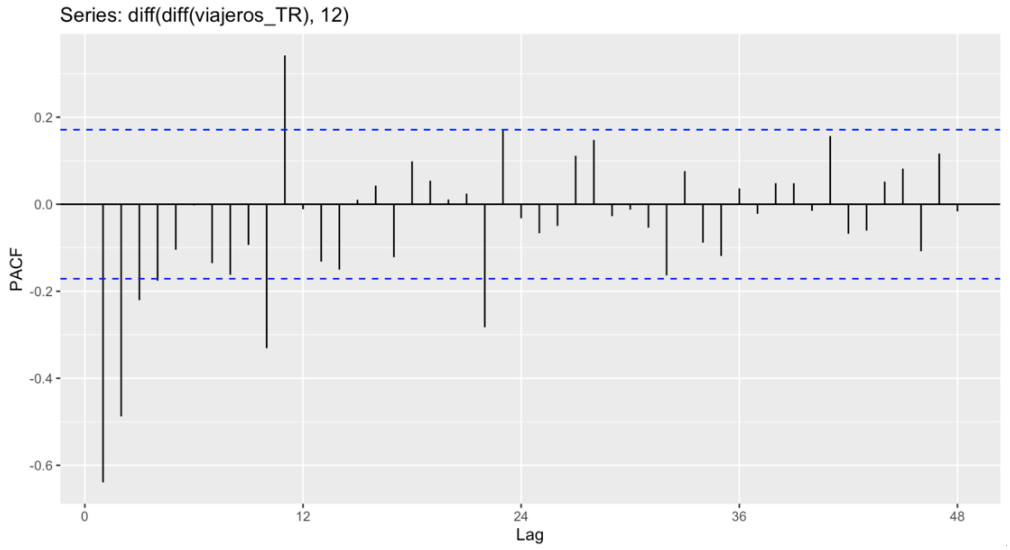
Ahora si vemos la seria diferenciada que nos ayudara a obtener los parámetros del modelo ARIMA. La gráfica resultante en esta serie que si se muestra estacionaria es la siguiente (líneas 184-188).



Función de autocorrelación simple ACF: esta es la gráfica resultante donde podemos comprobar que efectivamente si es estacionaria nuestra serie ya que no decrece lentamente (línea 191).



Función de autocorrelación parcial (PACF): esta gráfica mostrada a continuación representa la autocorrelación parcial de la serie, que junto a la de autocorrelación simple nos van a ayudar a determinar los ordenes de la parte ARIMA regular y ARIMA estacional (línea 172).



6. Escribir la expresión algebraica del modelo ajustado con los parámetros estimados. (1punto)

Con las gráficas ya obtenidas anteriormente vamos a obtener los coeficientes de la función ARIMA. La formula presenta el siguiente aspecto:

$$ARIMA(p,d,q)(P,D,Q)_s$$

Los primeros pertenecen a la parte regular y los segundos son de la parte estacional.

Parámetros:

- **Parte autorregresiva (p y P):** La función de autocorrelación parcial (PACF) muestra como en los retardos múltiplos de 12, al menos 2 son significativos, por lo tanto, para la parte estacional **P = 2**.

- **Parte de media móvil (q y Q):** Podemos ver para la parte regular como en la función de correlación simple (ACF) la función solo adquiere un valor significativo entre los primeros valores, sin embargo, en la de correlación parcial va decreciendo lentamente, por lo tanto, concluimos que **p = 0** y **q = 1**.

  Por otro lado, en la parte estacional fijándonos en los múltiplos de 12 para la función de autocorrelación simple (ACF) vemos como si para el retardo 12 es significativo y en los demás tienen un valor cercano a 0, como en el 24, o en el caso del 36 y 48 muy próximo a las bandas, por tanto, un valor próximo a cero. Por ello decimos que **Q = 1**.

- **Parte de integrada o de diferenciación (d y D):** estas será **d = 1** y **D = 1**, ya que te han hecho una diferenciación tanto en la parte estacional como la regular.

- Periodo (S): Se corresponde con el periodo de la serie temporal con la que estamos trabajando, es decir, 12.

Una vez estimados esto coeficientes obtenemos como resultado:

$$\boldsymbol{ARIMA}\ (\mathbf{0},\mathbf{1},\mathbf{1})(\mathbf{2},\mathbf{1},\mathbf{1})_{12}$$

Otra manera de conocer estos coeficientes en R sería aplicar la función *auto.arima().* Esta también la podemos utilizar de guía para saber si la estimación de estos coeficientes hecha con respecto a las gráficas de autocorrelación podría ser adecuada.

Con la función *auto.arima()* en las líneas 200 y 201 obtenemos:

$$\boldsymbol{ARIMA}\ (\mathbf{0},\mathbf{1},\mathbf{1})(\mathbf{2},\mathbf{1},\mathbf{1})_{12}$$

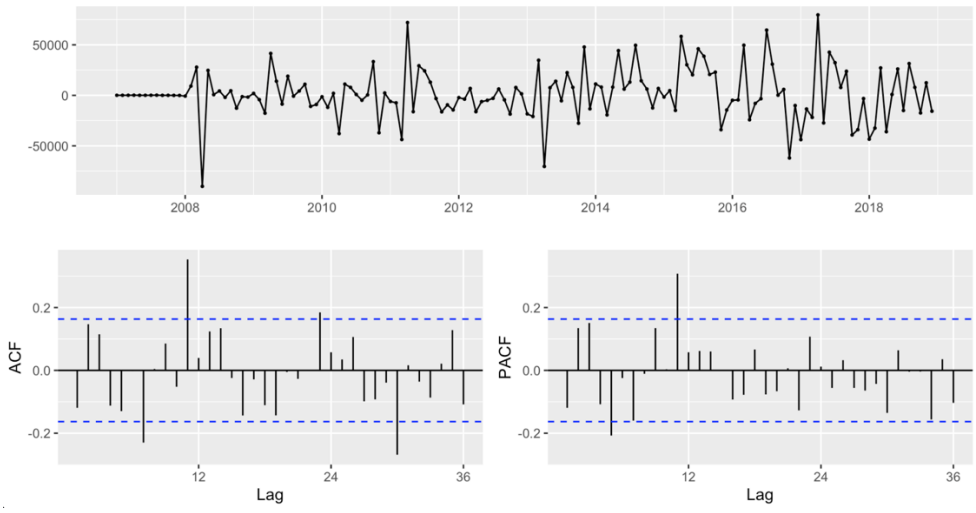
Efectivamente confirmamos que la estimación podría ser correcta, por lo tanto, ajustamos el modelo (líneas 204 - 205) y procedemos a su comprobar la idoneidad del modelo (línea 208). La salida de la función nos dice tras rechazar las hipótesis nulas que todos nuestros parámetros del modelo son significativos.

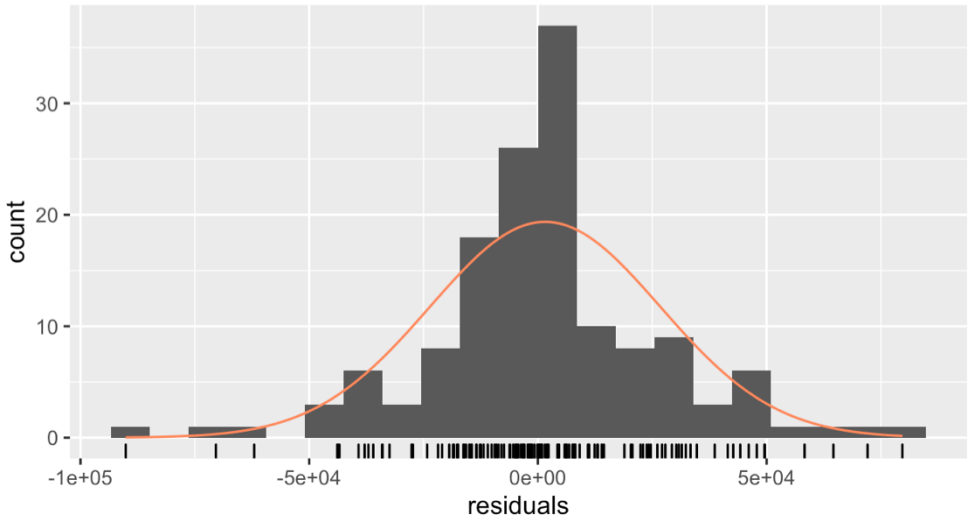
```
z test of coefficients:

     Estimate Std. Error  z value  Pr(>|z|)
ma1  -0.817757   0.048547 -16.8445 < 2.2e-16 ***
sar1 -1.206282   0.091611 -13.1674 < 2.2e-16 ***
sar2 -0.645169   0.076913  -8.3883 < 2.2e-16 ***
sma1  0.652121   0.097203   6.7089 1.962e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

7. **Calcular las predicciones y los intervalos de confianza para las unidades de tiempo que se considere oportuno (detallar el porqué de esas unidades), dependiendo de la serie, siguientes al último valor observado. Representarlas gráficamente. (1 punto)**

Graficamos el análisis de los residuos que nos permitirá saber si nuestra hipótesis es correcta (líneas 215 - 217). Los residuos como vemos en las siguientes gráficas, pese a ser significativos en el primer retardo de la parte estacional en ACF y PACF tienden a comportarse como un ruido blanco, por tanto, podemos afirmar que es una estimación adecuada.





El contraste Ljung-Box nos permite comprobar si los residuos son incorrelados. Siendo esta la hipótesis nula asociada a este contraste. Por tanto, tras el calculo de las funciones anteriores obtenemos, que p-value es demasiado pequeño, menor que 0.05, rechazando así la hipótesis nula y podemos afirmar los residuos que no son incorrelados, es decir tienen correlación, así estos se distribuyen de forma normal como podemos ver en la grafica anterior. Por todo ello, este modelo podría ser valido aun sin cumplir la hipótesis de correlación de residuos, aun que no creo que sea el modelo más adecuado para la predicción de nuestra serie temporal o incluso probar otras transformaciones de la serie original.

```
        Ljung-Box test

data:  Residuals from ARIMA(0,1,1)(2,1,1)[12]
Q* = 62.507, df = 20, p-value = 2.894e-06
```

Otra opción podría ser probar con otras estimaciones que nos dejaran aceptar esta hipótesis nula. La siguiente estimación de parámetros estarían dentro de los modelos que aceptarían la hipótesis nula.

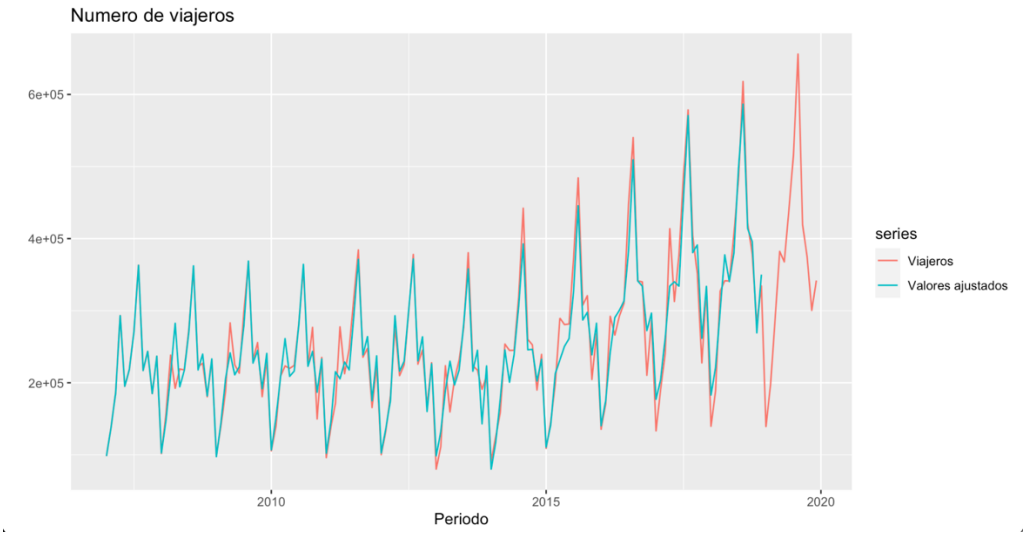
$$\boldsymbol{ARIMA}\ (\mathbf{0},\mathbf{1},\mathbf{2})(\mathbf{1},\mathbf{2},\mathbf{3})_{12}$$

Aun así, este modelo tras ver las graficas presentadas se ajustan menos a los valores reales, por ello he decidido descartarla(líneas 240- 267).

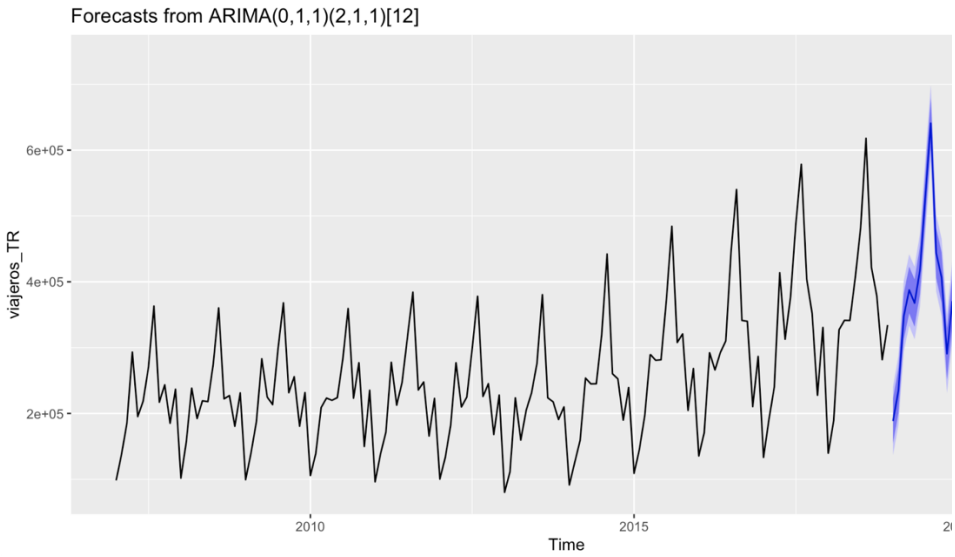
Vamos a calcular las predicciones para el año siguiente (2019), con la ayuda de la función *forecast()* dándonos como resultado las siguientes estimaciones junto a su intervalo de confianza (líneas 219 - 221).

	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
Jan 2019	188428.0	154644.9	222211.0	136761.2	240094.7
Feb 2019	235171.3	200831.8	269510.9	182653.6	287689.1
Mar 2019	347940.5	313053.5	382827.6	294585.4	401295.7
Apr 2019	387543.2	352117.1	422969.4	333363.6	441722.9
May 2019	367708.0	331750.8	403665.1	312716.3	422699.7
Jun 2019	418110.9	381630.5	454591.4	362318.9	473903.0
Jul 2019	530755.7	493759.4	567752.0	474174.8	587336.7
Aug 2019	640880.3	603375.2	678385.4	583521.2	698239.4
Sep 2019	443608.0	405601.0	481615.1	385481.2	501734.8
Oct 2019	406789.7	368287.2	445292.2	347905.2	465674.2
Nov 2019	290737.0	251745.4	329728.7	231104.5	350369.6
Dec 2019	369905.8	330431.1	409380.6	309534.5	430277.2

Antes de representar gráficamente esta estimación, también es interesante ver los valores estimados mediante el modelo escogido para cada uno de los valores observados, y como estos se parece bastante a los valores reales (líneas 223 - 227).



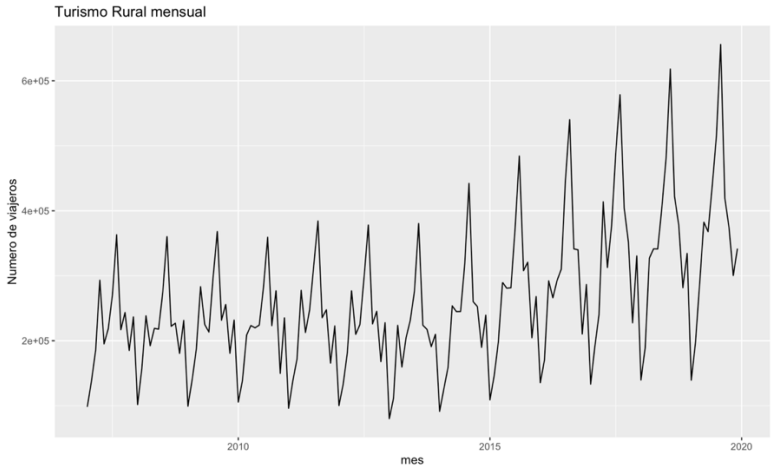
Y finalmente nos toca representar estos datos en una gráfica como se muestra a continuación (línea 230).



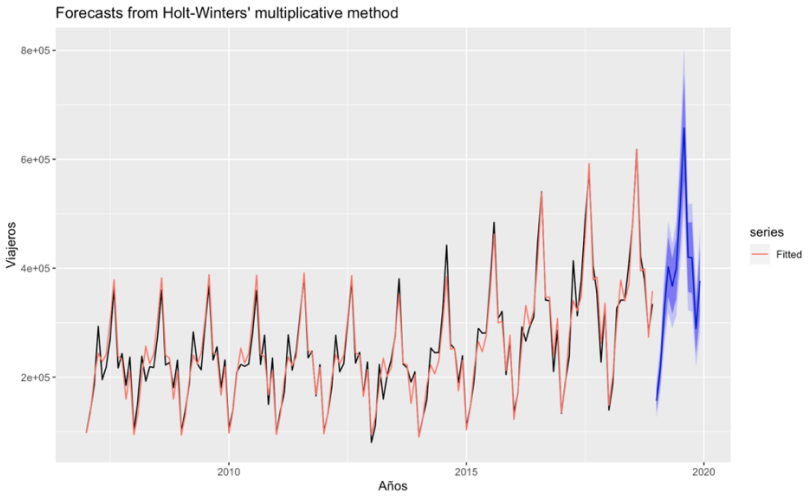
8. **Comparar las predicciones obtenidas con cada uno de los métodos para los valores que se reservaron en el apartado 3, tanto gráfica como numéricamente. Conclusiones. (0.75 puntos)**

   Por último, nos queda ajustar confrontar los modelos seleccionados en esta practica, tanto el método de alisado de Holt-Winters, como el ARIMA.

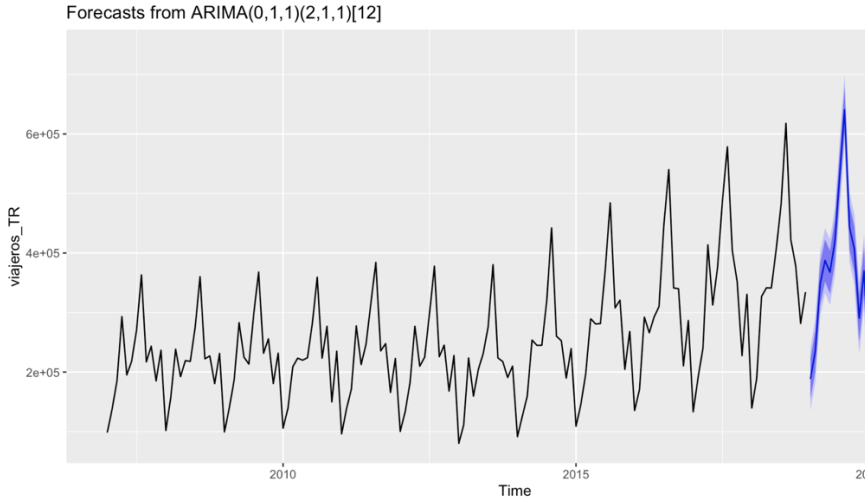
   Para ello primero vamos a observar la gráfica de datos originales:



   A continuación, la gráfica resultante con el método de Holt-Winters:



Y finalmente con el modelo ARIMA



Aun con todo lo dicho del modelo ARIMA, podemos ver claramente como este modelo es ganador ya que a simple vista vemos como tiene una ventana de confianza mas reducida e incluso la traza se parece mas a los datos originales.

Teniendo en cuenta esto vamos a analizar los parámetros numéricamente, es decir vamos a ver las bondades de cada modelo y compararlos.

```
> accuracy(fitARIMA)
                    ME     RMSE      MAE        MPE    MAPE      MASE        ACF1
Training set 1581.335 25142.49 17445.7 -0.8801576 7.33152 0.2362389 -0.1191929
> accuracy(viajeros_hw)
                   ME    RMSE      MAE        MPE     MAPE      MASE         ACF1
Training set 613.569 23455.2 18039.77 -0.3126888 7.482019 0.6742466 -0.06979627
```

Podemos apreciar como para el modelo ARIMA el error absoluto medio (MAE) es menor, sin embargo, la desviación estándar del error (RMSE) es también mayor.

Conclusión

Considero que el modelo ARIMA no es adecuado para la predicción de los viajeros vinculados al turismo rural, tampoco creo que el método de Holt-Winters sea el idóneo para nuestra serie. Existen otras alternativas, y es que podríamos entrar a un análisis mas profundo de los modelos como serían el análisis multivalente, heterocedasticidad estacional y cointegración de espacios.

Por último, tengo que destacar lo interesante que me parece el análisis de series temporales, ya que es una gran herramienta en el análisis de series temporales, porque mirando al pasado nos permite "predecir el futuro", salvo fenómenos excepcionales como por ejemplo esta pandemia que ha creado una serie de fenómenos anómalos en casi todas las series temporales como puede ser la bolsa, turismo, hostelería, compraventa de coches, etc. Sin embargo, es clave también para poder predecir de nuevo todas estas series en el caso de un fenómeno similar.