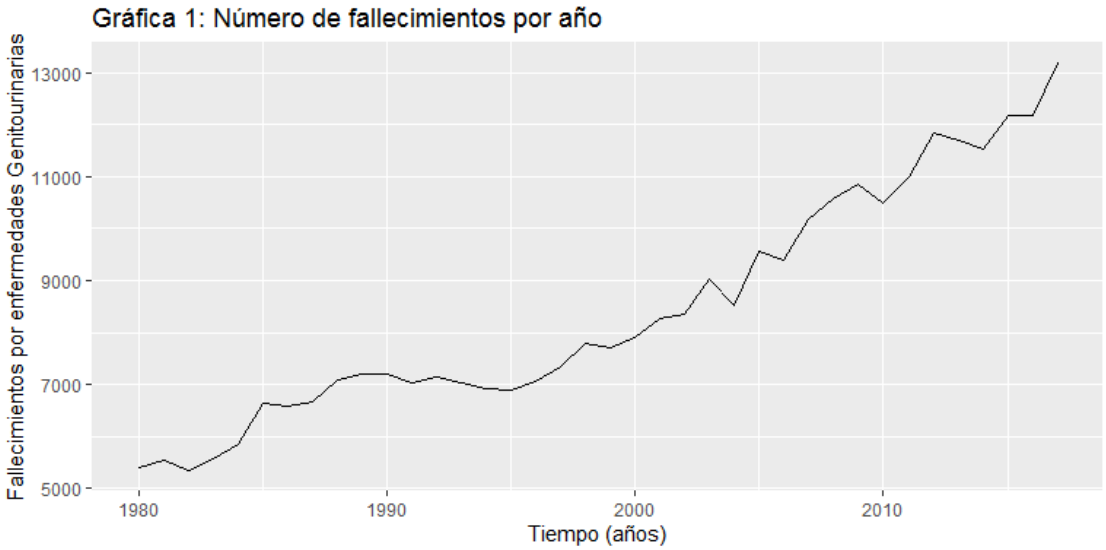
Práctica 4: Series Temporales

Irene Extremera Serrano

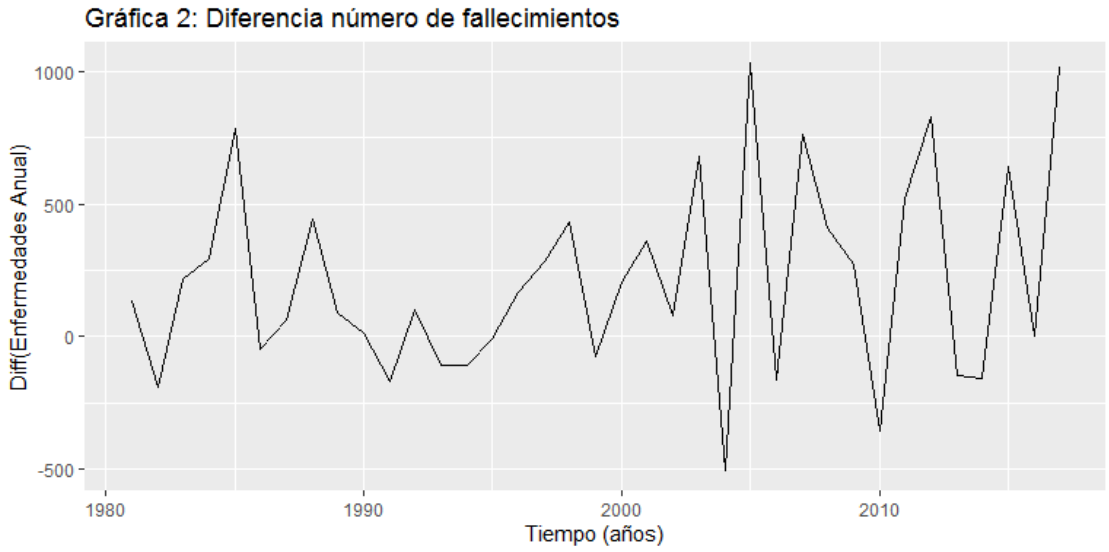
23/03/2020

La serie temporal con la que voy a trabajar recopila información sobre fallecimientos causados por enfermedades que afectan al sistema genitourinario, procede del INE y va desde enero de 1980 a diciembre de 2017. El objetivo de esta práctica es analizar la serie temporal sin estacionalidad por lo que haré uso de la serie anualizada y seguiré la metodología de Box y Jenkins para identificar qué modelo ARIMA genera unas mejores predicciones.

```
# Librerías a usar
library(forecast)
library(tseries)
library(aod)
setwd('D:/Desktop/Remember/Estudios/Educación Formal/Máster/Máster Valencia/Bioestadística/Curso 1/20 2-6Modelización Estadística/Series Temporales/Practicas/P4')
# Serie Anual
enf_gu <- read.table('Enfermedades_del_sistema_genitourinario.txt', header = TRUE)
enf_gu <- ts(enf_gu,start = c(1980,1), freq = 12)
enf_año <- aggregate(enf_gu,FUN = sum)

#Gráfica de las series
autoplot(enf_año,main='Gráfica 1: Número de fallecimientos por año',ylab='Fallecimientos por enfermedades Genitourinarias',xlab='Tiempo (años)')
```

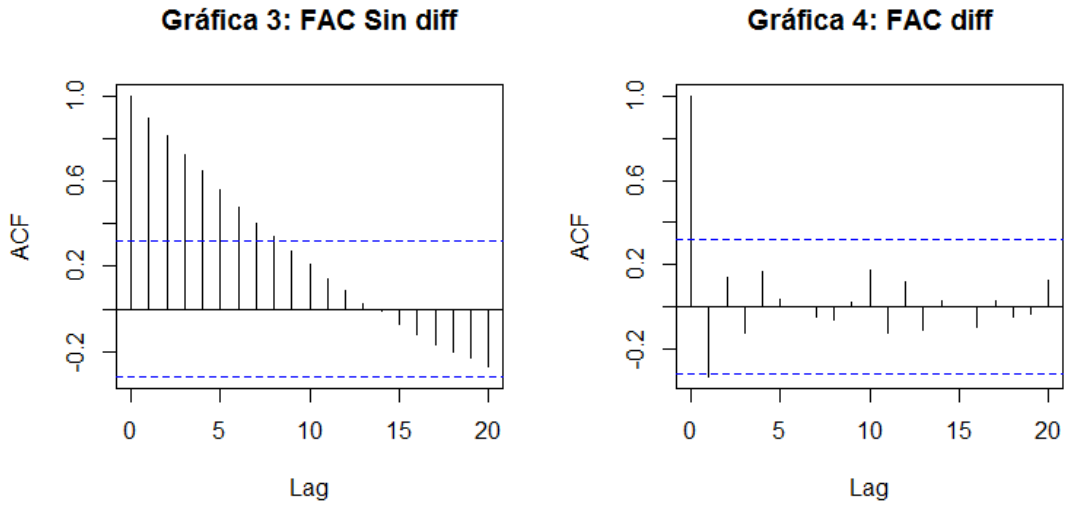


```
autoplot(diff(enf_año),main='Gráfica 2: Diferencia número de fallecimient
os', ylab='Diff(Enfermedades Anual)', xlab='Tiempo (años)')
```



En la gráfica uno se muestra la serie temporal anualizada sin estacionalidad y con tendencia. Como para realizar el siguiente análisis necesito que la serie sea estacionaria realizo una primera diferenciación (d=1) y tal y como muestra la gráfica dos efectivamente de esa manera la serie se torna estacionaria. Sin embargo, para asegurarme realizaré la función de autocorrelación de la serie anual sin diferenciar y diferenciada.

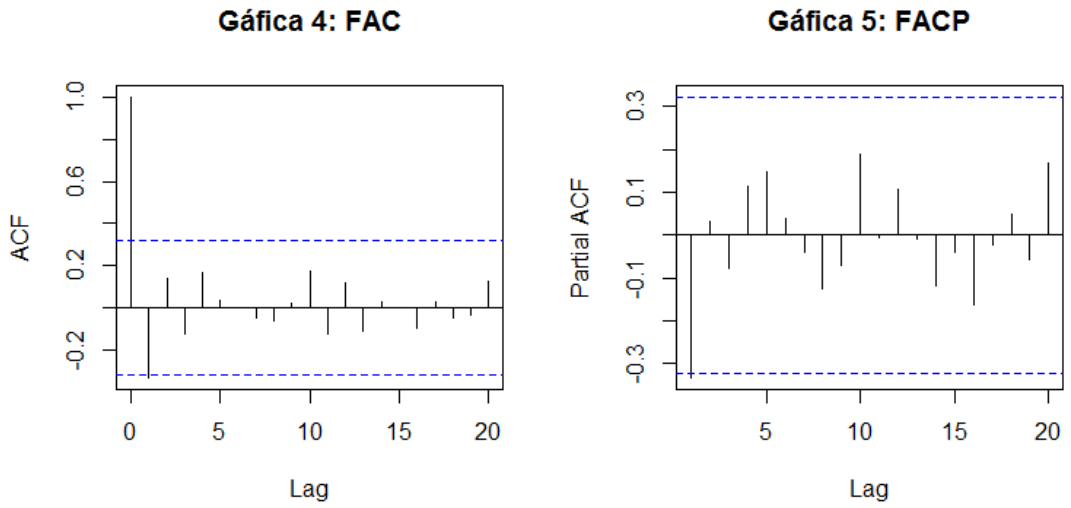
```
# Funciones de autocorrelación
par(mfrow=c(1,2))
acf(enf_año, main = "Gráfica 3: FAC Sin diff",lag = 20)
acf(diff(enf_año), main = "Gráfica 4: FAC diff",lag = 20)
```



Al comparar ambas funciones de autocorrelación, se ve enseguida que en la gráfica tres que corresponde a la función de autocorrelación de la serie anual sin diferenciar hay un decrecimiento progresivo, que por el contrario no ocurre en la gráfica 4 (serie anual diferenciada), en donde después del primer valor de ro (excluyendo ro 0) el resto de los valores quedan dentro del intervalo de confianza. Esto me indica que efectivamente voy a tener que diferenciar al menos una vez para que la serie sea estacionaria en tendencia. Además, prácticamente todos los valores se encuentran dentro del intervalo de confianza, lo cual es un indicativo de que la serie anual diferenciada es ergódica.

Una vez conseguida la estacionaridad observo el comportamiento de la gráfica de autocorrelación y la gráfica de autocorrelación parcial de la serie anual diferenciada una vez para ver qué modelo se ajustaría mejor a la serie.

```
par(mfrow=c(1,2))
acf(diff(enf_año), main = "Gáfica 4: FAC",lag = 20)
pacf(diff(enf_año), main = "Gáfica 5: FACP",lag = 20)
```



En la gráfica cuatro se muestra la función de autocorrelación, en la cual hay un valor en uno que supera las tres desviaciones típicas mientras que el resto de los valores quedan dentro de los intervalos de confianza. Lo mismo ocurre en la gráfica cuatro (función de autocorrelación parcial), hay un valor elevado en el primer valor y el resto son bastante pequeños (no superan las tres desviaciones típicas). Con respecto a si hay o no decrecimiento, no se ve muy claro en ninguna de las dos gráficas, o por lo menos no soy capaz de identificarlo.

Como no me han parecido muy claras las gráficas anteriores, voy a ver cuál es el modelo que me recomienda autoarima.

```
autoa <- auto.arima(enf_año, d=1)
autoa
```

```
## Series: enf_año
## ARIMA(1,1,0) with drift
##
## Coefficients:
##           ar1     drift
##       -0.3709  205.0287
## s.e.   0.1615   42.0268
##
## sigma^2 estimated as 127724:  log likelihood=-269.06
## AIC=544.13   AICc=544.85   BIC=548.96
```

Autoarima sugiere un ARIMA (1,1,0) y además recomienda incluir la constante, presenta un drift de 205.0287, con un coeficiente de valor -0.3709 y con un AIC de 544.85.

Una vez decidido como modelo el ARIMA (1,1,0) compruebo con el test de Wald si los coeficientes y el drift son significativos.

```
#ARIMA(1,1,0)
wald.test(b = coef(autoa), Sigma = vcov(autoa), Terms = 1)
```

```
## Wald test:
## ----------
##
## Chi-squared test:
## X2 = 5.3, df = 1, P(> X2) = 0.022
```

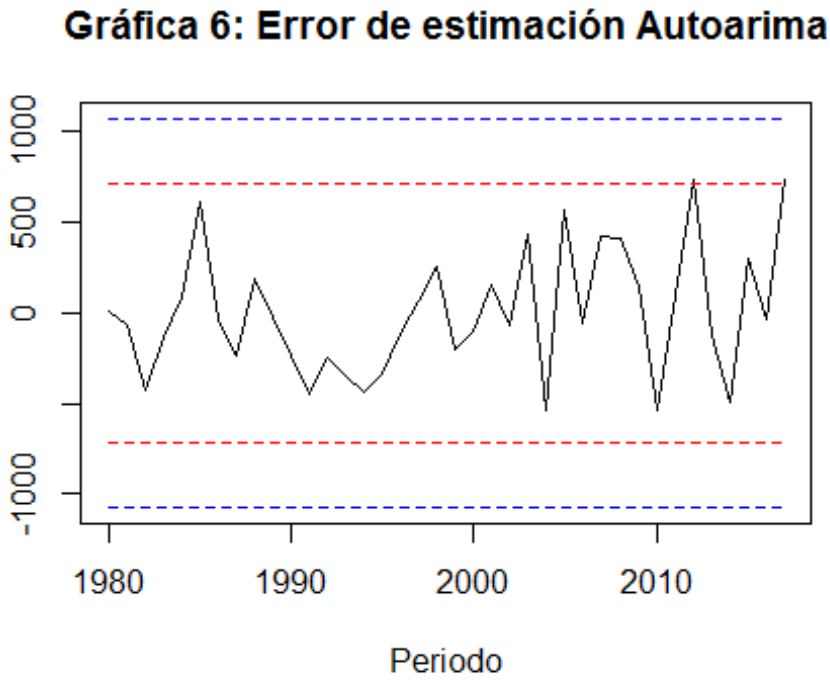
```
wald.test(b = coef(autoa), Sigma = vcov(autoa), Terms = 2)
```

```
## Wald test:
## ----------
##
## Chi-squared test:
## X2 = 23.8, df = 1, P(> X2) = 1.1e-06
```

Con esto se comprueba que tanto drift como el coeficiente salen significativos. Antes de escribir la ecuación del modelo a validar, comprobaré si hay valores atípicos (intervenciones) que puedan afectar a las predicciones e incluirlos si fuera necesario.

Para ello utilizo el residuo de la serie.

```
#Valores atípicos ARIMA (1,1,0)
esauto <- sqrt(autoa$sig)
ts.plot(autoa$residuals,2*esauto,-2*esauto,3*esauto,-3*esauto,xlab='Perio
do',plot.type='single',ylab='', main='Gráfica 6: Error de estimación Auto
arima',lty=c(1,2,2,2,2), col=c('black','red','red','blue','blue'))
```




```
#Miro el año
autoa$residuals > 2*esauto #2017 y 2012
```

```
## Time Series:
## Start = 1980
## End = 2017
## Frequency = 1
##  [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [13] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [25] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE
## [37] FALSE  TRUE
```

Se puede apreciar que hay dos valores que superan las dos desviaciones típicas, concretamente en los años 2017 y 2012. Al ser solamente dos valores que superan muy ligeramente las dos desviaciones no voy a considerar el incluirlos en el modelo final.

Una vez que me he decantado por el modelo ARIMA(1,1,0) doy paso a su validación comenzando con el análisis del error:

```r
#Calidad del modelo
accuracy(autoa)
```

```
##                      ME     RMSE      MAE        MPE     MAPE      MASE
## Training set -0.4230777 342.9881 275.0458 -0.3750701 3.256924 0.8576349
##                       ACF1
## Training set -0.0003781028
```

En ARIMA(1,1,0) el ME de -0.42 es muy cercano a 0, en cuanto al RMSE en media el modelo se equivoca 342.9881 casos, o visto de otra forma, el error porcentual medio MAPE es de 3.25%. Esto indica que el modelo con ese MAPE es bastante bueno. Aparte la capacidad de mejora de las predicciones por intervalo es bastante reducida, pues el ACF1 tiene un valor muy bajo, de -0.0004%.

A continuación, aplico el test de box para comprobar homocedasticidad:

```r
#Homocedasticidad
Box.test(autoa$residuals^2,lag=2, type='Ljung-Box')
```

```
##
##  Box-Ljung test
##
## data:  autoa$residuals^2
## X-squared = 2.5887, df = 2, p-value = 0.2741
```

Para un k=2 se aceptan que se cumple la hipótesis de homocedasticidad.

Para finalizar los test de validación compruebo la normalidad del residuo.

```r
#Normalidad
jarque.bera.test(autoa$residuals)
```

```
##
##  Jarque Bera Test
##
## data:  autoa$residuals
## X-squared = 1.6877, df = 2, p-value = 0.43
```

Para un p valor de 0.43 en ARIMA(1,1,0) se acepta la hipótesis de normalidad.

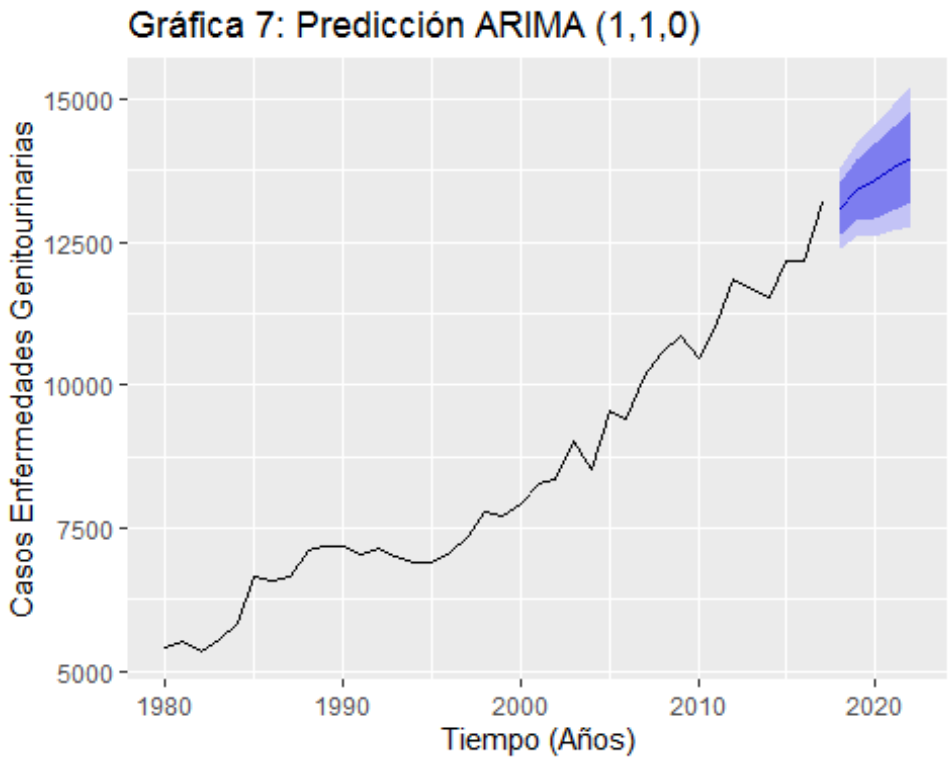
El siguiente paso va a ser realizar una predicción con el ARIMA(1,1,0) a 5 años vista.

```
pautoa <- forecast(autoa, h=5, level=c(80,95))

mediaa <- pautoa$mean
lowa1 <- pautoa$lower[,1]
lowa2 <- pautoa$lower[,2]
upa1 <- pautoa$upper[,1]
upa2 <- pautoa$upper[,2]
mata <- matrix(c(mediaa,lowa1,upa1,lowa2,upa2),nrow=5,ncol=5)
añosa <- c('2018','2019','2020','2021','2022')
ICma <- c('Media','Low 80','Hi 80', 'Low 95', 'Hi 95')
colnames(mata)<- ICma
row.names(mata) <- añosa
mata

##         Media   Low 80    Hi 80   Low 95    Hi 95
## 2018 13092.59 12634.58 13550.59 12392.12 13793.05
## 2019 13409.06 12867.96 13950.15 12581.52 14236.59
## 2020 13572.75 12927.71 14217.79 12586.24 14559.26
## 2021 13793.11 13069.57 14516.65 12686.55 14899.67
## 2022 13992.45 13194.52 14790.38 12772.12 15212.78
```

La predicción resultante muestra que la media de casos de fallecimientos por enfermedades que afectan al sistema genitourinario va aumentando de un año a otro progresivamente, en 2018 el número es de 13092.59 y asciende hasta fallecimientos 13992.45 en 2022, un incremento de 899.86 casos. Lo mismo ocurre con el intervalo de confianza, van incrementándose progresivamente a medida que pasan los años: el intervalo al 80 y 90 por ciento inferiores pasan de 12634.58 y 12392.12 a 13194.52 y 12772.12 respectivamente, y superiores pasan de 13550.59 y 13793.05 a 14790.38 y 15212.78 de 2018 a 2022.

Para poder ver las predicciones realizadas de forma gráfica me fijo en la gráfica 7.

```
autoplot(pautoa, main='Gráfica 7: Predicción ARIMA (1,1,0)', ylab='Casos
Enfermedades Genitourinarias', xlab='Tiempo (Años)')
```



En la gráfica 7 se ilustra lo comentado anteriormente, el aumento progresivo del número de casos de fallecimientos y con ello de su intervalo de confianza.

Todo lo anterior me indica que la ecuación teórica del modelo sería la siguiente:

$$y_t(1-\phi_1 L)(1-L)^d = c + \varepsilon_t$$

$$y_t = c + \phi_1(y_{t-1} - y_{t-2}) + y_{t-1} + \varepsilon_t$$

En el caso de querer hacer predicciones necesitaré la deriva, por ello utilizo las predicciones anteriores y sustituyo en la ecuación.

```
yt <- mediaa[3]
yt1 <- mediaa[2]
yt2 <- mediaa[1]
coef <- autoa$coef[1]
deriva <- -yt+(coef)*(yt1-yt2)+yt1
deriva

##       ar1
## -281.0798
```

La constante del modelo ARIMA(1,1,0) es -281.08 por lo que el modelo final tendrá la siguiente forma:

$$\bar{y}_t = -281.08 + (-0.3709)\,(y_{t-1} - y_{t-2}) + y_{t-1}$$

Por lo que el número de casos de fallecimientos debidos a enfermedades genitourinarias dependerá del número de casos del año anterior más el decremento de la diferencia de fallecimientos del año anterior con respecto a dos periodos antes menos un valor constante.

Para finalizar el análisis, voy a realizar una comparación con el modelo de alisado resultante de la práctica 2.

```
datos_año_ets<- ets(enf_año)
calidad <- summary(datos_año_ets)
```

```
## ETS(M,A,N)
##
## Call:
##  ets(y = enf_año)
##
##   Smoothing parameters:
##     alpha = 0.7407
##     beta  = 0.0013
##
##   Initial states:
##     l = 5168.4918
##     b = 184.221
##
##   sigma:  0.0426
##
##      AIC     AICc      BIC
## 588.7975 590.6725 596.9855
##
## Training set error measures:
##                    ME     RMSE      MAE         MPE     MAPE      MASE
## Training set 29.57535 348.7598 278.1385 0.008039696 3.231244 0.8672784
##                    ACF1
## Training set -0.1138204
```

El modelo obtenido por alisado es el modelo de alisado exponencial de Holt (M,A,N) con tendencia aditiva, error multiplicativo y sin estacionalidad.

Por un lado, el valor de MASE no es muy diferente, de 0.87% en alisado y en arima de 0.86% por lo que estos modelos no son mucho mejores con respecto al modelo más sencillo que podría aplicarse.

Por otro lado, el error medio del modelo de alisado es más alto, de 29.57 mientras que el obtenido por arima es muy cercano a 0, de -0.42.

Aparte, el ACIF también es similar, -0.0004% en ARIMA (1,1,0) (muy bajo) y de -0.11% (bajo) lo cual indica que la capacidad de mejora de las predicciones por intervalo es bastante reducida.

Además, los RMSE son bastante parecidos siendo ligeramente menor el de ARIMA (1,1,0), 348.76 casos en alisado y de 343 casos de fallecimientos en arima.

Lo mismo ocurre con AIC, eL de alisado tiene un valor de 588.80 y 544.13 el de arima.

Para finalizar, también tienen un valor de MAPE bastante parecido siendo mejor en alisado, de 3.23% frente al ARIMA(1,1,0) de 3,26%. Esto me indica que la calidad de predicción del modelo de alisado es ligeramente más alta que la del arima, aunque la del arima también es bastante buena.

Con esto compruebo que ambos modelos son muy parecidos y el de alisado tiene una calidad de predicción muy ligeramente mejor que la obtenida por ARIMA.

Aparte de comparar la calidad de ambos modelos comparo sus predicciones y observo lo siguiente:

```
prediccion <- forecast(datos_año_ets, h=5)
media <- prediccion$mean
low <- prediccion$lower[,1]
low2 <- prediccion$lower[,2]
up <- prediccion$upper[,1]
up2 <- prediccion$upper[,2]
mat <- matrix(c(media,low,up,low2,up2),nrow=5,ncol=5)
años <- c('2018','2019','2020','2021','2022')
ICm <- c('Media','Low 80','Hi 80', 'Low 95', 'Hi 95')
colnames(mat)<- ICm
row.names(mat) <- años
mat
```

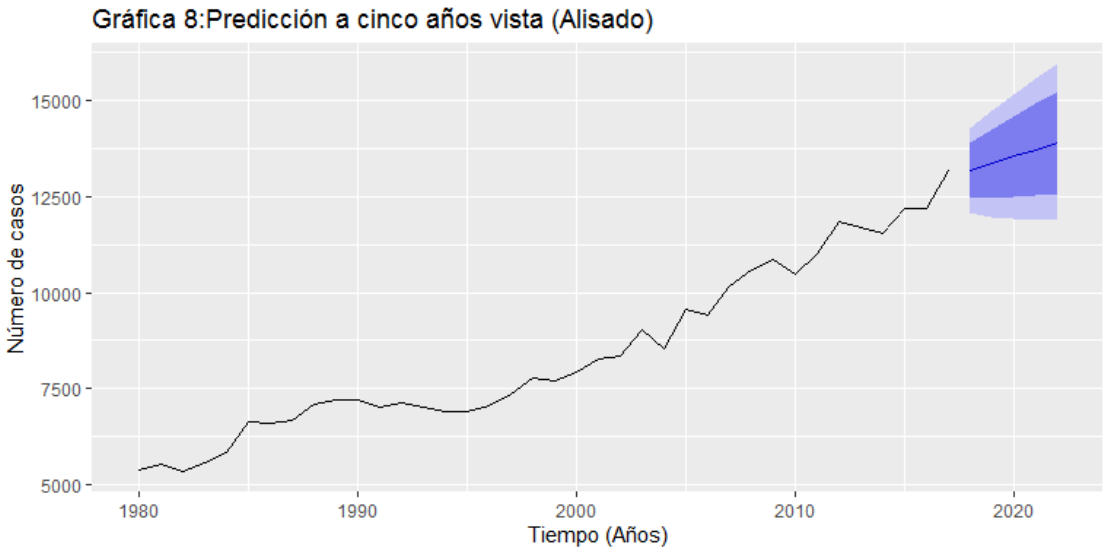
```
##         Media   Low 80    Hi 80   Low 95    Hi 95
## 2018 13164.36 12445.23 13883.49 12064.55 14264.17
## 2019 13350.06 12446.13 14254.00 11967.61 14732.52
## 2020 13535.77 12474.32 14597.21 11912.42 15159.11
## 2021 13721.47 12518.97 14923.97 11882.41 15560.53
## 2022 13907.17 12574.79 15239.56 11869.46 15944.88
```

```
#Diferencia del número de casos predichos con el modelo de alisado frente al arima.
media-mediaa

## Time Series:
## Start = 2018
## End = 2022
## Frequency = 1
## [1]  71.77308 -58.99511 -36.98362 -71.64250 -85.28072
```

La predicción con alisado muestra que la media de casos de fallecimientos por enfermedades que afectan al sistema genitourinario va aumentando de un año a otro progresivamente, en 2018 el número es de 13164.36 y asciende hasta fallecimientos 13907.17 en 2022, un incremento de 742.81 casos. En comparación con el anterior el número medio de casos es menor y además las predicciones en comparación al resto de años el número de casos predichos en ARIMA(1,1,0) es mayor.

Gráficamente tendría la siguiente forma.

```
autoplot(prediccion,xlab='Tiempo (Años)', ylab='Número de casos',main='Gráfica 8:Predicción a cinco años vista (Alisado)')
```



Cuando comparo la gráfica 8 con la gráfica 7 se observa que son bastante parecidas, de modo que cualquiera de ambos modelos sería útil a la hora de realizar predicciones. Sin embargo, todos los criterios que valoran la calidad de los modelos son mayores en alisado que en el modelo ARIMA(1,1,0) a excepción del MAPE, que es menor en el de alisado (3.23%) pero no muy diferente (3,26%). Esto hace muy difícil saber con cuál de los dos modelos decantarme, pero a la vista de que el modelo de alisado es más sencillo y tiene un MAPE mejor me decantaría por él, pues lo que me interesa en este caso en particular es que las predicciones a realizar sean buenas.