Práctica 2

Irene Extremera Serrano

04 de marzo de 2020

La serie temporal que voy a analizar contiene el número de fallecimientos a causa de enfermedades del sistema genitourinario por meses de enero de 1980 a diciembre de 2017 en España. La serie procede del INE (Instituto Nacional de Estadística) y tiene un total 456 datos de fallecimientos (uno por mes).

Lo primero que haré es cargar el fichero de datos y pasarlo a serie temporal, además de activar las librerías que voy a utilizar a lo largo del análisis.

```
Enf_GU <- read.csv("D:/Desktop/Remember/Estudios/Educación Formal/Máster/
Máster Valencia/Bioestadística/Curso 1/20 2-6Modelización Estadística/Ser
ies Temporales/Temas/T1/Enfermedades_del_sistema_genitourinario.txt")
enf_ts <- ts(Enf_GU, start = c(1980,1), freq = 12)
library(forecast)
```

```
## Warning: package 'forecast' was built under R version 3.6.2
```

```
## Registered S3 method overwritten by 'xts':
##   method     from
##   as.zoo.xts zoo
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

Ejercicio 1: Serie con el fechado anual.

El objetivo de este ejercicio es ajustar un modelo que explique cómo se han generado esos los datos de fallecimientos al año y generar una predicción a un año vista.

De modo que sumo el número de casos de los meses de cada año, así obtengo la serie temporal anual y en base a eso busco el modelo que mejor se ajuste.

```
datos_año <- aggregate(enf_ts,FUN = sum) #Agrego mi serie temporal en núm
ero de casos por año.
datos_año_ets <- ets(datos_año)
calidad <- summary(datos_año_ets) #Como no me sale ninguna phi es superio
r a 0.98 por lo que no tiene amortiguamiento.
```

```
## ETS(M,A,N)
##
## Call:
##  ets(y = datos_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 es el modelo de alisado exponencial de Holt (M,A,N). La tendencia es aditiva, que es lo que esperaba, y esta puede descomponerse en pendiente (b) y nivel (l). El error es multiplicativo, cosa que esperaba también ya que el esquema de la práctica uno salió multiplicativo, y al tratarse de una serie anual no tiene estacionalidad.

Por un lado, en cuanto a parámetros que componen el modelo, comenzaré explicando alpha y beta:

```
datos_año_ets$par[1:2]
```

```
##       alpha        beta
## 0.740711605 0.001318684
```

Estos parámetros estimados me dan la siguiente información del modelo:

- Alfa toma un valor de 0.740711605 , lo cual me indica que el nivel va a ir cambiando en el tiempo.
- Beta es de 0.001318684 (valor bajo), por lo que la pendiente se mantiene constante en el tiempo.

Por otro lado, la tendencia del modelo es aditiva, por lo que los últimos valores de nivel y pendiente me dan la siguiente información:

```
datos_año_ets$initstate

##        l        b
## 5168.492  184.221
```

El nivel de la serie toma el valor de 5168.492 y la pendiente tiene un valor de 184.221. El hecho de que la pendiente tenga un valor positivo me indica que el número de casos va a ir aumentando con los años. De modo que para la predicción del año siguiente la forma de calcularlo va a ser la suma de ambos parámetros.

Sin embargo, antes de realizar ningún tipo de predicción voy a observar cómo son los indicadores de calidad del modelo para valorar o no utilizarlo.

```
calidad

##                     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
```

Los distintos indicadores de calidad me dan la siguiente información:

- El valor de MAPE (error porcentual absoluto medio) es de 3.231244 % lo cual me indica que la calidad de predicción del modelo es muy alta.
- El valor de MASE es del 0.8672784% por lo que el modelo de alisado no es mucho mejor con respecto al modelo más sencillo que podría aplicarse.
- El valor de ACF1 de -0.1138204% me indica que la capacidad de mejora del modelo es bastante reducida.

Esto tiene sentido, ya que el error del modelo es bastante bajo y poco se va a poder mejorar para reducir ese error.

Una vez visto las distintas características del modelo me dispongo a realizar una predicción a tres años vista.

```
# Predicción con R.
prediccion <- forecast(datos_año_ets, h=3)

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=3,ncol=5)
años <- c('2018','2019','2020')
ICm <- c('Media','Low 80','Hi 80', 'Low 95', 'Hi 95')
colnames(mat)<- ICm
row.names(mat) <- años
mat #Lo meto en una matriz porque R no me da una salida.
```

```
##         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
```

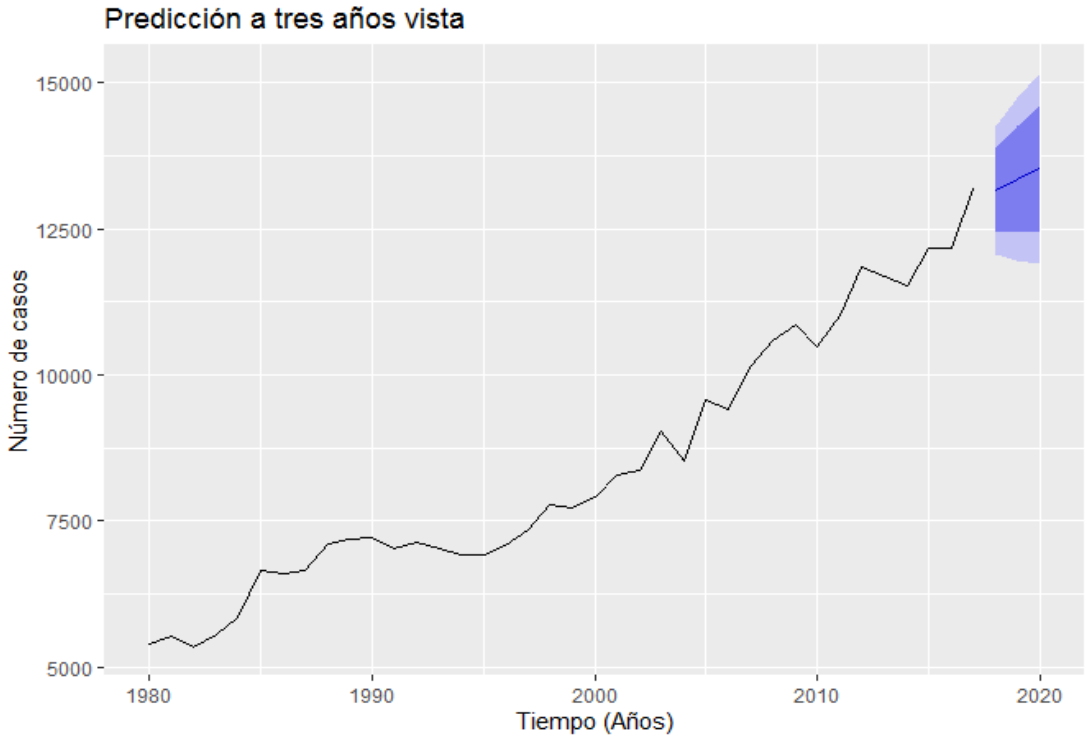
```
# Predicción a mano.
D <- nrow(datos_año_ets$states)
pred_mano <- (datos_año_ets$states[D,1]+(1:3)*datos_año_ets$states[D,2])
#Predicción extramuestral

pred_mano-mat[,1]
```

```
## 2018 2019 2020
##    0    0    0
```

Cuando realizo la predicción con el programa y a mano me dan el mismo valor, esto lo compruebo haciendo la diferencia de ambas estimaciones obteniendo como resultado es 0. Con esto abstraigo que el modelo obtenido es el correcto y se ajusta bastante bien a la serie temporal anual de fallecimientos por enfermedades que afectan al sistema genitourinario. De 2018 a 2020 el número de casos ha ido aumentando de 13164.36 en 2018, a 13350.06 en 2019, hasta 13535.77 en 2020, es decir, un incremento de 185.7 casos por año aproximadamente.

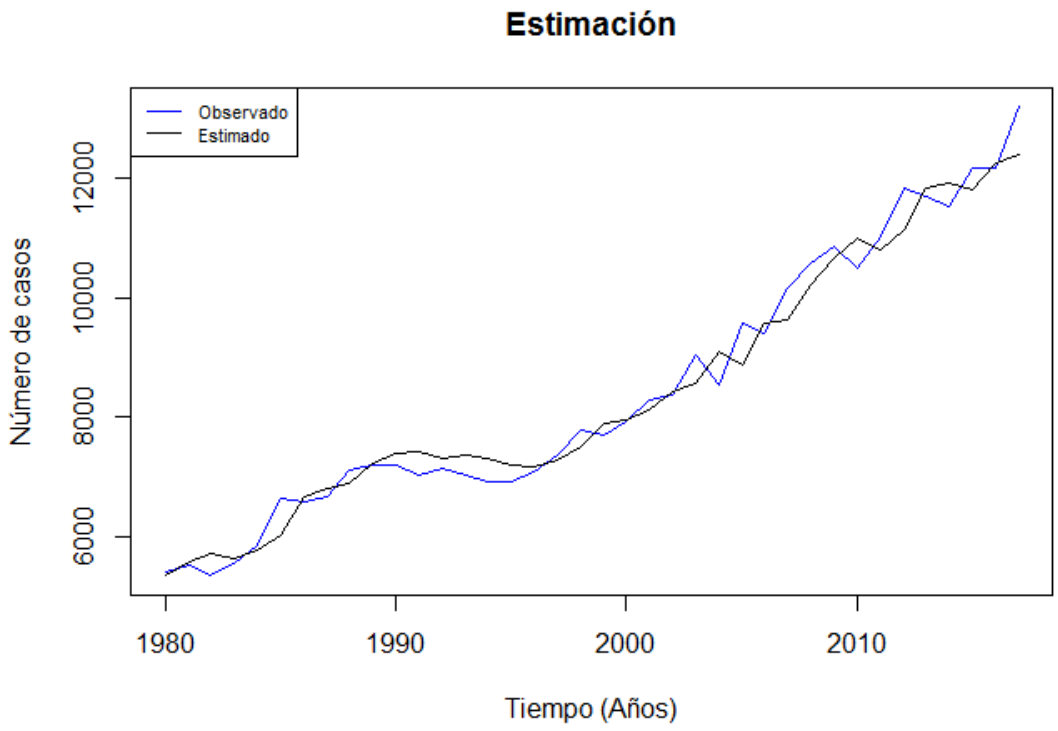
Para poder ver lo anterior de forma más visual esta realizo una gráfica de la predicción y de la estimación obteniendo lo siguiente:

```
autoplot(prediccion,xlab='Tiempo (Años)', ylab='Número de casos',main='Predicción a tres años vista')
```



En la gráfica de la predicción a tres años vista se ilustra la predicción a tres años vista del número de casos de fallecimientos debido a enfermedades del sistema genitourina rio. Según la predicción el número de casos en los próximos tres años parece que va a aumentar de forma constante llegando has los 13535.77 casos, por lo que se estima un aumento de 555 casos en ese periodo de tiempo.

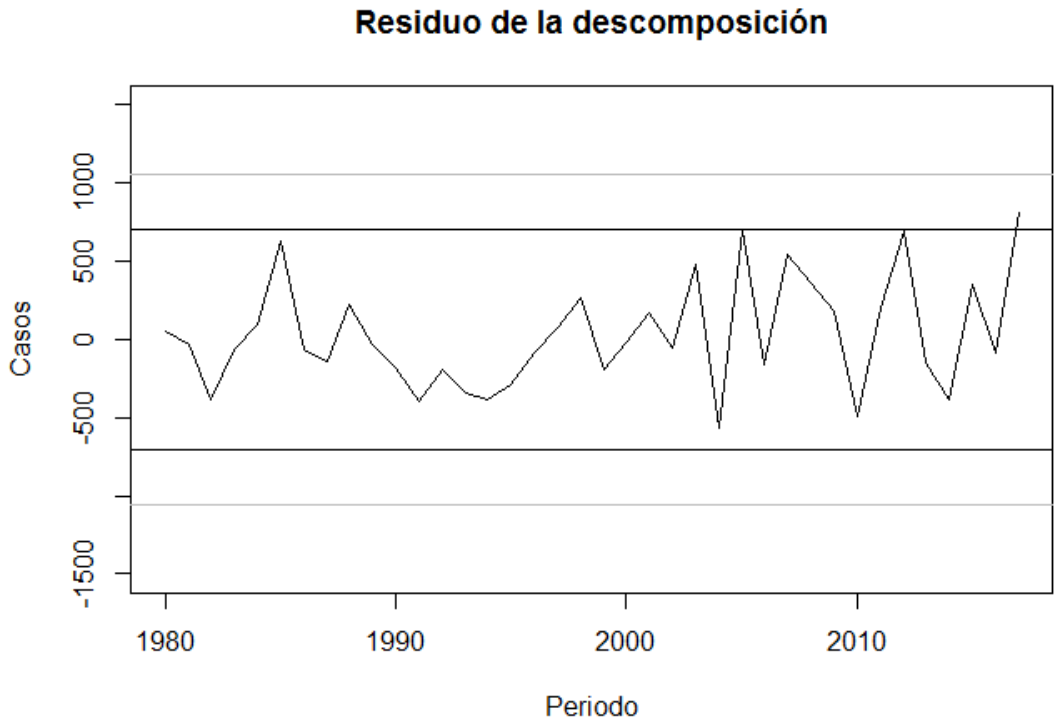
```
ts.plot(datos_año_ets$x,datos_año_ets$fitted,col=c('BLUE','BLACK'),main='
Estimación',
        xlab='Tiempo (Años)',ylab='Número de casos')
legend("topleft", col=c('blue','black'),legend = c("Observado","Estimado"
),lty=1,cex=0.7)
```



Con respecto a la gráfica de estimación aparecen dos líneas que ilustran los casos observados en azul y los casos estimados en negro. Se aprecia de esta manera que el número de casos estimados tiene una línea más suavizada que la línea que representa los casos reales que es mucho más abrupta. Ambas presentan una marcada tendencia ascendente y su correspondiente estacionariedad, que ya comenté en la práctica uno, entre los años 1989 y 1998.

Para finalizar el ejercicio analizaré el residuo de la serie para ver si hay algún valor atípico que pueda ser relevante.

```r
residuo <- datos_año_ets$x - datos_año_ets$fitted
se <- sd(residuo, na.rm = TRUE)

plot(residuo, xlab = 'Periodo', ylab = 'Casos',main = "Residuo de la desc
omposición", ylim=c(-1500,1500),col = "Black")
abline(h = c(-3 * se, -2 * se, 2 * se, 3 * se),col = c("Grey", "Black", "
Black", "Grey"))
```




```r
# residuo < -3*se De esta manera ver qué año es el valor atípico.
# residuo > 2*se
```

Se puede identificar que no hay ningún valor atípico que supere las tres deviaciones típicas, sin embargo, sí que se observan dos años que superan las dos desviaciones típicas: 2005 y 2017. Por lo que en la serie temporal anual de enfermedades que afectan al sistema genitourinario no se aprecian intervenciones significativas.

Ejercicio 2. Serie con el fechado original.

En este ejercicio realizaré el mismo procedimiento anterior, pero sin anualizar la serie temporal.

Primero miro el modelo que mejor se ajusta a la serie.

```
# library(forecast)
datos_ets <- ets(enf_ts)
calidad2 <- summary(datos_ets)
```

```
## ETS(M,A,M)
##
## Call:
##  ets(y = enf_ts)
##
##   Smoothing parameters:
##     alpha = 0.2026
##     beta  = 1e-04
##     gamma = 1e-04
##
##   Initial states:
##     l = 439.5583
##     b = 1.3845
##     s = 1.0986 0.9523 0.9181 0.8804 0.9851 1.0087
##            0.9091 0.9574 0.9637 1.0594 1.0631 1.204
##
##   sigma:  0.0672
##
##      AIC     AICc      BIC
## 6293.591 6294.988 6363.673
##
## Training set error measures:
##                      ME     RMSE     MAE        MPE     MAPE      MASE
ACF1
## Training set 0.6504983 48.41639 34.6059 -0.3850739 4.945122 0.6422966
0.1147674
```

El mejor modelo resultante tiene un error y una estacionalidad multiplicativos con tendencia aditiva, por lo que tengo un modelo de alisado de Hold-Winters multiplicativo (MAM). Como en la práctica anterior vi que el esquema era multiplicativo, era esperable que el error haya salido multiplicativo.

Sus parámetros estimados me dan la siguiente información:

```
datos_ets$par[1:2]

##        alpha         beta
## 0.2026127603 0.0001000071
```

- Alfa toma un valor de 0.2026127603 , lo cual me indica que el nivel va a ir no cambiando mucho en el tiempo.
- Beta es de 0.0001000071 (valor bajo), por lo que la pendiente se mantiene muy constante en el tiempo.

Por otro lado, los valores del nivel y la pendiente me dicen lo siguiente:

```
nfinal <- nrow(datos_ets$states) #meto en un objeto cuántas filas tiene s
tates
datos_ets$states[nfinal,1:2]

##           l           b
## 1120.958554    1.414518
```

El último valor de nivel que toma la serie es de 1120.9585541 y la pendiente tiene un valor de 1.414518. Como dije anteriormente, que la pendiente tenga un valor positivo indica que el número de casos va a ir aumentando con los años. Por lo tanto, para hacer la predicción de un mes determinado el procedimiento será la suma de ambos parámetros multiplicados por la estacionalidad de ese mismo mes del año anterior.

A continuación, me fijo en los indicadores de calidad para asegurarme de si vale o no la pena hacer predicciones usando este modelo:

```
calidad2

##                      ME     RMSE      MAE        MPE     MAPE      MASE
ACF1
## Training set 0.6504983 48.41639 34.6059 -0.3850739 4.945122 0.6422966
0.1147674
```

Y los distintos indicadores de calidad me dan la siguiente información: - El valor de MAPE (error porcentual absoluto medio) es de 4.945122% lo cual indica que la calidad de predicción del modelo es muy alta. - El valor de MASE es del 0.6422966% por lo que el modelo de alisado no es mucho mejor con respecto al modelo más sencillo que podría aplicarse. - El valor de ACF1 de 0.1147674% me indica que la capacidad de mejora del modelo es bastante reducida.

De modo que al ver que el modelo es bastante bueno me dispongo a realizar la predicción de un año completo.

```r
# Predicción con R.
prediccion2 <- forecast(datos_ets, h=12)

media <- prediccion2$mean
low <- prediccion2$lower[,1]
low2 <- prediccion2$lower[,2]
up <- prediccion2$upper[,1]
up2 <- prediccion2$upper[,2]
mat <- matrix(c(media,low,up,low2,up2),nrow=12,ncol=5)
años <- c('Enero','Febrero','Marzo','Abril','Mayo','Junio','Julio','Agost
o','Septiembre','Octubre','Noviembre','Diciembre')
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
## Enero      1351.2644 1234.9153 1467.613 1173.3238 1529.205
## Febrero    1194.6625 1089.7011 1299.624 1034.1379 1355.187
## Marzo      1192.0840 1085.3011 1298.867 1028.7736 1355.394
## Abril      1085.6675  986.5893 1184.746  934.1404 1237.195
## Mayo       1079.9426  979.6040 1180.281  926.4879 1233.397
## Junio      1026.8420  929.7736 1123.910  878.3888 1175.295
## Julio      1140.6896 1031.0449 1250.334  973.0024 1308.377
## Agosto     1115.4396 1006.4789 1224.400  948.7986 1282.081
## Septiembre  998.1514  899.1147 1097.188  846.6879 1149.615
## Octubre    1042.1923  937.2115 1147.173  881.6380 1202.747
## Noviembre  1082.3673  971.7308 1193.004  913.1634 1251.571
## Diciembre  1250.1558 1120.5393 1379.772 1051.9244 1448.387
```

```r
#A mano
# Predicción a mano.
D <- nrow(datos_ets$states)
pred_mano <- (datos_ets$states[D,1]+datos_ets$states[D,2])*datos_ets$stat
es[D,14:3]

pred_mano-mat[,1]
```

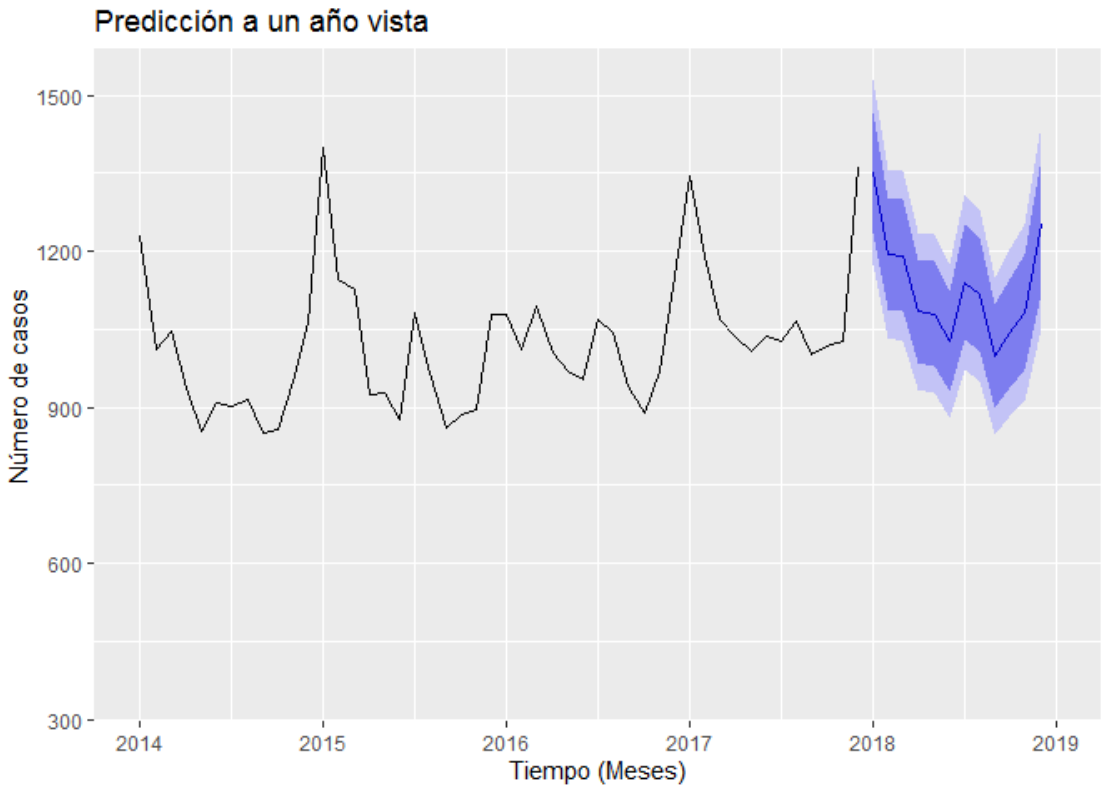
```
##           s12           s11           s10           s9            s8
##  2.273737e-13 -1.503729e+00 -2.997194e+00 -4.089315e+00 -5.416867e+00
##            s7            s6            s5            s4            s3
## -6.430087e+00 -8.560880e+00 -9.754407e+00 -9.963248e+00 -1.168862e+01
##            s2            s1
## -1.347122e+01 -1.709420e+01
```

La predicción con R y a mano sale aproximadamente 0 cuando hago la diferencia, por lo que efectivamente el modelo elegido es el correcto.

Puedo observar como se refleja la estacionalidad en el número de casos por mes, en donde en diciembre y enero, los meses más fríos, el número de fallecimientos es mayor (1351 en enero y 1250 en diciembre), mientras que en meses con unas temperaturas más suaves el número se reduce bastante (998 casos en septiembre). Aparte también se aprecia que el número de fallecimientos en julio (1141) es ligeramente mayor a junio (1027) y a agosto (1115), cosa que se observó en la práctica anterior.

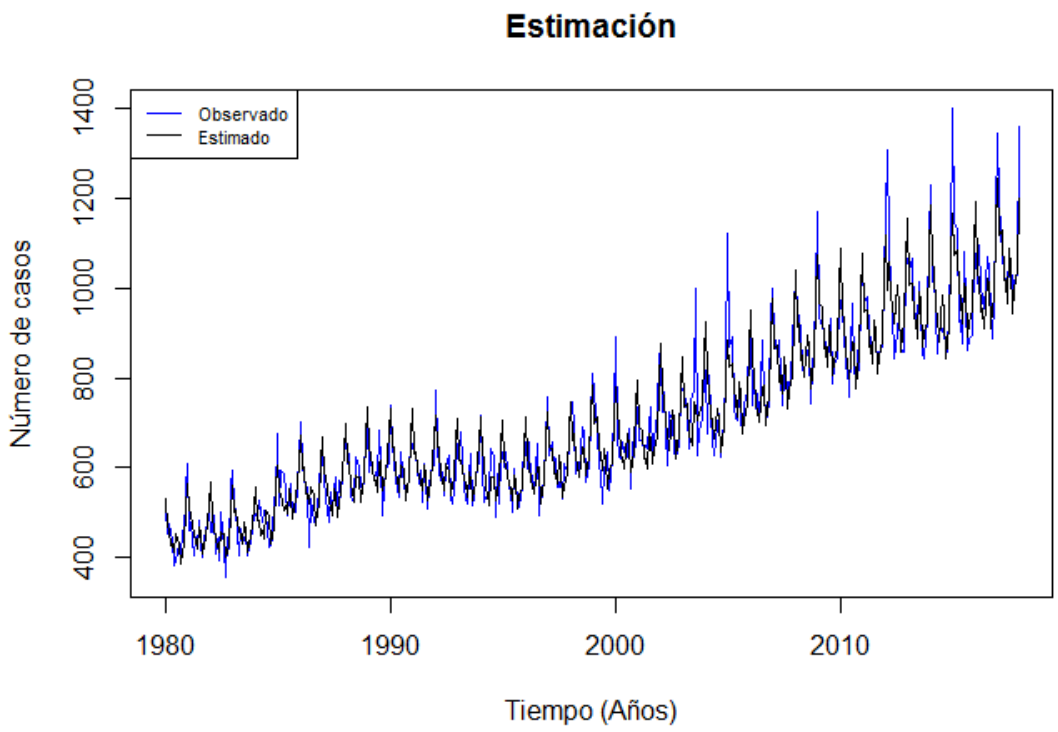
Para poder visualizarlo todo de una mejor manera posible realizo la gráfica de la predicción por un lado y de la estimación por otro.

```
autoplot(prediccion2, xlab='Tiempo (Meses)', ylab='Número de casos',main=
'Predicción a un año vista',xlim=c(2014,2019))

## Scale for 'x' is already present. Adding another scale for 'x', which
will
## replace the existing scale.

## Warning: Removed 408 rows containing missing values (geom_path).
```



Por un lado, en la gráfica de la predicción a un año vista (he cogido los años de 2014 a 2019 para poder visualizar bien la predicción) se observa que durante los primeros meses del año y finales los valores en cuanto al número de casos son más alto y hay un pequeño aumento en mitad del año, lo que se corresponderá probablemente, como vi en la práctica 1, al mes de julio.

```
ts.plot(datos_ets$x,datos_ets$fitted,col=c('BLUE','BLACK'),main='Estimación',
        xlab='Tiempo (Años)',ylab='Número de casos')
legend("topleft", col=c('blue','black'),legend = c("Observado","Estimado"),lty=1,cex=0.7)
```

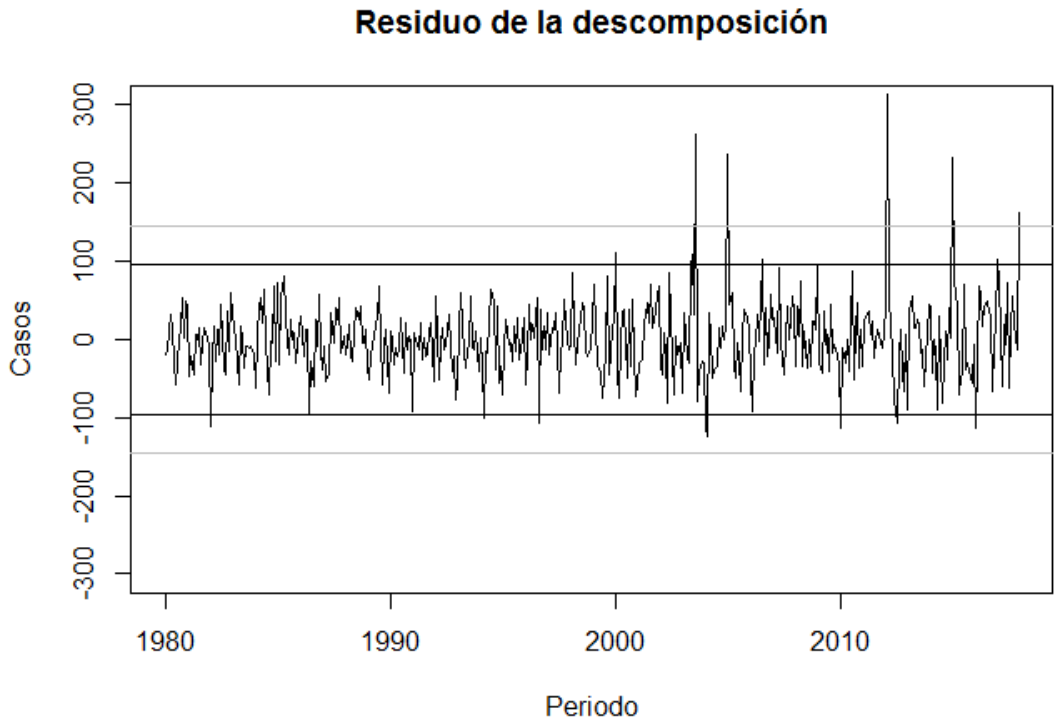


Por otro lado, en la gráfica de estimación se muestran en azul los valores de casos observados mientras que en negro los casos estimados por el modelo. Al igual que ocurría en la anterior gráfica, la línea de casos estimados está más suavizada en general, no presenta una estacionalidad tan marcada como la observada que es más abrupta. Esto se puede ver sobre todo entre los años 2000 y 2017 en donde hay una mayor oscilación en el número de casos observados ya que de 2000 hacia atrás el número de casos observados y estimados se ajustan bastante bien y no hay estas oscilaciones tan marcadas.

Estas oscilaciones probablemente se deban a posibles intervenciones, por lo que aparecerán como forma de valores atípicos en el residuo que analizaré a continuación.

La siguiente gráfica muestra el residuo del modelo que ha sido estimado de forma aditiva (restando a los valores originales los estimados por el modelo).

```r
residuo <- datos_ets$x - datos_ets$fitted #Residuo Aditivo
se <- sd(residuo, na.rm = TRUE)

plot(residuo, xlab = 'Periodo', ylab = 'Casos',main = "Residuo de la desc
omposición", ylim=c(-300,300),col = "Black")
abline(h = c(-3 * se, -2 * se, 2 * se, 3 * se),col = c("Grey", "Black", "
Black", "Grey"))
```




```r
# residuo < 3*se #De esta manera ver en qué mes y año es el valor atípico
.
# residuo > 2*se
# residuo < -2*se
```

Se muestran cinco periodos en los cuales el residuo supera las tres desviaciones típicas por la parte superior: en 2003 el mes de agosto, enero de 2005, febrero de 2012, en 2015 el mes de enero y en 2017 en diciembre. Todos estos meses coinciden con los meses en donde hay más casos de fallecimientos por estas enfermedades, lo cual podría atribuirse al frio. En agosto por otro lado, debió de ser un mes bastante caluroso ese año y ese pudo ser uno de los condicionantes de que el número de muertes aumentase.

Por otro lado, cabe mencionar que hay varios meses que superan las dos desviaciones típicas por debajo (enero de 1982, marzo de 1994, agosto de 1996, enero y febrero de 2004, enero de 2010, julio de 2012 y enero de 2016) aunque algunos los hay que las superan por la parte superior (enero del 2000, junio y agosto de 2003, enero de 2005, julio de 2006, febrero de 2012, enero de 2015 y enero y diciembre de 2017).

Tras todo este análisis, voy a usar la metodología de training set/test set con esta serie temporal pues anteriormente he observado que el modelo elegido comente un error de aproximadamente del 4.94% (valor de la estimación del error en la previsión intra-muestral a un periodo vista), de modo que con el fin de mejorar la capacidad predictiva estimo el error de previsión extramuestral a un periodo vista y a largo plazo.

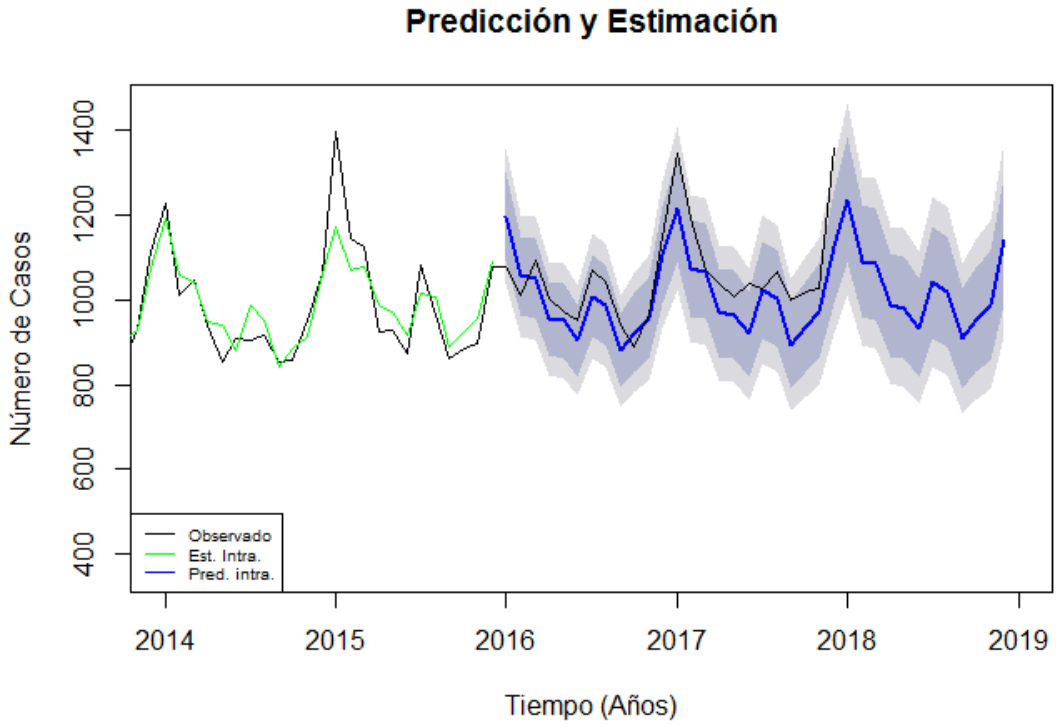
```
enf_intra <- window(enf_ts, end=c(2015,12))
enf_extra <- window(enf_ts,start=c(2016,1))

enf_ets_parcial <- ets(enf_intra,model = 'MAM')#Miro el mejor modelo para
una

enf_ets_test2 <- forecast(enf_ets_parcial,h=36) #Hago la predicción a un
año vista de la serie de varios años
accuracy(enf_ets_test2,enf_extra ) #Calidad del modelo en la previsión a
un periodo vista intramuestral y a largo plazo extramuestral
```

```
##                      ME     RMSE      MAE        MPE     MAPE      MAS
E
## Training set  0.3036609 47.63177 33.78956 -0.3661877 4.962537 0.652098
3
## Test set     50.5582454 83.83617 67.29686  4.5068096 6.138314 1.298749
3
##                   ACF1 Theil's U
## Training set 0.1328799        NA
## Test set     0.1836878 0.7345553
```

En los resultados obtenidos se aprecia que el error porcentual medio de la previsión intra-muestral a un periodo vista es de 4.96% (ha aumentado ligeramente con respecto al modelo anterior) mientras que la extra-muestral comete un error del 3.38%. Con esto se aprecia que ambas estimaciones a un año vista no aumentan demasiado el error, pero el error que comete el modelo intra-muestral es ligeramente mayor que el que comete el extra-muestral, por lo que sus predicciones serán menos precisas a un año vista. Por otro lado, con respecto a la predicción extramuestral a largo plazo se aprecia que el error de predicción aumenta considerablemente, 6.14% sin embargo no es un error muy alto por lo que las predicciones no serán del todo malas.

Para poder ver lo anterior a continuación se muestra la gráfica con los casos de fallecimientos reales por enfermedades genitourinarias, los valores estimados intra-muestrales y la predicción a un año vista (2018) intra-muestral.

```r
#Represento
plot(enf_ets_test2,xlim=c(2014,2019), main= 'Predicción y Estimación', yl
ab='Número de Casos', xlab= 'Tiempo (Años)')
#Predicción y observado.
lines(enf_ts) #Observado.
lines(enf_ets_parcial$fitted,col='GREEN') #Estimado intramuestral.
legend("bottomleft", col=c('black','green','blue'),
       legend = c("Observado", "Est. Intra.", 'Pred. intra.'),lty=1, cex=
0.6)
```



Los modelos generados parece que siguen el comportamiento de los datos originales, sin embargo, no terminan de ajustarse del todo a la realidad. De 2014 a 2016 las estimaciones presentan un comportamiento más suavizado que el comportamiento de los datos normales. En cuanto a la predicción intramuestral se observa que entre los años 2016 y 2018 queda ligeramente por debajo de los datos reales la mayor parte del tiempo, por lo que esperaría que cuando obtuviese los datos de 2018 a 2019 estos fuesen ligeramente más altos que las predicciones. Aun así, los valores quedan dentro de los intervalos de confianza generados por lo que no es una mala predicción.