Estimación suavizada del número de personas infectadas de COVID-19 en la CAPV

Alexander Olza Rodriguez

1/8/2021

1) Introducción

Desde Febrero de 2020 hasta hoy, se han recogido multitud de datos sobre la situación de la pandemia del COVID-19 en la CAPV. Sin embargo, el número de infectados reales es un aspecto inobservable que sólo puede estimarse mediante modelos. En este trabajo propondremos un modelo en espacio de estados para obtener una estimación suavizada. El modelo se ha entrenado con los datos de personas ingresadas en planta y en UCI, y con los datos estimados (por terceras personas, incluidos en el fichero de datos) del número reproductivo básico R_0. Los datos se han extraído de OpenData Euskadi, actualizados hasta el 04/01/2021.

2) Lectura y preparación de los datos

Primero cargamos todas las librerías necesarias para el trabajo:

```
library(readxl)
library(zoo)
library(xts)
library(dlm)
```

A continuación leemos el fichero de datos, y seleccionamos sólo las columnas que vamos a usar. Les asignamos nombres cortos, y transformamos R_0 a variable numérica con sus decimales correspondientes, ya que se ha leído erróneamente como character.

```
setwd("/home/alex/Downloads/master/series-temporales/covid")
covid <- read_excel("situacion-epidemiologica.xlsx",skip = 1)
covid <- as.data.frame(covid)
#names(covid) #Se omite esta línea en el trabajo porque tiene mucho output
cols <- colnames(covid)
ncols<-c("Fecha","TestTot","PCRPos","Ingresados","UCI","R0")
nloc<-c(1,2,7,18,19,20)
cbind(cols[nloc],ncols)
covid <- covid[,nloc]
colnames(covid) <- ncols
covid$R0 <- as.numeric(gsub(",", "\\.", covid$R0))
covid$Fecha <- as.Date(covid$Fecha)
```

Transformamos el objeto resultante en una serie temporal multivariante:

```
covid <- zoo(x=covid[,-1], order.by=as.Date(covid[,1], format="%d/%m/%Y"))
```

3) Modelización

3.1) Ecuaciones del modelo

Se propone el número de infectados como variable de estado α_t. Esta es la variable subyacente en todos los procesos (ingresos, altas, muertes...) relacionados con la enfermedad.

Como observaciones relacionadas con α_t se han elegido los ingresos en planta Ing_t, y en UCI, UCI_t. Podían haberse elegido otras que, a priori, podrían parecer intentos directamente dirigidos a estimar el número de infectados mediante medidas. Un ejemplo claro sería el número de PCR positivas por cada 100.000 habitantes. Sin embargo, se ha considerado que la muestra de personas a las que se hacen pruebas está demasiado sesgada como para ser valiosa. Al fin y al cabo, exceptuando los infrecuentes cribados, habitualmente sólo se han hecho pruebas a personas con síntomas o que hubiesen tenido contacto con positivos.

Por lo tanto, se ha considerado que los datos de atención hospitalaria representan mejor α_t. Teniendo en cuenta las características de la enfermedad, una determinada fracción de todos los casos serán graves (y necesitarán un ingreso en planta cierto tiempo después de la infección) y otra fracción serán críticos (y terminarán en la UCI tras evolucionar negativamente).

Según la revista Access Medicina, el 14% de los casos de COVID-19 son graves, y el 5% críticos. Se han tomado estas cifras para el trabajo, pero cabe mencionar que no están libres de sesgo. Para empezar, los casos asintomáticos sin registrar podrían hacer que la gravedad de la enfermedad estuviera sobreestimada. Además, estos datos no son específicos de la CAPV, aunque cabe esperar que la población de la CAPV no tenga particularidades que alteren estos datos, ya que se trata de una enfermedad nueva y la susceptibilidad de cualquier población será prácticamente la misma. La misma revista proporciona datos sobre el tiempo de evolución de la infección: afirma que la disnea (dificultad respiratoria) aparece en los casos graves con un retraso de 5-8 días, y que los casos críticos llegaron a serlo tras un tiempo de desarrollo de 8-12 días. Para este trabajo se han tomado retardos menores, con la intención de mantener un vector de estado de tamaño razonable y no tener que estimar demasiados parámetros. La decisión se ha tomado considerando que un modelo con tendencia lineal local para α_t es más adecuado que un paseo aleatorio.

```
fraccion_criticos<-0.05
fraccion_graves<-0.14
```

Por otra parte, se ha tenido en cuenta que, según la definición de R_0 propia de los modelos compartimentales (SIR y derivados), cada persona infectada contagia a su vez a otras R_0 personas en promedio. Este parámetro es variable en el tiempo, ya que depende tanto de la infectividad del virus como de factores demográficos y sociológicos (por ejemplo, las restricciones de movilidad).

Teniendo en cuenta estas consideraciones, las ecuaciones de observación propuestas son las siguientes:

$$\begin{aligned} Ing_t &= k_1\alpha_t + 0.14k_2R_0\alpha_{t-3} + 0.14k_3R_0\alpha_{t-4} + \epsilon_t^1 \\ UCI_t &= k_4\alpha_t + 0.05k_5R_0\alpha_{t-4} + 0.05k_6R_0\alpha_{t-5} + \epsilon_t^2 \end{aligned}$$

Se ha propuesto que algunas personas ingresan en UCI o planta inmediatamente. Otras ingresan con retraso, y en esos casos interviene el factor R_0.

Se imponen las siguientes restricciones adicionales, de forma que la suma de las contribuciones con distintos retardos sea efectivamente la fracción de casos graves o críticos respectivamente:

$$\begin{aligned} k_2 + k_3 &= 1 \\ k_5 + k_6 &= 1 \\ k_i \geq 0 \quad i &= 1, ..., 6 \end{aligned}$$

Estas restricciones llevan a la siguiente reformulación de las ecuaciones de observación, con ruido $\epsilon_t \sim N(0, V)$ y $0 \leq k_i \leq 1$:

$$Ing_t = k_1\alpha_t + 0.14R_{0,t}k_2\alpha_{t-3} + 0.14R_{0,t}(1-k_2)\alpha_{t-4} + \epsilon_t^1$$
$$UCI_t = k_3\alpha_t + 0.05R_{0,t}k_4\alpha_{t-4} + 0.05R_{0,t}(1-k_4)0.14\alpha_{t-5} + \epsilon_t^2$$

Para la ecuación de estado, se considerará que la pandemia tiene una dinámica propia de un modelo con tendencia lineal local, con ruido normalmente distribuido $\eta_t \sim N(0, W)$:

$$\alpha_t = \alpha_{t-1} + \delta_{t-1} + \eta_{\alpha,t-1}$$
$$\delta_t = \delta_{t-1} + \eta_{\delta,t-1}$$

Como se han considerado 5 retardos, el vector de estado será de dimensión 12. Tenemos un total de 8 parámetros a estimar:

- 4 para los coeficientes de las ecuaciones de observación
- 2 de la matriz de varianzas del ruido de las ecuaciones de estado
- 2 de la matriz de varianzas del ruido en las observaciones

Matricialmente, el modelo es el siguiente (marices de transición y de observación):

$$\begin{bmatrix} \alpha_t \\ \delta_t \\ \alpha_{t-1} \\ \delta_{t-1} \\ \alpha_{t-2} \\ \delta_{t-2} \\ \alpha_{t-3} \\ \delta_{t-3} \\ \alpha_{t-4} \\ \delta_{t-4} \\ \alpha_{t-5} \\ \delta_{t-5} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \alpha_{t-1} \\ \delta_{t-1} \\ \alpha_{t-2} \\ \delta_{t-2} \\ \alpha_{t-3} \\ \delta t-3 \\ \alpha_{t-4} \\ \delta_{t-4} \\ \alpha_{t-5} \\ \delta t-5 \\ \alpha_{t-6} \\ \delta_{t-6} \end{bmatrix} + \vec{\eta_t}$$

$$\begin{bmatrix} Ing_t \\ UCI_t \end{bmatrix} = \begin{bmatrix} k_1 & 0 & 0 & 0 & 0 & 0 & 0.14k_2R_0 & 0 & 0.14(1-k_2)R_0 & 0 & 0 & 0 \\ k_3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.05k_4R_0 & 0 & 0.05(1-k_4)R_0 & 0 \end{bmatrix} \begin{bmatrix} \alpha_t \\ \delta_t \\ \alpha_{t-1} \\ \delta_{t-1} \\ \alpha_{t-2} \\ \delta_{t-2} \\ \alpha_{t-3} \\ \delta_{t-3} \\ \alpha_{t-4} \\ \delta_{t-4} \\ \alpha_{t-5} \\ \delta_{t-5} \end{bmatrix} + \vec{\epsilon_t}$$

donde $k_i = \frac{1}{1+e^{-C_i}} \quad \forall i \neq 1, 3$, para asegurar que $0 \leq k_i \leq 1$. Las matrices de varianzas-covarianzas de los ruidos se consideran diagonales, con términos de tipo e^{C_i} donde $i = 5, 6$ para el ruido del estado y $i = 7, 8$ para las observaciones. También se da como nivel inicial del estado un vector de 12 ceros, y como varianza inicial una matriz C_0 a priori difusa.

3.2) Implementación

En la siguiente función se crea este modelo, y después se ajusta estimando los 8 parámetros de forma máximo-verosímil.

```
crearMod <- function(x) {
  modelo<-dlmModPoly(2)
  modelo$GG<-matrix(c(1,1,rep(0,10),
                      0,1,rep(0,10),
                      1,rep(0,11),
                      0,1,rep(0,10),
                      rep(0,2),1,rep(0,9),
                      rep(0,3),1,rep(0,8),
                      rep(0,4),1,rep(0,7),
                      rep(0,5),1,rep(0,6),
                      rep(0,6),1,rep(0,5),
                      rep(0,7),1,rep(0,4),
                      rep(0,8),1,rep(0,3),
                      rep(0,9),1,0,0),nrow=12,ncol=12,byrow=T)
  modelo$FF<-matrix(c(exp(x[3]),0,0,0,0,0,1,0,1,0,0,0,
                      exp(x[4]),0,0,0,0,0,0,0,1,0,1,0),nrow=2,ncol =12,byrow = T)
  modelo$JFF<-matrix(c(0,0,0,0,0,0,1,0,2,0,0,0,
                      0,0,0,0,0,0,0,0,3,0,4,0),nrow=2,ncol =12,byrow = T)
  modelo$X<-matrix(merge((1/(1+exp(-x[1])))*fraccion_graves*covid$R0,
                         (1-(1/(1+exp(-x[1]))))*fraccion_graves*covid$R0,
                         (1/(1+exp(-x[2])))*fraccion_criticos*covid$R0,
                         (1-(1/(1+exp(-x[2]))))*fraccion_criticos*covid$R0),
                   ncol=4,nrow=nrow(covid))
  modelo$m0<-matrix(rep(0,12),1,12)
  modelo$C0<-diag(1e7,12,12)
  modelo$W<-diag(c(exp(x[7]),exp(x[8]),rep(0,10)),
                 12,12)
  modelo$V<-diag(c(exp(x[5]),exp(x[6])),2,2)
  return(modelo)
}

ajuste<-dlmMLE(merge(covid$Ingresados,covid$UCI),
               parm=rep(0,8),build = crearMod)
```

Las matrices del sistema después de la estimación son las siguientes:

```
modelo_ajustado<-crearMod(ajuste$par)
modelo_ajustado
```

```
## $FF
##          [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12]
## [1,] 1.0664944    0    0    0    0    0    1    0    1    0    0    0
## [2,] 0.2562578    0    0    0    0    0    0    0    1    0    1    0
##
## $V
##          [,1]        [,2]
## [1,] 4474.751 0.000000e+00
## [2,]    0.000 8.745965e-09
##
## $GG
##       [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12]
##  [1,]    1    1    0    0    0    0    0    0    0     0     0     0
##  [2,]    0    1    0    0    0    0    0    0    0     0     0     0
##  [3,]    1    0    0    0    0    0    0    0    0     0     0     0
##  [4,]    0    1    0    0    0    0    0    0    0     0     0     0
##  [5,]    0    0    1    0    0    0    0    0    0     0     0     0
##  [6,]    0    0    0    1    0    0    0    0    0     0     0     0
##  [7,]    0    0    0    0    1    0    0    0    0     0     0     0
##  [8,]    0    0    0    0    0    1    0    0    0     0     0     0
##  [9,]    0    0    0    0    0    0    1    0    0     0     0     0
## [10,]    0    0    0    0    0    0    0    1    0     0     0     0
## [11,]    0    0    0    0    0    0    0    0    1     0     0     0
## [12,]    0    0    0    0    0    0    0    0    0     1     0     0
##
```

```
## $W
##           [,1]         [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11]
##  [1,] 143.4709 0.000000e+00    0    0    0    0    0    0    0     0     0
##  [2,]   0.0000 2.781281e-07    0    0    0    0    0    0    0     0     0
##  [3,]   0.0000 0.000000e+00    0    0    0    0    0    0    0     0     0
##  [4,]   0.0000 0.000000e+00    0    0    0    0    0    0    0     0     0
##  [5,]   0.0000 0.000000e+00    0    0    0    0    0    0    0     0     0
##  [6,]   0.0000 0.000000e+00    0    0    0    0    0    0    0     0     0
##  [7,]   0.0000 0.000000e+00    0    0    0    0    0    0    0     0     0
##  [8,]   0.0000 0.000000e+00    0    0    0    0    0    0    0     0     0
##  [9,]   0.0000 0.000000e+00    0    0    0    0    0    0    0     0     0
## [10,]   0.0000 0.000000e+00    0    0    0    0    0    0    0     0     0
## [11,]   0.0000 0.000000e+00    0    0    0    0    0    0    0     0     0
## [12,]   0.0000 0.000000e+00    0    0    0    0    0    0    0     0     0
##       [,12]
##  [1,]     0
##  [2,]     0
##  [3,]     0
##  [4,]     0
##  [5,]     0
##  [6,]     0
##  [7,]     0
##  [8,]     0
##  [9,]     0
## [10,]     0
## [11,]     0
## [12,]     0
##
## $JFF
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12]
## [1,]    0    0    0    0    0    0    1    0    2     0     0     0
## [2,]    0    0    0    0    0    0    0    0    3     0     4     0
##
## $X
##      [,1]  [,2]  [,3]  [,4]
## [1,]   NA    NA    NA    NA
## [2,]   NA    NA    NA    NA
## [3,] ...
##
## $m0
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12]
## [1,]    0    0    0    0    0    0    0    0    0     0     0     0
##
## $C0
##       [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  [,7]  [,8]  [,9] [,10] [,11] [,12]
##  [1,] 1e+07 0e+00 0e+00 0e+00 0e+00 0e+00 0e+00 0e+00 0e+00 0e+00 0e+00 0e+00
##  [2,] 0e+00 1e+07 0e+00 0e+00 0e+00 0e+00 0e+00 0e+00 0e+00 0e+00 0e+00 0e+00
##  [3,] 0e+00 0e+00 1e+07 0e+00 0e+00 0e+00 0e+00 0e+00 0e+00 0e+00 0e+00 0e+00
##  [4,] 0e+00 0e+00 0e+00 1e+07 0e+00 0e+00 0e+00 0e+00 0e+00 0e+00 0e+00 0e+00
##  [5,] 0e+00 0e+00 0e+00 0e+00 1e+07 0e+00 0e+00 0e+00 0e+00 0e+00 0e+00 0e+00
##  [6,] 0e+00 0e+00 0e+00 0e+00 0e+00 1e+07 0e+00 0e+00 0e+00 0e+00 0e+00 0e+00
##  [7,] 0e+00 0e+00 0e+00 0e+00 0e+00 0e+00 1e+07 0e+00 0e+00 0e+00 0e+00 0e+00
##  [8,] 0e+00 0e+00 0e+00 0e+00 0e+00 0e+00 0e+00 1e+07 0e+00 0e+00 0e+00 0e+00
##  [9,] 0e+00 0e+00 0e+00 0e+00 0e+00 0e+00 0e+00 0e+00 1e+07 0e+00 0e+00 0e+00
## [10,] 0e+00 0e+00 0e+00 0e+00 0e+00 0e+00 0e+00 0e+00 0e+00 1e+07 0e+00 0e+00
## [11,] 0e+00 0e+00 0e+00 0e+00 0e+00 0e+00 0e+00 0e+00 0e+00 0e+00 1e+07 0e+00
## [12,] 0e+00 0e+00 0e+00 0e+00 0e+00 0e+00 0e+00 0e+00 0e+00 0e+00 0e+00 1e+07
```

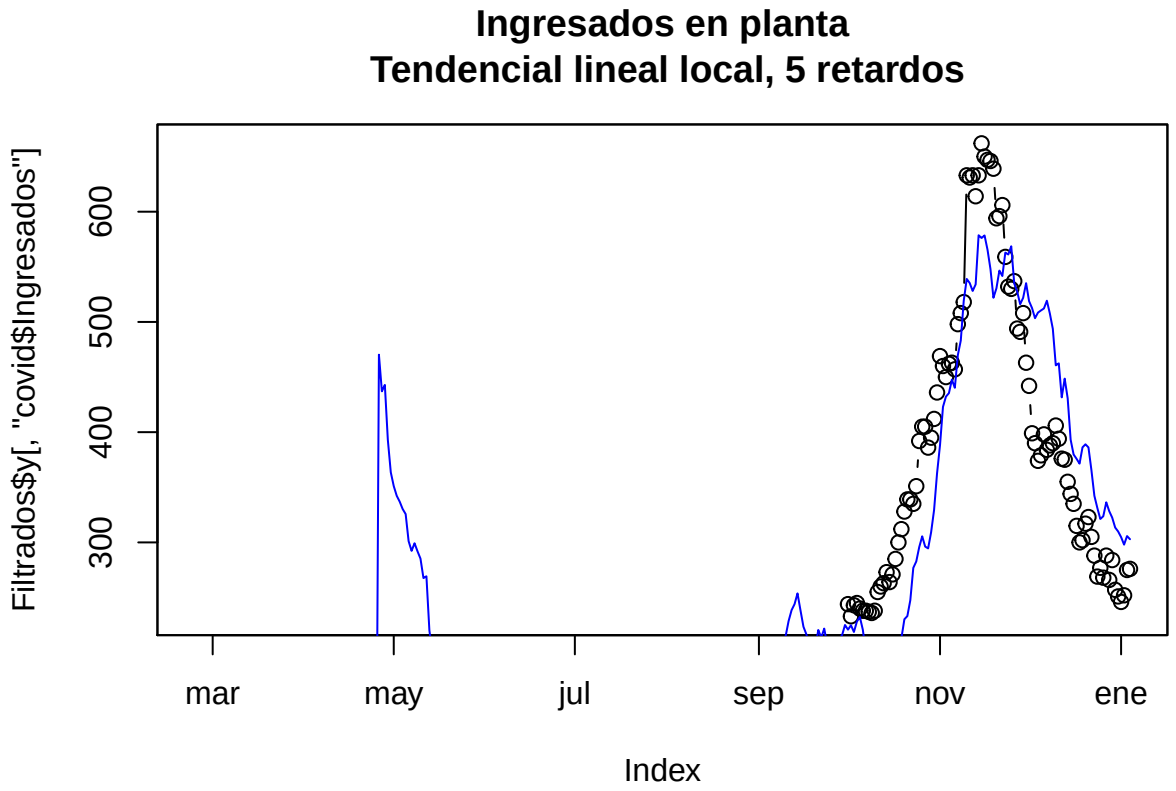
Según los datos hospitalarios de los que disponemos, si asumimos las ecuaciones que hemos planteado como correctas, lo más verosímil es que, entre el 14% de pacientes que ingresan con retraso en planta, el 76.3% lo hagan en el día $t-3$ y los demás en el día $t-4$. Así mismo, entre el 5% que ingresa con retraso en la UCI, el resultado de la estimación máximo-verosímil es que el 17.8% presenta una gravedad crítica tras 4 días, y el resto empeora más tarde $(t-5)$.

```
1/(1+exp(-ajuste$par[1:2]))
```

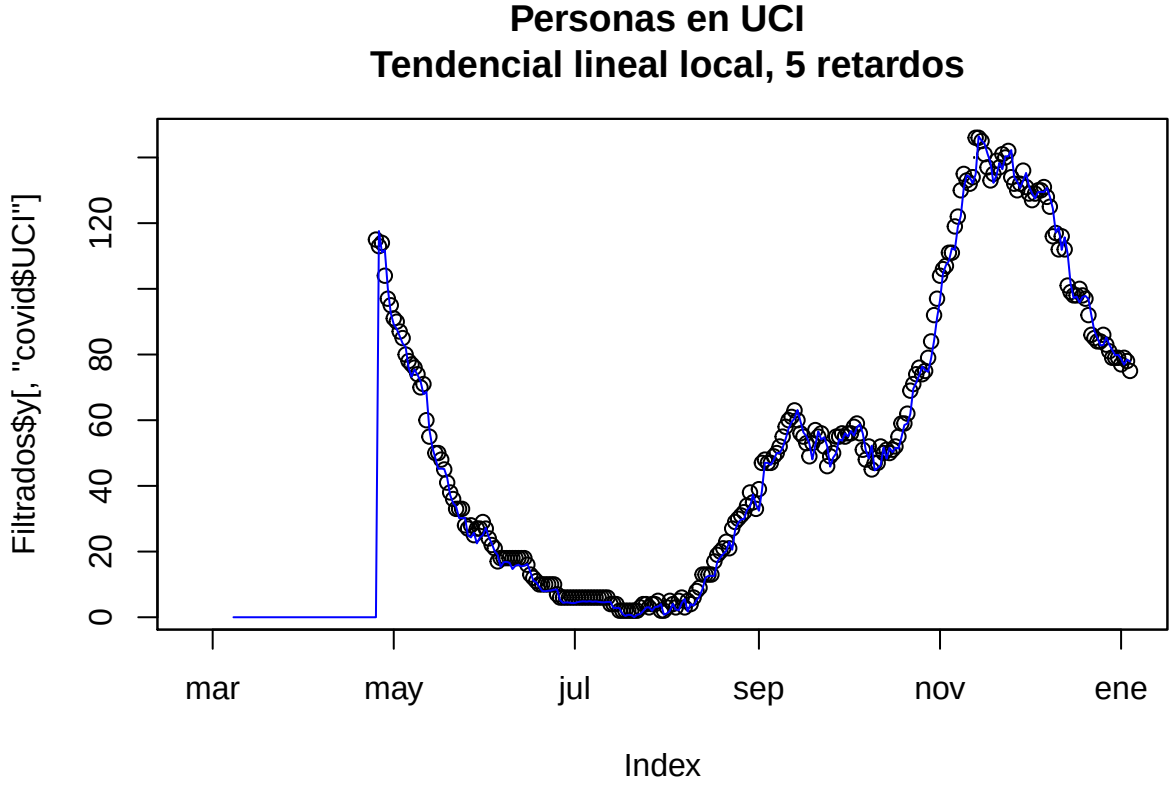
```
## [1] 0.76256314 0.09766103
```

A continuación procedemos al filtrado de las observaciones:

```
Filtrados<-dlmFilter(merge(covid$Ingresados,covid$UCI) ,modelo_ajustado)
plot(Filtrados$y[,"covid$Ingresados"],type = "b", main="Ingresados en planta
 Tendencial lineal local, 5 retardos")
lines(Filtrados$f[,"covid$Ingresados"],col="blue")
```




```
plot(Filtrados$y[,"covid$UCI"],type = "b",main = "Personas en UCI
 Tendencial lineal local, 5 retardos")
lines(Filtrados$f[,"covid$UCI"],col="blue")
```



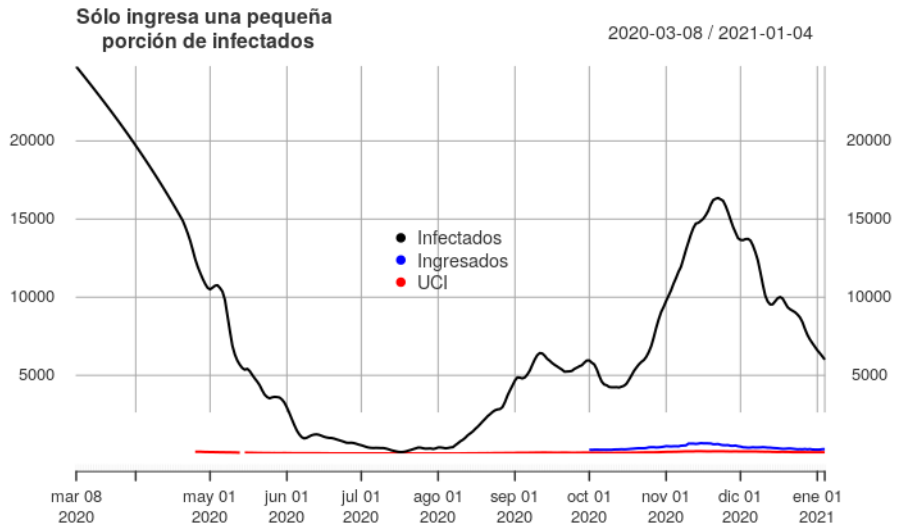
Vemos que el modelo se ajusta perfectamente a los ingresados en UCI, y no tanto a los de planta. Es común que, al plantear este tipo de modelos multivariantes, una de las series tenga mejor ajuste.

Ahora obtenemos una estimación suavizada del número de personas infectadas (variable de estado). `dlmSmooth` devuelve las componentes de la serie (tendencia, estacionalidad y perturbaciones) por separado y en formato matricial, así que las sumamos con `rowSums` y las guardamos como serie temporal:

```
Suavizados<-dlmSmooth(Filtrados ,modelo_ajustado)
infectados<-as.xts(rowSums(dropFirst(Suavizados$s)))
index(infectados)<-time(covid)
```

A continuación comparamos gráficamente la estimación de personas infectadas con los datos hospitalarios:

```
plot(window(merge(infectados,covid$Ingresados,covid$UCI),start=as.Date('2020-03-08')),
     plot.type="single",col=c("black","blue","red"),
     main = "Sólo ingresa una pequeña
     porción de infectados")
addLegend(legend.names = c("Infectados","Ingresados","UCI"),
          col=c("black","blue","red"),pch=19,
          legend.loc ="center")
```

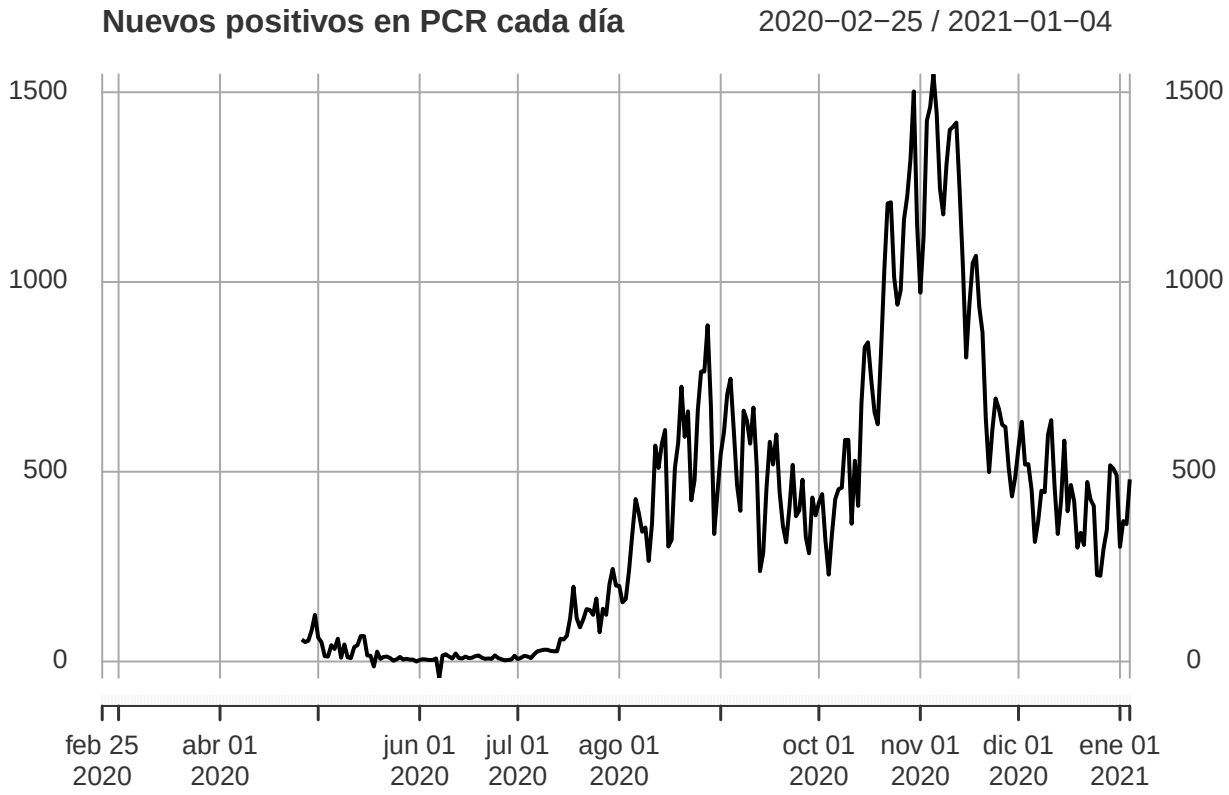


Vemos que los infectados presentan la estructura de olas propia de esta pandemia. La primera ola presenta una curvatura extraña, ya que en esas fechas sólo tenemos datos de R_0.

3.3) Comentarios sobre los resultados

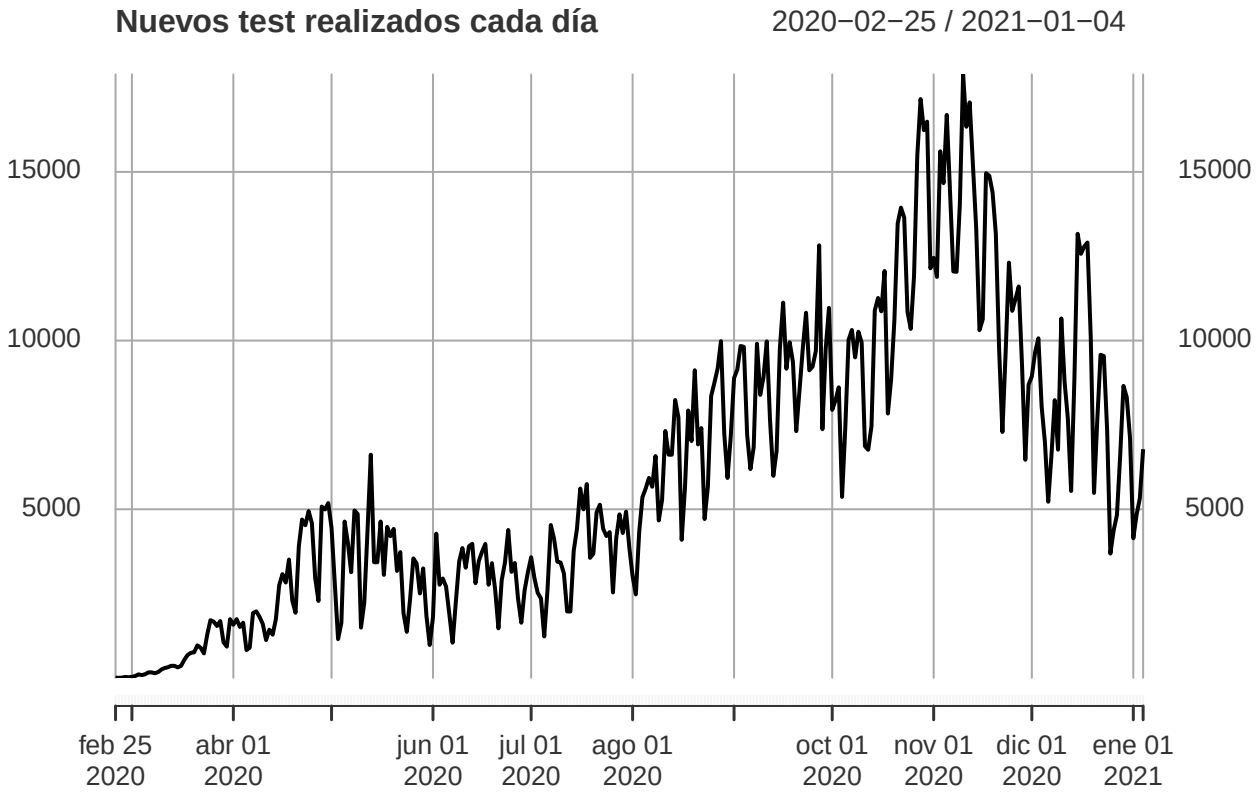
Como ya hemos comentado, el número de infectados es imposible de medir. Sin embargo, en la siguiente figura podemos ver que el perfil de nuevas PCR positivas sigue una estructura similar a nuestra estimación suavizada. Reiteramos que hay que resistir cualquier tentación de inferir el número de infectados a partir del porcentaje de PCR positivas porque daría una cifra sobreestimada. Sin embargo la cantidad de positivos detectados sigue guardando cierta relación con el número de infectados reales, y que los perfiles se parezcan es una buena señal.

```
par(mfrow=c(1,1))
plot(as.xts(diff(covid$PCRPos)),
     main="Nuevos positivos en PCR cada día")
```



No obstante, la relación no es directa, ya que está influida por el número de test diarios realizados. Esta cifra presenta variabilidad temporal, ilustrada en la figura de abajo. Además, se ve que a partir de Noviembre de 2020 el número de test diarios se redujo, y esto hubo de tener un impacto en que el número de PCR positivas cayese a partir de esas fechas, cuando nuestra estimación suavizada muestra que las infecciones aún estaban subiendo.

```
plot(as.xts(diff(covid$TestTot)),
     main="Nuevos test realizados cada día")
```



Más allá de la forma del perfil de los infectados, conviene recordar que el nivel de dicho perfil está sujeto a las cifras que hemos tomado como indicativas de la gravedad de la enfermedad. Como ya se ha dicho, estas cifras están posiblemente sobreestimadas. Esto implica que la cantidad de personas infectadas es probablemente mayor que la que hemos estimado, aunque con el mismo perfil.