Un modelo epidemiológico de series temporales para la COVID-19

Víctor de Buen Remiro. victor.debuen@inverence.com

10 de mayo de 2020

INVERENCE

Resumen

En todo el mundo ha quedado patente que existen multitud de problemas asociados a la recopilación de los datos diarios relativos a la pandemia del COVID-19. Las causas son múltiples, desde cambios metodológicos a problemas administrativos o falta de medios. Como consecuencia, a menudo se producen importantes distorsiones en las series temporales que describen la epidemia en una zona geográfica concreta, haciendo bastante complicado aplicar de forma directa la teoría epidemiológica determinista, como los modelos SIR y sus variantes.

El enfoque adoptado aquí es el de la anidación de modelos. En este caso se parte de un primer modelo determinista para la previsión de casos confirmados de contagio, basado en una modificación de la curva de Gompertz, expresado como una regresión lineal bajo cierta transformación y ampliado con análisis de intervención para el tratamiento de disrupciones contables. A los errores de este primer modelo se les aplica un modelo ARMA, lo cual permite al sistema ir aprendiendo de la evolución temporal de la epidemia y de los errores cometidos por el propio modelo, para adaptarse así de forma muy dinámica. Para la estimación de fallecidos y recuperados se utiliza un modelo de transferencia dinámica de Box-Jenkins sobre los casos confirmados, seguido de sendos modelos ARMA.

Esta mezcla de tecnologías cooperando entre sí produce claramente mejores previsiones que con cada una de ellas por separado.

1. Introdución

En este documento se describe la metodología aplicada por la empresa Inverence en el análisis de datos y la estimación de previsiones de la pandemia del COVID-19 en diferentes áreas geográficas.

El primer problema a la hora de estimar los parámetros de cualquier enfoque epidemiológico es que en realidad el número de contagiados es básicamente inobservable, y menos aún de un modo coherente a lo largo del tiempo y con un nivel geográfico mínimamente detallado.

Cualquier ataque debe tener en cuenta además los problemas de contabilización en condiciones tan luctuosas. Los datos se conocen con retraso, el cuál varía con el tiempo y es particularmente más alto los fines de semana y festivos, muy probablemente por falta de personal administrativo.

No se publica el número de tests realizados cada día por lo que es difícil saber qué parte de la evolución se debe al contagio real y cuánto al tamaño del muestreo, pues el número de tests disponibles censura la información de contagios confirmados de forma no homogénea en el tiempo.

La falta de facultativos disponibles impide registrar todos los fallecimientos causados fuera de los hospitales, especialmente en las residencias de ancianos.

Las distintas medidas de confinamiento surten efecto con un retardo dependiente del periodo larvario de la infección que podría estar entre 5 y 15 días, complicando mucho la medición de sus efectos.

La llegada súbita de nuevos materiales como respiradores, la experiencia acumulada por los equipos sanitarios, pueden causar un descenso de fallecidos que no se debe a las propiedades intrínsecas de la epidemia.

Por si esto fuera poco, todas estas circunstancias varían en cada país y en cada región dependiendo de factores políticos y socio-económicos, y probablemente también por las circunstancias meteorológicas.

Los modelos SIR y sus variantes, y en general todos los de corte determinista, resuelven ecuaciones en derivadas parciales de un modo en cierta forma similar a los modelos meteorológicos que tan buenos resultados están dando. Sin embargo los modelos epidemiológicos de corte determinista no siempre funcionan muy bien en la práctica porque se requiere un nivel de información muy bueno y detallado, cosa que no está ocurriendo, que está de hecho a años luz del nivel de detalle de los datos de los cientos de miles o millones de estaciones meteorológicas que toman datos muy fiables de multitud de magnitudes físicas por todo el mundo cada hora o menos. En una epidemia no puedes tener ese nivel de detalle informativo y hay que entenderlo, porque lo que hay que hacer es salvar a la gente, y no se emplea todo el tiempo que sería necesario en contabilizar adecuadamente todo cada día.

Para poder adaptarse a todas estas fluctuaciones se hace imprescindible un enfoque muy dinámico a la par que suficientemente robusto, tanto desde el punto de vista teórico como del cálculo numérico. La solución adoptada es un conjunto de modelos anidados por fases, en el que las salidas de los modelos de una fase es utilizada como entrada en los de la siguiente fase, mejorando así los resultados que se podrían obtener de cada tipo de modelos por separado.

2. Antecedentes

Existe una gran variedad de modelos matemáticos que tratan de explicar los fenómenos epidemiológicos [Vynnycky(2016), Grassly and Fraser(2008)], algunos de los cuales tienen en cuenta la naturaleza estocástica no lineal de los procesos subyacentes, aunque la mayoría son de corte determinista.

Muchos de estos últimos (ej. [Ke Wu and Sornette(2020)]) se basan en ajustar curvas suaves (continuas y derivables múltiples veces) al número de contagiados en función del tiempo $y(t)$. En general se trata de soluciones de un sistema de ecuaciones en diferencias parciales, como los modelos compartimentales [Fred Brauer(2008)] , o bien se simplifican a una única ecuación diferencial, como por ejemplo la ecuación de Richards [RICHARDS(1959)]

$$\frac{y'(t)}{y(t)} = \alpha\left(1 - \left(\frac{y(t)}{K}\right)^{\nu}\right) \tag{1}$$

Particularmente, una que se utiliza muy frecuentemente (ej. [Zhao(2020)]) es la curva de Gompertz [Gompertz(1825)] que cumple esta ecuación de forma asintótica cuando $\nu \to 0^+$, la cual se reduce a la expresión

$$k^{\frac{y'(t)}{y(t)}} \propto \frac{1}{y(t)} \tag{2}$$

donde a $\frac{y'(t)}{y(t)}$ se le llama tasa de crecimiento continuo del contagio y k es una constante arbitraria positiva. Tomando logaritmos se obtiene la igualdad

$$\frac{y'(t)}{y(t)} = k_0 - k_1 \ln\left(y(t)\right) \ \wedge \ k_1 > 0 \tag{3}$$

que en su forma discreta de diferencias finitas se puede escribir como

$$\frac{y_t - y_{t-1}}{y_{t-1}} = k_0 - k_1 \ln\left(y_{t-1}\right) \tag{4}$$

Esto nos permite obtener k_0 y k_1 mediante una simple regresión lineal que nos da una forma relativamente robusta de estimar la tasa de crecimiento durante el periodo inicial de la epidemia, cuando todavía hay pocos casos y la tasa observada es muy inestable.

De hecho ésta puede pasar de ser cero a estar cerca de la unidad de un día para otro y al revés, tal y como se observa en la figura 1.

Así pues, esta regresión no nos sirve para prever sino tan sólo para mejorar la calidad de los datos en la fase inicial.

La solución general de la curva de Gompterz es

$$y\left(t\right) = a\mathrm{e}^{-b\mathrm{e}^{-ct}} \ \wedge \ a, b, c > 0 \tag{5}$$

y su derivada

$$y'\left(t\right) = abc\mathrm{e}^{-ct}\mathrm{e}^{-b\mathrm{e}^{-ct}} \tag{6}$$

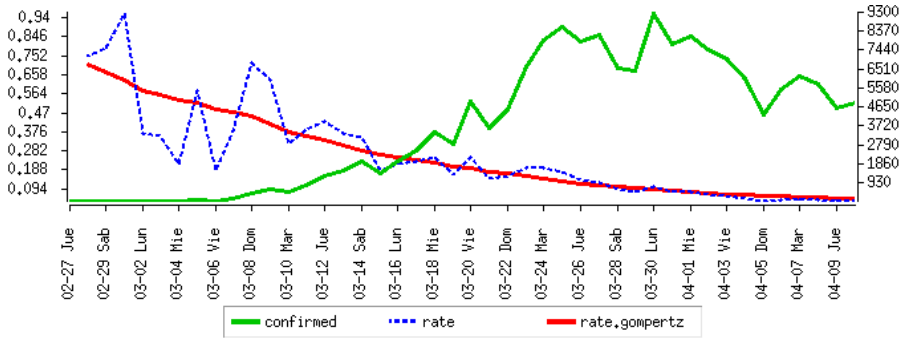

Figura 1: Tasa de crecimiento de casos confirmados en España (en azul) y la aproximación de Gompertz (en rojo) usando la fórmula 4 . Cuando hay pocos confirmados (en verde) es una estimación de la tasa mucho más robusta que la propia tasa observada.

La tasa de crecimiento continuo es por tanto

$$\frac{y'(t)}{y(t)} = cb\mathrm{e}^{-ct} \tag{7}$$

y su logaritmo resulta decrecer linealmente en el tiempo

$$\ln\left(\frac{y'(t)}{y(t)}\right) = \ln(cb) - ct \tag{8}$$

En diferencias finitas podría escribirse así

$$-\ln\left(\frac{y_t - y_{t-1}}{y_{t-1}}\right) = \alpha + \beta t \ \wedge \ \beta > 0 \tag{9}$$

y llamando tasa de crecimiento discreto a la cantidad

$$\tau_t = \frac{y_t - y_{t-1}}{y_{t-1}} \tag{10}$$

podemos establecer una relación lineal en el tiempo sobre la transformación logarítmica de su inversa

$$\zeta_t = \ln\left(\frac{1}{\tau_t}\right) = -\ln(\tau_t) = \alpha + \beta t \ \wedge \ \beta > 0 \tag{11}$$

que sí es perfectamante válida a efectos predictivos de la tasa tanto hacia el futuro como hacia el pasado, pues no es otra cosa que una simple función exponencial en el tiempo:

$$\tau_t = e^{-\zeta_t} = e^{-(\alpha+\beta t)} \ \wedge \ \beta > 0 \tag{12}$$

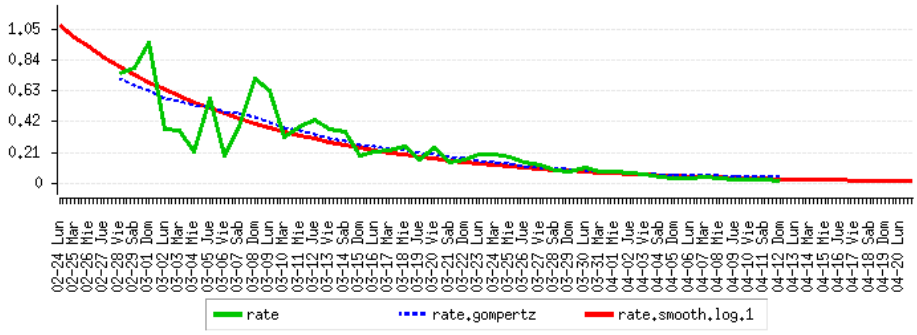

Figura 2: Tasa de crecimiento, aproximación de Gompertz (eq 4) y previsión suavizada exponencialmente (eq 12)

3. Propuestas

Sin embargo, esta transformación exponencial negativa en el tiempo implica que en el instante inicial de la epidemia la tasa tendría valor finito, cuando en realidad los contagiados habrían pasado de 0 a un valor positivo. Teniendo en cuenta esto, parece más razonable emplear una transformación de la tasa que además de ser asintótica en el eje de abscisas lo sea también en el de ordenadas, es decir, que la tasa tienda a infinito cuando el tiempo tiende a cero, además de tender a cero cuanto el tiempo tiende a infinito. La curva de la tasa tendría que tener por lo tanto forma hiperbólica.

Una familia paramétrica de funciones que tiene dicha forma, y es bastante similar a la exponencial negativa cuando el tiempo tiende a infinito (ver figura 3), es la cosecante hiperbólica de las potencias supraunitarias de la variable linealizada en el tiempo

$$\tau_t = \frac{1}{\sinh\left(\zeta_t^{\nu}\right)} = \frac{2}{e^{\zeta_t^{\nu}} - e^{-\zeta_t^{\nu}}} \ \wedge \ \nu \geq 1 \tag{13}$$

Si definimos la variable auxiliar $z_t = e^{\zeta_t^{\nu}}$ nos queda

$$\tau_t = \frac{2}{z_t - \frac{1}{z_t}} \tag{14}$$

$$\left(z_t - \frac{1}{z_t}\right)\tau_t = 2 \tag{15}$$

$$z_t^2\tau_t - 2z_t - \tau_t = 0 \tag{16}$$

$$z_t = \frac{2 \pm \sqrt{4 + 4\tau_t^2}}{2\tau_t} = \frac{1 \pm \sqrt{1 + \tau_t^2}}{\tau_t} \tag{17}$$

Puesto que $z_t > 0$ sólo se admite la solución positiva y así se obtiene la inversa de la transformación cosecante hiperbólica

$$\zeta_t = \left(\ln \left(\frac{1 + \sqrt{1 + \tau_t^2}}{\tau_t} \right) \right)^{\frac{1}{\upsilon}} \tag{18}$$

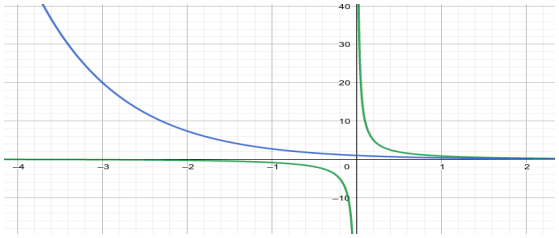

Figura 3: Comparación de la función exponencial negativa (azul) y la cosecante hiperbólica (verde)

Se podría criticar este nuevo enfoque aduciendo que este juega con ventaja al incorporar el parámetro ajustable υ, pero lo mismo se puede hacer con el enfoque de Gompertz ampliándolo a la familia paramétrica de transformaciones exponenciales negativas

$$\tau_t = e^{-z_t^{\nu}} = \frac{1}{e^{z_t^{\nu}}} \ \wedge \ \nu \geq 1 \tag{19}$$

Si ajustamos las potencias que dan el mejor ajuste en cada familia vemos que dan resultados muy parecidos durante el periodo de datos observados y futuros, debido a que el término $e^{-\zeta_t^{\nu}}$ tiende a cero con el tiempo, pero la familia exponencial da resultados absurdos para la tasa en el pasado, como puede apreciarse en la figura 4.

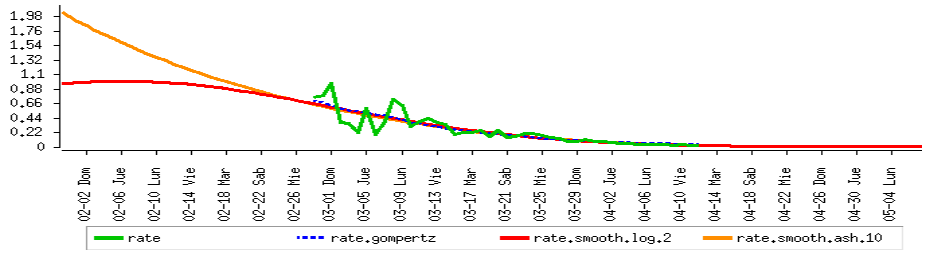

Figura 4: Aproximación de Gompertz en azul. Ajustes óptimos de la tasa (en verde) con las familias exponencial (en rojo) y cosecante hiperbólica (en naranja)

A efectos de predicción futura no se aprecian diferencias significativas entre ambos enfoques (ver figura 5)

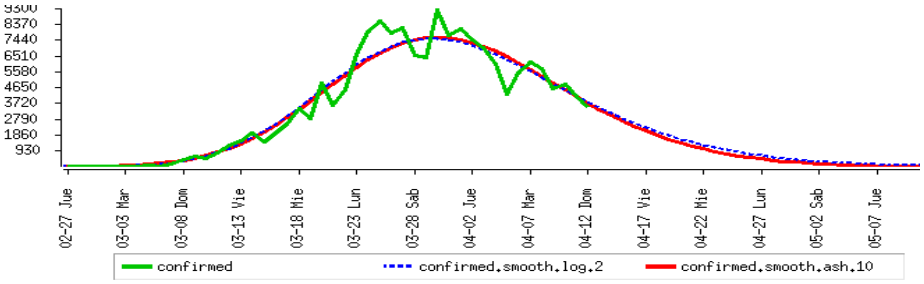

Figura 5: Comparación del ajuste de casos confirmados (en verde) con las familias exponencial (en azul) y cosecante hiperbólica (en rojo)

Este nuevo enfoque nos permite definir las curvas de contagio mediante modelos de regresión lineal sobre el tiempo de distintas transformaciones de la tasa de contagio. Esto cobra una gran importancia cuando el modelo debe adaptarse a cambios estructurales como el que ha ocurrido en España con la entrada masiva de nuevos tests que han permitido hacer muchos más de los que se hacían anteriormente.

En la métrica linealizada mediante la transformación cosecante hiperbólica inversa se observa que efectivamente hay un punto de ruptura, un antes y un después (figura 6)

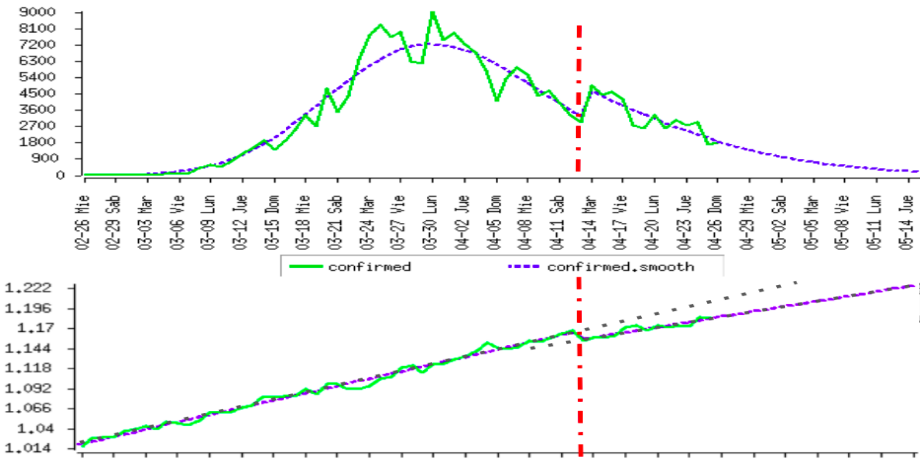

Figura 6: A mediados de abril el número de confirmados (arriba) aumenta pero no de debe a un rebrote dela epidemia, sino a se están haciendo muchos más tests. Esto se traduce en un cambio de tendencia y nivel en la métrica linealizada (abajo).

Este tipo de cambios se manejan de una forma muy natural en esta métrica linealizada, simplemente duplicando los parámetros para adaptarse tanto al

comportamiento previo como al posterior al instante de ruptura T_0, dando como resultado una nueva regresión lineal con 4 parámetros en lugar de 2

$$\begin{aligned} \zeta_t &= \left(\ln \left(\frac{1+\sqrt{1+\tau_t^2}}{\tau_t} \right) \right)^{\frac{1}{\upsilon}} \\ \hat{\zeta}_t &= \begin{cases} \alpha_0 + \beta_0 t & \forall t < T_0 \\ \alpha_1 + \beta_1 t & \forall t \geq T_0 \end{cases} \\ \zeta_t - \hat{\zeta}_t &= \epsilon_t \sim N\left(0, \varrho^2\right) \end{aligned} \tag{20}$$

4. Definiciones

Llamaremos zona geográfica a un país, un conjunto de países o una región administrativa que sea parte de un país concreto.

Para cada zona geográfica $g = 1, 2, \ldots, G$ observamos las siguientes series temporales

- $c_{g,t}$: Casos nuevos confirmados diariamente
- $d_{g,t}$: Defunciones diarias
- $r_{g,t}$: Altas médicas diarias (Recuperados)

En cada región se construyen las series acumuladas[1]

- $C_{g,t} = \sum_{t=1}^{T} c_{g,t}$: Casos confirmados acumulados
- $D_{g,t} = \sum_{t=1}^{T} d_{g,t}$: Defunciones acumuladas
- $R_{g,t} = \sum_{t=1}^{T} r_{g,t}$: Altas médicas acumuladas (Recuperados)
- $A_{g,t} = C_{g,t} - D_{g,t} - R_{g,t}$: Casos confirmados activos
- $a_{g,t} = A_{g,t} - A_{g,t-1}$: Casos activos nuevos

Los correspondientes totales agregados geográficamente, cuando dicha agregación sea pertinente, por ejemplo por tratarse de niveles administrativos de un mismo país, los notaremos de manera general

$$x_{0,t} = \sum_{g=1}^{G} x_{g,t} \, \forall \, x = c, d, r, a \tag{21}$$

$$X_{0,t} = \sum_{g=1}^{G} X_{g,t} \, \forall \, x = C, D, R, A \tag{22}$$

[1] En realidad, los datos publicados suelen ser los casos y defunciones acumulados y unas veces se dan los recuperados acumulados y otras los casos activos. Sea como sea, basta con tres de estas magnitudes para poder reconstruir el resto.

5. Modelo de la tasa de contagios confirmados

En todos los países y regiones el número de total de contagiados es desconocido en tiempo real, únicamente se observan los casos confirmados oficialmente, y sólo pasado el tiempo se puede llegar a tener una estimación del total de casos, nunca de su evolución diaria. Así que lo único que se puede hacer es pensar que los confirmados son una proporción del total que permanece aproximadamente constante en el tiempo. Hay que tener en cuenta por tanto que si el número de tests realizados se incrementa ostensiblemente, como realmente ha ocurrido en muchas partes, habría que hacer las oportunas correcciones para normalizar los datos tal y como se ha visto en el apartado anterior.

La tasa de contagio confirmado se define como el cociente entre los nuevos confirmados y los casos acumulados hasta el día previo

$$\tau_{g,t} = \frac{c_{g,t}}{\sum_{k=1}^{t-1} c_{g,k}} = \frac{C_{g,t} - C_{g,t-1}}{C_{g,t-1}} = \frac{C_{g,t}}{C_{g,t-1}} - 1 \tag{23}$$

Obviamente, bajo este supuesto de proporcionalidad de los contagios observados sobre el total, la tasa de confirmados coincide exactamente con la tasa de contagios y se puede modelar como una regresión lineal sobre la transformación inversa de la cosecante hiperbólica definida en la ecuación 20, teniendo en cuenta que en cada región pueden variar tanto los parámetros de la regresión y su varianza como el instante de disrupción

$$\begin{gathered} \zeta_{g,t} = \left(\ln \left(\frac{1+\sqrt{1+\tau_{g,t}^2}}{\tau_{g,t}} \right) \right)^{\frac{1}{v}} \\ \hat{\zeta}_{g,t} = \begin{cases} \alpha_{g,0} + \beta_{g,0} t & \forall t < T_{g,0} \\ \alpha_{g,1} + \beta_{g,1} t & \forall t \geq T_{g,0} \end{cases} \\ \epsilon_{g,t} \sim N\left(0, \varrho_g^2\right) \end{gathered} \tag{24}$$

De esta forma se obtiene una previsión suavizada de la tasa de contagio con un punto de disrupción

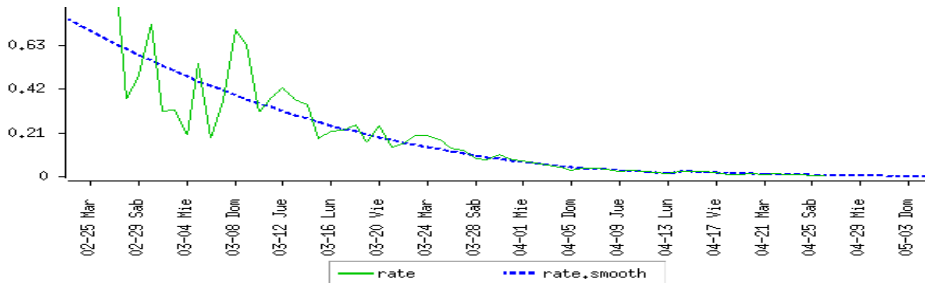

Figura 7: Previsión suavizada de la tasa de contagios confirmados en España.

6. Modelo de contagios confirmados

Conocida la previsión de la tasa de crecimiento $\hat{\tau}_{g,t}$ podemos calcular la previsión suavizada de los casos acumulados $\hat{C}_{g,t}$ de este modo

$$1 + \hat{\tau}_{g,t} = \frac{\hat{C}_{g,t}}{\hat{C}_{g,t-1}} \tag{25}$$

$$\ln\left(1 + \hat{\tau}_{g,t}\right) = \ln \hat{C}_{g,t} - \ln \hat{C}_{g,t-1} \tag{26}$$

$$\ln \hat{C}_{g,t} = \sum_{k=1}^{t} \ln\left(1 + \hat{\tau}_{g,k}\right) \tag{27}$$

$$\hat{C}_{g,t} = \exp\left(\sum_{k=1}^{t} \ln\left(1 + \hat{\tau}_{g,k}\right)\right) \tag{28}$$

El número de nuevos casos confirmados también se puede expresar en función de las previsiones de las tasas de este modo

$$\hat{c}_{g,t} = \hat{C}_{g,t} - \hat{C}_{g,t-1} = \exp\left(\sum_{k=1}^{t-1} \ln\left(1 + \hat{\tau}_{g,k}\right)\right) \hat{\tau}_{g,t} \tag{29}$$

Este primer enfoque de modelación nos dará una curva muy suave en el tiempo, de hecho es continua e infinitamente derivable[2], alrededor de la cual se moverá el número efectivamente observado de casos confirmados. De esa diferencia entre el valor observado y el previsto se ha comprobado que se puede aceptar su normalidad bajo cierta transformación instantánea de Box-Cox [Box and Cox(1964)], la cual puede variar en cada región. También se ha comprobado empíricamente que aunque sí es estacionaria[3] no es independiente del pasado, como puede verse en su función de autocorrelación temporal (figura 8).

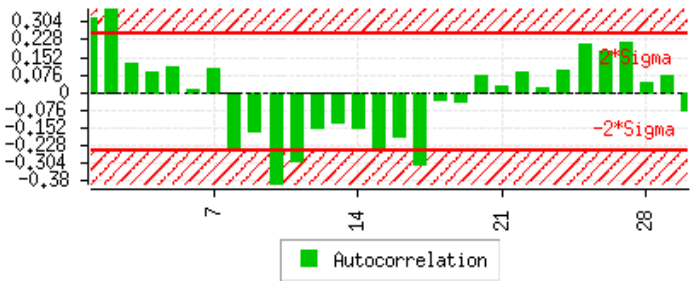

Figura 8: Función de correlación (ACF) de la diferencia entre casos confirmados y su previsión suavizada

[2]Salvo las disrupciones exógenas como las ya explicadas anteriormente

[3]La distribución de cualquier vector de un número prefijado de valores consecutivos es independiente del instante de inicio elegido

Por este motivo se ha desarrollado un modelo de tipo ARMA expresado con la notación de Box-Jenkins [Box and Jenkins(1970)] sobre la discrepancia $w^{\mathbf{c}}_{g,t} = c_{g,t} - \hat{c}_{g,t}$

$$\begin{gathered} w^{\mathbf{c}}_{g,t} - \phi^{\mathbf{c}}_{g,1} w^{\mathbf{c}}_{g,t-1} - \cdots - \phi^{\mathbf{c}}_{g,p} w^{\mathbf{c}}_{g,t-p} = a^{\mathbf{c}}_{g,t} - \theta^{\mathbf{c}}_{g,1} a^{\mathbf{c}}_{g,t-1} - \cdots - \theta^{\mathbf{c}}_{g,q} a^{\mathbf{c}}_{g,t-q} \\ a^{\mathbf{c}}_{g,t} \sim N\left(0, \sigma^2_{\mathbf{c},g}\right) \end{gathered} \tag{30}$$

Esto se puede expresar de forma más compacta utilizando polinomios en el operador de retardo $Bx_t = x_{t-1}$

$$\phi^{\mathbf{c}}_g(B)\left(c_{g,t} - \hat{c}_{g,t}\right) = \theta^{\mathbf{c}}_g(B)\, a^{\mathbf{c}}_{g,t} \tag{31}$$

siendo

$$\begin{aligned} \phi^{\mathbf{c}}_g(B) &= 1 - \phi^{\mathbf{c}}_{g,1} B - \cdots - \phi^{\mathbf{c}}_{g,p} B^p \\ \theta^{\mathbf{c}}_g(B) &= 1 - \theta^{\mathbf{c}}_{g,1} B - \cdots - \theta^{\mathbf{c}}_{g,q} B^q \end{aligned} \tag{32}$$

los polinomios autoregresivo (AR) y de media móvil (MA) respectivamente, los cuales han de tener todas sus raíces fuera del círculo unitario complejo.

Este modelo nos proporciona unas previsiones más ajustadas a los valores observados como se puede observar en la figura 9

$$\check{c}_{g,t} = \hat{c}_{g,t} + \frac{\theta^{\mathbf{c}}_g(B)}{\phi^{\mathbf{c}}_g(B)} \check{a}^{\mathbf{c}}_{g,t} \tag{33}$$

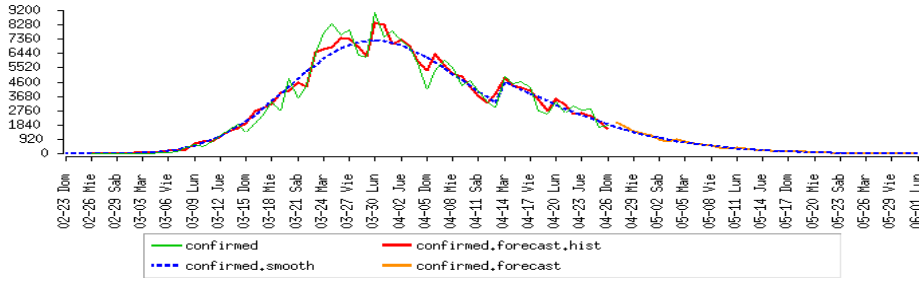

Figura 9: Previsión X-ARMA sobre la curva de contagiados con disrupción

Se trata pues de un modelo anidado que ensambla la salida de un modelo de corte determinista[4] (eq. 29) con un modelo estocástico con estructura dinámica lineal (eq.30) y diferentes grados autoregresivo p y de media móvil q según cada región, los cuáles han sido identificados automáticamente mediante un algoritmo basado en el criterio de información bayesiano y una batería de tests diagnósticos acerca de las autocorrelaciones, normalidad, significación y estacionariedad.

[4] Además de lo ya explicado, esta parte determinista incorpora efectos compensatorios de suma cero para ajustar mejor los efectos contables relativos a los fines de semana en los que por falta de personal administrativo no se contabilizan los casos correctamente hasta el lunes o el martes.

Dada la gran incertidumbre que afecta a todos estos procesos, es inevitable que la suma de las previsiones de confirmados en cada región de forma independiente no coincida exactamente con el total agregado (con $g = 0$), aunque tal y como se observa en la figura 10 tampoco hay una gran discrepancia.

De todos modos es conveniente combinar estadísticamente las previsiones para incrementar la robustez del sistema y que los resultados sean congruentes entre sí.

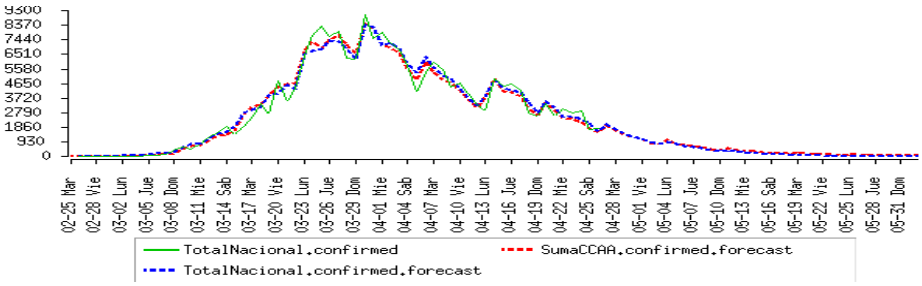

Figura 10: Comparación de la previsión del modelo sobre el total agregado del país (en azul) con la suma de las previsiones de los modelos regionales (en rojo)

El procedimiento seguido es simplemente ponderar cada previsión de forma inversamente proporcional a la varianza de los errores cometidos por cada modelo.

$$c_{g,t} \sim N\left(\check{c}_{g,t}, s^2_{g,t}\right) \tag{34}$$

$$c_{0,t} = \sum_{g=1}^{G} c_{g,t} \tag{35}$$

$$\sum_{g=1}^{G} c_{g,t} \sim N\left(\sum_{g=1}^{G} \check{c}_{g,t}, \sum_{g=1}^{G} s^2_{g,t}\right) \tag{36}$$

$$\tilde{c}_{0,t} = \frac{1}{\left(\sum_{g=1}^{G} s^2_{g,t}\right)^{-1} + s^{-2}_{0,t}} \left(\frac{1}{\sum_{g=1}^{G} s^2_{g,t}} \sum_{g=1}^{G} \check{c}_{g,t} + \frac{1}{s^2_{0,t}} \hat{c}_{o,t} \right) \tag{37}$$

$$\tilde{c}_{g,t} = \check{c}_{g,t} \left(\frac{\tilde{c}_{0,t}}{\sum_{g=1}^{G} \check{c}_{g,t}} \right) \tag{38}$$

7. Modelo de transferencia de defunciones y altas

Todos los fallecidos han sido confirmados en un momento anterior o en el propio día si el caso se confirma en el propio parte de defunción. Así pues, existe una serie de números $0 \leq \gamma^{\mathbf{d}}_{g,k} \leq 1$ a los que llamaremos coeficientes de transferencia de mortalidad, tales que

$$d_{g,t} = \sum_{k=0}^{t} \gamma^{\mathbf{d}}_{g,k} c_{g,t-k} \tag{39}$$

Para los recuperados existe otra serie de números $0 \leq \gamma^{\mathbf{r}}_{g,k} \leq 1$ análoga a la anterior, a los que llamaremos coeficientes de transferencia de recuperación. Teniendo en cuenta que para poder ser dado de alta hay que pasar un periodo de cuarentena mínimo, digamos de Q días, en este caso nos queda que

$$r_{g,t} = \sum_{k=Q}^{t} \gamma^{\mathbf{r}}_{g,k} c_{g,t-k} \tag{40}$$

Una vez que un caso ha sido confirmado sólo son posibles dos situaciones finales, o el paciente es dado de alta o desgraciadamente fallece.

$$\sum_{t=1}^{\infty} c_{g,t} = \sum_{t=1}^{\infty} d_{g,t} + \sum_{t=1}^{\infty} r_{g,t} = \sum_{t=1}^{\infty} \sum_{k=0}^{t} \gamma^{\mathbf{d}}_{g,k} c_{g,t-k} + \sum_{t=q}^{\infty} \sum_{k=Q}^{t} \gamma^{\mathbf{r}}_{g,k} c_{g,t-k} \tag{41}$$

Esto nos impone una restricción unitaria sobre la suma de los coeficientes de transferencia

$$1 = \sum_{k=1}^{\infty} \gamma^{\mathbf{d}}_{g,k} + \sum_{k=Q}^{\infty} \gamma^{\mathbf{r}}_{g,k} \tag{42}$$

No es posible estimar explícitamente estos infinitos coeficientes de transferencia, ni siquiera estableciendo un tiempo máximo tras el cual se puedan considerar ceros. En ningún caso habría suficiente superficie de contraste. Pero sí podemos hacer algunas suposiciones al respecto, como que pueden ser aproximados mediante una función continua con un número relativamente pequeño de parámetros. Una familia particularmente adecuada para representar este tipo de fenómenos son las de funciones de transferencia de Box-Jenkins [Box and Jenkins(1970)] expresadas como un cociente de polinomios de retardo de grado finito del que resulta un polinomio de retardos con grado infinito

$$\gamma(B) = \frac{\omega(B)}{\delta(B)} = \sum_{k=0}^{\infty} \gamma_k B^k \tag{43}$$

Concretamente, en este caso se ha usado la subfamilia con numerador instantáneo y denominador formado por hasta cuatro raíces reales positivas ordenadas $0 \leq \delta_1 \leq \delta_2 \leq \delta_3 \leq \delta_4 < 1$

$$\gamma^{\mathbf{d}}_{g}(B) = \omega^{\mathbf{d}}_{g,0} \prod_{j=1}^{4} \frac{1 - \delta^{\mathbf{d}}_{g,j}}{1 - \delta^{\mathbf{d}}_{g,j} B} \tag{44}$$

$$\gamma_g^{\mathbf{r}}(B) = \omega_{g,0}^{\mathbf{r}} B^q \prod_{j=1}^{4} \frac{1-\delta_{g,j}^{\mathbf{r}}}{1-\delta_{g,j}^{\mathbf{r}} B} \tag{45}$$

que como se puede ver en la figura 11 es una familia muy flexible de curvas

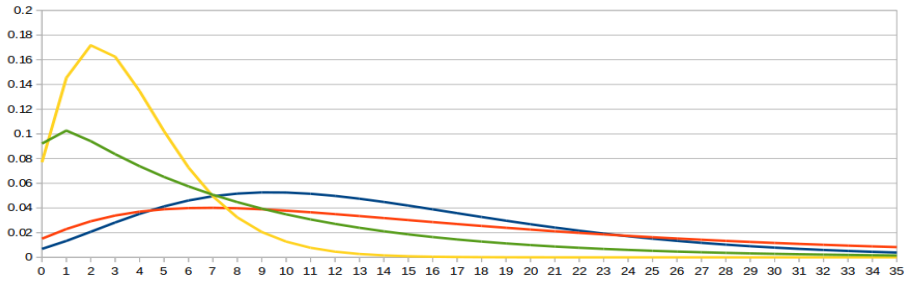

Figura 11: Ejemplos de funciones de transferencia construibles bajo la fórmula 44

Cada uno de los factores de la forma monomio inverso es la suma de una progresión geométrica de retardos

$$d(B) = \frac{1-\delta}{1-\delta B} = (1-\delta)\sum_{k=0}^{\infty} \delta^k B^k \tag{46}$$

que es unitario por construcción ($d(1) = 1$), por lo que la restricción 42 se reduce a

$$1 = \omega_{g,0}^{\mathbf{d}} + \omega_{g,0}^{\mathbf{r}} \tag{47}$$

Una vez estimados los parámetros que mejor ajustan los datos observados teniendo en cuenta la restricción anterior, obtenemos las previsiones suavizadas de las curvas de defunciones y recuperados en función la previsión suavizada de contagios confirmados

$$\hat{d}_{g,t} = \gamma_g^{\mathbf{d}}(B)\, \hat{c}_{g,t-k} \tag{48}$$

$$\hat{r}_{g,t} = \gamma_g^{\mathbf{r}}(B)\, \hat{c}_{g,t-k} \tag{49}$$

Al igual que ocurría con los casos confirmados las discrepancias entre estas curvas suavizadas y los valores observados resultan ser estacionarias pero no independientes en el tiempo por lo que se les aplica sendos modelos ARMA

$$\begin{gathered} \phi_g^{\mathbf{d}}(B)\left(d_{g,t} - \hat{d}_{g,t}\right) = \theta_g^{\mathbf{d}}(B)\, a_{g,t}^{\mathbf{d}} \\ a_{g,t}^{\mathbf{d}} \sim N\left(0, \sigma_{\mathbf{d},g}^2\right) \end{gathered} \tag{50}$$

$$\begin{gathered} \phi_g^{\mathbf{r}}(B)\left(r_{g,t} - \hat{r}_{g,t}\right) = \theta_g^{\mathbf{r}}(B)\, a_{g,t}^{\mathbf{r}} \\ a_{g,t}^{\mathbf{r}} \sim N\left(0, \sigma_{\mathbf{r},g}^2\right) \end{gathered} \tag{51}$$

También cabe utilizar aquí el proceso de combinación de previsiones para obtener agregaciones geográficas congruentes.

El modelo ha resultado finalmente algo más complicado porque hay un brusco descenso en el nivel de fallecidos que no parece guardar relación con la evolución de los casos sino con alguna otra causa exógena al proceso epidemiológico, ya sea de naturaleza contable o por mejoras en los servicios hospitalarios.

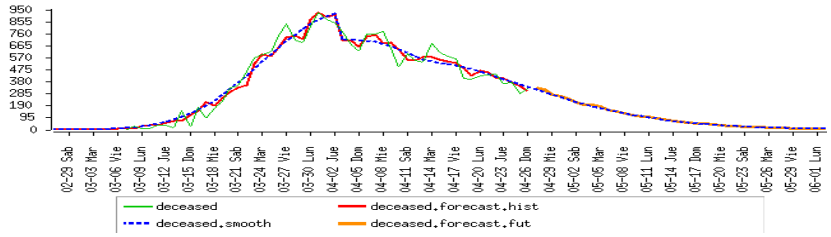

Figura 12: Previsiones de fallecimientos en España

Por otra parte, en la serie de altas se ha observado (ver figura 13) que la función de transferencia ha evolucionado aplanándose a lo largo del tiempo, quizás por la disponibilidad de camas o por el aprendizaje de los equipos sanitarios que han ido alargando cada vez más los periodos de baja.

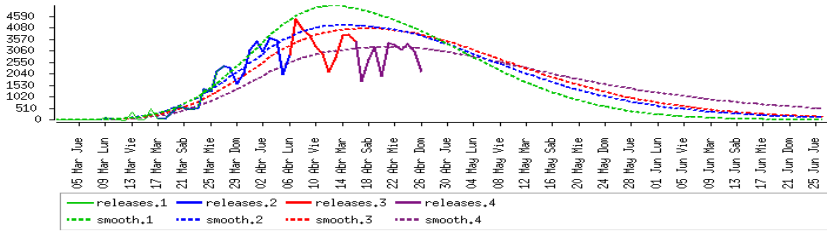

Figura 13: Evolución de la función de transferencia de confirmados a recuperados

La previsión determinista en este caso no es tan suave ni tan ajustada como se esperaría, pero lo cierto es que esta serie contiene una enorme cantidad de ruido. La parte ARMA mejora ostensiblemente la previsión como puede verse en la figura 14.

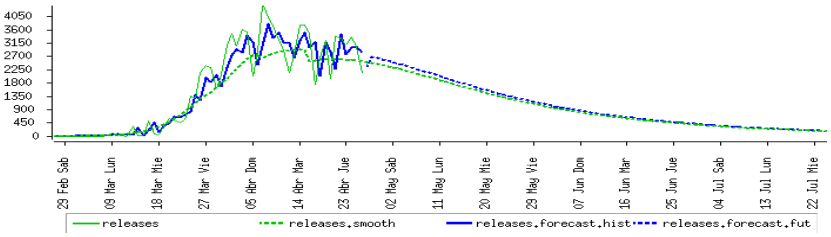

Figura 14: Previsiones de recuperados en España

8. Conclusiones

No cabe duda de que la cooperación entre distintas ramas científicas y tecnológicas es una las claves para el avance de la ciencia y la sociedad en los últimos tiempos.

El sistema de previsión presentado en este artículo ha sido desarrollado con recursos muy limitados por Inverence, un equipo de analistas estadísticos bayesianos expertos en series temporales sin conocimientos específicos de epidemiología, pero aún así, y partiendo de uno de los modelos epidemiólogicos más sencillos, como es la curva de Gompertz, se consigue mejorar ostensiblemente los resultados, como puede verse en la web `https://covid19.inverence.com/`.

El paradigma bayesiano consiste básicamente en incorporar a los modelos estadísticos todo el conocimiento a priori disponible acerca de los fenómenos analizados, y este sistema pretende ser sólo una pequeña muestra de la potencia de este modo de ver la ciencia.

La pregunta que nos hacemos es ¿hasta dónde podría llegar un equipo multidisciplinar altamente experimentado compuesto por epidemiólogos, estadísticos y científicos de datos? Deseamos ferviertemente que no haya que esperar a una nueva pandemia o al rebrote del COVID19 antes de poder responderla.

Referencias

[Box and Cox(1964)] George E.P. Box and David R. Cox. An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 211–252, 1964.

[Box and Jenkins(1970)] George.E.P. Box and Gwilym M. Jenkins. *Time Series Analysis: Forecasting and Control*. Holden-Day, 1970.

[Fred Brauer(2008)] Jianhong Wu Fred Brauer, Pauline van den Driessche. Compartmental models in epidemiology. In *Mathematical Epidemiology*, pages 19–79. Springer Berlin Heidelberg, 2008. URL `https://doi.org/10.1007/978-3-540-78911-6`.

[Gompertz(1825)] Benjamin Gompertz. On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies. 1825. URL `http://www.med.mcgill.ca/epidemiology/hanley/c609/material/Gompertz-1825.pdf`.

[Grassly and Fraser(2008)] Nicholas C. Grassly and Christophe Fraser. Mathematical models of infectious disease transmission. *Nature Reviews Microbiology*, 6(6):477–487, May 2008. URL `https://doi.org/10.1038/nrmicro1845`.

[Ke Wu and Sornette(2020)] Qian Wang Ke Wu, Didier Darcet and Didier Sornette. Generalized logistic growth modeling of the covid-19 outbreak in 29 provinces in china and in the rest of the world. 2020. URL `https://arxiv.org/pdf/2003.05681.pdf`.

[RICHARDS(1959)] F. J. RICHARDS. A flexible growth function for empirical use. *Journal of Experimental Botany*, 10(2):290–301, 1959. URL `https://doi.org/10.1093/jxb/10.2.290`.

[Vynnycky(2016)] Richard G. Vynnycky, Emilia; White. *An introductory book on infectious disease modelling and its applications.* Editorial Mundial, 2016. URL `http://anintroductiontoinfectiousdiseasemodelling.com`.

[Zhao(2020)] Lin Jia. Kewen Li.Yu Jiang. Xin Guo. Ting Zhao. Prediction and analysis of coronavirus disease 2019. 2020. URL `https://arxiv.org/pdf/2003.05447.pdf`.