Modelado y análisis de la evolución de una epidemia vírica mediante filtros de Kalman: el caso del COVID-19 en España

Antonio Gómez Expósito, Catedrático de Ingeniería Eléctrica, Universidad de Sevilla
José A. Rosendo Macías, Catedrático de Ingeniería Eléctrica, Universidad de Sevilla
Miguel A. González Cagigal, Becario FPU, Universidad de Sevilla

Nota: Los resultados mostrados en este documento reflejan los datos disponibles a 24 de marzo de 2020. Se pretende actualizar el documento diariamente en el mismo repositorio (idUS) con datos del día anterior.

Resumen

Este trabajo presenta una metodología original para el tratamiento de los datos reportados de positivos y fallecidos por una epidemia vírica. El objetivo principal es caracterizar la evolución de la progresión del número de infectados reales, y en consecuencia poder predecir en qué momento se alcanzará el pico de la epidemia en un caso de estudio concreto, en este caso la del Covid-19 en España. Los resultados obtenidos muestran claramente el efecto beneficioso de las medidas de confinamiento adoptadas, y prevén que el pico se producirá aproximadamente a finales de marzo o principios de abril.

1. Introducción

A pesar de los espectaculares avances médicos del siglo XX, y la práctica erradicación de enfermedades víricas que en el pasado causaron gran mortandad, como la viruela, las sociedades modernas siguen siendo muy vulnerables a la aparición repentina de nuevos virus, como el llamado Covid-19, para los que no existe vacuna. Por añadidura, la globalización de la economía y el turismo de masas provocan que, una vez iniciado un episodio vírico en una región de un país (en el caso del Covid-19 la región china de Wuhan), éste se extienda casi inevitablemente y de forma veloz al resto del mundo.

En ausencia de un tratamiento eficaz, una vez pasado cierto umbral, el principal y casi único remedio contra la extensión de la enfermedad a toda la población es el confinamiento, cuyo objetivo es reducir al máximo los contactos entre personas, y por tanto la morbilidad. Esto obliga a paralizar buena parte de la actividad económica, con los consiguientes perjuicios para las sociedades afectadas.

Por ello, todos los agentes involucrados (gobiernos, organismos internacionales, instituciones, empresas e individuos) tienen el máximo interés en saber cómo evolucionará en el tiempo el número de afectados y fallecidos, con vistas por un lado a constatar los efectos beneficiosos del confinamiento, y por otro a planificar los ya de por sí saturados recursos sanitarios y tomar las medidas económicas que palíen en la medida de lo posible los devastadores efectos de una epidemia como la del Covid-19.

Los científicos, ingenieros, economistas, sociólogos, etc. disponen de numerosas herramientas matemáticas y estadísticas (recientemente rebautizadas como "data analytics") para el tratamiento y filtrado de series temporales, con vistas a extraer información útil de los datos disponibles, por definición inciertos, tales como tendencias, patrones, periodicidad, valores medios o esperados, varianzas, etc. En el caso concreto de los datos que se reportan ante una epidemia como la del Covid-19, existen básicamente dos categorías de modelos para procesar la información:

1. Modelos que intentan caracterizar la realidad "física" que genera dichos datos. En el caso de una epidemia vírica, estos modelos [1],[2] consideran por ejemplo qué fracción de personas están en el

trabajo, en la enseñanza o en la calle, cuánto tarda en manifestar síntomas una persona contagiada, cuál es la tasa de mortalidad según franjas de edad, etc., etc. Este tipo de modelado es muy utilizado en ingeniería, porque la dinámica de los sistemas o dispositivos con los que se trabaja está generalmente bien caracterizada, mediantes relaciones matemáticas que se obtienen de las leyes físicas que los gobiernan (tal es el caso por ejemplo de las redes eléctricas o de un satélite artificial).

2. Modelos que intentan determinar parámetros o variables explicativas desde un punto de vista puramente matemático (caja negra), sin entrar en cuáles son las causas o interacciones entre componentes que explican los datos resultantes. Dados unos datos inciertos, que entran al sistema regularmente (en nuestro caso, cada día), se trata de caracterizar su evolución temporal ajustando los parámetros de un modelo matemático, de forma que las diferencias entre lo observado y lo estimado se reduzcan al máximo. Dentro de esta categoría pueden considerarse dos variantes:

   2.a) Modelos matemáticos que no prejuzgan cuál será exactamente la forma o curva de la evolución temporal de las magnitudes, sino que utilizan una ecuación de transición de estados, la cual intenta captar la dinámica del problema en cuestión relacionando las variables en un instante en función de las variables en el instante anterior. En este caso, se trata de determinar cómo evolucionan en el tiempo los coeficientes que definen dicha ecuación. En el caso de las epidemias, uno de los modelos más populares quizá sea el conocido como SIR (Susceptibles-Infectados-Recuperados), utilizado por ejemplo en [3] para analizar la evolución del Covid-19 en España[1].

   2.b) Modelos matemáticos basados en que la evolución de afectados, fallecidos, etc. obedece a una curva predeterminada (basada en la experiencia de epidemias anteriores), cuyos coeficientes se estiman en base a la serie temporal de datos reportados. Por ejemplo, la evolución del número acumulado de afectados se puede aproximar satisfactoriamente mediante una curva sigmoide, tal como han supuesto en [4], donde se utiliza la curva propuesta por Gompertz [5].

La metodología utilizada en este trabajo pertenece a la segunda categoría. Como se explica en la siguiente sección, hemos adoptado una variante del modelo SIR que considera que el número de susceptibles de ser infectados es suficientemente grande, y relativamente constante, como para no tener que ser considerado explícitamente en el modelo, sino que puede ser embebido en otros parámetros más significativos del mismo, como la tasa de progresión geométrica de los afectados. Por otro lado, el modelo propuesto distingue expresamente entre personas que han dado positivo en un test, y las personas realmente infectadas, que son muchas más y sobre las que no hay información fiable.

La otra gran diferencia de este trabajo respecto a los demás es el empleo de lo que se conoce como un Filtro de Kalman, para procesar tanto el modelo dinámico supuesto como la información disponible en cada momento. El Filtro de Kalman fue propuesto para sistemas dinámicos lineales a comienzos de los años 60, y está considerado como una de las herramientas fundamentales que permitieron al hombre pisar la luna, al ser empleado con éxito en el guiado de las misiones espaciales del programa Apolo [6]. Este filtro, que constituye una generalización de la técnica conocida como "mínimos cuadrados recursivos", estima la evolución de máxima verosimilitud (o sea, la más probable estadísticamente, de acuerdo a las incertudumbres supuestas) del estado de un sistema dinámico, y puede ser generalizado al caso no lineal así como incorporar la estimación de los propios parámetros del modelo.

[1] En el momento de escribir este documento, los autores de [3] han manifestado que dejan de hacer sus informes diarios al observar un comportamiento anómalo del modelo, debido posiblemente a cambios en el número y forma de hacer los tests. El último informe fue del día 22 de marzo.

2. Metodología e hipótesis

El lector no especialista, o no interesado en los modelos matemáticos, puede ir directamente a la sección de resultados.

Partimos de un modelo SIR, descrito matemáticamente como:

$$S(n+1) = S(n) - \beta \cdot S(n) \frac{I(n)}{n_t} \quad (1)$$

$$I(n+1) = I(n) + \beta \cdot S(n) \frac{I(n)}{n_t} - \gamma \cdot I(n) \quad (2)$$

$$R(n+1) = R(n) + \gamma \cdot I(n) \quad (3)$$

donde n es el tiempo en días, $S(n)$, $I(n)$ y $R(n)$ son las poblaciones susceptible, infectada y recuperada, respectivamente, β y γ son las tasas de transmisión y de recuperación respectivamente, y n_t es la población total de la región estudiada.

En este trabajo se aborda el problema de estimar dichas magnitudes y parámetros a partir de los datos reportados por el Ministerio de Sanidad (disponibles en referencias como [7]), especialmente, casos testeados y dados como positivos, P(n), y casos reportados como fallecidos, F(n), en base a un modelo de progresión de variables ocultas (PVO).

En un primer análisis puede simplificarse la ecuación (2) a una progresión geométrica

$$I(n+1) = \left(1 + \beta \cdot \frac{S(n)}{n_t} - \gamma\right) \cdot I(n) = r(n) \cdot I(n) \quad (4)$$

de razón variable en el tiempo $r(n) = 1 + \beta \cdot \frac{S(n)}{n_t} - \gamma$, que de poder ser estimada haría innecesaria la ecuación (1). Realmente, salvo que el número de infectados sea una fracción significativa del total de la población, el número de susceptibles puede considerarse razonablemente como constante. En todo caso, su impacto quedaría embebido en $r(n)$.

Para poder usar las cifras de fallecidos y de positivos en este problema, se proponen las siguientes relaciones:

$$\Delta F(n) = m(n) \cdot I(n) \quad (5)$$

$$P(n) = a(n) \cdot I(n) \quad (6)$$

donde $\Delta F(n) = F(n) - F(n-1)$ es la variación o incremento en el número de fallecidos, $m(n)$ es una tasa de mortalidad, y $a(n)$ es una tasa que modela la fracción de infectados que son sometidos a test, para considerar la posibilidad real de que haya más infectados que los reportados como positivos, como será claramente el caso de buena parte de la población asintomática.

A partir de (4) y (5):

$$\Delta F(n) = m(n) \cdot I(n) = m(n) \cdot r(n-1) \cdot I(n-1)$$
$$\Delta F(n-1) = m(n-1) \cdot I(n-1)$$

y dividiendo:

$$r(n-1) \cdot K_m(n-1) = r_F(n-1) \quad (7)$$

donde $K_m(n-1) = \frac{m(n)}{m(n-1)}$ es un ratio de variación de la mortandad y $r_F(n-1) = \frac{\Delta F(n)}{\Delta F(n-1)}$ es un ratio de variación de fallecidos.

Análogamente, a partir de (4) y (6):

$P(n) = a(n) \cdot I(n) = a(n) \cdot r(n-1) \cdot I(n-1)$
$P(n-1) = a(n-1) \cdot I(n-1)$

y dividiendo:

$$r(n-1) \cdot K_a(n-1) = r_P(n-1) \quad (8)$$

donde $K_a(n-1) = \frac{a(n)}{a(n-1)}$ es un ratio de variación del coeficiente $a(n)$ y $r_P(n-1) = \frac{P(n)}{P(n-1)}$ es un ratio de evolución de positivos.

Con este enfoque, puede formarse un sistema de ecuaciones en variables de estado que permite estimar mediante filtro de Kalman la secuencia $r(n)$. Este tipo de filtros es capaz de estimar la evolución dinámica tanto de parámetros como de variables de estado, y aunque sólo tiene probada su máxima verosimilitud en problemas lineales, se utiliza con gran éxito en problemas no lineales, como el que nos ocupa. En este caso, el vector de estado lo constituyen las variables $r(n)$, $K_m(n)$ y $K_a(n)$, que pueden ser estimadas a partir del vector de pseudomedidas $[r_F(n-1) \quad r_P(n-1)]^t$. Así, el modelo para el filtro del Kalman quedaría:

Ecuación de estado:

$$x(n) = [r(n-1) \quad 1 \quad 1]^t + w(n)$$

siendo $x = [r \quad K_m \quad K_a]^t$ el vector de estado, y $w = [w_r \quad w_m \quad w_a]^t$ un vector de ruidos gaussianos que modela el error de modelo y que tiene matriz de covarianzas Q.

Ecuación de medida:

$$h(n) = [r(n) \cdot K_m(n) \quad r(n) \cdot K_a(n)]^t + v(n)$$

siendo $v(n)$ un vector de ruidos gausianos que modela el error de medida y que tiene matriz de covarianzas R. En este estudio las covarianzas adoptan los siguientes valores: Q=diag(0.1, 0.01, 0.01) y R=diag(20, 1).

Iterando con el filtro de Kalman puede encontrarse la secuencia de estados $x(n)$, cuya primera componente es la estimación de máxima verosimilitud de $r(n)$, más fiable que ratios de datos en bruto, como $r_F(n)$ o $r_P(n)$.

Una vez obtenida la estimación de la secuencia $r(n)$, pueden estimarse adicionalmente dos magnitudes de gran interés:

1) El día del futuro en que se alcanzará el pico de infección, n_p, que ocurrirá cuando $r(n_p) = 1$, y que supone que la población infectada deja de crecer. Para ello, la secuencia $r(n)$ debe ser ajustada, por ejemplo mediante un ajuste polinomial que permita predecir su comportamiento futuro a partir de su historia. Es importante destacar en este punto que estamos hablando del pico de infectados, siendo el pico de positivos un *proxi* del mismo.
2) La población infectada, $I(n)$. Dicha magnitud puede expresarse a partir de los infectados en un día previo, p.e., el día 0 del estudio, así

$$I(n) = \prod_{i=0}^{n-1} r(i) \cdot I(0)$$

Nótese que, para este cálculo es necesario disponer con precisión de la cantidad de infectados el primer día en que empieza a hacerse el análisis, $I(0)$, lo cual no es realista. Para nuestros resultados, adoptaremos la hipótesis conservadora $I(0) = P(0)$, que por tanto sólo proporcionará una cota mínima a la estimación de infectados. Si la realidad de infectados fuese un factor k de los positivos, $I(0) = k \cdot P(0)$, entonces la cantidad real de infectados crecerá con dicho factor: $k \cdot I(n)$.

3. Resultados y Discusión

Con el modelo presentado anteriormente y los datos obtenidos de [7] se ha utilizado el estimador dinámico de estado basado en filtros de Kalman para identificar la evolución de la secuencia $r(n)$, a la cual se ha aplicado en paralelo un ajuste polinómico de segundo orden con el fin de extrapolar la información disponible y poder predecir su evolución futura.

En primer lugar, se comparan los resultados del proceso de estimación de $r(n)$ con los valores de los ratios de variación $r_F(n)$ y $r_P(n)$ presentados en las ecuaciones (7) y (8) respectivamente, obtenidos a partir de los datos publicados diariamente de fallecidos y positivos. La Figura 1 muestra dicha comparación, donde puede apreciarse claramente que el filtro de Kalman (línea zaul) es capaz de filtrar de forma eficiente las grandes oscilaciones que pueden apreciarse en las ratios obtenidas a partir de datos *crudos*, permitiendo así realizar un mejor ajuste para la extrapolación.

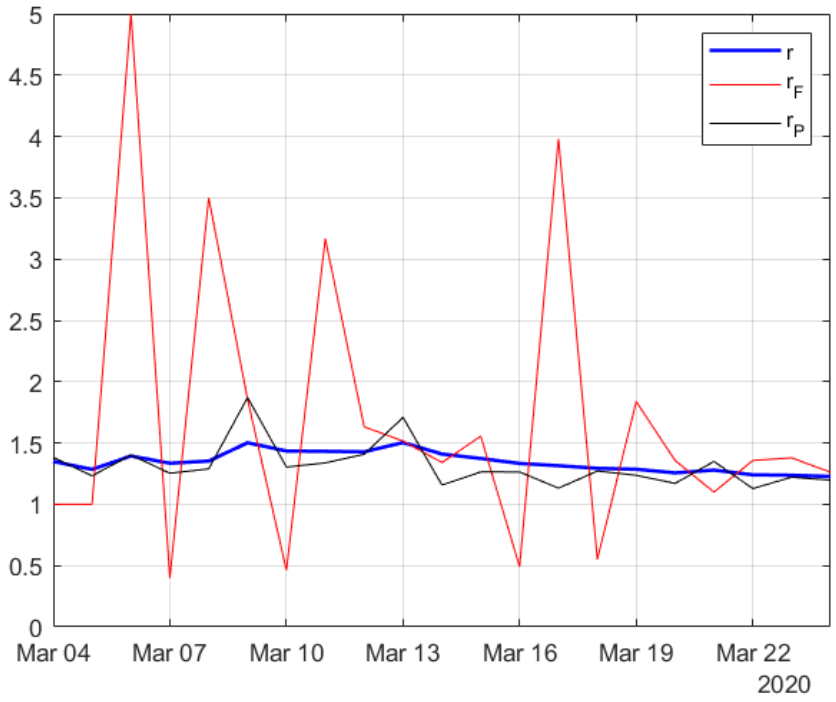

Figura 1. Comparación de las tasas de crecimiento obtenidas con datos publicados y mediante filtro de Kalman

En la figura 2 vuelve a mostrarse los resultados de la estimación de $r(n)$, incluyendo esta vez una banda de incertidumbre de $\pm 3\sigma$, junto con la predicción para los próximos 8 días. Al tratarse de una progresión geométrica, el máximo de infectados se dará cuando la razón $r(n)$ valga 1, que a la vista de los resultados mostrados ocurrirá previsiblemente entre el 30 de marzo y el 1 de abril, si se mantienen las condiciones actuales. No se olvide que el valor estimado tiene una incertidumbre delimitada por las líneas rojas. Si $r(n)$ se acercase a los valores superiores de la banda, entonces el pico se produciría más tarde. Y lo contrario si el valor real de $r(n)$ estuviese en la parte inferior de la banda.

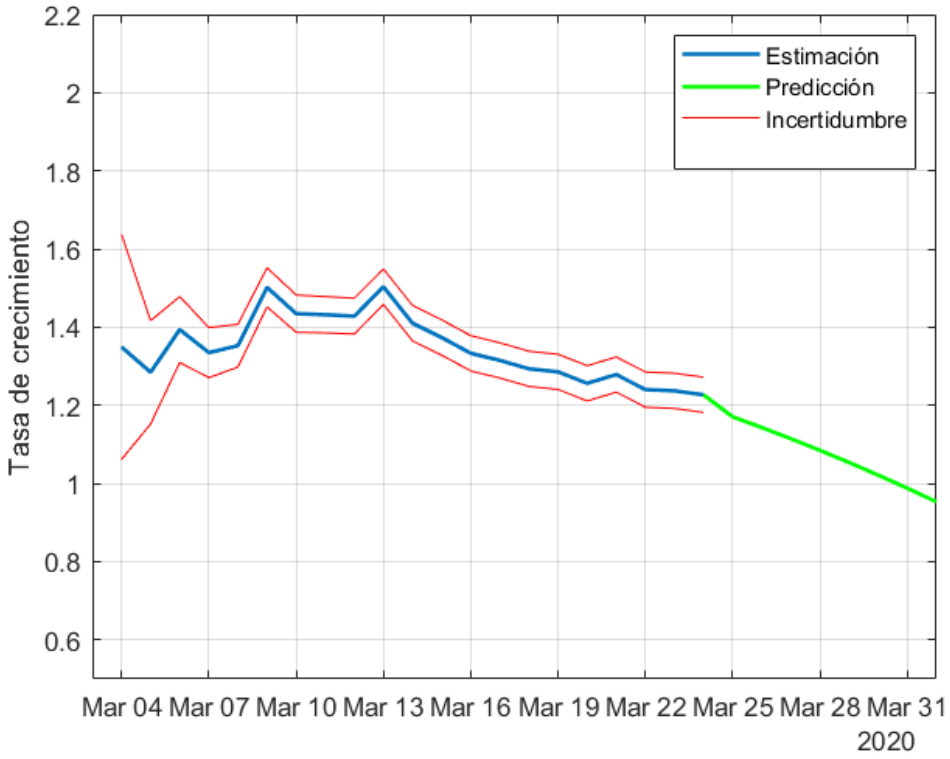

Figura 2. Estimación y predicción de la tasa de crecimiento

En la figura 2 puede apreciarse claramente que $r(n)$ comenzó a decrecer de forma sostenida, desde picos máximos de 1,5, a partir de la fecha en que se tomaron las medidas de confinamiento (14 de marzo). De no haberse tomado ninguna medida, y caso de haberse mantenido la $r(n)$ en el entorno de 1,5, la mayor parte de la población se habría visto afectada y el pico se habría alcanzado entre 7 y 10 días después. Recíprocamente, si el aislamiento de la población se hubiese implantado antes, el pico habría sido menor y se hubiera producido antes.

Con estos valores, suponiendo, como ya se ha comentado, un número de infectados inicial igual al de positivos reportados, se puede obtener la predicción de la cota inferior para los infectados reales, tal y como se muestra en la Figura 3, donde se incluye así mismo los valores correspondientes a los próximos 8 días.

Se ha incluido también en la representación el número de positivos reportados cada día, pudiendo apreciarse de esta manera la diferencia entre éstos y los afectados reales. Se puede observar que el máximo número de personas afectadas se aproxima a los 160.000 y que este valor se alcanza a fecha del 30 de Marzo, donde se predice el pico de la pandemia de acuerdo al modelo presentado en este informe. Es importante volver a

remarcar que el valor mostrado, tanto en la etapa de estimación como en la de predicción, de infectados por el Covid-19 se corresponde con una cota inferior de los mismos, al haber supuesto que el número de positivos coincide con el de afectados al iniciar las simulaciones. De esta manera, en caso de suponer que el número de afectados en el origen de tiempos era 5 veces el número de positivos, el máximo esperado estaría por encima de los 800.000 (cifra del mismo orden que la estimada en [4] y bastante inferior a las primeras estimaciones de [3]). De forma genérica, siendo k el ratio entre el número de personas infectadas respecto a reportadas en el inicio, se establecería una predicción correspondiente a $160.000 \cdot k$ en el pico de la curva de la Figura 3.

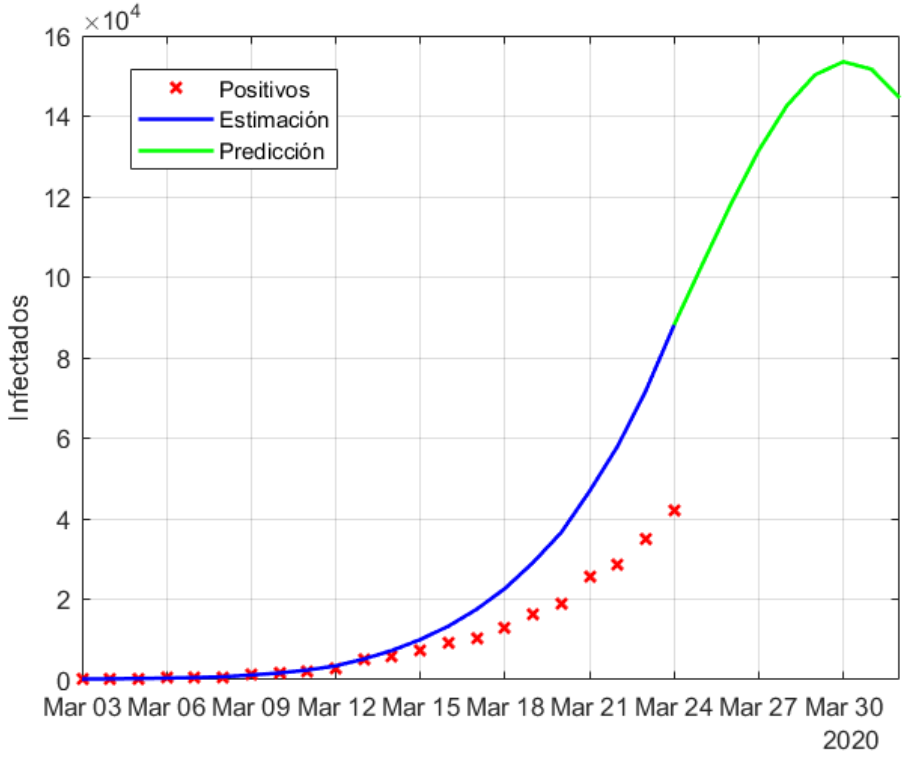

Figura 3. Estimación y predicción de infectados

A continuación, para la tasa $a(n)$ definida en (6), es preciso evaluar el cociente entre los positivos reportados y los infectados estimados mediante filtro de Kalman, suponiendo que estos valores se pueden ajustar a un polinomio cuadrático, que permite establecer una predicción del número de positivos reportados en los 8 próximos días. Este resultado se muestra en la Figura 4, donde se incluyen tanto los datos de positivos y fallecidos como la predicción de positivos. No nos ha parecido pertinente mostrar la predicción de fallecidos, que al ser un dato acumulativo no mostraría un pico, sino una tendencia asintótica hacia un valor constante (curva sigmoidal).

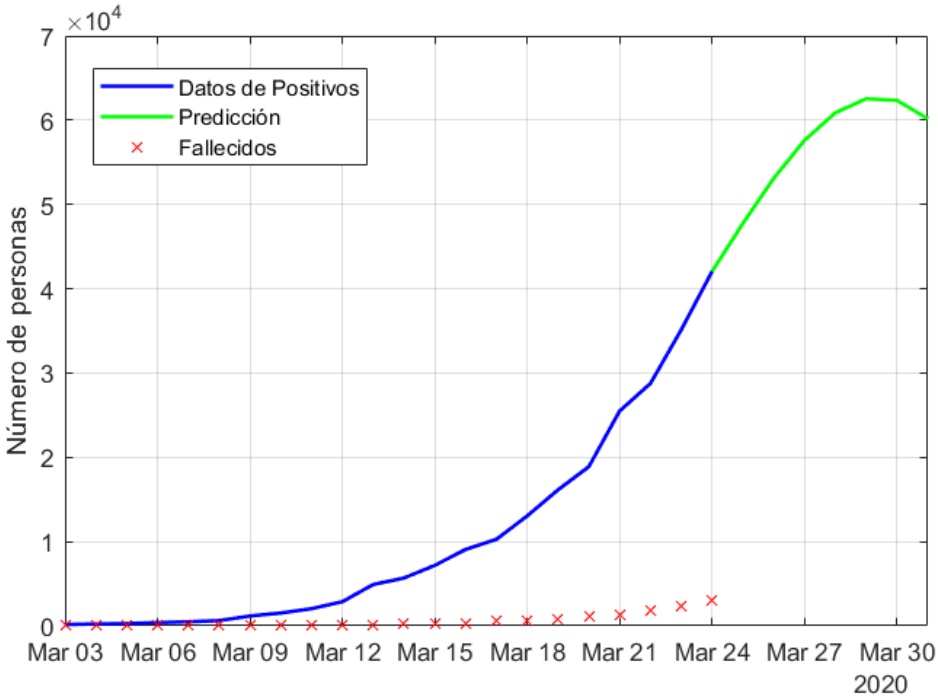

Figura 4. Predicción de los positivos reportados

Para resumir estos resultados, se muestran en la tabla siguiente las predicciones de positivos en los próximos días y las de infectados para $k = 3$.

Fecha	Predicción de los positivos reportados	Predicción de infectados ($k = 3$)
25 de Marzo	47755	309433
26 de Marzo	53101	353775
27 de Marzo	57611	394352
28 de Marzo	60878	427820
29 de Marzo	62544	450844
30 de Marzo	62348	460554
31 de Marzo	60176	455043
1 de Abril	56103	433797

Puede comprobarse que la predicción que ofrece este modelo resulta en un número máximo de infectados muy inferior a los que inicialmente se reportaban con el modelo SIR empleado en la referencia [3], que sobrepasaba los 2,5 millones de personas.

4. Conclusiones

En este trabajo se ha presentado una metodología capaz de aprovechar los datos oficiales de positivos y fallecidos para hacer una estimación de la progresión real de infectados, y en consecuencia un pronóstico de cuándo se alcanzará el pico de la epidemia y a cuánta gente como mínimo afectará. El modelo utilizado es una variante original del conocido como modelo SIR, que no ofrece buenos resultados cuando el número de afectados es muy pequeño en comparación con la población total, y la herramienta adoptada para el tratamiento

de datos es un filtro de Kalman no lineal (conocido como Unscented Kalman Filter), junto a ajustes polinómicos de las variables estimadas por el filtro.

Los resultados muestran una tendencia de decrecimiento sostenido de la constante de la progresión, que desde valores máximos de 1,5 se acerca casi linealmente a 1, lo cual ocurrirá previsiblemente antes del 2 de abril. En el pico, el número de positivos reportados estará en torno a los 55.000, salvo que el número de tests realizados aumente respecto a lo que se venía haciendo hasta ahora, como anuncian las autoridades. Consideramos conveniente no dar una cifra concreta de fallecidos, pero en todo caso estimamos que estará entre dos y tres veces la cifra alcanzada en China.

Como en todo estudio de estas características, basado en datos tan volátiles e inciertos como el número de positivos reportados, los resultados deben tomarse con cautela. Es intención de los autores repetir el análisis con datos diarios, y actualizar los resultados y conclusiones en caso necesario. En todo caso, estando ya tan relativamente cerca el pico de la epidemia, no se prevén cambios drásticos en las tendencias de las variables, salvo que se tomen medias más extremas de aislamiento, que posiblemente ya no se justifiquen. Lo que sí resulta imprescindible es mantener como mínimo el confinamiento actual.

Referencias

[1] T. Pueyo, “Coronavirus: Why you must act now”, en: https://medium.com/@tomaspueyo/coronavirus-act-today-or-people-will-die-f4d3d9cd99ca

[2] Imperial College COVID-19 response team, “Impact of non-pharmaceutical interventions (NPIs) to reduce COVID19 mortality and healthcare demand”, en: https://www.imperial.ac.uk/media/imperial-college/medicine/sph/ide/gida-fellowships/Imperial-College-COVID19-NPI-modelling-16-03-2020.pdf?referringSource=articleShare

[3] Universitat Politècnica de València, “Modelización epidemiológica del Covid-19”, en: http://covid19.webs.upv.es/

[4] Computational biology and complex systems (BIOCOMSC), UPC, “Analysis and prediction of COVID-19 for different regions and countries” en: https://biocomsc.upc.edu/en/covid-19/daily-report

[5] Madden LV. Quantification of disease progression. Protection Ecology 1980; 2: 159-176.

[6] D. Simon, "Optimal State Estimation: Kalman, H In_nity and Nonlinear Approaches". ISBN: 13978-0-471-70858-2

[7] https://www.worldometers.info/coronavirus/country/spain/