

Estimación y previsión de fallecidos por Covid-19 en diferentes países

Inverence

26 de abril de 2020

1. Introducción

La determinación del número de muertes causadas por Covid-19 y su previsión, incluso en el mismo curso de la epidemia, es el indicador más importante y duro de la severidad de ésta en cada país. Es por ello que en algunos casos los gobiernos tienden a sentirse incómodos con fallecimientos elevados y, por idénticas razones, los partidos opositores ponen el énfasis en la calidad de la información proporcionada por las autoridades sanitarias.

El número de fallecidos por Covid-19 en cierto país es una cuestión de hecho y debe ser medido de la manera más objetiva posible. Sin embargo, ésta tarea no será sencilla, ni siquiera cuando la pandemia haya acabado. Mucho más difícil, por tanto, cuando la epidemia está en proceso.

Este documento presenta la forma en la que Inverence estima el número de muertos en cada momento del desarrollo de la epidemia en diferentes países.

2. Definición de muertes incrementales por Covid-19

El objetivo de nuestro trabajo consiste en medir las muertes incrementales debidas a la epidemia de Covid-19 en cada país. Las muertes incrementales están dadas por la diferencia entre las muertes observadas en los registros administrativos y el nivel normal de mortalidad, esperado en ausencia de la epidemia, en el periodo de duración de esta. Este nivel normal es, de hecho, una previsión del número de muertes que posee cierto grado de error como después se verá. Ello implica que toda estimación de muertes incrementales es probabilística, lo que en modo alguno reduce su valor.

La medida de la mortalidad incremental elimina la ambigüedad inherente a otras definiciones (muertos por Covid-19, o con Covid-19) y es operacionalmente más adecuada que medir el número de muertos restringiéndolos a los diagnosticados mediante un test. Esta última definición deja fuera a todas las personas fallecidas sin test positivo por Covid-19 y a las que no se les ha hecho autopsia y depende, por tanto, del número de tests/autopsias realizados/as. Debe

enfatzarse que todos los procedimientos de conteo directo generan problemas de medida, incluyendo posibles contabilidades basadas en criterios facultativos u otras que han sido propuestas, como el uso de datos provenientes de empresas funerarias.

Las muertes incrementales contienen muertes indirectas tales como el aumento del número de suicidios generados por la situación de estrés causada bien por la propia epidemia o por las medidas de confinamiento adoptadas por los gobiernos; el presumible aumento de muertes derivadas de violencia en el hogar, que incluye violencia de género, parental y de otras categorías; incluyen también el incremento de muertes de personas que requieren tratamiento de urgencia y no lo han tenido debido a la congestión hospitalaria.

Estos incrementos indirectos pueden ser también negativos —menor número de muertes— como las que se derivan de menores problemas respiratorios debido a la disminución de la contaminación, menores muertes por accidentes de automóvil y derivadas de accidentes laborales.

En conjunto, todos estos efectos indirectos son relativamente pequeños en relación a las muertes directas por Covid-19 y cabe esperar, por tanto, que las muertes incrementales aproximen bien las muertes directas por Covid-19. Pero, en cualquier caso, las muertes incrementales son de interés en sí mismas ya que describen los efectos a corto plazo de la epidemia en cada país.

Ciertamente, otros incrementos futuros de muertes causados por la interrupción de tratamientos y, sobre todo, por el retraso en diagnósticos que son esenciales para prevenir muertes futuras pueden existir también y no pueden calibrarse en el estudio de las muertes a corto plazo. Naturalmente, los efectos directos e indirectos de la epidemia en el largo plazo pueden ser importantes ya que incluso los hábitos sociales y la propia estructura económica se verán modificados por ella. El estudio epidemiológico de estos efectos de largo plazo queda fuera del ámbito de este documento técnico, pero podrá ser abordado en el futuro mediante generalizaciones de las líneas que aquí se proponen y otros aportes.

3. Terminología

Existen, fundamentalmente, dos tipos de fuentes de datos de la mortalidad Covid-19:

- 1) El conteo directo realizado por las autoridades sanitarias en cada país, que en algunos casos ha dado en llamarse *cifra oficial* y
- 2) Los registros administrativos de fallecimientos que se poseen a efectos civiles (registros civiles en España).

En lo que resta del presente documento, la primera fuente será indicada mediante la clave *conteo directo*; así nos referiremos a muertes o fallecidos según el conteo directo, o datos de conteo directo o expresiones similares. Es cierto que pueden existir fuentes de conteo directo distintas a los recuentos de las autoridades sanitarias pero, por razones de simplicidad, esas otras fuentes no son consideradas en este documento y por tanto podemos reservar el término

conteo directo para hacer referencia a esta fuente.

La segunda fuente está siendo estructurada de acuerdo a un proyecto europeo llamado EuroMoMo. En España el Instituto de la Salud Carlos III recoge los fallecimientos de 3820 municipios que suponen el 93 % de la población. Por tanto, la mortalidad indirecta por Covid-19 se infravalora levemente según estas cifras pero el hecho de que los 4291 municipios restantes contengan tan sólo el 7 % de la población indica que se trata de municipios rurales en los que la epidemia ha penetrado mucho menos. Por tanto, el sesgo existente es probablemente de un orden de magnitud menor que la población no representada. Esta fuente será identificada en los que sigue como *datos MoMo*, *fallecidos MoMo* o expresiones similares que contendrán siempre el término *MoMo*.

4. La medida de las muertes incrementales

El principal problema para medir las muertes incrementales es que las cifras MoMo se actualizan con retraso respecto al momento en el que se producen los fallecimientos. En el caso de España la consolidación completa de las cifras no ocurre hasta después de cuatro semanas, aunque se dispone de actualizaciones parciales. En otros países este lapso de tiempo es aún mayor: en Bélgica no existen ni siquiera datos parciales de fallecimientos MoMo en las últimas tres semanas.

Por tanto, la medición del incremento de mortalidad por todas las causas en el momento mismo del desarrollo de la epidemia implica dos tipos de estimación independiente:

- 1) La de mortalidad esperada y su distribución de probabilidad en el periodo de estimación y
- 2) La de mortalidad real diaria (o semanal, de acuerdo a la periodicidad base de los datos) que será observada en el futuro, tomando en consideración la experiencia histórica de actualización y la información de contorno disponible, como son las cifras de conteo directo.

La primera tarea es relativamente simple, especialmente si se dispone de la mortalidad esperada a largo plazo, como es el caso de España. En el caso del estado de Nueva York, Irlanda, Inglaterra o Escocia no disponemos de dicha estimación. Pero aún en estos casos la estimación de la mortalidad esperada es relativamente simple.

Más complicado es el problema de prever la mortalidad que será observada en el futuro cuando la información tiene un grado de actualización limitado y posiblemente irregular a lo largo del tiempo.

Trataremos cada uno de los problemas separadamente.

5. La mortalidad esperada y su distribución de probabilidad

Vamos a suponer en esta sección que poseemos una serie temporal que proporciona la estimación de largo plazo de la mortalidad esperada por todas las causas en cada momento temporal. Sea dicha serie temporal m_t .

La mortalidad observada por todas las causas, MoMo, se observa a través de la serie $d_{i,t}$ con $i > t$, que no es otra cosa que la observación de la mortalidad MoMo del día t en la fecha, naturalmente posterior, i . Denotamos la mortalidad real como d_t y en general para $i - t$ suficientemente grande $d_{i,t} = d_t$ mientras que para valores moderados o pequeños de $i - t$ tenemos que $E(d_{i,t} - d_t) < 0$.

La serie temporal $v_t = d_t - m_t$ que registra la diferencia entre muertes observadas y muertes esperadas deambula alrededor de 0, como puede verse en el gráfico 1, referido a España en el periodo pre-Covid. Este comportamiento es consistente con el hecho de que m_t representa la estacionalidad anual y la evolución tendencial de la mortalidad, es decir el componente de bajas frecuencias de d_t . A corto plazo, existen circunstancias que modifican dicha mortalidad de forma autocorrelacionada, sea por variabilidad metereológica o por otras causas no identificadas y esto es lo que representa el comportamiento deambulante de v_t .

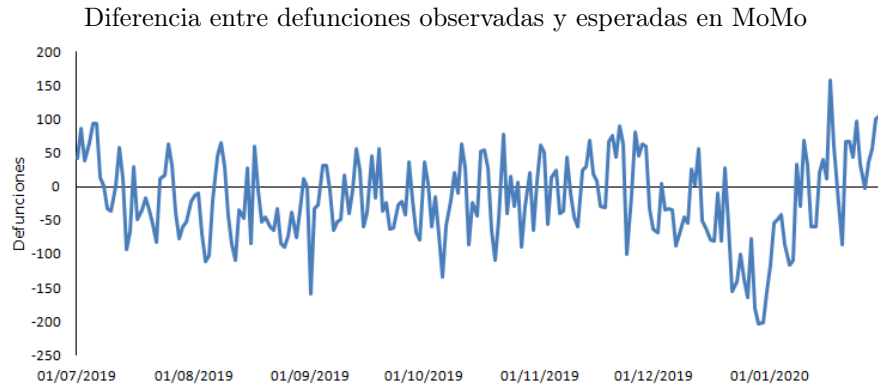


Figura 1: Se muestra la serie temporal $v_t = d_t - m_t$ que contiene la diferencia entre defunciones observadas y defunciones esperadas en MoMo, en un periodo previo al inicio de la epidemia (de julio 2019 a enero 2020). Puede observarse su comportamiento deambulante alrededor del valor 0, con claras muestras de autocorrelación.

La serie v_t es representable mediante un proceso ARMA estacionario. Y por tanto, podemos prever d_t como la suma de m_t y v_t . Dadas las características de ambas series, en horizontes largos la previsión de d_t tiende a ser idéntica a m_t , mientras que con pocos días por delante la previsión de v_t será en general diferente de cero, y por tanto la previsión en horizontes cortos tenderá a ser ligeramente distinta de m_t . Este modelo, además, proporciona la estructura de

varianzas y covarianzas del error de previsión pre-Covid. Las muertes incrementales son una traslación sobre dicha línea base y tienen, por tanto, la misma distribución de probabilidad conjunta que v_t .

Supongamos, por un momento, que observamos las muertes reales por todas las causas en cierto periodo, desde el comienzo de la epidemia hasta cierto punto temporal y que el número de días de esa observación perfecta es h . Supongamos también que podemos establecer el momento, s , previo al que se produce la primera muerte por Covid-19. En tal caso, el incremento diario de muertes en el periodo $[s + 1, s + h]$ es el vector probabilístico $[\Delta_{s+j}] = [d_{s+j} - m_{s+j} - v_{s+j}]$, con $1 < j \leq h$, y valor esperado $E[\Delta_{s+j}] = [d_{s+j} - m_{s+j} - \hat{v}_{s+j|s}]$ (donde $\hat{v}_{s+j|s}$ indica la previsión para el día $s + j$ condicional a la información existente respecto a v_t hasta el día s). Puesto que v_t tiene una representación ARMA, su previsión está definida, así como la matriz de covarianzas de sus errores. Por tanto, condicional al conocimiento de d_{s+j} en el periodo $[s + 1, s + h]$, conocemos el valor incremental esperado de las muertes causadas por Covid-19, $\sum_{j=1,h} d_{s+j} - m_{s+j} - \hat{v}_{s+j|s}$ y la varianza de su error.

6. Observación consolidada, estimación y previsión

El principal problema para la estimación de los efectos sobre la mortalidad de Covid-19 es que los datos MoMo tienen cierta estructura de retardo en su actualización. El gráfico 2 muestra dicha actualización en la Comunidad de Madrid en tres momentos diferentes del tiempo. Se aprecia que nuevas actualizaciones del número de fallecimientos presentan mayor número de éstos para las fechas recientes mientras que para fechas lejanas la situación es mucho más estable. De hecho, en España los investigadores del Instituto de la Salud Carlos III (ISCIII) proporcionan ciertas fechas de corte para las que los datos MoMo, anteriores a tales fechas, pueden considerarse consolidados, para cada comunidad autónoma de dicho país. Por tanto supondremos que el periodo en el que se registra mortalidad por Covid-19 está determinado y que los datos MoMo pueden ser considerados como consolidados antes de cierta fecha. Ello puede introducir pequeños sesgos en el cálculo, pero, una vez más, tales sesgos no son sustanciales.

Así pues, desde el punto de vista de la medida de la mortalidad, la epidemia Covid-19 puede dividirse en tres etapas: el periodo de observación consolidada, el periodo intermedio que llega hasta el momento presente, en el que sólo tenemos datos MoMo no consolidados junto a datos de conteo directo, que tienen las limitaciones expresadas en la sección 2, y el periodo futuro hasta el fin de la epidemia, o al menos del presente ciclo de la epidemia.

En el periodo de observación nuestra incertidumbre puede expresarse del modo descrito en la sección 5 de este documento.

Para medir la mortalidad en el segundo periodo, que hemos llamado periodo de estimación ya que los datos no se observan directamente, requerimos un

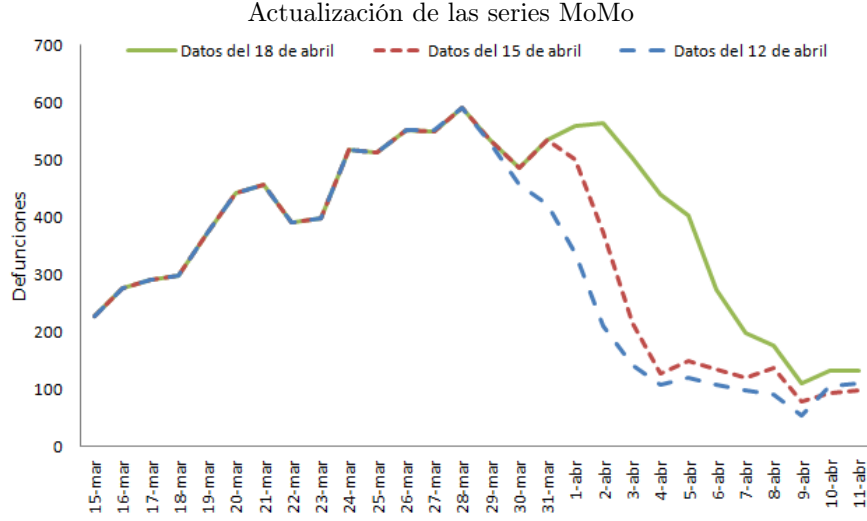


Figura 2: Se muestra la serie de defunciones diarias registradas en MoMo para la Comunidad de Madrid, publicadas en 3 fechas distintas. Se aprecia cómo los datos de las últimas 2 semanas sufren actualizaciones parciales, mientras que los datos más antiguos (consolidados) se mantienen invariantes.

modelo que ligue los datos MoMo y los datos de conteo directo. Estos últimos, aún con sus limitaciones son, por tanto, esenciales y deberían poseer el mayor grado de homogeneidad histórica, algo que no siempre ocurre. Tal importancia no elimina el hecho de que son sesgados por definición y no pueden ser tomados, por tanto, como medida de la mortalidad real.

La vinculación entre datos MoMo y de conteo directo puede hacerse a través del modelo de regresión

$$E(\Delta_t) = \alpha f_t + e_t \quad (1)$$

dónde f_t es el dato de conteo directo en el momento t relativo a la mortalidad por Covid, $E(\cdot)$ es el operador esperanza, Δ_t es la variable aleatoria muerte incremental en el periodo t , y e_t es la serie de error,

Es importante observar que del análisis efectuado en la sección 5, se deriva que la matriz de covarianzas del error es $\Sigma_h + \sigma^2(u_t)I_h$, dónde

I_h es la matriz unidad de orden h , $\sigma^2(u_t)$ es la varianza del error, no observado, que la regresión (1) tendría si la variable dependiente fuese exactamente Δ_t y Σ_h es la matriz de covarianzas de orden h de los errores de previsión, de 1 hasta h periodos por delante del modelo ARMA de v_t .

Es relativamente simple estimar el modelo (1) mediante MCMC e incluso mediante un Gibbs Sampler (un caso particular de MCMC) en el que cada paso de muestreo embute una estimación GLS condicional al parámetro $\sigma^2(u_t)$ de la iteración previa.

El modelo (1) puede criticarse sobre la base de la falta de constancia del parámetro α . Es natural pensar que en los primeros momentos de la epidemia, con un nivel de información más bajo que en los momentos posteriores y sobre todo cuando el sistema sanitario se somete a un mayor nivel de tensión, podemos encontrar un parámetro α mayor que en periodos posteriores. Esto es fundamental a la hora de estimar la mortalidad en el periodo de estimación pero también en las proyecciones que deben hacerse fuera de este periodo. Concretamente (1) podría ser sustituido por

$$E(\Delta_t) = \alpha_t f_t + e_t \quad (2)$$

dónde α_t toma niveles distintos en distintos periodos. Es difícil prescribir de forma general una especificación para α_t que tenga sentido en cada situación. Una posibilidad simple que los métodos bayesianos permiten es especificar dos regiones temporales. La primera está formada por el periodo que va desde el comienzo de la epidemia hasta el momento en el que las unidades de cuidados intensivos alcanzan su máximo. El segundo periodo correspondería a la parte última de la epidemia. Pueden suponerse, por tanto, dos constantes, una para cada periodo con la restricción de que $\alpha_1 > \alpha_2 > 1$.

No es necesario restringir el número de regímenes en la relación datos de conteo directo-MoMo a 2. A costa de introducir más información a priori podemos ampliar ese número a más periodos. Algo que puede hacerse en esta situación es exigir que las variaciones entre parámetros sean pequeñas, lo que puede parametrizarse haciendo que dicha diferencia siga una distribución de probabilidad normal de media cero y varianza relativamente pequeña, truncada precisamente en cero. Aún más, las segundas diferencias entre parámetros que deben ser positivas si la curva que describen los parámetros α_t es decreciente y convexa, pueden parametrizarse de un modo semejante.

Otra alternativa que también debe considerarse en estas situaciones es la introducción directa de una función determinista de amortiguamiento dentro de cierta familia, seleccionando ulteriormente el elemento de la familia que proporciona mejor ajuste.

En cualquier caso dichas especificaciones deben asegurar resultados con sentido epidemiológico y práctico, lo que no puede hacerse si no introducimos información a priori en la estimación.

7. Previsión

El tercer periodo, que corresponde a la previsión de fallecimientos en lo que resta de la epidemia, es el más difícil. La previsión del número de fallecimientos requiere de dos cosas:

- a) Un modelo sobre la evolución futura de f_t y
- b) Una hipótesis sobre la evolución de α_t que bien puede ser la constancia del último valor estimado.

En una entrega posterior describiremos el modelo con el que hemos estado previendo los fallecimientos por conteo directo f_t . Solo añadiremos aquí que

nuestro modelo se concibió como un modelo adaptativo, lo que le ha permitido sobrevivir más que otros modelos más rígidos en una situación de baja calidad de datos que a su vez han sufrido numerosas roturas, difíciles incluso de rastrear.