

Informe breve

30-06-2020

## Análisis de los retrasos en la actualización de las series históricas de casos en España

Martí Català, Sergio Alonso, Enric Álvarez, Daniel López, Miquel Marchena,

David Conesa, Pere-Joan Cardona, Clara Prats

*Comparative Medicine and Bioimage Centre of Catalonia; Institute for Health Science Research Germans Trias i Pujol  
Computational Biology and Complex Systems; Universitat Politècnica de Catalunya - BarcelonaTech*

**Con la colaboración de:** Guillem Álvarez, Oriol Bertomeu, Laura Dot, Lavínia Hriscu,

Helena Kirchner, Miquel Marchena, Daniel Molinuevo, Pablo Palacios, Sergi Pradas,

David Rovira, Xavier Simó, Tomás Urdiales

---

Contacto: clara.prats@upc.edu

With the financial support of:



## Introducción

Este informe breve tiene como objetivo evaluar los retrasos presentes en las actualizaciones de las series históricas publicadas por el ISCiii<sup>1</sup>. Desde el 10 de junio de 2020, dicho organismo publica periódicamente una revisión actualizada de las series históricas de casos de Covid-19 para las provincias y comunidades autónomas españolas. En estas series, la **fecha de referencia para los casos es la de aparición de los síntomas**. En los casos asintomáticos, se asignan a 6 días antes de la fecha de diagnóstico. Es de esperar que este criterio conduzca a un **retraso aparente** en las notificaciones con respecto a los datos publicados con la fecha de diagnóstico o con la propia fecha de notificación. El motivo es que **los casos que han empezado a mostrar síntomas en los últimos días probablemente aún no hayan sido diagnosticados, registrados o validados debido a los retrasos propios del proceso**. Dichos casos serán asignados al día correspondiente en actualizaciones posteriores.

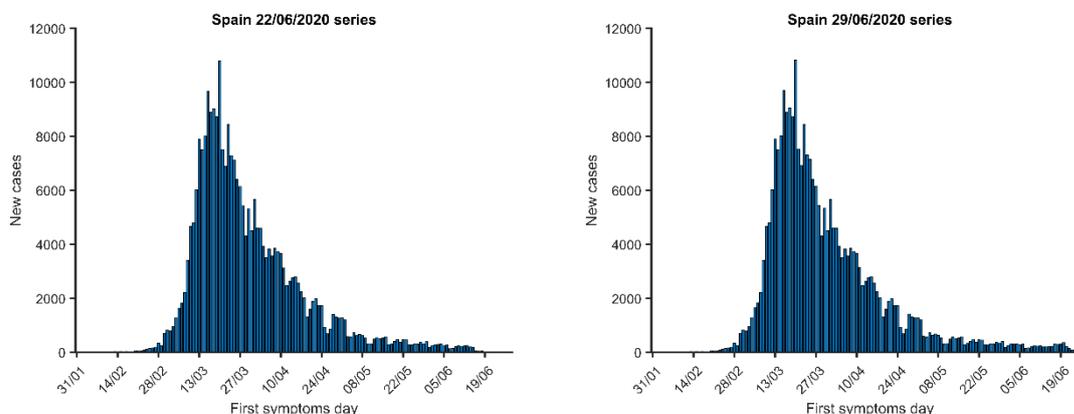
Este retraso en la actualización de los datos correspondientes a los últimos días, probablemente inevitable en muchos de los casos, tiene que ser tenida en cuenta a la hora de analizar dichas series. Por ejemplo, la estimación de la IA7 o de la Rt sólo podrá hacerse hasta el día en que los datos estén más o menos consolidados, esperándose pocas variaciones en actualizaciones sucesivas. En cambio, los datos de los últimos días estarán infraestimando el número total de casos.

El 17 de junio presentamos un primer informe donde comparábamos los datos publicados el 11 de junio con los datos publicados el 16 de junio. En aquel informe pudimos estimar correctamente los retrasos de algunas comunidades autónomas, pero los resultados aún estaban distorsionados por la revisión en profundidad de las series históricas. A nivel de país, los resultados apuntaban a un retraso de entre 4 y 7 días. En este momento nos emplazamos a hacer un estudio similar cuando estas series ya estuvieran en una fase de consolidación más avanzada, para poder identificar el retraso característico en el diagnóstico y notificación.

Hoy presentamos los resultados de la **comparación de la serie publicada el 22 de junio con la publicada el 29 de junio**. El hecho de comparar dos series que han sido publicadas con una semana de diferencia responde a los resultados preliminares, donde se apuntaba el retraso característico de 4-7 días. De esta manera, nos aseguramos de que los datos de los días 15 a 21 de junio están cerca de su valor definitivo.

## Análisis global a nivel de país

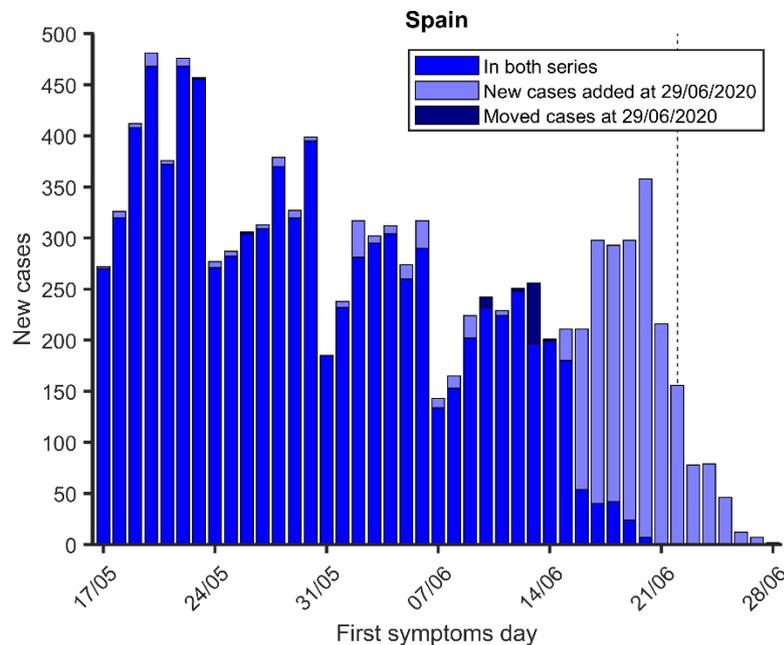
En la siguiente figura (Fig. 1) se muestra cómo se distribuyen en el tiempo los casos nuevos, por fecha de diagnóstico, a nivel de España. A la izquierda, serie publicada el 22 de junio de 2020, con datos que llegan hasta el 21 de junio de 2020 (**primera serie**, de aquí en adelante). A la izquierda, serie publicada el 29 de junio de 2020, con datos que llegan hasta el 28 de junio de 2020 (**segunda serie**, de aquí en adelante).



**Fig. 1.** Casos por fecha de aparición de síntomas en España según datos del ISCiii publicados el 22/06 (izq.) y el 29/06 (derecha).

<sup>1</sup> <https://cnecovid.isciii.es/covid19/>

La segunda serie, además de aportar una semana más de datos, **modifica la primera en algunos puntos**. En la siguiente gráfica (Fig. 2), mostramos los valores presentes en ambas series para las últimas seis semanas, así como los nuevos casos añadidos por la segunda serie (azul claro) y los casos que han sido reubicados por la segunda serie (azul oscuro). Se muestran sólo las últimas seis semanas, para poder percibir correctamente los cambios. La línea punteada muestra el día en que se publica la primera serie (22 de junio de 2020).



**Fig. 2.** Comparación de los casos nuevos, situados por fecha de aparición de síntomas, presentes en ambas series (22/06 y 29/06) y añadidos o resituados por la segunda (29/06).

Vemos que la segunda serie modifica algunos valores de días intermedios, en menor o mayor medida según el día. En particular, observamos que los últimos 6 días de la primera serie (16 a 21 de junio) son modificados de forma muy significativa por la segunda serie, y el día 15 de junio (7 días atrás respecto el último punto) es modificado en menor medida. Por último, vemos que los casos nuevos recientes aportados por la segunda serie (última semana, después de la línea punteada, del 22 al 28 de junio) son pocos. De hecho, esperamos que estos valores se vean incrementados con actualizaciones sucesivas. Se reafirma una vez más que **indicadores habitualmente utilizados en el análisis de la situación como el número reproductivo o la incidencia acumulada los últimos 7 días tienen que evaluarse con precaución**, ya que *de facto* estarían subestimando el nivel de riesgo epidemiológico real.

## Análisis a nivel de comunidad autónoma

Los cambios detectados a nivel de país varían según la comunidad autónoma. La siguiente figura (Fig. 3, diagrama de cajas) muestra dónde añade los nuevos casos la segunda serie, con respecto a la primera, para cada comunidad autónoma. Si no hubiera retrasos, se esperaría que todos los casos que aporta de más la segunda serie correspondieran a la última semana (22 a 28 de junio). No obstante, vemos que se añaden casos en días anteriores, es decir, **se están modificando datos reportados por la primera serie**.

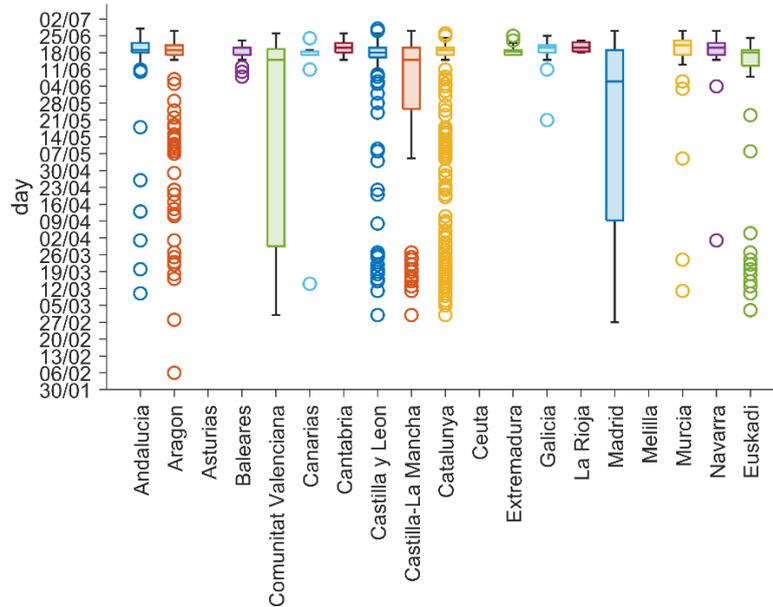


Fig. 3. Distribución en el tiempo de los casos aportados por la segunda serie (29/06) para cada comunidad autónoma.

En la mayoría de comunidades los casos nuevos se acumulan en los últimos días (parte superior de la gráfica), y se modifican de forma puntual algunos valores históricos (círculos aislados). Esto es **compatible con el retraso característico en el diagnóstico y registro de los casos, puesto que no es verosímil que los casos se diagnostiquen el mismo día de aparición de los síntomas**. Sin embargo, hay dos comunidades que destacan por seguir modificando de forma sustancial los valores históricos, Comunidad de Madrid y Comunitat Valenciana. En estas comunidades, de acuerdo a lo que observamos, las series históricas aún no se habrían consolidado. Castilla-La Mancha también estaría corrigiendo valores históricos, pero en menor medida.

Es interesante comparar cómo ha variado la situación en este sentido desde el último informe, donde el efecto de las correcciones históricas estaba muy presente. La siguiente figura (Fig. 4) muestra la comparativa de los dos informes. Se puede ver cómo la mayoría de comunidades autónomas han reducido significativamente el porcentaje de datos antiguos corregidos, mostrando pues un buen grado de consolidación de las series.

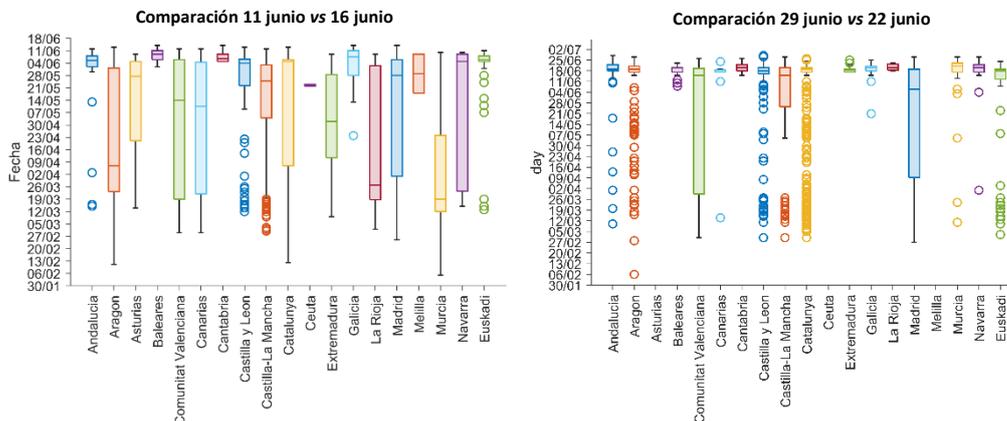


Fig. 4. Distribución de casos nuevos en las dos comparativas realizadas.

En el apéndice A se muestra la comparación temporal de las dos series de datos comunidad a comunidad para las últimas 6 semanas. En general, se observa que las comunidades autónomas presentan cambios significativos en los últimos 5 a 7 días, en la mayoría de ellas. Se observan con claridad los cambios históricos en las gráficas de Castilla-La Mancha, Comunidad de Madrid y Comunitat Valenciana.

## Evaluación del error acumulado

Para evaluar el error cometido, se ha utilizado la metodología siguiente. Se toma como inicio el 21 de junio y se estudia la serie en sentido inverso, desde el final hasta el principio. Para cada día se evalúa el porcentaje de error acumulado como:

$$\text{Error acumulado}(t_f - t) = \frac{\sum_{t_f}^{t_f-t} \text{Diferencia de casos diarios entre ambas series}}{\sum_{t_f}^{t_f-t} \text{Casos nuevos reportados por la segunda serie}}$$

Este error se evalúa para toda la serie histórica empezando a  $t_f = 21 \text{ de junio}$ . Para cada región (comunidad autónoma y país), se evalúa si el error acumulado está por debajo del 10 % en toda la serie. Si no es así, se va un día atrás y se hace el mismo análisis empezando a  $t_f = 20 \text{ de junio}$ . Si hay regiones con errores iguales o superiores al 10 %, se vuelve a repetir para  $t_f = 19 \text{ de junio}$ , y así sucesivamente. Finalmente, se mira cuántos días hacia atrás hemos tenido que ir para cada región, con tal de obtener un error que esté, de forma consistente, por debajo del 10 %. La siguiente tabla muestra los resultados de comparar las dos series en estudio, es decir, error cometido en la serie del 22 de junio con respecto a la actualización del 29 de junio. Esta comparación nos da, como resultado, los días que se estiman como no fiables porque se espera que sus datos sean modificados significativamente en actualizaciones sucesivas. En consecuencia, **podemos evaluar cuál sería el último día fiable en los datos actuales, es decir, en la serie del 29 de junio.**

	Días no fiables (error acum. $\geq 10\%$ )	Último día fiable de la serie de 29 de junio
Andalucía	>7	21 de junio*
Aragón	6	22 de junio
Asturias	0	28 de junio
Baleares	7	21 de junio
Comunitat Valenciana	>7	21 de junio*
Canarias	5	23 de junio
Cantabria	7	21 de junio
Castilla y León	7	21 de junio
Castilla-La Mancha	>7	21 de junio*
Catalunya	6	22 de junio
Ceuta	0	28 de junio
Extremadura	5	23 de junio
Galicia	7	21 de junio
La Rioja	4	24 de junio
Madrid	6	22 de junio
Melilla	0	28 de junio
Murcia	>7	21 de junio*
Navarra	6	22 de junio
Euskadi	>7	21 de junio*
España	7	21 de junio

\*Pendiente de ser revisado en informes sucesivos, ya que un valor superior a 7 días en el número de datos no fiables puede indicar correcciones significativas en serie histórica. En consecuencia, no podríamos hablar de datos no fiables debido al retraso sino, en parte, debido a la revisión.

## Conclusiones

---

De este análisis se desprende que, efectivamente, hay un período de entre 4 y 7 días, en función de la comunidad autónoma, en el cual los datos deberían tomarse con cautela, ya que las cifras estarían subestimadas y serían corregidas al alza en actualizaciones posteriores. El factor de la revisión de la serie histórica sigue enmascarando ligeramente el efecto del retraso de los últimos días, aunque en menor medida. En cualquier caso, los resultados confirman que **los datos de los últimos 5-7 días no deberían ser tenidos en cuenta a la hora de analizar la situación actual**. Aunque puede parecer que es un intervalo grande, no hay que perder de vista que se está trabajando con fecha de inicio de síntomas, que de forma natural es unos días anterior a la fecha de registro.

La mayoría de fuentes de datos oficiales trabajan con fecha de diagnóstico o con fecha de notificación, de manera que **a la hora de comparar hay que tener en cuenta el retraso natural entre los tres tipos de registros**. Por ejemplo, si entre la aparición de los síntomas y su diagnóstico puede haber un periodo de unos 2-4 días, se espera que en las series que trabajan con fecha de diagnóstico el intervalo de días no fiables sea entre 2 y 4 días menor al de las series por inicio de síntomas.

## Apéndice A. Comparación de las dos series para cada comunidad autónoma

