Modelización empírica de la epidemia COVID-19

Marta Ginovart*, Mónica Blanco
Universitat Politècnica de Catalunya
marta.ginovart@upc.edu

El presente documento complementa al artículo publicado por las autoras con el mismo tituloes de en la versión en papel de la revista *Suma,* n.º 94, septiembre de 2020. En el artículo se cita en varias ocasiones este documento

ANEXO en formato de tabla. Datos COVID-19

Fecha	Día	Casos totales	Casos en China	Casos en otros países	Muertes totales	Recuperados totales	Recuperados totales
		(Fuente1)	(Fuente2)	No China (Fuente2)	(Fuente1)	(Fuente1)	(Fuente2)
15/01/2020	1						
16/01/2020	2						
17/01/2020	3						
18/01/2020	4	201					
19/01/2020	5	222					
20/01/2020	6	282	278	4			
21/01/2020	7	314	326	6			25
22/01/2020	8	581	547	8			28
23/01/2020	9	846	639	14			30
24/01/2020	10	1320	916	25		38	36
25/01/2020	11	2010	2000			49	49
26/01/2020	12	2800	2700	57		56	54
27/01/2020	13	4590	4400			68	63
28/01/2020	14	6070	6000	87		103	110
29/01/2020	15	7820	7700	105		135	133
30/01/2020	16	9830	9700	118		211	141
31/01/2020	17	12000	11200	153		243	220
01/02/2020	18	14600	14300	173		328	284
02/02/2020	19	17400	17200	183		487	487
03/02/2020	20	20600	19700	188		632	621
04/02/2020	21	24600	23700	212		909	899
05/02/2020	22	28300	27400	227	565	1030	1100
06/02/2020	23	31500	30600	265	638	1560	1500

07/02/2020	24	34900	34100	317	724	2080	2000
08/02/2020	25	37600	36800	343	813	2690	2600
09/02/2020	26	40600	39800	361	910	3320	3200
10/02/2020	27	43100	42300	457	1020	3750	3900
11/02/2020	28	45200	44300	476	1120	4740	4700
12/02/2020	29	47000	44700	523	1260	5900	5200
13/02/2020	30	64400	59800	538	1380	6680	6300
14/02/2020	31	67100	66300	595	1530	7720	8100
15/02/2020	32	69300	68300	685	1670	8200	9400
16/02/2020	33	71400	70400	780	1780	10600	10900
17/02/2020	34	73300	72400	896	1870	12700	12600
18/02/2020	35	75200	74100	1000	2010	13500	14400
19/02/2020	36	75700	74500	1100	2130	14500	16100
20/02/2020	37	76800	75000	1200	2250	18400	18200
21/02/2020	38	77800	75500	1400	2360	20900	18900
22/02/2020	39	78800	76900	1700	2460	22600	22900
23/02/2020	40	79300	76900	2000	2620		23400
24/02/2020	41	80200	77200	2400	2700		25200
25/02/2020	42	81100	77700	2800	2760		27900
26/02/2020	43	82300	78100	3300	2800		30400
27/02/2020	44	83700	78500	4300	2860		33300
28/02/2020	45	85200	78800	5300	2920	39449	36700
29/02/2020	46	87100	79300	6800	2980		39800
01/02/2020	47	89100	79800	8500	3040	45202	42700
02/02/2020	48		80000	10300			45600
03/02/2020	49		80200	12700			48200
04/02/2020	50		80300	14900			51200

TEXTO PARA INCORPORAR EN LA WEB DE SUMA, Y A LA QUE SE HACE REFERENCIA EN EL MANUSCRITO

Algunos aspectos sobre el manejo del Excel que pueden ser convenientes de recordar antes de proceder a la realización de las tareas COVID-19 de manera ágil y autónoma:

1) Para la representación gráfica de los datos temporales se puede utilizar la opción del diagrama de dispersión.
2) Para ajustar una función matemática que describa el comportamiento de un conjunto de datos se puede utilizar una herramienta que ya tiene incorporada la misma hoja de cálculo. Para escoger el tipo de modelo que se ajustará a los datos representados, existe la opción de "añadir tendencia", la cual proporciona opciones para elegir el tipo de función matemática que se utilizará en el

ajuste, según sea la evolución temporal que estos datos muestren: modelo lineal, logarítmico, polinómico, potencial, o exponencial. No obstante, es muy importante insistir en la existencia de otros tipos de programas que permiten trabajar el ajuste o modelización de datos de forma más completa.

3) Para valorar la bondad o calidad del modelo escogido que tiene que explicar la variable respuesta (y) a partir de la variable explicativa (x) se puede obtener el valor del coeficiente de determinación R^2 (que se estudiará con más detalle en la asignatura de Estadística, en segundo curso).
4) Para conseguir una representación gráfica en escala logarítmica se puede modificar en el mismo gráfico original el formato del eje de ordenadas ("Escala logarítmica" en la base 10). También se puede crear una nueva variable aplicando la transformación logarítmica en base 10 (o en cualquier otra base) sobre la variable respuesta, y representar después la variable transformada.
5) Para explorar o ver con más detalle partes específicas de una gráfica, se puede hacer un "zoom" de estas partes, modificando directamente sobre misma gráfica el formato de las escalas cambiando los mínimos y máximos de los ejes.
6) Para complementar la respuesta a algunas de las tareas que se proponen, es necesario indicar que las opciones de funciones matemáticas que de forma directa ofrece esta hoja de cálculo para ajustar datos no son suficientes para poder modelizar de forma conveniente el comportamiento de todas las evoluciones temporales que muestran estos datos. Es oportuno presentar las funciones sigmoides, también conocidas como funciones tipo S, como por ejemplo, la función Gompertz, la función logística y la función loglogística.

El primer conjunto de13 tareas COVID-19 a realizar con los datos se encuentran en el Anexo en relación a modelos lineales o linealizables:

Tarea 1. Representa la evolución temporal que se observa para la variable "Casos totales (Fuente 1)". Explica cuál es el comportamiento de estos datos de forma global. ¿Hay alguna función matemática de las que proporciona la opción "añadir tendencia" a los datos representados en la hoja de cálculo Excel que permita un ajuste adecuado para estos datos?

Tarea 2. Considera únicamente los primeros 15 días de los datos observados para la variable "Casos totales (Fuente 1)". Explica cuál es el comportamiento de estos datos, e identifica la función matemática que podría generar un ajuste adecuado para los datos de la primera etapa de la evolución temporal de la epidemia en China.

Tarea 3. Representa en un mismo gráfico el que sería la evolución de un modelo exponencial para la propagación de una epidemia utilizando la función CasModelExp (t) = 34 exp (0.366 t) y los datos reales publicados "Casos totales (Fuente 1)" para el período temporal que va desde el día 1 (15 de enero de 2020) hasta el día 25 (8 de febrero). ¿Qué se observa si comparas las dos evoluciones en este periodo temporal de 28 días? En el caso de utilizar en este mismo gráfico una escala logarítmica para el eje de ordenadas, ¿que se observa?

Tarea 4. Considera únicamente los últimos 15 días de los datos observados para la variable "Casos totales (Fuente 1)". Explica cuál es el comportamiento de estos datos.

Tarea 5. ¿Qué se observa si utilizas una escala logarítmica en la representación gráfica de todos los datos correspondientes a la variable "Casos totales (Fuente 1)"? Si primero realizas la transformación logarítmica (en base 10 o en otra base) para esta variable y después representas la variable transformada, ¿qué es lo que observas? ¿Crees que es conveniente hacer uso de una transformación logarítmica en este contexto? ¿Consideras que las opciones que has visto y utilizado anteriormente en "añadir tendencia" son suficientes y/o convenientes para explicar el comportamiento que se observa para esta variable?

Tarea 6. Representa gráficamente los datos correspondientes a la variable "Casos en China (Fuente 2)", que son casos infectados (acumulados) en el país en que se inició la enfermedad COVID-19. ¿Puedes identificar alguna fecha (observación) que tenga una relevancia especial en esta evolución temporal? ¿Intuyes cuál puede ser la razón? Busca en la prensa que sucedió en China el día 13 de febrero en relación a la identificación de personas infectadas.

Tarea 7. ¿Qué tipo de evolución temporal muestra la variable "Casos en China (Fuente 2)"? Comenta la evolución o comportamiento de la epidemia a partir de la forma que muestra esta serie temporal. Compara esta evolución temporal con la evolución temporal de los casos totales (Font1) representada anteriormente y discute semejanzas y disimilitudes. ¿Consideras que el ajuste de estos datos encaja adecuadamente en algunas de las opciones "añadir tendencia" que proporciona el Excel?

Tarea 8. ¿Podríamos argumentar o justificar si en el conjunto de países que no son China, la propagación de la epidemia se está conteniendo? ¿Qué tipo de evolución muestran los casos detectados fuera de China?

Tarea 9. Representa gráficamente la variable "Muertos totales (Font 1). ¿Qué tipo de evolución temporal observas? ¿Qué tipo de modelo podrías ajustar a estos datos? Asumiendo que la evolución temporal puede reflejar dos etapas distintas en la propagación del virus, ¿cuál sería la mejor manera de explicar el comportamiento de esta variable?

Tarea 10. ¿Qué puedes comentar sobre la evolución del porcentaje de muertes que provoca este coronavirus?

Tarea 11. ¿Qué tipo de evolución temporal muestra la variable "Recuperados totales (Fuente 1)? ¿Y la de variable "Recuperados totales (Fuente 2)? ¿Cómo puedes valorar la concordancia o la discrepancia de la información que proporcionan estas dos fuentes de información sobre la misma variable? Considerando ahora otra de las variables para la que tienes dos fuentes de información, la del número de casos (infectados) detectados por el mundo, que proporciona de forma directa la Fuente 1, como totales, y de forma indirecta la Fuente 2, como casos de China y casos de otros países que no son China, ¿cómo puedes valorar su concordancia o su discrepancia? ¿Qué puedes decir sobre estas dos fuentes de información?

Tarea 12. Genera en la hoja de cálculo la variable derivada a partir de datos originales que sea el número de casos "nuevos" diarios (Casos(día)-Casos(día-1)): el número diario de casos "nuevos" en el mundo, en China, y en el conjunto de países sin China. Inspecciona el comportamiento de estas tres nuevas variables. ¿Se pueden identificar períodos de tiempo que muestren un comportamiento diferente para estas variables generadas? ¿Se observa el mismo comportamiento en China que en el conjunto del resto de países?

Comentario 1: El propósito de los datos que se publicaron en ese contexto era saber el número de casos acumulados en la fecha y, por lo tanto, no se puede concluir que la diferencia entre un día y el día anterior sea realmente el número de casos "nuevos", ya que casos que aparecen reportados en un día podrían haber surgido en fechas anteriores (y haber sido recuperados con posterioridad a su aparición). Cualquier inferencia que se haga sobre las diferencias de un día respecto al anterior debe hacerse con precaución atendiendo a esta consideración previa. Ante la situación excepcional de desconocimiento de la situación real de la población en el momento de la publicación de los datos que utilizamos, se debe dar a esta variable derivada el valor que le corresponde, siendo una forma alternativa de explicar lo que se va conociendo sobre la propagación del coronavirus (calculando el incremento diario que informa sobre la velocidad, en analogía a la derivada de una función).

Comentario 2: Se puede proceder a la eliminación de los datos correspondientes a las fechas de los días 13 y 14 de febrero (observaciones 30 y 31), afectados por un cambio de criterio en la contabilidad de casos que las instituciones sanitarias hicieron en China en ese momento.

Comentario 3: Una alternativa a los casos "nuevos" diarios para comparar lo que pasa en China con lo que pasa en el conjunto de los otros países durante este periodo temporal, es generar una nueva variable que informe sobre el "número de casos "nuevos" por caso detectado". Esta nueva variable ((Casos(día)-Casos(día-1))/Casos(día-1)) se puede expresar como porcentaje.

Tarea 13. ¿Qué puedes decir sobre la evolución del número de muertos diarios? Con los datos disponibles, ¿puedes relacionar esta variable con las variables que has ido explorando anteriormente? ¿Qué se observa?

El segundo conjunto de tareas a realizar con los datos se encuentran en el Anexo en relación a modelos no lineales:

Tarea 14. Investiga con un programa estadístico la modelización a partir de funciones sigmoides de los datos correspondientes a la variable "Casos en China (Fuente 2)".

Tarea 15. Ajusta los modelos de Gompertz, logístico y loglogístico a un subconjunto de datos compuesto por los casos de infectados (acumulados) de los primeros días de la epidemia en China, no más allá del día 25 de los 50 disponibles. Esto garantiza que con estos datos aún no se pueda visualizar con claridad la curva tipo S. Utiliza los modelos que hayas obtenido con estos ajustes para este subconjunto de datos, y predice valores futuros de casos infectados en fechas posteriores a las utilizadas en el ajuste. Compara estas predicciones futuras conseguidas con los valores reales ya observados y conocidos de China.

Tarea Opcional. A partir de lo que has trabajado y aprendido con los datos analizados de la evolución de la epidemia COVID-19 en China:

a) Analiza los datos publicados hasta la fecha de hoy (era primeros de abril de 2020) de la evolución de la epidemia COVID-19 en:
   - Catalunya (http://aquas.gencat.cat/ca/actualitat/ultimes-dades-coronavirus, http://governobert.gencat.cat/ca/dades_obertes/, https://analisi.transparenciacatalunya.cat/ca/)
   - España (https://www.mscbs.gob.es/profesionales/saludPublica/ccayes/alertasActual/nCov-China/home.htm, https://cnecovid.isciii.es/, https://cnecovid.isciii.es/covid19/#documentaci%C3%B3n-y-datos).

b) Investiga el uso de los modelos Gompertz, logístico y loglogístico, para conseguir predicciones del número de casos infectados (acumulados) futuros en Catalunya y en España. ¡Tus predicciones las podrás comparar con los datos observados al cabo de unos días! ¡Y recuerda…, se han impuesto estas funciones porque "sospechamos" que estos primeros datos que vemos son sólo la primera parte de una evolución que resultará de tipo S!

La modelización empírica es utilizada por grupos de investigación para estudiar el comportamiento de la epidemia COVID-19 en los diversos países del mundo y en zonas geográficas limitadas.

Son muchos y diversos los modelos que han surgido y que se están mejorando día a día a partir de los datos que se van recogiendo, almacenando y procesando continuamente (ver por ejemplo, https://www.medrxiv.org/content/10.1101/2020.05.13.20101329v1, https://biocomsc.upc.edu/en).

Al mismo tiempo, considerando la dinámica de contagios producida por el SARS-CoV-2, datos de movilidad urbana y accediendo al censo de los municipios de España se elaboraron modelos espacio-temporales (https://www.medrxiv.org/content/10.1101/2020.03.21.20040022v1) con probabilidades de contagio que permitieron motorizar la evolución de este riesgo por zonas específicamente delimitadas (http://deim.urv.cat/~alephsys/COVID-19/spain/es/index.html).

Es de destacar, entre otras muchos proyectos surgidos en esta alerta sanitaria, "Acción matemática contra el coronavirus" del Comité Español de Matemáticas (http://matematicas.uclm.es/cemat/covid19/) que posibilitó la construcción de un meta-predictor o "predictor cooperativo" para facilitar a las autoridades información del comportamiento a corto plazo de variables de interés en la expansión de la COVID-19, basándose en combinaciones optimizadas de predicciones de diferentes modelos/algoritmos y desagregadas por comunidades autónomas españolas (https://covid19.citic.udc.es/).

Estas referencias a proyectos que trabajan con datos de la COVID-19 son únicamente para ilustrar entornos de trabajo distintos, y son unos pocos ejemplos de los muchos proyectos que existen (y que existirán) al entorno de esta epidemia.

Son múltiples las bases de datos que se pueden consultar, y que son consultadas para la elaboración de modelos COVID-19, como por ejemplo (a más a más de las mencionadas previamente):

- WHO, https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports
- CDC, https://www.cdc.gov/coronavirus/2019-ncov/index.html
- ECDC, https://www.ecdc.europa.eu/en/geographical-distribution-2019-ncov-cases
- DXY, https://ncov.dxy.cn/ncovh5/view/en_pneumonia?from=dxy&source=&link=&share=
- JHU, https://coronavirus.jhu.edu/map.html