Análisis estadístico de la evolución del COVID-19

Carla Albo González
Anaís de los Ángeles Hernández Pereda

Análisis estadístico de la evolución del COVID-19

En este proyecto se han realizado siete sesiones de una hora impartidas cada martes desde el 16 de marzo hasta el 27 de abril. El grupo ha estado integrado por dos estudiantes de 3º de la ESO del Colegio Fuentelarreyna de Madrid, http://www.colegiofuentelarreyna.org/. Las sesiones se han desarrollado online y han sido dirigidas por su tutora Stella García Moreno en colaboración con Beatriz González Pérez del Departamento de Estadística e Investigación Operativa de la Universidad Complutense de Madrid, madre de una de las alumnas. Durante las sesiones, las alumnas han aprendido a construir en tiempo real sus propios modelos matemáticos basados en ajustes exponenciales que explican y predicen la evolución del COVID-19 en su provincia de residencia. Para ello, se han utilizado los datos disponibles en la web del Instituto de Salud Carlos III. Las prácticas han sido realizadas con Excel y no ha sido necesario ningún conocimiento previo del programa. Al finalizar las sesiones las alumnas han elaborado un informe con todos los análisis que han realizado, que se incorpora a continuación a esta memoria. También han realizado una práctica libre aplicando dichos conocimientos a la cuarta ola, que también se incorpora a esta memoria. Finalmente, se ha creado un repositorio de acceso público asociado al proyecto: https://github.com/bgonzalez380/Proyecto-Covid-2021, donde se pueden descargar los ficheros Excel de las alumnas con el contenido de cada práctica.

Índice

Análisis estadístico de la evolución del COVID-19. Caso de Estudio: Madrid

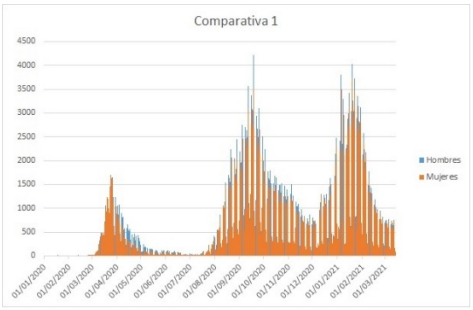
Te puedes descargar las prácticas de cada día en nuestro repositorio público:

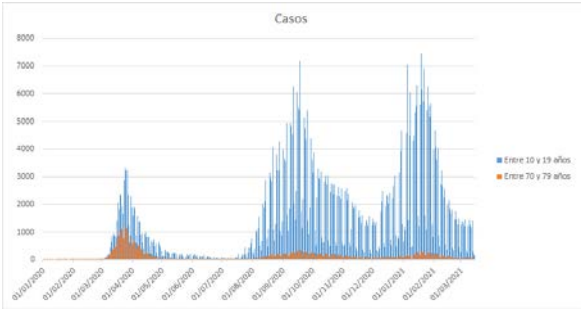
https://github.com/bgonzalez380/Proyecto-Covid-2021

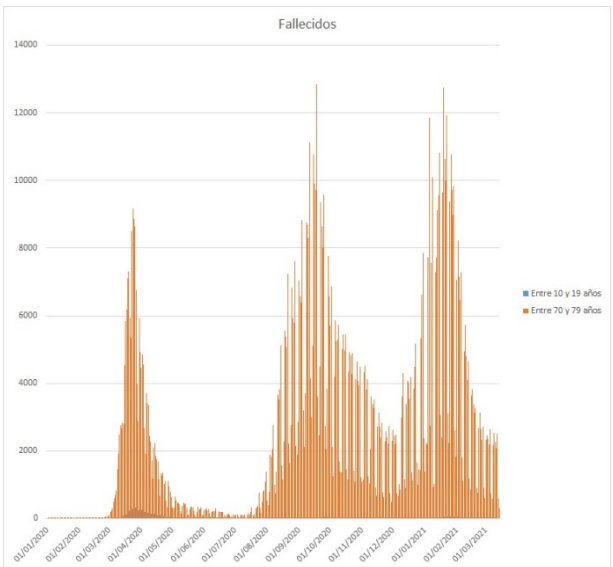
Datos 16/03/2021:

Los datos utilizados para este proyecto han sido sacados de la página web oficial del Instituto Nacional Carlos III (https://cnecovid.isciii.es/covid19/) Dichos datos estaban en formato "csv". Para insertarlos en Excel, lo que hicimos fue abrir una nueva hoja de cálculo, y presionar la opción "datos" en el menú de arriba y, a continuación, "desde texto".
Después, seleccionamos los datos descargados e hicimos un filtro (función filtro de Excel) para quedarnos solo con los de la comunidad de Madrid, representada con una M mayúscula en los datos.
Estos datos estaban separados por grupo de edad y sexo así que previamente tuvimos que sumar los valores de todas las categorías del mismo día para calcular el total de ese día.

Una vez tuvimos los datos filtrados y preparados, hicimos dos gráficos en dos distintas hojas (comparativa 1 y 2). En el primero comparamos el número de casos entre los hombres y las mujeres, y en el segundo el número de fallecidos y casos entre los contagiados con edades comprendidas entre los 10 y 19 años y los 70 y 79 años.
Como podemos observar, (colores) el número de casos es mayor entre los jóvenes, sin embargo, los fallecidos ancianos superan los de los jóvenes.





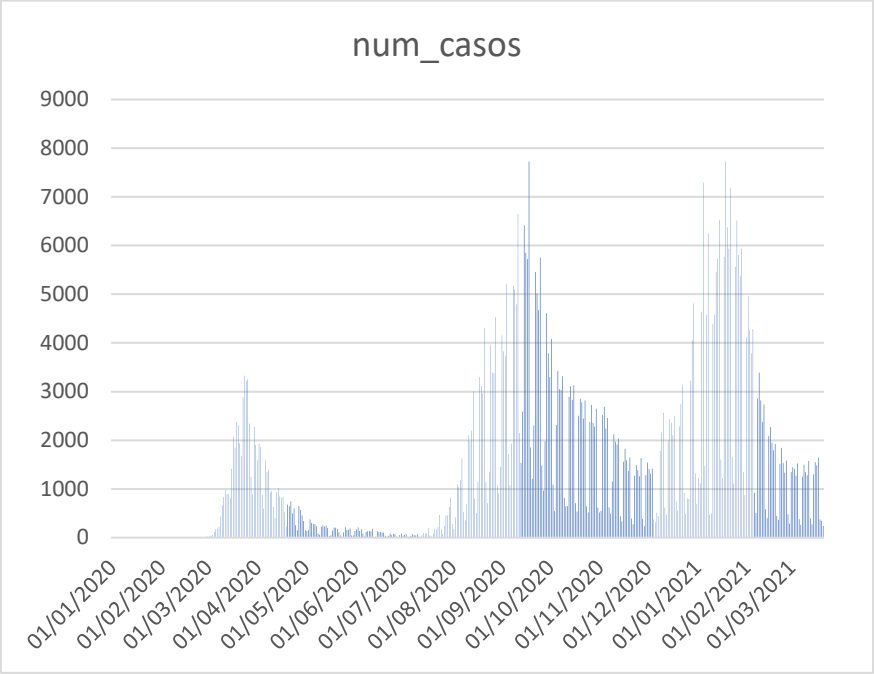


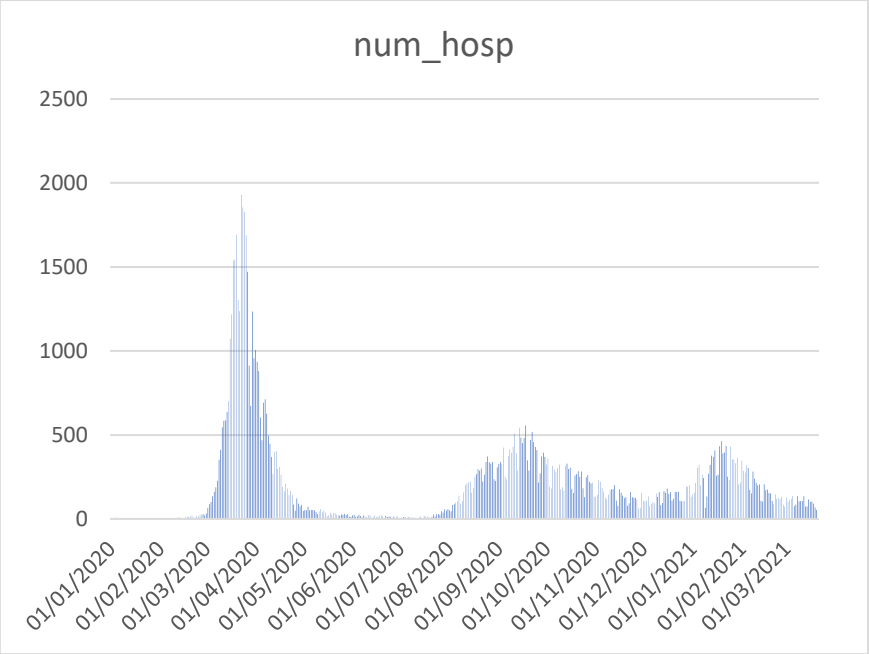
Datos 23/03/2021:

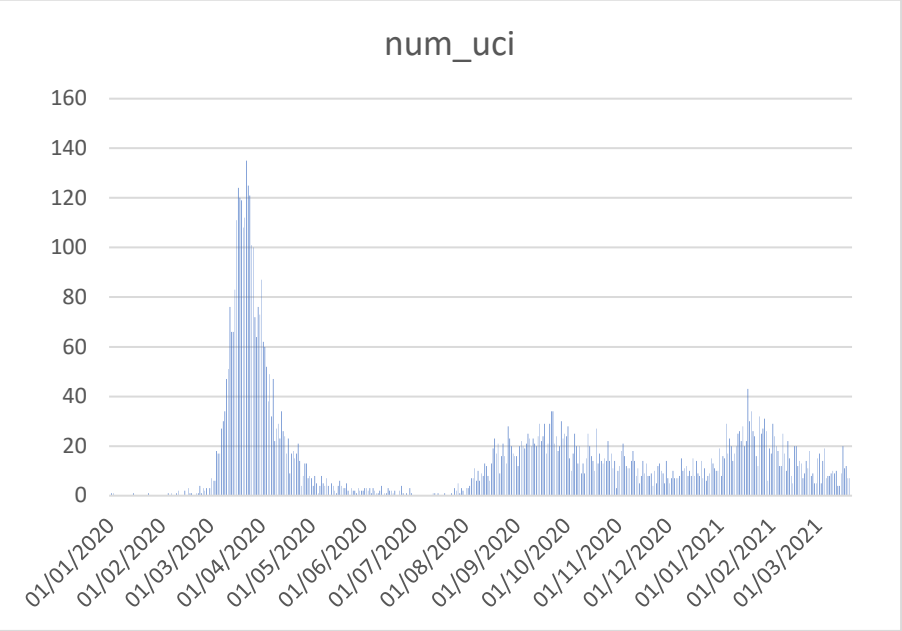
El 23 de marzo actualizamos los datos, descargándonos los que subieron ese día. Estos contenían el número de contagiados, hospitalizados, ucis y defunciones hasta el 21 de marzo.

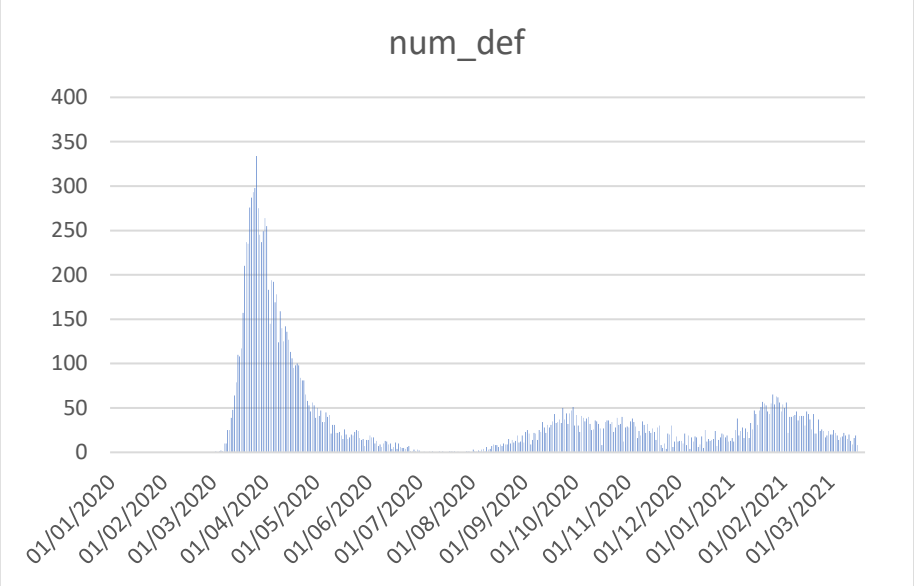
Nosotros hicimos el mismo filtro que hicimos con los datos anteriores, cogiendo solo los de Madrid, y los pegamos en una hoja a parte (hoja 3). A continuación, cogimos nuestra selección, y la copiamos y pegamos en una hoja nueva (hoja 3). Después de esto calculamos la frecuencia acumulada (se expresa con el siguiente símbolo: F_i) del número de casos, ucis, defunciones y hospitalizados. Para calcularla, sumamos el número actual del valor que estemos tratando con el anterior, y así con todos los valores, para obtener el número total de ucis, casos, defunciones y hospitalizados que había en todos los días.
El último día en el que teníamos los datos disponibles, el 21 de marzo de 2021, la frecuencia acumulada para cada variable era la siguiente:
Casos: 611.587
Ucis: 6.641
Hospitalizados: 90877
Defunciones: 14465
También hicimos una tabla con el pico de cada ola con los casos, hospitalizados, ucis y defunciones, ayudándonos de un gráfico para encontrar el valor máximo más fácilmente. Este es el resultado:

	Primera ola	Segunda Ola	Tercera Ola
casos	24/03/2020	18/09/2020	19/01/2021
hospitalizados	23/03/2020	18/09/2020	20/01/2021
ucis	23/03/2020	22/09/2020	18/01/2021
defunciones	27/03/2020	30/09/2020	26/01/2021
casos	3328	7724	7729
hospitalizados	1930	556	391
ucis	135	34	43
defunciones	334	51	65




num_hosp
2500
2000
1500
1000
500
0
01/01/2020
01/02/2020
01/03/2020
01/04/2020
01/05/2020
01/06/2020
01/07/2020
01/08/2020
01/09/2020
01/10/2020
01/11/2020
01/12/2020
01/01/2021
01/02/2021
01/03/2021

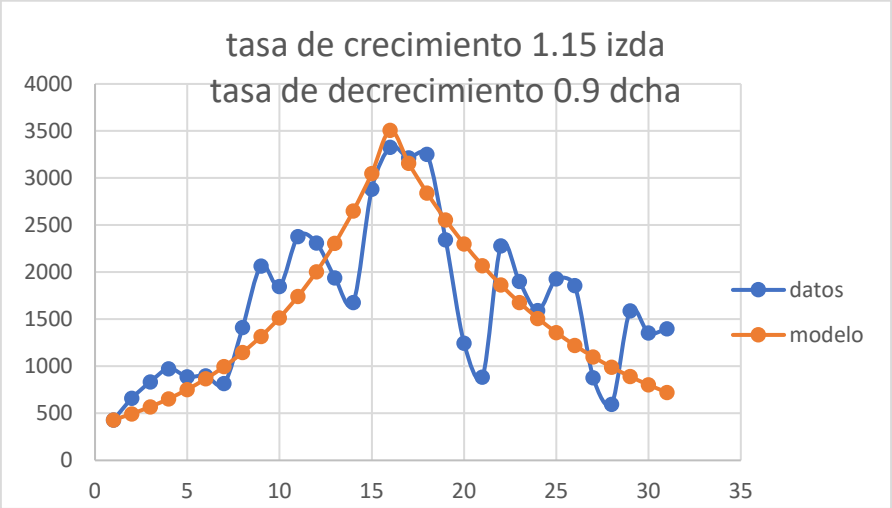

num_uci
160
140
120
100
80
60
40
20
0
01/01/2020
01/02/2020
01/03/2020
01/04/2020
01/05/2020
01/06/2020
01/07/2020
01/08/2020
01/09/2020
01/10/2020
01/11/2020
01/12/2020
01/01/2021
01/02/2021
01/03/2021


num_def
400
350
300
250
200
150
100
50
0
01/01/2020
01/02/2020
01/03/2020
01/04/2020
01/05/2020
01/06/2020
01/07/2020
01/08/2020
01/09/2020
01/10/2020
01/11/2020
01/12/2020
01/01/2021
01/02/2021
01/03/2021

30/03/2021

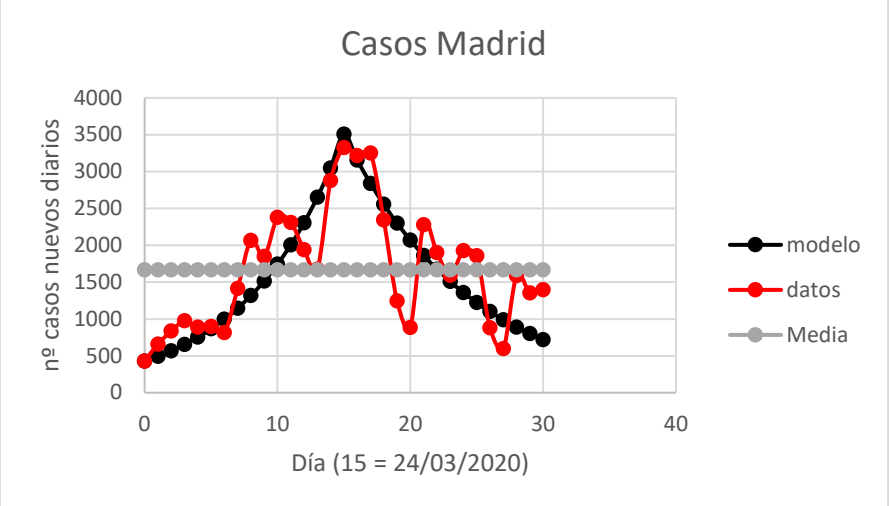
El día 30 de marzo, en un nuevo libro de Excel, pusimos los números del 0 al 30 en una columna, siendo estos los días de un mes. A continuación, adjuntamos los datos de N.º de casos de Madrid sólo de los días 09/03/2020 al 08/04/2020, ya que el 24/03/2020 fue el pico de la primera ola y esos son los días previos y posteriores.

Después, elaboramos el número reproductivo de estos días, poniendo en la primera mitad un número mayor que 1 (crecimiento de los días 0 al 15) y menor que 1 en la otra (decrecimiento en los días 16 al 30). Para ello, los datos oficiales los insertamos en un gráfico de dispersión. En una nueva columna, escribimos en la primera celda el número de casos del día 0 de los datos oficiales (431), y en la siguiente pusimos una fórmula que consiste en multiplicar el dígito de la celda anterior por el número reproductivo correspondiente a la nueva. Esta fórmula es aplicada a todas las celdas de la columna hasta el día 30. Los datos de esta columna los añadimos en una nueva serie en el gráfico de dispersión creado anteriormente, para poder contrastarlos. El objetivo principal fue hacer coincidir el pico de los datos de nuestro modelo con el de los datos oficiales, por lo que editamos el número reproductivo hasta que fue muy similar.

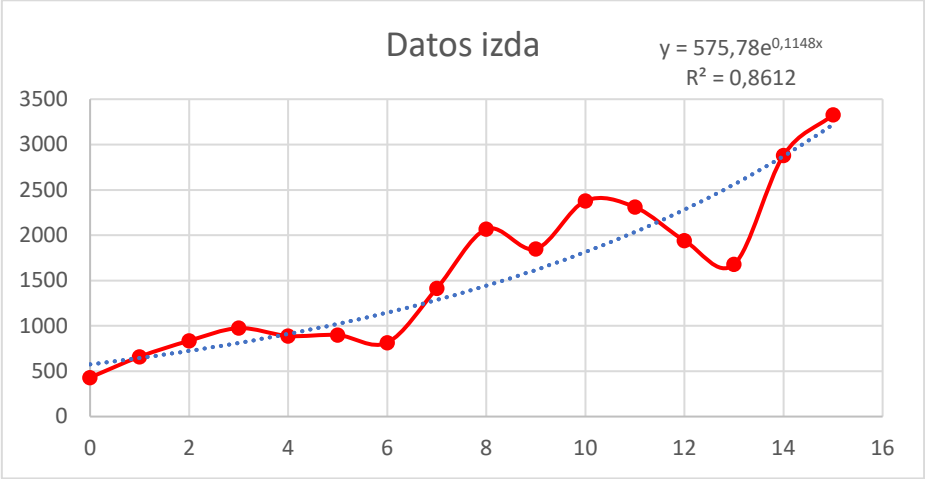


06/04/2021

Calculamos el error cuadrático medio entre el modelo y los datos oficiales, restándole al día 0 del modelo el correspondiente de los datos oficiales, y elevándolo al cuadrado, para después aplicar la fórmula al resto de los días. Calculamos también el error cuadrático medio para el peor modelo (asociado a la constante que pasa por la media de los datos oficiales) que resultó ser del orden de tres veces superior al del modelo ajustado.

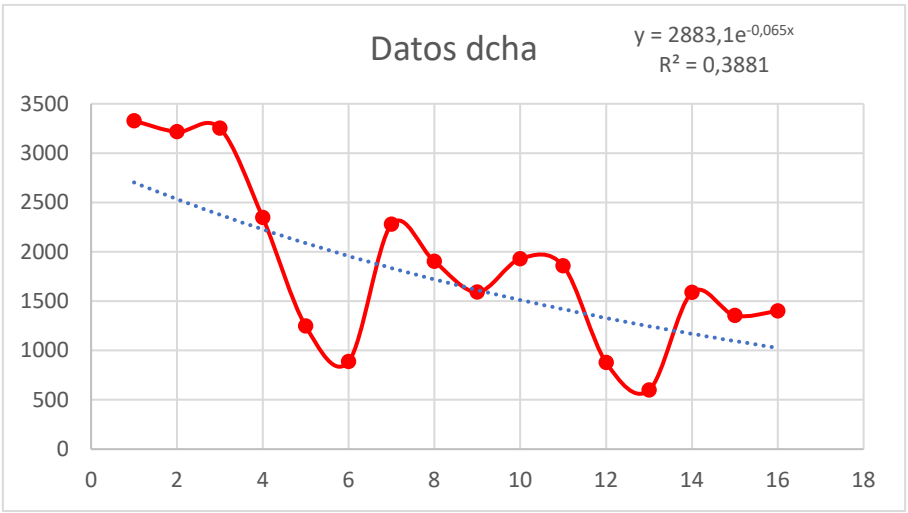


Luego, añadimos un nuevo gráfico de dispersión con los datos de los primeros 15 días de los datos oficiales, y le adjuntamos una línea exponencial desde el menú agregar línea de tendencia.



13/04/2021

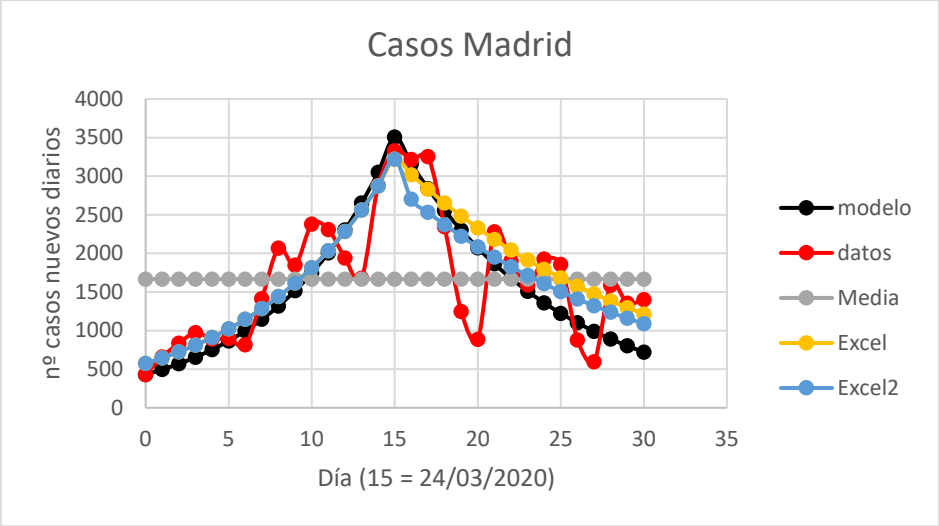
Hicimos lo mismo después del pico.



Aprendimos entonces lo que era el número de Euler “e” y su interpretación como crecimiento y decrecimiento exponencial.

2,71828182845904523536028747135 27 (y sigue...)

Con los ajustes automáticos del Excel empalmamos los modelos exponenciales de la izda y de la dcha de dos formas distintas: empalme con las fórmulas de las líneas exponenciales generadas automáticamente con el excel y empalme con las tasas de crecimiento y decrecimiento calculadas a partir de dichas exponenciales para activar el procedimiento manual. La idea es que a partir de estas dos líneas exponenciales, produjimos un nuevo modelo. Para ello, hicimos el producto de los datos de la misma: 575,78 por el número de Euler “e”, elevado a 0,1148 que multiplica al N.º de día del mes. Luego, aplicamos esta fórmula al resto de días hasta el pico. Sacamos, en una columna nueva, el número reproductivo de estos datos mediante una división, la segunda celda del modelo entre la primera. Lo mismo hicimos con la línea exponencial de la derecha que tiene exponente negativo. Observamos que las tasas de crecimiento y decrecimiento del proceso automático se parecían a las que nosotros habíamos encontrado manualmente.



20/04/2021

Aunque podíamos comparar todos los modelo con el error cuadrático medio, aprendimos a expresar con números cuán bueno es el modelo restando a 1 el cociente entre el error cuadrático medio de un modelo ajustado dividido por el peor error cuadrático medio asociado a la constante (coeficiente de determinación R2). Si multiplicamos la cifra resultante por 100, nos da el porcentaje de fiabilidad de nuestro modelo. El mejor modelo es el azul con un 0.67 aunque el modelo negro que habíamos ajustado manualmente no se queda lejos con un 0.6. Al modelo asociado a la Media le corresponde el valor mínimo de 0. Al modelo que pasa por todos los datos le corresponde el valor máximo de 1.

ECM	0,39341898	1	0,37385546	0,33069727
R^2	0,60658102	0	0,62614454	0,66930273

27/04/2021

Lo primero que hicimos fue coger el número de casos (sólo los de Madrid) y lo pegamos en Excel. Sanidad dice que un infectado empieza contagiar 5 días después de su infección (hay margen de error, por ello hicimos los cálculos a partir de tres días consecuentes); por tanto, contamos desde el día que hay un contagiado cinco días, así que en el 5º día se muestra el N.º de contagiados a partir del contagiado del día 1 (aplicamos esto a todos los valores).

> directa. En nuestro grupo, atendiendo a que una persona infectada tarda aproximadamente 5 días a ser infecciosa, definimos el *número reproductivo empírico*, p, como el cociente de los casos nuevos dividido por los casos nuevos de hace 5 días[2]. Utilizamos [1]

[1] Uso de modelos para el análisis y la predicción de la dinámica epidemiológica de la COVID-19 Martí Català, Sergio Alonso, Enric Álvarez, Pere-Joan Cardona, Daniel López, Clara Prats. X Jornada de Enfermedades Emergentes, https://www.uitb.cat/event/x-jornada-de-enfermedades-emergentes-2020/.

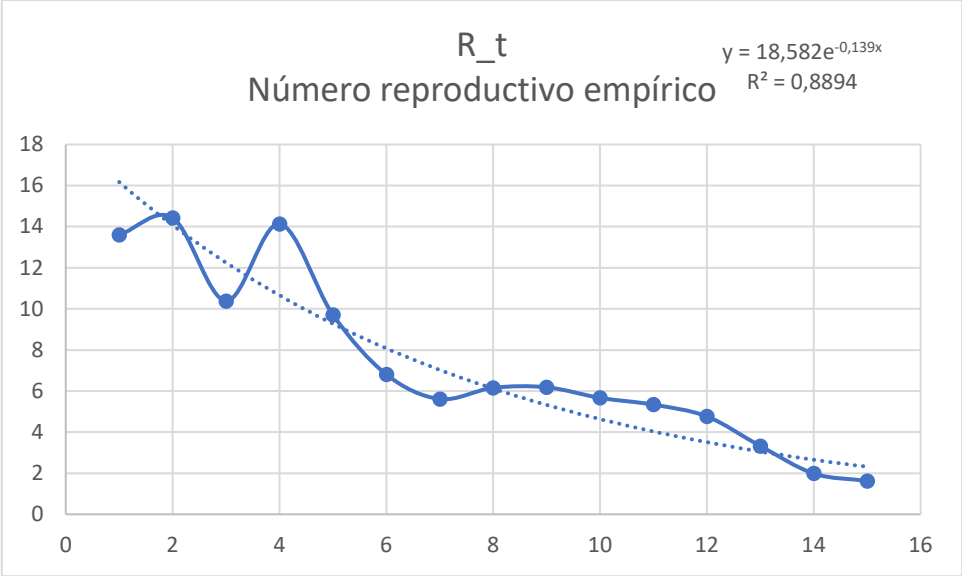
Indicadores de la situación, la tendencia y el riesgo epidemiológico

En los datos epidemiológicos se utilizan parámetros como el número reproductivo básico (R_0, valor esperado al inicio de una epidemia) o el número reproductivo específico (R_t, variable a lo largo de la epidemia). Estos números determinan el número promedio de nuevos contagios por cada caso. Si es mayor que 1, el número de casos nuevos diarios va en aumento; si es menor que 1, el número de casos nuevos diarios disminuye. Dada las dificultades de utilizar los modelos clásicos y de determinar valores característicos de la epidemia, discutidas en la Introducción, se utiliza un indicador empírico que puede mostrar una información similar. Definimos Rho (ρ_t) como el número reproductivo empírico:

$$\rho_t = \frac{N_{new}\,(t-1)+N_{new}\,(t)+N_{new}\,(t+1)}{N_{new}\,(t-6)+N_{new}\,(t-5)+N_{new}\,(t-4)} \quad (2)$$

donde N_{new} son los casos nuevos en un cierto día. Para mi-

A partir de estos 6 datos que escogimos como muestra, calculamos el N.º reproductivo empírico del virus. Después, añadimos un gráfico de dispersión para ver la evolución de la tasa de contagios. A continuación, insertamos una línea exponencial para observar hacia la tendencia del N.º de infectados. Calculamos la ecuación en el gráfico, numeramos los días desde el primer valor del N.º reproductivo empírico hasta el pico del N.º de contagiados real; utilizamos esa numeración como la variable dependiente de la ecuación del gráfico, y con ello obtenemos el N.º reproductivo empírico de nuestro modelo.
Luego, buscamos el pico de nuestro modelo; este ocurría cuando la unidad del N.º reproductivo pasa por el valor 1. En este caso, el pico sería entre el 21 de marzo y el 22 de marzo (día 21 y 22 del eje horizontal). El margen de error entre el pico real (24 de marzo) y el pico del modelo es muy pequeño.



Práctica Libre:

Prediciendo la cuarta ola

Resumen: En este proyecto se explica cómo funciona de forma matemática una pandemia. También se hace y se presenta la línea que tiende a seguir la cuarta ola del SARS-CoV-2, y recogiendo esos datos se realiza un modelo que predice la variación de infectados en Madrid después del pico de la cuarta ola. Por último, se comprueba su fiabilidad. El modelo en cuestión está realizado solo utilizando valores matemáticos, es decir, no se tendrán en cuenta factores como las vacunas, las vacaciones, el toque de queda o el aumento de viajes.

Nota: Para una mejor comprensión, se recomienda leer a la par el documento Excel que acompaña a este documento.

Palabras clave: pandemia, ola (pandemia), pico (pandemia) número de Euler, número reproductivo, media, error cuadrático medio, línea de tendencia.

Antecedentes: Una **pandemia** es la propagación mundial de una nueva enfermedad (OMS).

"El término **'ola'** no es un término científico en sí," explica María Elisa Calle […] "sino la vulgarización de un incremento de la incidencia sobre un periodo previo […]". "Es algo mucho más periodístico que técnico", añade Pedro Gullón Tosio, […]. (maldita.es)

R0 [**número reproductivo**] es el número con el que los epidemiólogos miden cómo un patógeno se multiplica a través de una población. (gizmodo.com)

"En el área de las matemáticas, podemos definir el **número e** como la base de la función exponencial natural […]. El número es […] irracional debido a que […] sus decimales son infinitos." (eustron96.com)

Una línea de tendencia es una recta que marca la tendencia de un activo. (economipedia.com)

Discusión: Ya hemos sufrido tres olas de la pandemia del SARS-CoV-2, y ha pasado más de un año desde que por primera vez oímos en el telediario el nombre de este extraño virus. A pesar de esto, la pandemia no ha acabado todavía, y ahora estamos en la llamada "cuarta ola". Muchos son los que se preguntan cuándo el número de contagiados no será lo suficientemente relevante para aparecer en las noticias. En este proyecto, se intentará hallar de forma matemática la respuesta a esta pregunta que tanto nos abruma.

1) Se seleccionan los datos (Excel: hojas 1 y 2)

Lo primero que tenemos que hacer es coger datos oficiales sobre el número de contagios con los que poder trabajar. Los que se han utilizado en este proyecto son los del Instituto Nacional Carlos III, cuya última actualización fue el 27 de abril. A continuación se abren los datos en Excel.

Como solo vamos a trabajar con los contagiados de la Comunidad de Madrid, se hace un filtro (función Filtro de Excel) y se escogen solo los datos de esa comunidad. Después se copian y pegan en otra hoja (hoja 2), para facilitar el manejo de los datos. El siguiente paso es mirar a partir de qué día el número de contagios va a ser considerado parte de la 4ª ola. Para recibir apoyo visual, se construye un gráfico con los datos (Gráfico 1). El punto de partida escogido en este proyecto es el 13 de marzo.

2) **Se hace el modelo** (Excel: hoja 3)

Para hacer el modelo, primero se cogen los datos del 13 de marzo en adelante y se pegan en una nueva hoja (hoja 3). A continuación, se hace un gráfico (Gráfico 2) para ayudarnos a encontrar el pico. Como se puede observar, la ola no es muy pronunciada, sino que tiene forma de semicírculo. Por esta razón, el pico no es muy evidente y parece estar formado por el número de contagios desde el 3 hasta el 19 de abril. Por ello, hacemos la media de los días, y nos queda que el pico es el día 12,5, es decir, el 13 de abril.

Ahora se numeran los días antes y después del pico, como se puede ver en las columnas D y E, y se hace un gráfico con los casos a la derecha y a la izquierda del pico. En las opciones del gráfico se selecciona “agregar línea de tendencia exponencial”, “presentar ecuación en el gráfico” y “presentar el valor R cuadrado en el gráfico” (gráficos 3 y 4).

La ecuación en el gráfico es la ecuación que expresa el valor de x e y en la línea de tendencia que se ha añadido. R^2 es el margen de error de la línea de tendencia respecto a la ola real. Como se puede observar, el margen de error de antes del pico es mayor, porque este tiene más variaciones que los casos después del pico.

Para calcular el número de casos según la ecuación de la recta de Excel, se tiene que escoger un valor para x, y el valor que le corresponda a y serán las personas contagiadas.

El valor de x utilizado serán los días que hemos numerado antes. Los resultados se pueden ver en las columnas G y H. A continuación, se ha calculado el número reproductivo del virus, es decir, a cuánta gente contagia una persona según nuestro modelo. Esto se obtiene dividiendo el número de casos actual entre el anterior. Si el numerador (nº de casos actual) es mayor que el anterior, significa que los casos están creciendo y el número resultante será mayor que cero. Por el contrario, si sale un número menor que cero, significa que los casos están bajando.

En este caso, antes del pico el número reproductivo es 1,02, por lo que la ola sube de forma muy lenta, y baja de la misma forma después del pico, con un 0,96 como número reproductivo.

Si seguimos añadiendo más días y calculamos el valor de y (número de contagiados) para cada uno de ellos, nos sale que el día número 173 después del pico, es decir, el dos de octubre, los contagiados en Madrid habrán alcanzado el valor cero.

3) **Se comprueba su fiabilidad** (Excel: hoja 4)

Si comparamos los casos del modelo con los de la ola real, se ve que no son exactamente iguales. Para calcular el margen de error del modelo entero, no basta con sumar el R^2 de las dos ecuaciones de la línea de tendencia y hacer la media. En este caso, lo que hay que hacer es restar el número de casos reales que hay cada día menos los casos equivalentes a ese día del modelo de Excel, y todo ello elevarlo al cuadrado. Una vez tengamos el error cuadrático medio de todos los días, se hace la media, y el número resultante es 78.695,2.

Este dato en sí no nos aporta mucha información. Si queremos saber cuán bueno es el modelo, tendremos que compararlo con otro modelo, que o bien sea el mejor de todos o bien el peor.

En este proyecto se ha decantado por el peor. Para ello necesitamos un modelo en el cual la x solo tome un valor, que será la media de los casos reales (1576,6).

Ahora se calcula el error cuadrático medio del peor modelo posible de la misma forma que se calculó el error del modelo de Excel, y hacemos la media. El número que queda es 166.847,4. Como es lógico, el error resultante es muchísimo mayor que el de nuestro modelo inicial.

Por último, tenemos que comparar estos dos errores para saber la fiabilidad de nuestro modelo. Para ello hacemos el error cuadrático medio de los errores. El resultado es un 47% (se multiplica la cifra que nos ha salido, un 0,47 por 100 para averiguar el porcentaje), pero como la comparación es con el peor modelo posible, este porcentaje nos indica lo "malo" que es nuestro modelo. Para saber el porcentaje de "acierto", le restamos a 0,47 la unidad, y nos da 0,53. Por lo tanto, este modelo tiene una tasa de acierto del 53%.

Conclusión: La intención de este proyecto es, como ya he dicho antes, explicar cómo es una pandemia de forma matemática, y con ello ver por qué es tan esencial limitar los contagios. Con modelos matemáticos o sin ellos, es importante saber que todavía queda bastante tiempo hasta que esta pandemia se quede solo como un mal recuerdo y un virus más en la lista de enfermedades. Por ello, aunque las vacunas puedan acelerar el proceso, debemos de intentar que el número reproductivo del virus baje lo máximo posible siguiendo las recomendaciones sanitarias. Si cada uno pone su granito de arena, seremos capaces de hacer que este valor no vuelva a subir del cero, y por lo tanto, poder volver a respirar tranquilos.

Anexos:

Gráfico 1:

Gráfico 2:

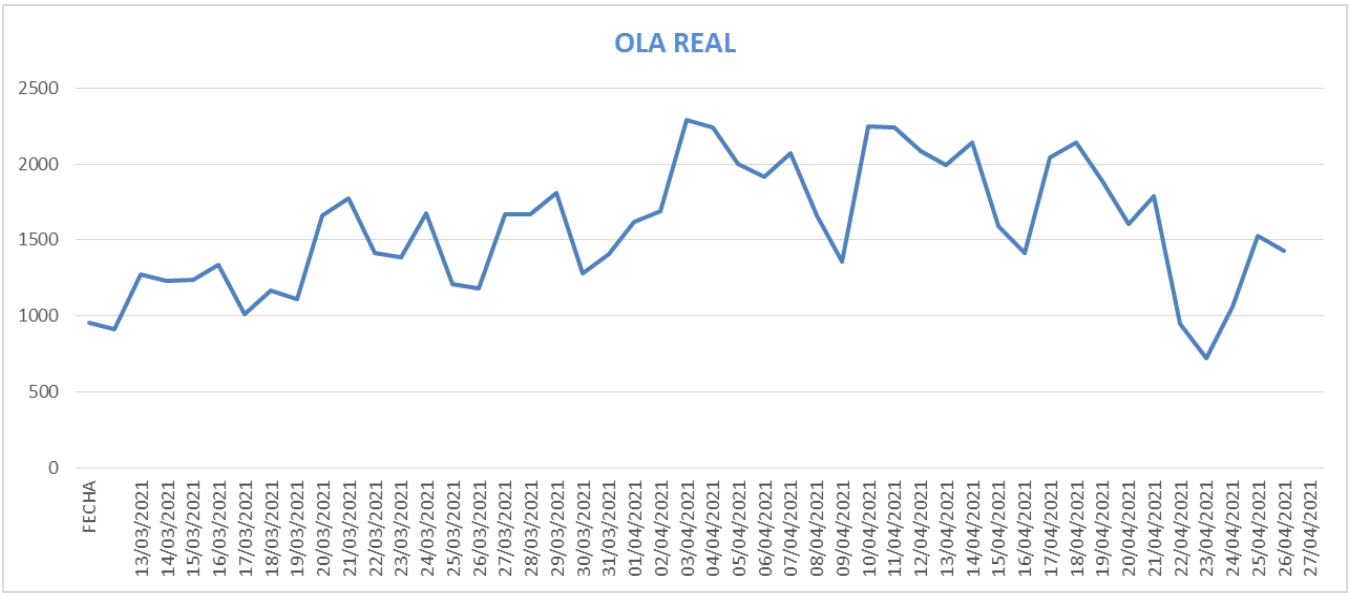


Gráfico 3:

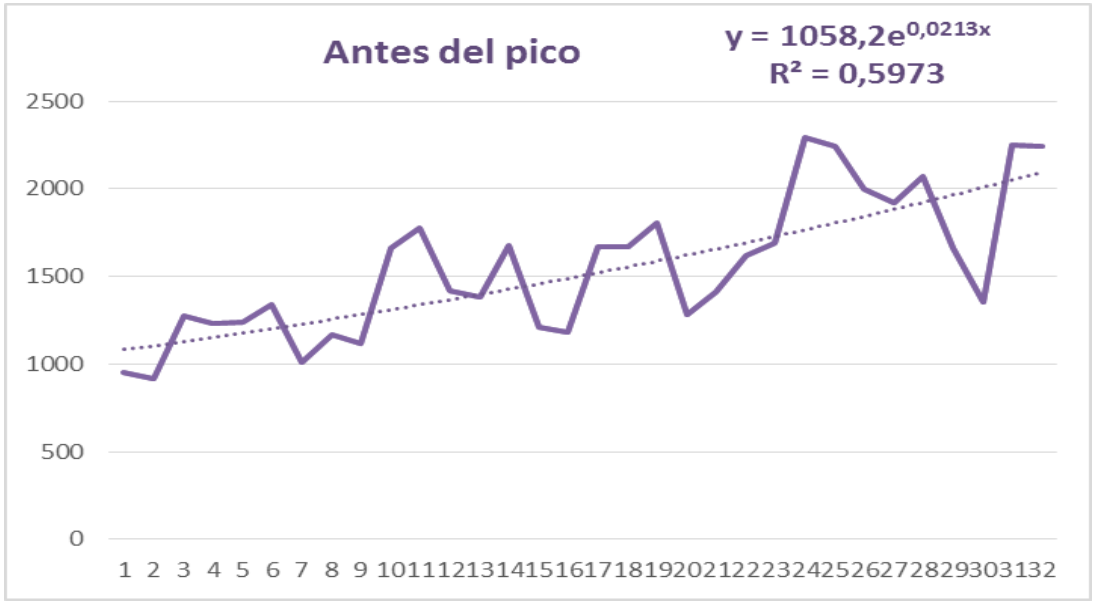


Gráfico 4:

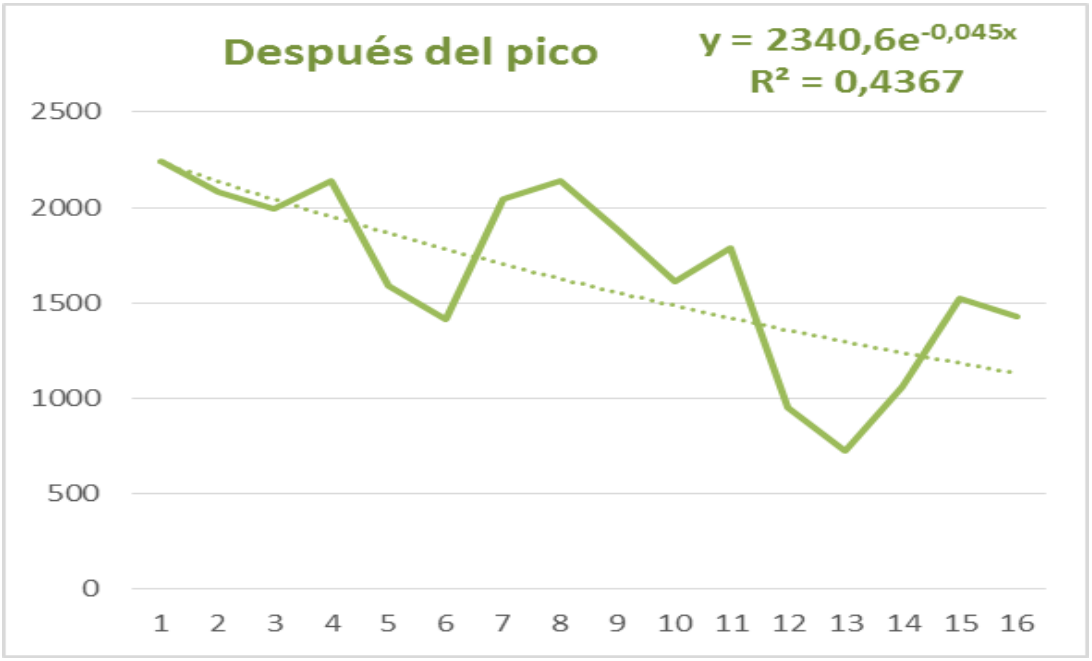
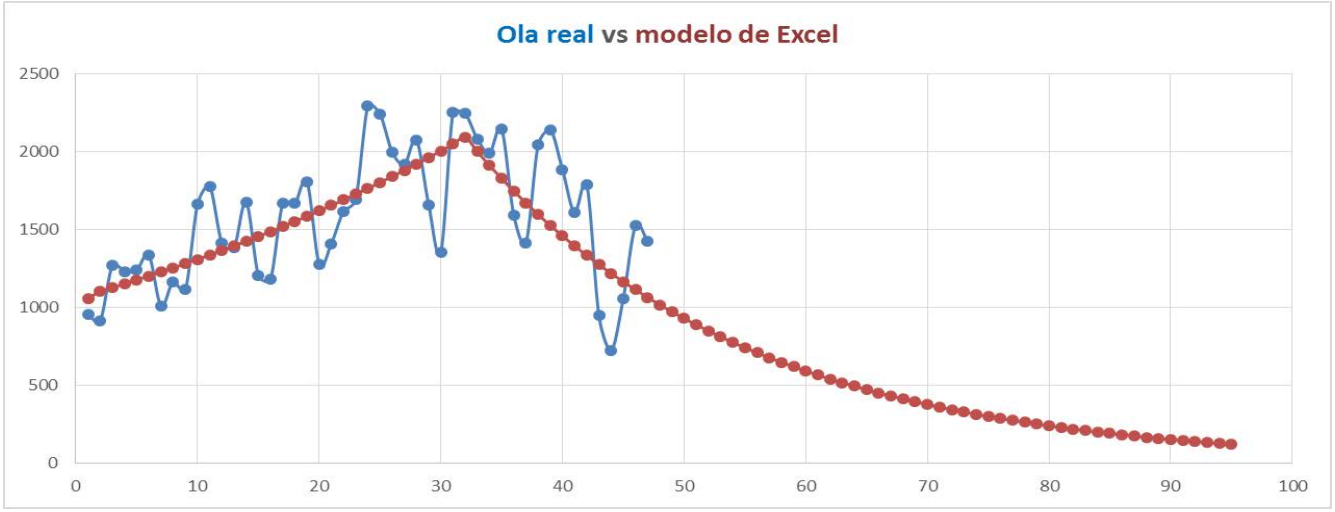


Gráfico 5



Bibliografía:

Datos (Instituto Nacional Carlos iii):

https://cnecovid.isciii.es/covid19/

https://cnecovid.isciii.es/covid19/#documentaci%C3%B3n-y-datos

Antecedentes:

https://www.who.int/csr/disease/swineflu/frequently_asked_questions/pandemic/es/

https://maldita.es/malditateexplica/20210118/ola-covid19-tercera-coronavirus/

https://es.gizmodo.com/que-es-el-numero-r0-de-un-virus-y-por-que-es-tan-import-1841299256

https://www.euston96.com/numero-e/

https://economipedia.com/definiciones/linea-de-tendencia.html

Discusión:

https://mx.hola.com/actualidad/2020021231040/coronavirus-nombre-oficial-enfermedad/