

Estimación de población contagiada por Covid-19 usando regresión Logística generalizada y heurísticas de optimización

Villalobos Arias, Mario Alberto *

6 de abril de 2020

Resumen

En este trabajos se presenta una propuesta para la estimación de la poblaciones usando ajuste de curvas del tipo logística. Este tipo de curvas se utilizan para el estudio de crecimiento de poblaciones, en este casos población de personas infectadas por el virus Covid-19; y también se puede utilizar para aproximar la curva de supervivencia que se utiliza en estudios actuariales y otras similares.

Keywords: Heurísticas de optimización, regresión logística generalizada, ajuste de curvas, covid-19.

1. Introducción

Las curvas de crecimiento de poblaciones siguen el conocido comportamiento logístico como se muestra en la figura 1.

Un modelo que se utiliza para el ajuste de curvas de poblaciones es el Logístico, que se utiliza la siguiente ecuación

$$P(t) = \frac{1}{1 + e^{-at+b}} \quad (1.1)$$

Como se ve este modelo tiene varios problemas entre ellos que los datos están en $[0, 1]$, y no es flexible, la ventaja es que este modelo es que se puede obtener una aproximación de la solución óptima al transformar los datos y utilizar regresión lineal.

Como se ve en el gráfico 2 al aplicar regresión logística y al dividir todos los datos entre el máximo valor (66818) se obtiene una curva que ajusta bien, con un $R^2 = 0,99955$, que

*Universidad de Costa Rica, CIMPA y Escuela de Matemática, San José, Costa Rica, mario.villalobos@ucr.ac.cr

Instituto tecnologico de Costa Rica, Escuela de Matemática, Cartago, Costa Rica, marvillalobos@itcr.ac.cr

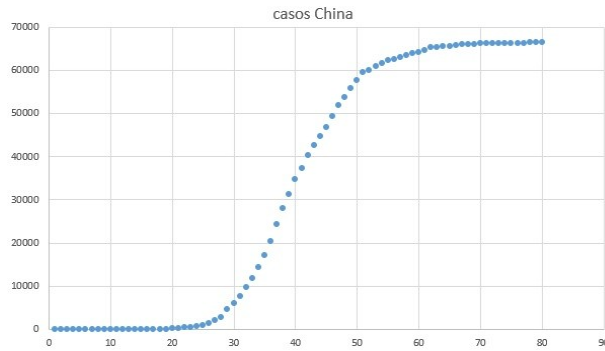


Figura 1: casos totales de contagiados en China

es muy bueno, desde el punto vista estadístico, pero como se ve en la figura hay muchos valores que no ajustan muy bien en las curvas y no es bueno para la predicción, como veremos más adelante.

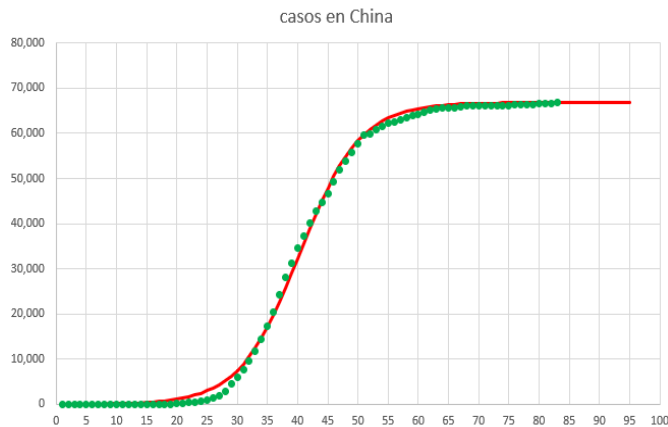


Figura 2: Ajuste con regresión logística totales de contagiados en China

2. El modelo modelo SIR

La versión clásica para el estudio de epidemias es el modelo modelo SIR en el cual se divide a la población en tres grupos: los susceptibles, los infectados y los recuperados (SIR), esto en el caso más simple y que la población va cambiando de susceptible a infectados y luego recuperados.

Se supone que el decrecimiento es proporcional a la cantidad de infectados multiplicado por la misma cantidad de susceptibles. el cambio en los recuperados es igual a cierto porcentaje

de los infectados y finalmente la cantidad de infectados va a cambiar aumentando por los susceptibles y luego le quitamos la cantidad de los infectados que pasa recuperados. De esta manera se obtienen las siguientes ecuaciones

$$\begin{aligned}\frac{dS}{dt} &= -\frac{\beta IS}{N}, \\ \frac{dI}{dt} &= \frac{\beta IS}{N} - \gamma I, \\ \frac{dR}{dt} &= \gamma I,\end{aligned}$$

con

$$\frac{dS}{dt} + \frac{dI}{dt} + \frac{dR}{dt} = 0,$$

que nos da

$$S(t) + I(t) + R(t) = \text{constant} = N,$$

El problema con el modelo SIR es que como vemos se necesitan los tres parámetros que se tienen en las ecuaciones diferenciales pero con pocos datos o digas en pocos días es muy difícil determinar esos parámetros

3. Regresión Logística generalizada

Por lo anterior se proponen la utilización de un modelo un poco más complejo a la regresión logística y que sea más fácil de determinar que el SIR

Una primera versión es:

$$P(t) = \frac{M}{1 + e^{-at+b}} \quad (3.1)$$

que agrega un parámetro más, por determinar, que es la población límite M , ya con esta modificación no se puede resolver mediante transformación y regresión lineal. Para resolver este problema se debe utilizar técnicas no lineales de optimización para resolverlo, además tiene el problema de rigidez, esto es que no se ajusta suficientemente a ciertas curvas, este modelo se puede utilizar para predicción, pero desde el punto de ajuste no es tan preciso como veremos más adelante.

Para flexibilizar la curva se agrega un parámetro extra α de la siguiente forma

$$P(t) = \frac{M}{(1 + e^{-at+b})^\alpha}. \quad (3.2)$$

Este parámetros α agrega flexibilidad en el ajuste, recuérdese las gráficas de $y = x^{1/3}$, $y = x^{1/2}$, $y = x$, $y = x^2$, $y = x^3$.

Para esta función el punto de inflexión es cuando

$$P''(t) = a^2 c e^{at+b} (e^{at+b} + 1)^{-c-2} (c e^{at+b} - 1) = 0$$

por lo que se tiene que el punto de inflexión se obtiene para

$$t = -\frac{\ln(\alpha) + b}{a}$$

Este es el mismo modelo conocido como la curva de Richards que se utiliza para modelar crecimientos de poblaciones.

$$Y(t) = A + \frac{K - A}{(C + Qe^{-Bt})^{1/\nu}} \quad (3.3)$$

Note que con algunos cálculos y $A = 0$ se da la igualdad con la logística generalizada. El caso:

$$Y(t) = \frac{K}{(1 + Qe^{-\alpha\nu(t-t_0)})^{1/\nu}}$$

que es solución de la ecuación diferencial:

$$Y'(t) = \alpha \left(1 - \left(\frac{Y}{K} \right)^\nu \right) Y$$

3.1. Función de Gompertz

debe su nombre a Benjamin Gompertz, el primero en trabajar en este tipo de función es un caso particular de la de Richards, y tiene por ecuación la siguiente:

$$G(t) = ae^{-be^{-ct}},$$

Además su segunda derivada es:

$$G''(t) = bc^2 e^{be^{cx} + cx} (1 + be^{cx}) = 0$$

que nos da el punto de inflexión se alcanza en $t = -\log(-b)/c$ lo que en el caso de epidemias nos dice en que punto el crecimiento de los casos diario va a iniciar a descender. Esta es una función más simple pues solo tiene 3 parámetros en vez de la LG, que tiene 4, y por ende va a tener menos óptimos locales.

3.2. Otras versiones

Otras versiones más complejas son

$$P(t) = \frac{M + ct}{(1 + e^{-at+b})^\alpha}. \quad (3.4)$$

Al momento de escribir este trabajo, los datos de covid-19, de Korea del sur se tiene que ajustar con una curva como esta.

Y la siguiente que se utilizó para ajustar las curvas de supervivencia

$$P(t) = \frac{M}{(1 + e^{-at^2+bt+c})^\alpha}, \quad (3.5)$$

4. Hipótesis y propuesta

La propuesta de este trabajo es, primero utilizar la curva logística generalizada o la curva de Gompertz para hacer un ajuste de los datos en los que se se tengan la curva casi completa, por ejemplo datos de China y de Corea del Sur, (30 de marzo)

Por otro lado cuando se tengan todos los datos existe una curva del tipo de regresión logística generalizada o de Gompertz que ajusta los datos.

Por lo que la hipótesis aquí es: Si yo tengo la parte bajo de la Curva, es decir los primeros valores de la curva, yo puedo obtener los parámetros de la curva, y obtener la curva completa. Y con esto se puede utilizar para predecir el crecimiento de la población, en este caso, por ejemplo, el número total de casos por el covid-19 en un país o región y cuando se llega al punto de inflexión, es decir, momento en que el número de casos diarios empieza a descender.

Más específicamente si se tiene los datos de unos 20 a 30 ó 35 días la pregunta es:

¿se puede determinar los parámetros de la curva que ajusta los datos completos?

Si esto se logrará como vemos tendríamos una forma de predecir el comportamiento del crecimiento de poblaciones con sólo tener unos pocos días, es decir podemos determinar los parámetros de la curva con pocos días, que es muy difícil con el modelo SIR determinar los parámetros de las ecuaciones diferenciales.

5. Ajuste de curvas

Para este trabajo se están utilizando los datos provistos por el European Centre for Disease Prevention and Control en la página para bajar los datos diaios ver [1].

Lo primero que se presenta es verificar, como se sabe, que la curva LG ajusta muy bien los datos de covid-19. Para el caso de China y Korea del Sur, que son los que se tiene la curva casi completa.

Utilizando un algoritmo de optimización no lineal se tienen los siguientes resultados.

5.1. China: datos original

Para los datos de China se obtiene los siguientes datos y el gráfico en la figura 3.

<i>fecha</i>	<i>M</i>	<i>a</i>	<i>b</i>	<i>α</i>
31/03/2020	81149	-0,2348563	9,89996092	0,8809852

Con $R^2 = 0,998878$. Como se observa se obtiene una muy buena aproximación, pero como sabemos y se ve en el la figura 3, hay un salto en esos datos, por lo que se decidió corregir ese salto poniendo del dato del día 13/02/2020 igual al día anterior y el del día 14/02/2020 igual al día siguiente.

Con estas correcciones se obtiene los siguientes datos y el gráfico de la figura 4

<i>fecha</i>	<i>M</i>	<i>a</i>	<i>b</i>	<i>α</i>
31/03/2020	66871	-0,15032863	4,21103499	4,436336

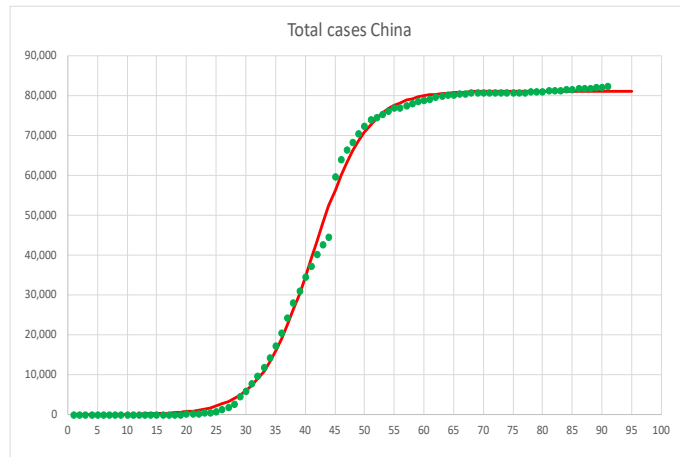


Figura 3: Ajuste con regresión logística totales de contagiados en China

En este caso se obtiene un $R^2 = 0,999885$, que es mejor al de los datos sin corrección. Observe que en los últimos datos se ve que hay una tendencia lineal, esto se puede mejorar como se ve en el caso de Corea del Sur, como se ve en la siguiente subsección.

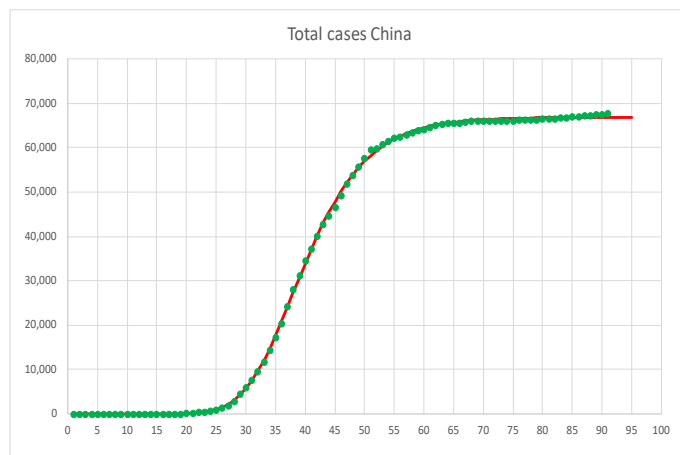


Figura 4: Ajuste con regresión logística generalizada en China **datos corregidos**

5.2. Corea del Sur: Datos Originales

Al aplicar el método a los datos de Corea del Sur se obtiene los datos siguientes y el gráfico de la figura 5.

<i>fecha</i>	<i>M</i>	<i>a</i>	<i>b</i>	<i>α</i>
31/03/2020	10079	-0,16416438	0,18353146	874,692503

Y se obtiene un $R^2 = 0,99875$, que no es muy bueno como se ve en la figura 5 que el ajuste no es muy bueno.

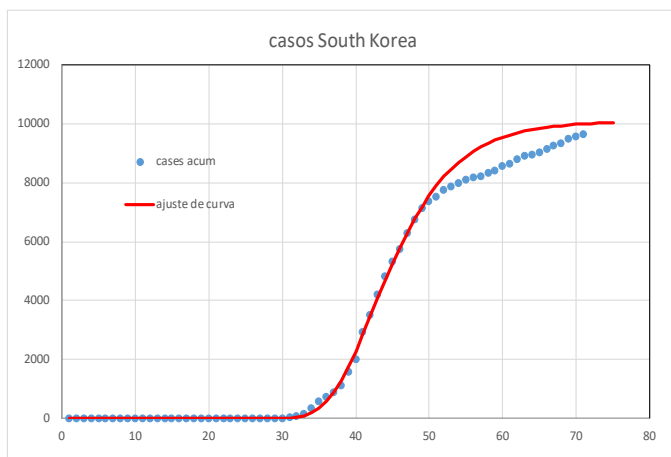


Figura 5: Ajuste con regresión logística generalizada Corea del Sur

Para mejorar este ajuste se propone utilizar la curva de ajuste (3.4), a saber:

$$P(t) = \frac{M + ct}{(1 + e^{-at+b})^\alpha}.$$

Con esta curva se logra mejorar el ajuste al obtener los siguientes datos:

<i>fecha</i>	<i>c</i>	<i>M</i>	<i>a</i>	<i>b</i>	<i>α</i>
31/03/2020	92,63407116	3046	-0,36922771	15,5617475	0,9691940884

Y ahora el $R^2 = 0,99988$ que es mejor al resultado anterior y como se ve en el gráfico de la figura 6.

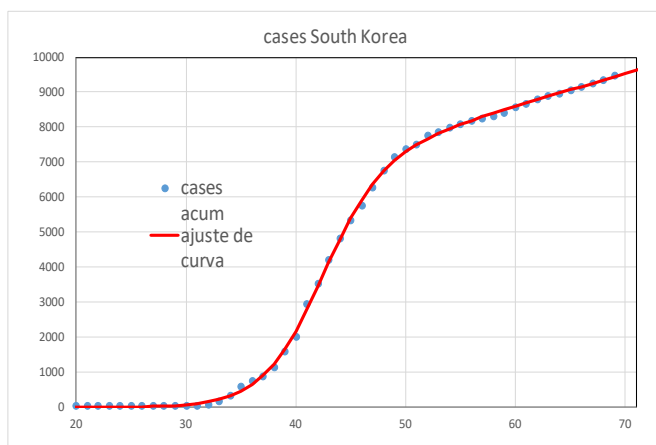


Figura 6: Ajuste con regresión logística generalizada de contagiados en Corea del Sur **con tendencia final lineal**.

Como se ve el ajuste con este tipo de curvas logísticas generalizadas es muy bueno obteniendo R^2 superiores a 0.9999.

6. Pruebas de Predicción

Como se vio en la sección anterior las curvas logísticas aproximan muy bien, como se ha demostrado en trabajos anteriores (por ejemplo [2, 4]).

En esta sección vamos a tratar de medir el porcentaje de error relativo que se comete al utilizar las curvas propuestas, se utilizará una curva de prueba y los datos de China y Corea del Sur

Curva de prueba

Para tratar de validar la hipótesis vamos a generar una curva con valores

$$\begin{array}{cccc} M & a & b & \alpha \\ \hline 2000 & -0,05 & -2 & 10 \end{array}$$

y al utilizar el algoritmo de optimización se obtiene lo siguientes cotas superiores para el error relativo al usar la función LG, con los días indicados:

PORCENTAJE DE ERROR RELATIVO EN CURVA DE PRUEBA

USANDO FUNCION LOGÍSTICA GENERALIZADA

número de días	máx error rel.
25 días	10.451 %
30 días	4.801 %
35 días	1.825 %
40 días	1.188 %

Al utilizar la función de Gompertz se mejora notablemente la predicción, como se había mencionado por tener menos parámetros, al utilizar el ejemplo con

$$\begin{array}{ccc} a & b & c \\ \hline 2000 & -2 & 10 \end{array}$$

Al ejecutar el método se obtiene los siguientes porcentajes de error relativo

PORCENTAJE DE ERROR RELATIVO EN CURVA DE PRUEBA

USANDO GOMPERTZ

número de días	máx error rel.
25 días	1.6697 %
30 días	0.7492 %
35 días	0.8928 %
40 días	0.3234 %

Como se ve con la función de Gompertz, al parecer, se puede garantizar mejor predicciones.

Datos de China

Veamos ahora el caso de China, en el que se tiene la curva casi completa. Como vemos al utilizar los días del 16 al 45 (20 días), observe que los primeros 16 días casi no hubo cambios en los datos por lo que no se utilizaran esos datos, con estos datos con LG se obtiene:

Gompertz

Al utilizar los días del 16 al 45 (20 días) se obtiene un error máximo de 9.91 % (promedio de 3 corridas, básicamente da el mismo valor)

Al utilizar los días de 15 al 55 (30 días) se obtiene el gráfico de la figura 7, al medir los errores relativos se obtiene un valor máximo de 4.49 % (promedio de 3 corridas), en la predicción de los datos hasta el 31/03/2020.

PORCENTAJE DE ERROR RELATIVO PREDICCIÓN CHINA USANDO GOMPERTZ	
número de días	máx error rel.
20 días	9.91 %
25 días	4.49 %

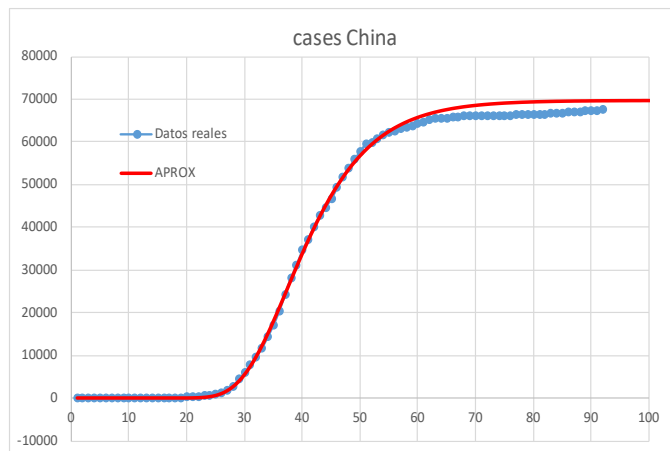


Figura 7: Predicción con Gompertz, China, 30 días

Datos de Corea del Sur

Para Corea del sur vamos a utilizar los datos a partir del día 28 desde el primer caso (16/02/202), este caso recuérdese que estos datos tiene una tendencia lineal en los últimos días, ver figura 8.

PORCENTAJE DE ERROR RELATIVO PREDICCIÓN SOUTH KOREA

USANDO GOMPERTZ	
número de días	máx error rel.
20 días	17.83 %
25 días	6.49 %

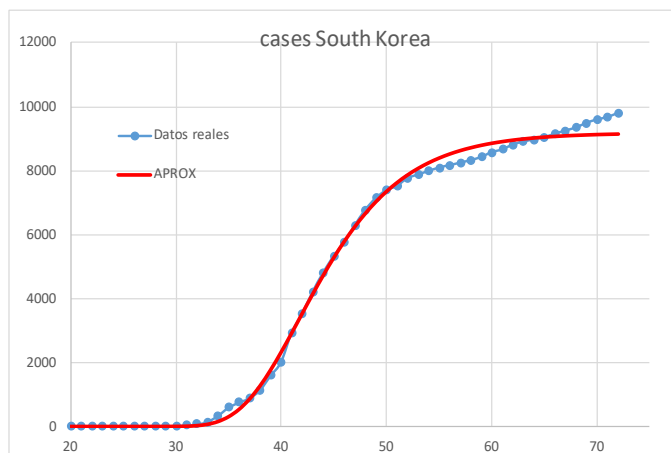


Figura 8: Predicción con Gompertz, South Korea, 25 días

7. Resultados

Como resultados de este trabajo y como hemos visto se utilizará este método para dar una predicción para los datos de algunos países empezando por Costa Rica y luego se presentaran los datos de Italia, España, y con los ajustes que tenemos ya se tiene el de China y Corea del Sur.

7.1. China y Corea del Sur

Para el caso de China y Corea del Sur se puede usar los resultados obtenidos en la subsecciones 5.1 y 5.2.

7.2. Costa Rica

Para Costa Rica se tienen los siguientes datos a partir del día 6 de marzo en que se detectó el primer caso.

DATOS SOBRE COVID-19 EN COSTA RICA

Fecha	casos diarios	Total de casos	Fecha	casos diarios	Total de casos	Fecha	casos diarios	Total de casos
06/03/20	1	1	15/03/20	8	35	24/03/20	19	177
07/03/20	4	5	16/03/20	6	41	25/03/20	24	201
08/03/20	4	9	17/03/20	9	50	26/03/20	30	231
09/03/20	4	13	18/03/20	19	69	27/03/20	32	263
10/03/20	4	17	19/03/20	18	87	28/03/20	32	295
11/03/20	5	22	20/03/20	26	113	29/03/20	19	314
12/03/20	1	23	21/03/20	4	117	30/03/20	16	330
13/03/20	3	26	22/03/20	17	134	31/03/20	17	347
14/03/20	1	27	23/03/20	24	158	01/04/20	28	375
06/03/20	1	1	15/03/20	8	35	24/03/20	19	177
07/03/20	4	5	16/03/20	6	41	25/03/20	24	201
08/03/20	4	9	17/03/20	9	50	26/03/20	30	231
09/03/20	4	13	18/03/20	19	69	27/03/20	32	263
10/03/20	4	17	19/03/20	18	87	28/03/20	32	295
11/03/20	5	22	20/03/20	26	113	29/03/20	19	314
12/03/20	1	23	21/03/20	4	117	30/03/20	16	330
13/03/20	3	26	22/03/20	17	134	31/03/20	17	347
14/03/20	1	27	23/03/20	24	158	01/04/20	28	375

Con estos datos al hacer el ajuste con la función LG se obtiene los parámetros siguientes, ya con esta cantidad de datos básicamente se tiene un sólo óptimo:

PARÁMETROS OBTENIDOS PARA COSTA RICA

USANDO LOGÍSTICA GENERALIZADA

M	a	b	c
886	-0.0780135	-4.91608521	1023.90189

Este ajuste tiene un $R^2 = 0,99858343$

En el gráfico de la figura 9 se ve la predicción de casos, se puede ver el número de casos diarios en el recuadro superior y el ajuste de los datos reales en el recuadro inferior.

Se incluyen 2 líneas de cota superior e inferior para la predicción, que se supone un 7% de error en los datos tomado a partir de los resultados de la sección 6.

Se ve además que de seguir las situaciones como están al día de hoy habr un tope de aproximadamente 1000 casos, esto es desde el punto de vista matemático o ajuste de curvas.

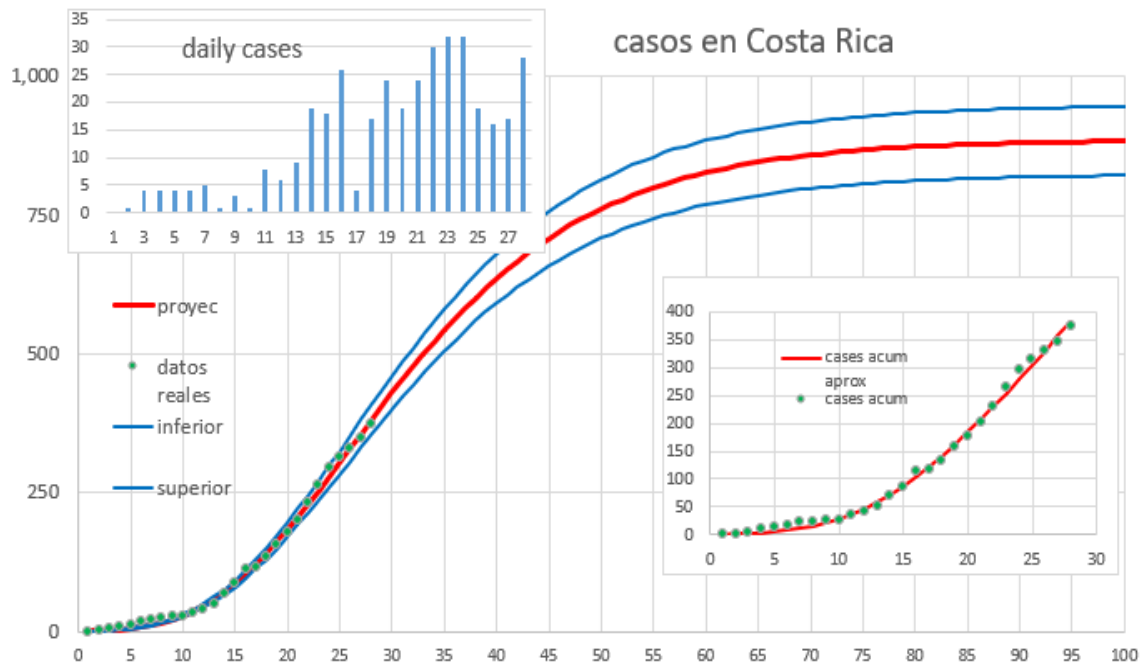


Figura 9: Predicción con LG, Costa Rica

Gompertz

Con la curva de Gompertz se obtiene básicamente el mismo resultado como se ve a en lo que sigue:

PARÁMETROS OBTENIDOS PARA COSTA RICA		
USANDO GOMPERTZ		
a	b	c
887	-6.923348908	-0.077891632

Este ajuste tiene un $R^2 = 0,99853504$, en la figura 10 se puede ver el gráfico resultante. Debe notarse que los resultados de utilizar estas 2 funciones se han ido aproximando entre ellas conforme pasan los días y para muestra se presenta el valor máximo entre ellas en la siguiente tabla.

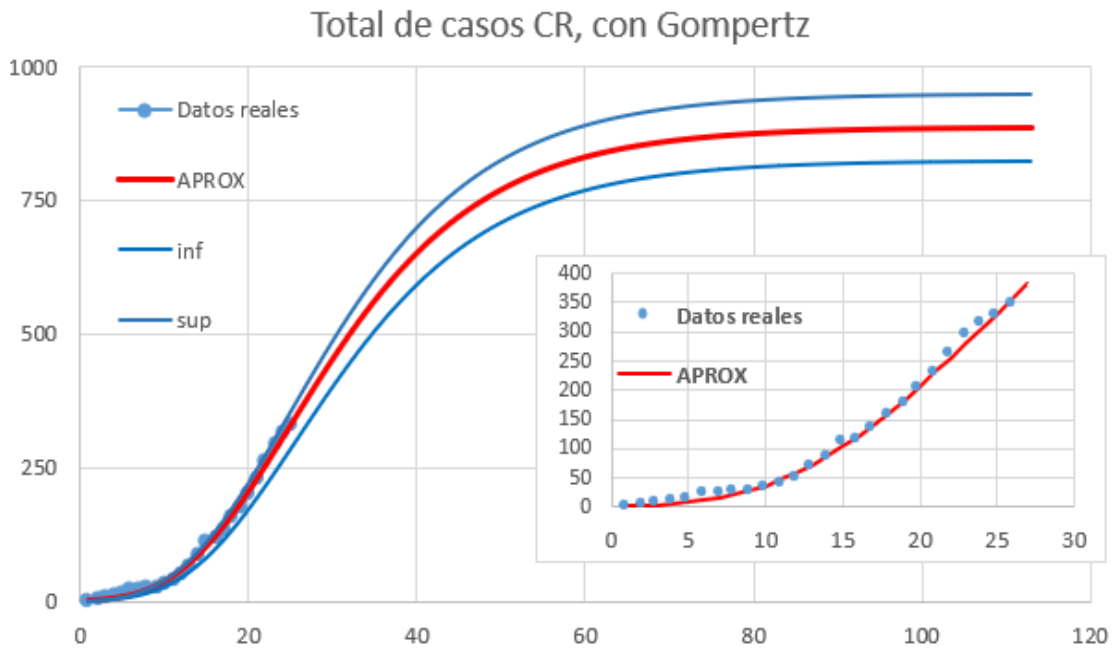


Figura 10: Predicción con Gompertz, Costa Rica

VALOR MÁXIMO ALCANZADO POR LAS CURVAS
SEGÚN DÍA PARA COSTA RICA

Fecha	Gompertz	LG
23/03/20	2162	1983
24/03/20	1445	1359
25/03/20	1310	1274
26/03/20	1583	1505
27/03/20	2048	1974
28/03/20	2426	2343
29/03/20	1743	1730
30/03/20	1193	1185
31/03/20	924	922
01/04/20	887	886

7.3. Italia

Para Italia se presenta los resultados por Gompertz y con los datos al 30/3/2020.

PARÁMETROS OBTENIDOS PARA ITALIA USANDO GOMPERTZ		
a	b	c
261 052	-43.77482253	-0.063450536

Este ajuste tiene un $R^2 = 0,99968$, en la figura 11 se puede ver el gráfico resultante.

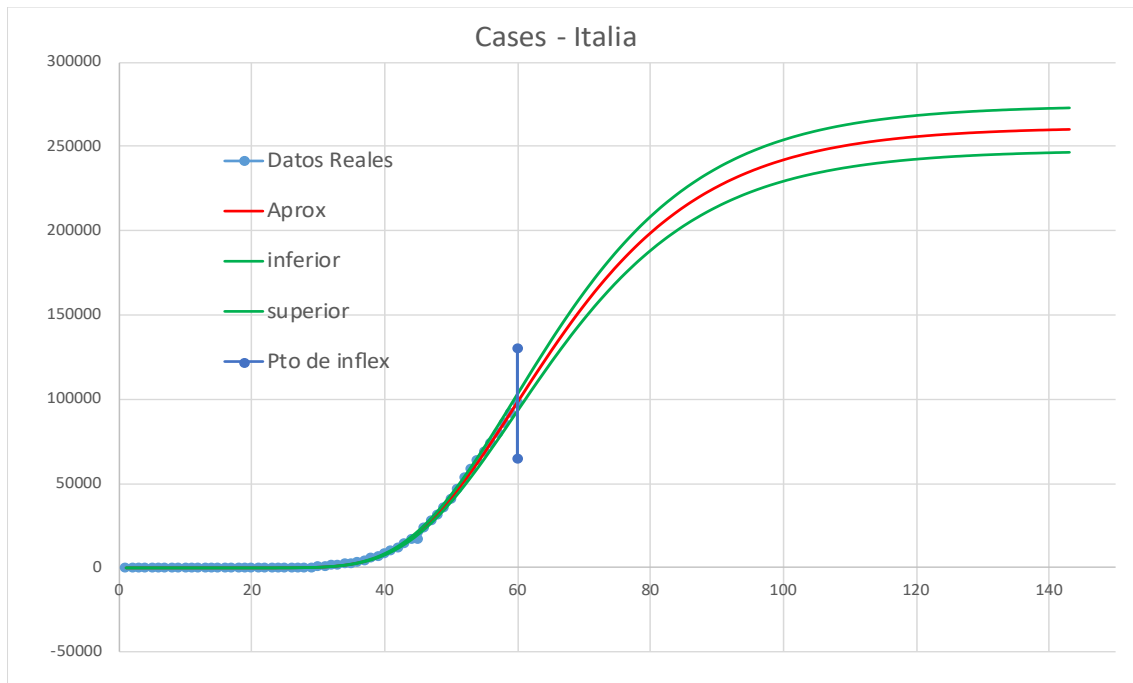


Figura 11: Predicción con Gompertz, Italia

7.4. España

Para España se presenta los resultados por Gompertz y con los datos al 31/3/2020.

PARÁMETROS OBTENIDOS PARA ESPAÑA		
USANDO GOMPERTZ		
a	b	c
468 495	-69.20043418	-0.061995389

Este ajuste tiene un $R^2 = 0,999410067$, en la figura 12 se puede ver el gráfico resultante.

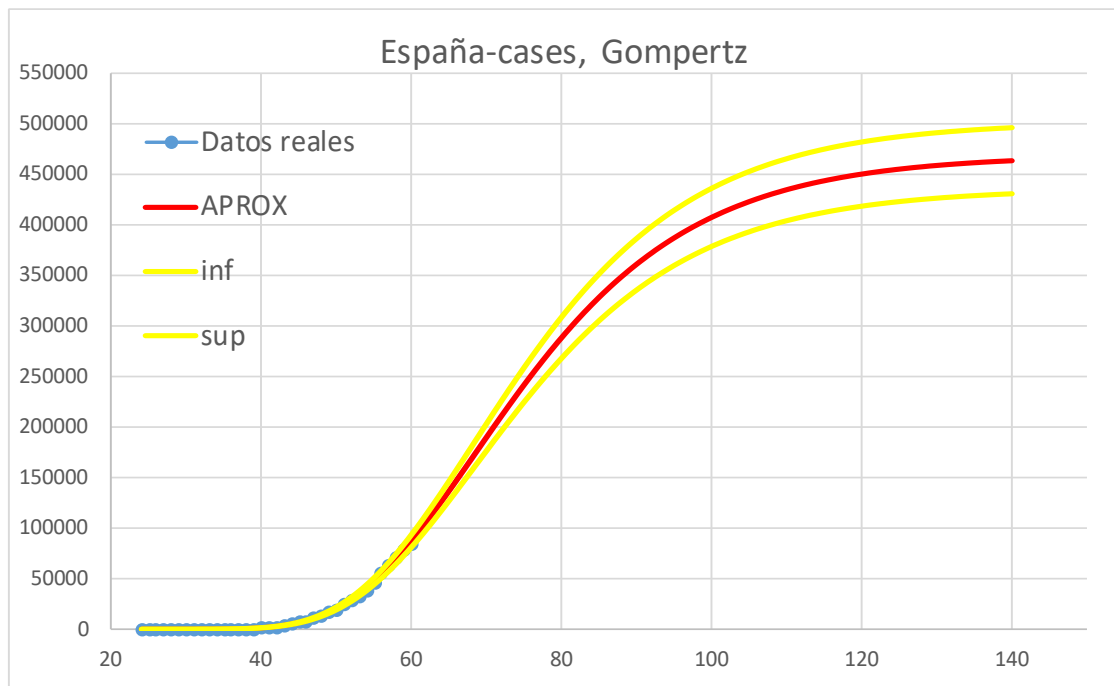


Figura 12: Predicción con Gompertz, España

7.5. Otros países

Se tienen los resultados de otros países que se van a ir incorporando más adelante

8. Conclusiones y trabajo futuro

Con este trabajo se muestra como se puede ajustar curvas de crecimiento de poblaciones, utilizando las funciones LG y Gompertz, inclusive en el caso más general como en los datos de Corea del Sur, sección 5.2, donde se está haciendo de la logística con recta. Se ve que estos métodos podrían ser utilizados para predecir el crecimiento de poblaciones, en este caso de personas infectadas por el Covid-19, y podrían ayudar a los expertos en pandemias para tomar las medidas necesarias.

8.1. Trabajo futuro

Se necesita hacer más estudios para afinar los resultados de este trabajo

Ver la posibilidad de usar curvas semejantes para la aproximación de otro tipo de datos.

Referencias

- [1] "Download today's data on the geographic distribution of COVID-19 cases worldwide", European Centre for Disease Prevention and Control, <https://www.ecdc.europa.eu/en/publications-data/download-todays-data-geographic-distribution-covid-19-cases-worldwide>
- [2] Fekedulegn, Desta; Mairitin P. Mac Siurtain; Jim J. Colbert (1999). "Parameter Estimation of Nonlinear Growth Models in Forestry" (PDF). *Silva Fennica*. 33 (4): 327-336. Archived from the original (PDF) on 2011-09-29.
- [3] Kirckpatrick, S.; Gelatt, J. Vecchi, P. 1983. "Optimization by simulated annealing", *Revista Science*. Volumen 220, número 4598.
- [4] Pella, J. S.; Tomlinson, P. K. (1969). "A Generalised Stock-Production Model". *Bull. Inter-Am. Trop. Tuna Comm.* 13: 421-496.
- [5] Richards, F. J. (1959). "A Flexible Growth Function for Empirical Use". *Journal of Experimental Botany*. 10 (2): 290-300. doi:10.1093/jxb/10.2.290