ISSN 1870-4069

Optimización de los coeficientes del modelo predictivo del número de casos diarios de coronavirus Covid-19 en México

Ana C. Zenteno, María del C. Santiago, Yeiny Romero, Judith Pérez, Gustavo T. Rubín, Antonio E. Álvarez

Benemérita Universidad Autónoma de Puebla,
Facultad de Ciencias de la Computación,
México

{ana.zenteno, marycarmen.santiago, yeiny.romero, judith.perez, gustavo.rubin }@correo.buap.mx, antonio.alvarez@alumno.buap.mx

Resumen. El coronavirus covid-19 ha sorprendido a todos, millones de personas en todos los países y regiones han cambiado sus vidas con consecuencias inimaginables además de las lamentables muertes que ha cobrado. Al ser una pandemia muy grande, se requieren modelos matemáticos para simular escenarios y proporcionar a los tomadores de decisiones información más precisa basada en las variables de comportamiento actuales. En este trabajo partimos del análisis y procesamiento de los reportes oficiales diarios del número de casos positivos confirmados, cada uno se aproxima a un polinomio, se optimizan los coeficientes de éste y se retroalimenta el polinomio inicial. Con esta metodología se encontró un mecanismo para predecir el número de casos positivos hasta por 3 días con un error menor de 19.9%, con el fin de mejorar esta estimación se aplicará metodología para procesar el número de casos en estados y municipios, y contar con un modelo que brindará una predicción de casos hasta 2 semanas después del último reporte de casos y se actualizará diario incrementando la base de datos, disminuyendo el error.

Palabras clave: Covid-19, Modelo, Estadística.

Optimization of the Coefficients of the Predictive Model of the Number of Daily Cases of Coronavirus Covid-19 in Mexico

Abstract. The coronavirus covid-19 has surprised everyone, millions of people in all countries and regions have changed their lives with unimaginable consequences in addition to the unfortunate deaths it has claimed. Being a very large pandemic, mathematical models are required to simulate scenarios and provide decision makers with more accurate information based on current behavioral variables. In this work we start from the analysis and processing of the daily official reports of the number of confirmed positive cases, each one

approaches a polynomial, its coefficients are optimized and the initial polynomial is fed back. With this methodology, a mechanism was found to predict the number of positive cases for up to 3 days with an error of less than 19.9%, in order to improve this estimate, a methodology will be applied to process the number of cases in states and municipalities, and have a model that will provide a prediction of cases up to 2 weeks after the last case report and will be updated daily increasing the database, decreasing the error.

Keywords: Covid-19, Model, Statistics.

1. Introducción

El surgimiento de casos de virus en china en diciembre de 2019, particularmente en Wuhan (Hubei) y vinculado a un mercado mayorista de marisco, pescado y animales vivos ha propiciado una ola de investigaciones del desarrollo de la pandemia en todos los países. El 31 de diciembre de 2019, la Organización Mundial de la Salud (OMS) recibió reportes de varios casos de neumonía de etiología desconocida. A fecha de 3 de enero de 2020, las autoridades nacionales de China notificaron a la OMS que, en total, existían 44 pacientes con neumonía de etiología desconocida, de entre cuales, 11 pacientes estaban gravemente enfermos, mientras que los 33 pacientes restantes se encontraban en situación estable.

Según informaciones difundidas en los medios de comunicación, el mercado implicado en Wuhan se cerró el 1 de enero de 2020 por saneamiento y desinfección ambiental [1]. El día 8 de enero se Tailandia detectó un primer caso (fuera de China), siendo el 10 de enero el día que se presenta el primer fallecimiento causado por el virus. El incremento de los casos que aparecen en China y en otros países pone en evidencia la gravedad de la situación y la OMS el 10 de enero publica orientaciones técnicas y recomendaciones para todos los países sobre el modo de detectar casos, realizar pruebas de laboratorio y gestionar los posibles casos. Para el 30 de enero la OMS señala la existencia de un total de 7818 casos confirmados en todo el mundo, la mayoría de ellos en China y 82 en otros 18 países. La OMS evalúa el riesgo en China como muy alto y el riesgo mundial como alto [2].

Diferentes esfuerzos se están llevando a cabo en la intención de modelar el número de casos sospechosos, confirmados, decesos, ocupación de instalaciones hospitalarias, formas y patrones de contagio, entre otras variables. Un equipo de trabajo del Centro de Investigación y Docencia Económicas (CIDE) liderado por el profesor Alarid-Escudero desarrollaron un modelo matemático de proyecciones sobre los efectos de las distintas prácticas de mitigación sobre COVID-19. El modelo denominado SC-COSMO (Stanford-CIDE COronavirus Simulation MOdel) es un modelo matemático epidemiológico de cómo evoluciona la enfermedad y modela también los mecanismos en los que los individuos interactúan entre sí [3]. Incorpora análisis demográficos para considerar a los individuos susceptibles, a los expuestos, los infectados y los recuperados, los analiza conforme a los patrones de contacto para la transmisión. A partir de este análisis, se puede calcular la evolución de cómo las personas se contagian y desarrollan la enfermedad.

Al revisar las cifras de casos confirmados, se miden día a día los casos reales de infectados y la capacidad para detectarlos por parte de los gobiernos [4]. El tiempo es

un factor que permite observar los casos que mejoran, los decesos y la evolución en duplicación y las tendencias que siguen los casos a nivel local y global.

El "Covid-19 Modelo numérico de casos de infección y estimaciones epidémicas modelo asimétrico -Gompertz" [5] que por medio de la ecuación Gompertz obtiene una estimación de la demanda hospitalaria incluso para casos de terapia intensiva en España. Es un claro ejemplo de los beneficios del confinamiento y que permite a la sociedad tener una mejor atención médica en caso de ser contagiado.

Sitios en internet como Worldometer analizan, validan y agregan datos de miles de fuentes en tiempo real y proporcionan estadísticas globales COVID-19 para una amplia audiencia en todo el mundo [6]. Los análisis que se pueden realizar generan proyecciones distintas para casos locales.

En Korea, un estudio pretende identificar el patrón de transmisión local de COVID-19 utilizando modelos matemáticos para predecir el tamaño de la epidemia y el momento del final de la propagación, teniendo como resultados una estimación de que el número de transmisiones por paciente infectado era aproximadamente 10 veces mayor en el área de Daegu / Gyeongbuk que el promedio de todo el país [7].

Predecir el desarrollo de una epidemia no es tarea fácil, y menos cuando se tiene un virus emergente. La validez de la mayoría de los modelos predictivos se basa en numerosos parámetros, que involucran características biológicas y sociales a menudo desconocidas o altamente inciertas [8].

El análisis de los datos es relevante ya que ayuda a los gobiernos de los países a tomar decisiones sobre el proceso de transmisión de la Enfermedad COVID-19 por medio de predicción hacia adelante y de inferencias hacia atrás de la situación epidémica [9].

Existen fuentes que confirman que el aislamiento de personas expuestas e infectadas, y la reducción de la transmisión y la tasa de retorno de la estadía en el hogar pueden mitigar las pandemias [10]. Al igual que con todos los modelos matemáticos, la capacidad predictiva del modelo está limitada por la precisión de los datos disponibles y por el llamado nivel de abstracción utilizado para modelar el problema [11].

En el inicio de la pandemia en México la información del aumento de casos avanzaba lentamente lo cual no permitía generar modelos que describieran el comportamiento estatal y nacional, a partir de que a nivel nacional se alcanzaron los 500 casos confirmados positivos fue posible aplicar diversas estrategias metodológicas para extraer de la información reportada diariamente, modelos y tendencias de la información.

Aunque el número de casos acumulados aumenta diariamente se observaron comportamientos característicos en los reportes presentados, además se analizó el porcentaje de casos distribuido con respecto a las fechas de inicio de síntomas, las de ingreso y defunciones, de donde se encuentra que un gran porcentaje de los casos confirmados fallecen pocos días después de su ingreso a los servicios hospitalarios y más alarmante aún es que de estas lamentables defunciones la mayoría presentó síntomas mucho tiempo antes de su ingreso. Esta información es muy importante para nosotros porque reafirma el hecho de que el número de casos acumulados se alimenta con casos ocurridos varios días antes del reporte.

En este trabajo se propone una metodología para generar predicciones del número de casos diarios confirmados positivos mediante una doble regresión polinomial, hasta

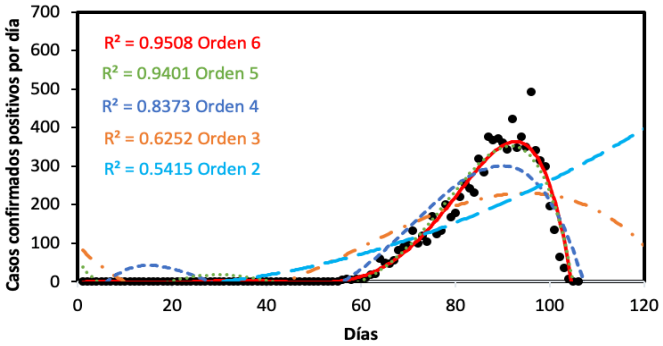

Fig. 1. Comparación de las curvas y coeficientes R^2 de los polinomios de ordenes 6,5,4,3 y 2 para describir los casos diarios confirmados positivos en el reporte oficial del dia 20 de abril de 2020.

Tabla 1. Ecuaciones de ajuste y coeficientes de correlación.

Polinomio de ajuste	Coeficiente de correlación
$y = -3E\text{-}08x^6 + 8E\text{-}06x^5 - 0.0006x^4 + 0.0241x^3 - 0.4045x^2 + 2.5168x - 3.3367$	0.9508
$y = -3E\text{-}06x^5 + 0.0006x^4 - 0.0466x^3 + 1.5115x^2 - 18.575x + 55.462$	0.9401
$y = -1E\text{-}04x^4 + 0.0196x^3 - 1.1642x^2 + 23.104x - 102.98$	0.8373
$y = -0.0017x^3 + 0.3007x^2 - 12.113x + 93.385$	0.6252
$y = 0.0336x^2 - 0.6276x - 11.42$	0.5415

la fecha no se conoce ningún trabajo de esta naturaleza aplicado al análisis de la pandemia COVID-19.

El modelo utiliza los reportes oficiales publicados diariamente y genera los modelos matemáticos para cada reporte, de los cuales se obtienen sus coeficientes o variables numéricas que se optimizan y a partir de ellos generar una predicción del número de casos confirmados positivos diarios y de esta forma generar un modelo que describa mejor los escenarios futuros.

2. Metodología

Los datos que se obtuvieron de los informes diarios que provee el gobierno de México a través de la Secretaría de Salud y publicados en el portal oficial.

2.1. Modelo matemático

La información obtenida del reporte diario se filtra para considerar solamente los casos confirmados positivos, aunque la metodología se aplicara posteriormente a los casos negativos y a los casos pendientes de confirmación.

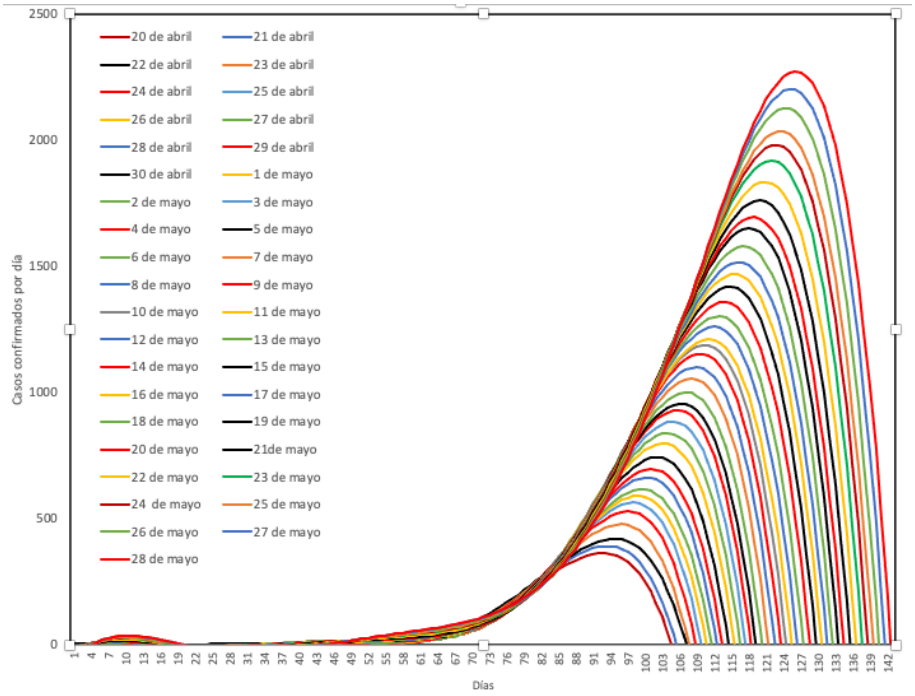

Fig. 2. Simulación de la curva reportada para cada día con el polinomio del orden que maximiza el coeficiente de correlación.

2.2. Metodología

A partir de un reporte oficial se filtra la información de casos confirmados positivos por fecha de inicio de síntomas y con una regresión polinomial se obtienen las ecuaciones que maximicen el coeficiente de correlación que se calcula a partir de la ecuación 1. Este proceso se realizó a partir del reporte del 20 de abril con un total de 8772 casos confirmados positivos y 108 días transcurridos desde el primer caso, como se muestra en la Tabla 1 y la figura 1:

$$R^2 = \frac{\sum_{t=1}^{T}(\hat{Y}-\bar{Y})^2}{\sum_{t=1}^{T}(Y_t-\bar{Y})^2}. \quad (1)$$

Una vez que se identifica que la regresión polinomial de orden 6 es el menor orden que maximiza el coeficiente de correlación en el mayor número de reportes se aplica este orden a todos los reportes y se obtiene la figura 2.

Los coeficientes de las curvas de la figura 2 se procesan y se obtiene ahora el polinomio que mejor describe su comportamiento, como tenemos polinomios de orden 6, serán 7 coeficientes que se procesan como 7 vectores de información correspondiente a las fechas de reporte. En este punto no es suficiente contar con el máximo coeficiente de correlación vector, ya que como se observa en la Figura 3 aunque la ecuación de mayor orden describe mejor la cantidad de puntos suministrada, sin embargo, esta puede tener cambios de pendiente muy fuertes después del último punto reportado y

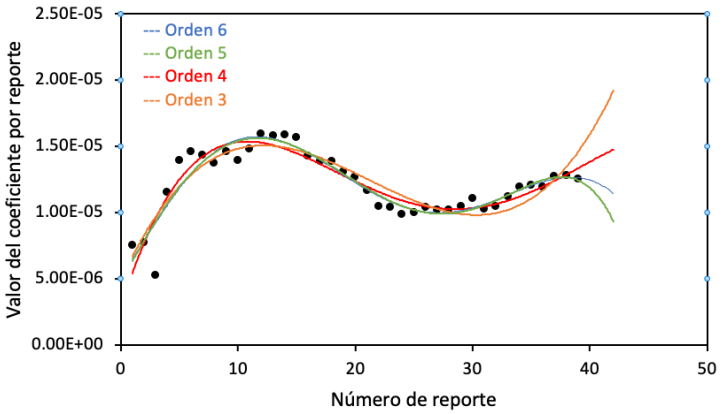

Fig. 3. Elementos del vector de coeficientes A_5 de los polinomios de orden 6 de la figura 2 y sus respectivos ajustes polinomiales de orden 6, 5, 4 y 3. Se puede observar que después del último punto reportado los polinomios de orden 4 y 5 presentan una pendiente similar a la que se encuentra en la zona de información reportada.

Tabla 2. Ecuaciones de ajuste y coeficientes de correlación por coeficiente.

Polinomio de ajuste	Coeficiente de correlación
$y = 9E\text{-}14x^6 - 1E\text{-}11x^5 + 8E\text{-}10x^4 - 2E\text{-}08x^3 + 1E\text{-}07x^2 + 1E\text{-}06x + 5E\text{-}06$	0.8502
$y = -4E\text{-}12x^5 + 3E\text{-}10x^4 - 8E\text{-}09x^3 - 7E\text{-}09x^2 + 2E\text{-}06x + 5E\text{-}06$	0.8487
$y = -4.96E\text{-}11x^4 + 5.73E\text{-}09x^3 - 2.15E\text{-}07x^2 + 2.89E\text{-}06x + 2.72E\text{-}06$	0.821
$y = 2E\text{-}09x^3 - 1E\text{-}07x^2 + 2E\text{-}06x + 5E\text{-}06$	0.7703
$y = 0.0336x^2 - 0.6276x - 11.42$	0.5415

adoptar un comportamiento abrupto. Para resolver esto se debe analizar la pendiente de los polinomios de regresión obtenidos en puntos posteriores a los reportados a fin de describir escenarios futuros más confiables es decir menor error en la predicción.

Como se observa en la figura 3, los polinomios de regresión obtenidos tienen una pendiente en los puntos desconocidos similar a la pendiente en los puntos conocidos son los de orden 5, 4, 6 y 3 respectivamente, lo cual no se puede concluir del coeficiente de correlación, en cuyo caso nos indicaría 6, 5, 4 y 3, para el orden del polinomio más confiable.

De esta forma se obtiene una ecuación para calcular cada uno de los coeficientes de acuerdo con el día de reporte y utilizar la ecuación obtenida al inicio que maximizaba el coeficiente de correlación, con los nuevos coeficientes.

Finalmente, este proceso se valida en días previos con la reproducibilidad de la información inicial, no se espera que esta sea idéntica, porque ha pasado por una optimización en los coeficientes de la familia de curvas a cada reporte diario, como se muestra en los resultados.

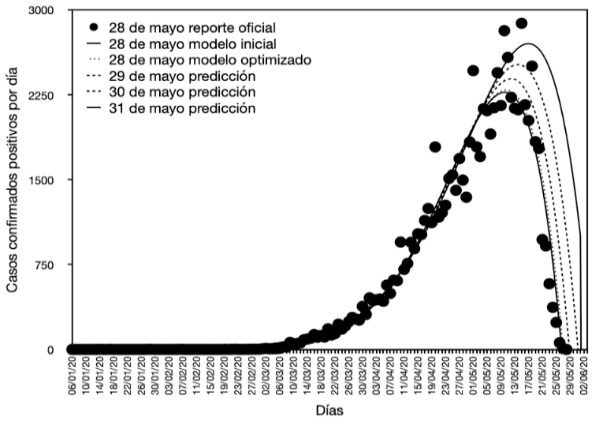

Fig. 4. Simulación con los coeficientes optimizados para días posteriores al reporte oficial del 28 de mayo.

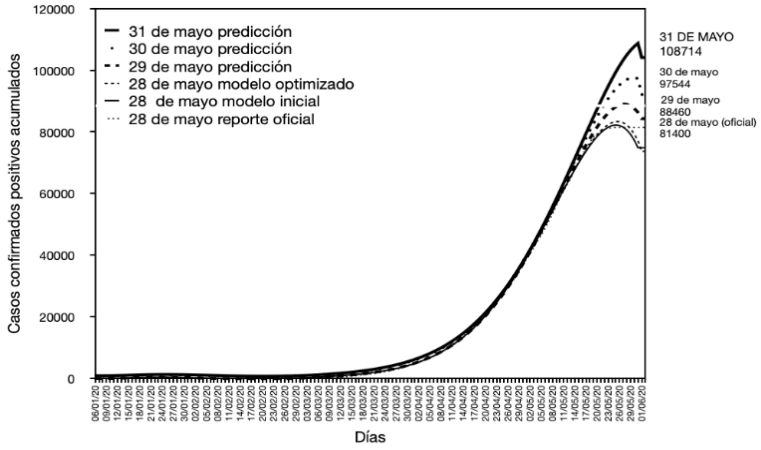

Fig. 5. Simulaciones utilizando los coeficientes de orden 5 optimizados con la pendiente más cercana a la presentada en los puntos reportados.

3. Resultados

Uno de los primeros resultados es la validación de los coeficientes calculados para cualquier día de los reportados, lo cual, como es de esperarse, no muestra diferencia notable con respecto a los previamente obtenidos y mostrados en la figura 2, posterior a este proceso se obtienen del mismo modelo los días posteriores a los reportados, como se muestran a continuación en las figuras 4 y 5.

Como se observa en las simulaciones de la figura 4 la validación con el modelo sin los coeficientes optimizados es muy buena y para días posteriores muestra una buena confiabilidad en la predicción las cuales son utilizadas para generar la gráfica de casos

Tabla 3. Porcentaje de error en las estimaciones a 3 días con Regresiones Polinomiales de orden 6 en el reporte diario y de orden 5 en los vectores de coeficientes.

FECHA	ERROR
28 de mayo de 2020	2.345%.
29 de mayo de 2020	5.118 %
30 de mayo de 2020	11.463 %
31 de mayo de 2020	19.9%

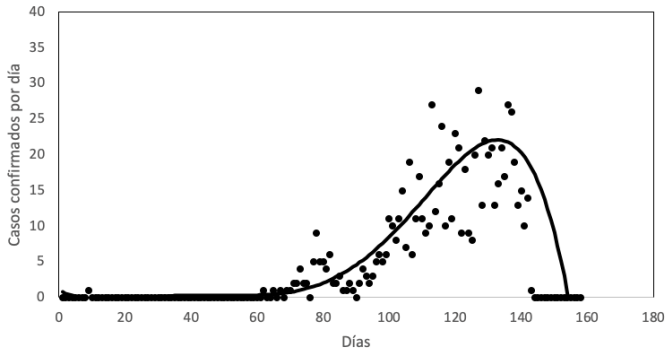

Fig. 6. Modelo para simular los casos confirmados positivos por día en Aguascalientes.

acumulados de la figura 5, en la que la validación presenta un error con respecto al reporte oficial de 0.971% el modelo simulado inicial y el modelo optimizado con 2.345%.

De acuerdo con los reportes de días posteriores se tienen los siguientes porcentajes de error de la tabla 3.

Los resultados encontrados al optimizar los coeficientes muestran un error creciente debido al ajuste en la optimización de los coeficientes. Sin embargo, se ha encontrado que muchos estados presentan comportamientos matemáticos distintos lo cual origina que el error empiece a crecer conforme se busca información en más días posteriores al reporte.

Este análisis por estados ya se está realizando debido a la alta complejidad de aplicar esta metodología a 32 estados por al menos 15 días, en cuyo caso la solución implica un sistema aproximado de 500X500 ya que el algoritmo lleva a cabo 5 regresiones polinomiales a cada estado y posteriormente otras 5 a cada uno de los 6 coeficientes de las anteriores así como el cálculo de los respectivos R^2 y las pendientes, lo cual ya está generando resultados que retroalimentarán el sistema aquí presentado.

En seguida mostramos algunos resultados, en la figura 6 se muestra el caso de Aguascalientes que muestra como el modelo de orden 4 maximiza el coeficiente de correlación, pero aun así la alta dispersión de la información reportada genera que el modelo presente porcentajes de error altos en los modelos optimizados. En la figura 7 se observa el modelado con polinomios de orden 5 de casos diarios para Baja California, Ciudad de México y Estado de México.

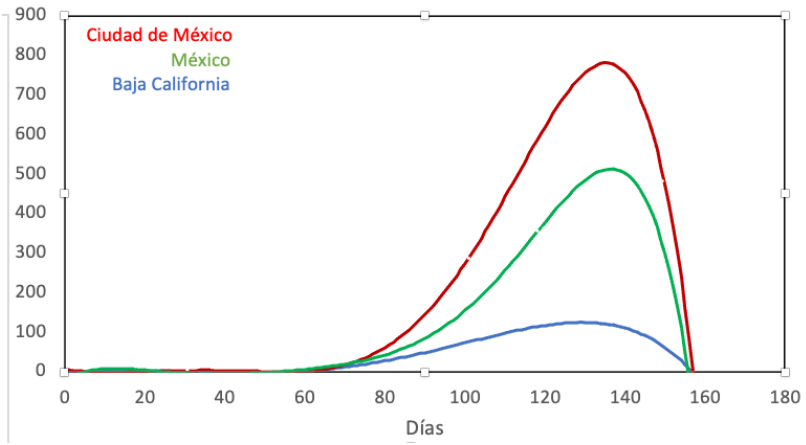

Fig. 7. La optimización de los coeficientes en estados como Ciudad de México soporta hasta un orden 6 a diferencia de estados como Aguascalientes donde el máximo orden es 4.

4. Conclusiones

Se desarrolla una propuesta metodológica mediante una regresión polinomial de orden 6 para cada reporte diario de casos confirmados positivos seguida de otra de orden 5 para cada uno de los 7 coeficientes de los polinomios anteriores, donde el orden 6 es el menor orden que maximiza R^2 en la mayoría de los reportes y el orden 5 se obtiene en base al análisis de la pendiente en los coeficientes. La aplicación de esta metodología brindó resultados con un porcentaje de error del 5.1% a un día, 11.4% a dos días y 19.9% a tres días en el número de casos confirmados positivos diarios y aunque el error aumenta, esta metodología brinda muy buenos resultados en la mayoría de los estados y a nivel nacional, ya se está desarrollando el análisis por municipios para retroalimentar esta matriz a fin de localizar a tiempo los puntos de crecimiento acelerado y su contribución en el crecimiento global de casos confirmados positivos a nivel nacional. Debido a las características de la metodología utilizada, se utilizará también para generar el escenario de defunciones, casos negativos y pendientes a fin de integrar todos en un análisis con modelos más confiables para aplicar otras técnicas de pronóstico.

Referencias

1. OMS: Neumonía de Causa desconocida – China. https://www.who.int/csr/don/05-january-2020-pneumonia-of-unkown-cause-china/es/ (2020)
2. Novel Coronavirus (2019-nCoV): Situation Report-10. https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200130-sitrep-10-ncov.pdf?sfvrsn=d0b2e480_2 (2020)
3. CIDE: CIDE y Stanford desarrollan modelo matemático de proyecciones sobre COVID-19. https://www.cide.edu/saladeprensa/cide-y-stanford-desarrollan-modelo-matematico-de-proyecciones-sobre-covid-19/ (2020)

4. El País: Así evoluciona la curva del coronavirus en México, Colombia, Chile, Argentina y el resto de Latinoamérica. https://elpais.com/sociedad/2020/04/07/actualidad/1586251 212_0 90043.html (2020)
5. Burgos, P.: COVID-19 Modelo numérico de casos de infección y estimaciones epidémicas Modelo Asimétrico -GOMPERTZ. 10.13140/RG.2.2.19440.40969 (2020)
6. Worldometers: COVID-19 Coronavirus Pandemic. https://www.worldometers.info/ coronavirus/ (2020)
7. Prediction of COVID-19 transmission dynamics using a mathematical model considering behavior changes in Korea. https://pesquisa.bvsalud.org/portal/resource/ es/mdl-32375455 (2020)
8. Chaos theory applied to the outbreak of COVID-19: an ancillary approach to decision making in pandemic context. https://pesquisa.bvsalud.org/portal/resource/es/mdl-32381148 (2020)
9. Propagation analysis and prediction of the COVID-19. https://www.sciencedirect.com/science/article/pii/S2468042720300087 (2020)
10. Model the transmission dynamics of COVID-19 propagation with public health intervention. https://www.medrxiv.org/content/10.1101/2020.04.22.20075184v1 (2020)
11. Mathematical Modeling of Epidemic Diseases: A Case Study of the COVID-19 Coronavirus. https://arxiv.org/pdf/2003.11371.pdf (2020)