

Propuesta metodológica para la predicción a corto plazo de contagios de COVID-19

María del Carmen Santiago Díaz, Ana Claudia Zenteno Vázquez,
Yeiny Romero Hernández, Judith Pérez Marcial,
Gustavo T. Rubín Linares,
Antonio Eduardo Álvarez Núñez

Benemérita Universidad Autónoma De Puebla,
México

{marycarmen.santiago, ana.zenteno, yeiny.romero,
judith.perez, gustavo.rubin}@correo.buap.mx, eduard-
alvarez@live.com.mx

Resumen. El coronavirus covid-19 es una pandemia muy grande y se requieren modelos matemáticos para simular escenarios y proyecciones que brinden 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 N , cada reporte se divide por estados y se calcula el orden "n" de una regresión polinomial $P(x)$ que maximice R^2 , a los coeficientes A_i de ésta también se les aplican regresiones polinomiales de diversos órdenes para maximizar R^2 para cada estado y cada reporte, enseguida se aplica la derivada a estas regresiones polinomiales de los coeficientes y con ella se determina el orden más adecuado para el polinomio de los coeficientes el cual se utiliza para generar el polinomio de cada estado nuevamente. Con esta metodología se encontró un mecanismo para predecir a corto plazo el escenario estatal y nacional en días posteriores al último reporte oficial.

Palabras clave: Covid-19, modelo, predicción, contagios.

Methodological Proposal for the Short-Term Prediction of COVID-19 Infections

Abstract: The covid-19 coronavirus is a very large pandemic and mathematical models are required to simulate scenarios and projections that provide 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 N , each report is divided by states and the order "n" of a polynomial regression $P(x)$ is calculated that maximizes R^2 , to the coefficients A_i of this, polynomial regressions of various orders are also applied to maximize R^2 for each state and each report, then the derivative is applied to these polynomial regressions of the coefficients and with it the most suitable order for the polynomial of the coefficients is determined, which is used to generate the polynomial of each state again. With this methodology, a mechanism was found

to predict in the short term the state and national scenario in days after the last official report.

Keywords: Covid-19, model, prediction, contagion.

1. Introducción

La sociedad a lo largo de la historia ha tenido que combatir diversos tipos de enfermedades que han ocasionado altas tasas de mortalidad y graves consecuencias económicas y de salud. Los virus que causan enfermedades siempre han existido, se han mutado y han aparecido en diferentes momentos del mundo. En diciembre de 2019, surgen casos de virus en China, particularmente en Wuhan (Hubei). Estos casos están vinculados a un mercado mayorista de mariscos, pescado y animales vivos, y han propiciado una ola de investigaciones del desarrollo de la pandemia en todos los países. La Organización Mundial de la Salud (OMS) recibió reportes de varios casos de neumonía de etiología desconocida. El 3 de enero de 2020, las autoridades de China notificaron a la OMS que existían 44 pacientes con neumonía de etiología desconocida, de entre los cuales, 11 pacientes estaban gravemente enfermos. Según informaciones difundidas en los medios de comunicación, el mercado implicado en Wuhan se cerró el 1 de enero de 2020 para realizar acciones de saneamiento y desinfección ambiental [1]. El día 8 de enero en 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 y gestionar casos y para realizar pruebas de laboratorio. 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]. Para el mes de septiembre, a nivel mundial se contabilizan más de 34 millones de contagios y aproximadamente un millón de decesos. El 1 de octubre, en México se confirman 748,315 casos y 78,078 decesos [3].

2. Estado del arte

Diferentes investigaciones se están llevando con el propósito de modelar el número de casos sospechosos, confirmados, decesos, ocupación de instalaciones hospitalarias, formas y patrones de contagio, entre otras variables. En los modelos epidemiológicos se parte del supuesto de que los individuos se encuentran en uno de varios estados posibles, Susceptible (S), Infectado (I) o Recuperado (R).

Los modelos matemáticos SI, SIS, SIR y sus derivados se emplean para predecir el impacto de las pandemias en las sociedades, en los que la interacción de los individuos es crucial para la propagación del virus y por eso es recomendación el confinamiento para minimizar los estragos en las sociedades. Son usados para tratar enfermedades que afectan a poblaciones grandes y a menudo surgen representados a través de ecuaciones diferenciales[4].

En el Centro de Investigación y Docencia Económicas (CIDE) el equipo del profesor Escudero desarrolló un modelo matemático de proyecciones sobre los efectos de las distintas prácticas de mitigación sobre COVID-19. Otro modelo, denominado SC-COSMO por las siglas Stanford-CIDE CORonavirus Simulation MOdel se encarga de analizar cómo evoluciona la enfermedad y modela también los mecanismos en los que los individuos interactúan entre sí [5]. Este modelo realiza análisis demográficos para considerar a los individuos susceptibles, a los expuestos, los infectados y los recuperados, conforme a los patrones de contacto que fueron clave para la transmisión de covid-19.

Los casos confirmados, sospechosos y decesos se miden diariamente brindando información a los gobiernos sobre la capacidad de atención a la población [6]. En este punto, el tiempo es un factor que permite observar las tendencias de los casos cuando hay mejora, los decesos y la evolución en duplicación en los casos a nivel local y global. El “Covid-19 Modelo numérico de casos de infección y estimaciones epidémicas modelo asimétrico - Gompertz” [6] obtiene una estimación de la demanda hospitalaria para casos graves que necesitan terapia intensiva en España y muestra los beneficios de mantenerse en casa, para que, en caso de ser contagiado se cuente con una mejor atención médica.

En México, la información del aumento de casos fluía lentamente y no permitía generar modelos que brindaran información concreta del desarrollo de la pandemia a nivel estatal y nacional. Cuando los casos confirmados comenzaron a incrementar, a partir de 500 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 aumentaba diariamente se observaron comportamientos característicos en los reportes presentados, además se analizó el porcentaje de casos distribuidos con respecto a las fechas de inicio de síntomas, las de ingreso y defunciones, y 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. Por eso, este modelo utiliza los reportes oficiales publicados diariamente y genera los modelos matemáticos para cada uno, los cuales mediante un proceso de análisis similar se descomponen en sus coeficientes o variables numéricas para optimizarse y a partir de ellos mejorar la estimación del comportamiento del número de casos confirmados positivos reportados diariamente y generar un modelo que describa mejor los escenarios futuros.

3. Metodología

La información reportada por la Dirección General de Epidemiología del gobierno de México y publicados en el portal oficial brindan un panorama muy completo de la evolución de la pandemia, sin embargo la interpretación no es trivial en el sentido de que la información requiere ser procesada con extremo cuidado. Utilizamos para este trabajo sólo la información correspondiente al número de casos confirmados positivos

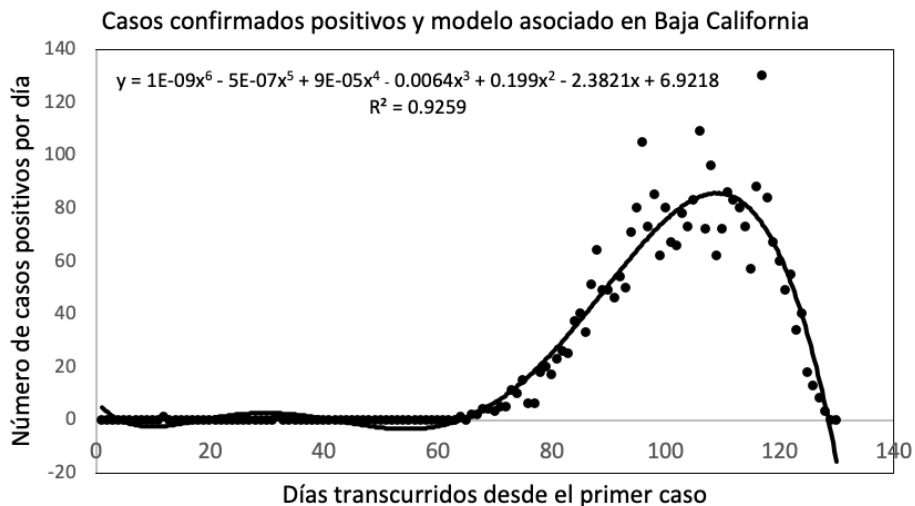


Fig. 1. Gráfica que muestra los casos confirmados positivos por día en Baja California, el polinomio de ajuste y el coeficiente de correlación R^2 .

para la prueba de SARS-COV2 por día, por estado y por municipio. En un trabajo relacionado procesamos a nivel nacional una metodología basada en los coeficientes de los polinomios de regresión aplicados a cada día de reporte nacional [7]. En este trabajo mostramos una variante a esta metodología, ya que el llevar el modelo polinomial a la base de datos de casos confirmados por estado, implica estados donde hay una alta dispersión en los datos, es decir, no siguen una tendencia totalmente homogénea. El modelo polinomial se justifica de acuerdo a las siguientes evidencias obtenidas al procesar la información:

1. Brinda un Coeficiente de Correlación R^2 cercano a 1 al utilizarlo para describir el número de casos diarios reportados.
2. Genera en el número acumulado de casos una aproximación menor al 5%.
3. A partir del orden 6 los cambios en el Coeficiente de Correlación son menores a 0.01, sin importar el día que se analice.
4. El pico máximo del número de casos diarios sigue el mismo comportamiento en el modelo polinomial que en el reportado.

Entonces realizamos un filtrado e inspección de la información por estado y aplicamos regresiones polinomiales encontrando que los coeficientes de correlación para cada estado presentan el valor más próximo a 1 para polinomios de orden 6, con lo cual utilizamos los 7 coeficientes por cada estado en la información reportada por día y realizamos una regresión polinomial a cada familia de coeficientes. Posteriormente aplicamos el coeficiente de correlación para analizar cual polinomio nos brinda una mejor aproximación y en este punto el sistema nos muestra lo obvio, que el orden más alto nos va a generar el coeficiente R^2 más alto, sin embargo una inspección nos revela que este hecho nos lleva a que una vez que el polinomio genera una curva muy similar a la información de los coeficientes, al pasar por el último dato tiene un cambio significativo en su comportamiento, por lo cual, nos surge de forma

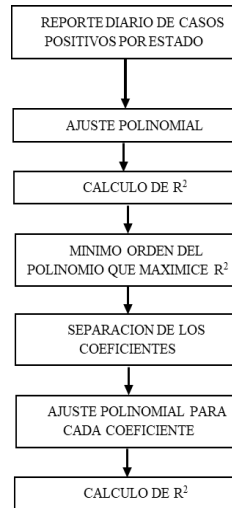


Fig. 2. Diagrama a bloques que muestra la metodología de análisis y procesamiento para la información de casos positivos por día para cada estado.

natural la necesidad de aplicar la derivada a la curva posterior a su entorno conocido y así quedarnos con el que tenga la derivada más suave y un coeficiente de correlación en un rango aceptable. En seguida mostraremos las etapas discutidas previamente.

3.1. Modelo matemático

A la información filtrada por estados le aplicamos el modelo polinomial para obtener una representación analítica del comportamiento de los casos confirmados positivos con lo cual determinamos el mínimo orden con el mayor coeficiente R^2 , este modelo es del mismo orden para todos los estados, tanto los que concentran el 70% de los casos a nivel nacional como aquellos con solo el 1%, como se muestra en la figura 1.

Se llevó a cabo el análisis de todos los estados desde el 15 de mayo hasta el 24 de junio y como puede mostrarse en la figura 1 en el caso de Baja California, hay una gran dispersión en la información, sin embargo, el modelo tiene un buen coeficiente R^2 para orden 6. Partiendo de estos hechos, se establece una ecuación que determina el comportamiento de cada uno de los 7 coeficientes del polinomio de orden 6, para cada estado. Es decir, para un estado tenemos en este rango de fechas 41 ecuaciones de orden 6 de las cuales obtenemos 41 elementos para cada coeficiente y que enseguida aplicamos regresiones de orden 2, 3, 4, 5 y 6 y sus respectivos coeficientes de correlación. Como se muestra en el siguiente diagrama a bloques.

Una vez que realizamos el procesamiento hasta la separación de los coeficientes de la figura 2, encontramos que cada uno de estos coeficientes del polinomio de orden 6 de la ecuación 1 tienen características como las que se muestran en la figura 3 para el caso de Aguascalientes, donde ya se concentra la información del coeficiente A_6 para los 41 reportes:

$$y = A_6x^6 + A_5x^5 + A_4x^4 + A_3x^3 + A_2x^2 + A_1x^1 + A_0 . \quad (1)$$

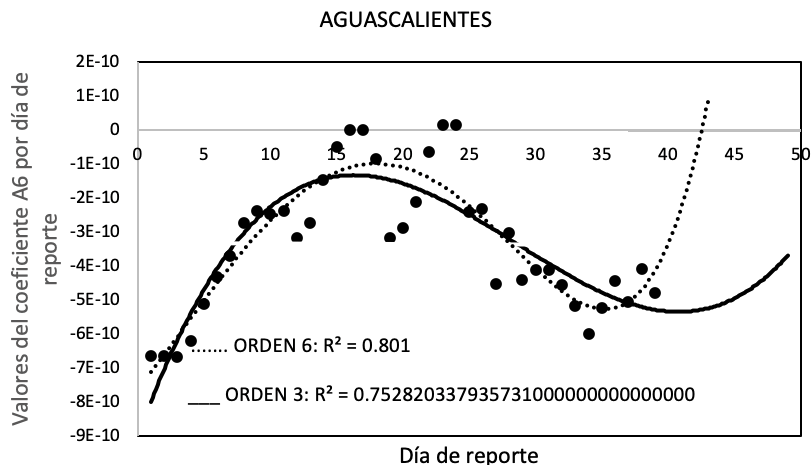


Fig. 3. Gráfica del coeficiente A_6 obtenido del polinomio de orden 6 del reporte diario para el número de casos positivos en Aguascalientes y las curvas polinomiales de orden 6 y 3, respectivamente.

La figura 3 muestra un resultado muy frecuente que se presenta al llevar a cabo la determinación del modelo de un conjunto de información, el aumentar el orden de ajuste polinomial nos lleva a mejorar el coeficiente R^2 , sin embargo la curva de orden mayor puede tener un comportamiento asintótico o demasiado brusco comparado con el comportamiento de la información que está modelando, por ello aplicamos la derivada a la curva y con esto obtenemos el orden que posea un balance entre el coeficiente R^2 y la suavidad de cambio que mide la derivada.

Enseguida aplicamos las regresiones polinomiales de orden 2, 3, 4, 5 y 6 a los 7 coeficientes de cada estado para los 41 días reportados. Y con ayuda de la derivada que implementamos de forma discreta, es decir, a partir de las diferencias entre pares de puntos, disminuimos el orden del R^2 máximo a un orden con $R^2 > 0.7$, este hecho disminuye el orden a 5 y cuatro y en algunos casos hasta 3, esta decisión se lleva a cabo de forma automática por el sistema, obteniendo fluctuaciones de 2 y 3, a diferencia de lo que se obtendría únicamente con el coeficiente R^2 , en cuyo caso obtenemos que predominan los órdenes 5 y 4, y de forma poco significativa 2 y 3.

Estos ordenes se promedian para cada estado, es decir en cada estado se promedian los órdenes de los 7 coeficientes y este promedio se establece como el orden de todos los coeficientes para ese estado.

4. Resultados

Una vez que se determina el orden del polinomio para cada coeficiente de cada estado se utiliza este polinomio para generar el nuevo polinomio de orden 6 del estado, pero ahora con el ajuste de los coeficientes.

En la figura 4 se muestra la información reportada los días 3 y 16 de junio, es notoria la dispersión en la información sin embargo el ajuste después del tratamiento

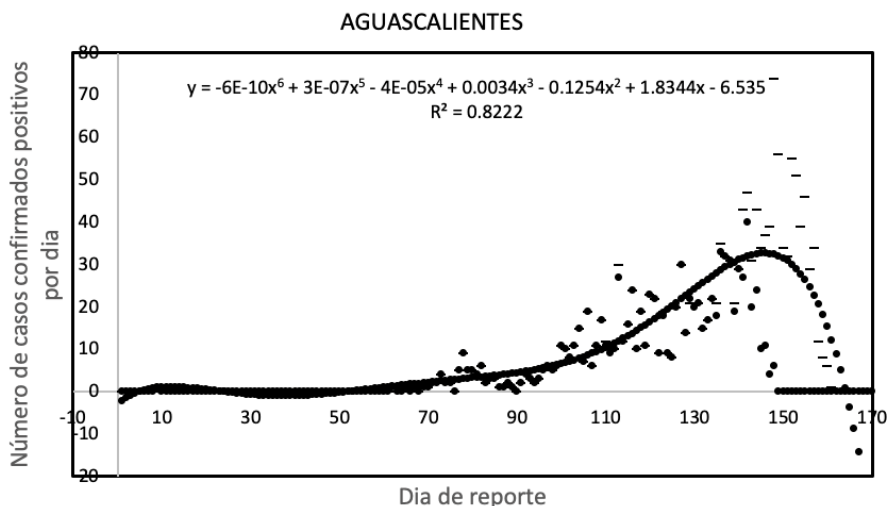


Fig. 4. Gráfica de casos positivos para los días 3 y 16 de junio (- - , . . . , respectivamente) y la simulación polinomial de orden 6 con los coeficientes ajustados en orden 4.

Tabla 1. Validación y proyecciones a 2 días de las simulaciones.

	No. de casos reportados día 41	No. de casos simulados día 41	No. de casos simulados día 42	No. de casos simulados día 43
Nacional	129184	131801.65	134085.74	138567.29
Baja California	6436	6343.87	6475.17	6616.65
Puebla	4742	4700.12	4878.48	5091.51

mencionado anteriormente obtiene una curva de orden 6 para el estado de Aguascalientes con un coeficiente $R^2=0.82$, utilizando un coeficiente de orden 4 para los coeficientes, siendo que el coeficiente sugerido sin la derivada es de orden 5.

Esta información obtenida para todos los estados nos permite determinar el número de casos confirmados positivos a nivel nacional y estatal, para los días reportados y días posteriores, como se observa en la siguiente tabla.

5. Conclusiones

El sistema presenta errores con respecto a los reportes oficiales en un rango promedio del 4.91%, aunque se obtuvieron errores cercanos a cero en general, también se tienen estados donde la gran dispersión origina errores del 30%, esto es debido a que el sistema polinomial ajusta buscando el mejor polinomio de los casos positivos y posteriormente hacemos otro ajuste polinomial en los coeficientes del polinomio

anterior, pero los dos conjuntos de información presentan una alta dispersión, es decir, el costo por buscar una mayor precisión en el número de casos resulta en algunos estados en una excelente aproximación, sin embargo hay otros en los cuales se debe considerar un criterio distinto debido a que internamente la información en cada estado presenta diferentes velocidades de contagio por día y en la información estatal solo se observa como una alta dispersión. La estrategia que ya estamos desarrollando es el análisis por municipio, el cual tendrá mucho más detalle de la velocidad de propagación de la pandemia, pero sin duda el problema será que las regresiones polinomiales deberán aplicarse de forma más cuidadosa, lo cual en el caso por estados ya se utilizaron algunas consideraciones, como restringir el dominio de la información útil que modela el polinomio, es decir, no considerando en algunos estados el mismo inicio, ya que no todos los estados tienen el primer caso el mismo día y esto afecta considerablemente el modelo, ya que al ser de orden alto es natural que posea diversas oscilaciones aun cuando la información se mantenga constante.

6. Trabajo futuro

Los resultados obtenidos han demostrado que la metodología es adecuada ya que brinda resultados cercanos a los casos reportados, así que ahora se debe integrar la información de municipios, estados y país, además de la información de casos negativos, sospechosos y defunciones, lo cual, junto con las características de la población que ha sido reportada nos permitirá conocer más detalles del comportamiento matemático. Aunque no se tiene ninguna referencia de alguna metodología parecida si se está comparando con aquellas que utilizan el número de casos acumulados, a fin de encontrar mejoras en el modelo.

Referencias

1. OMS: Neumonía de Causa desconocida China (2020)
2. Novel: Coronavirus (2019-nCoV) Situation Report-10 (2020)
3. Gobierno de México: Datos Coronavirus (2020)
4. Sánchez-Villegas, P., Daponte-Codina, A.: Modelos predictivos de la epidemia de COVID-19 en España con curvas de Gompertz. *Gaceta Sanitaria* (2020)
5. CIDE: CIDE y Stanford desarrollan modelo matemático de proyecciones sobre COVID-19 (2020)
6. Borja, A., Grasso, D., Llaneras, K., Galindo, J.: Así evoluciona la curva del coronavirus en México, Colombia, Chile, Argentina y el resto de Latinoamérica. *El País* (2020)
7. Zenteno, A.C., Santiago, M.C., Romero, Y., Pérez, J., Rubín, G.T., Álvarez, A.E.: Optimización de los coeficientes del modelo predictivo del número de casos diarios de coronavirus Covid-19 en México. In: *Simposio Nacional de Inteligencia Artificial e Industria 4.0* (2020)
8. Li, L., Yang, Z., Dang, Z., Meng, C., Huang, J., Meng, H., Wang, D., Chen, G., Zhang, J., Peng, H., Shao, Y.: Propagation analysis and prediction of the COVID-19. *Infectious Disease Modelling*, 5, pp. 282–292 (2020)
9. Ketema-Mamo, D.: Model the transmission dynamics of COVID-19 propagation with public health intervention. *Applied Mathematics* (2020)

10. Sameni, R.: Mathematical modeling of epidemic diseases: A case study of the COVID-19 (2020)