Proyecto 2: Predicción de casos de COVID-19 usando procesos de regresión Gaussiana

Miguel Angel Soto Hernandez

Centro de Investigación en Computación, Instituto Politécnico Nacional
Ciudad de México, México
msotoh2021@cic.ipn.mx

Resumen. En este proyecto se realiza una predicción de casos de COVID-19 basado en un modelo Gaussiano para el que se utilizaron los datos públicos mundiales referentes a los casos de la enfermedad. Los datos específicos utilizados como prueba para este modelo corresponden a los países de México, Nueva Zelanda y Taiwan. Sin embargo, el conjunto de datos comprende información mundial. El alcance de este artículo es la predicción de número de nuevos casos para los 3 países que se consideran dentro de él, con una mira de 90 días en el futuro a partir de la fecha estipulada como la última fecha de registro de casos. De este modo se podrán identificar los distintos escenarios posibles que puedan suceder.

Palabras clave: Procesos de Regresión Gaussianos, Estadística, COVID-19

I. Introducción

El virus SARS-COV-2 descubierto a finales del año 2019 que genera la enfermedad COVID-19 fue declarado como pandemia mundial por la Organización Mundial de la Salud (OMS) en Marzo de 2020. Esto causó una fuerte crisis en los distintos ámbitos de la sociedad. Sin embargo, la comunidad científica de todo el mundo comenzó a trabajar en diferentes áreas con la finalidad de entender como es que la enfermedad se propagaba, cuales eran los medios más fuertes de propagación, así como modelos que predecían con que rapidez se propagaría, y como se combatiría con la finalidad de tener control sobre esta.

Actualmente y a pesar de las medidas de salubridad alrededor del mundo para combatir la pandemia causada por el la enfermedad coronavirus (COVID-19) causada por el virus SARS-COV-2 se sigue viviendo una gran incertidumbre acerca del futuro que le depara a muchos países respecto la enfermedad, ya que distintas organizaciones de salud de los distintos países han alertado sobre una tercera ola de contagios más fuerte si no se siguen cumpliendo las normas sanitarias que se llevan hasta ahora.

Hasta ahora se han realizado diversas aproximaciones de predicciones para observar el comportamiento de la enfermedad para los proximos meses o incluso años, los cuales se describen de una manera resumida y formal en la sección II, trabajos previos.

Este trabajo pretende crear un modelo que a través de un proceso Gaussiano sea capaz de predecir el comportamiento de la enfermedad COVID-19 para los próximos 90 días de los países de México, Nueva Zelanda y Taiwan. De esta manera se podrán comparar los escenarios de la enfermedad alrededor del mundo.

Un proceso Gaussiano es un método bayesiano no paramétrico, dónde normalmente nos interesa inferir una distribución sobre los parámetros de una segunda distribución que se utiliza para ajustar los datos, es decir, la probabilidad. En este caso, estamos infiriendo directamente una distribución sobre las funciones, y se define de la siguiente manera:

$$p(f|x) \sim N(f|\mu, K)$$

dónde $\mu = (m(x_1), ..., m(x_N))$ son las funciones medias, $K_{ij} = k(x_i, x_j)$ es la función kernel que define la matriz de covarianza y $f = (f(x_1), ..., f(x_N))$ son las realizaciones de los valores de las funciones.

El fundamento del proceso Gaussiano es la distribución gaussiana multivariante, que es la generalización multidimensional de la distribución Gaussiana. En el caso multivariante, la distribución está definida por un vector de media μ y una matriz de covarianza simétrica y positiva definida Σ. Dónde μ representa el valor esperado de cada variable aleatoria, mientras que Σ describe dos fenómenos: su diagonal que expresa la varianza de cada dimensión y la covarianza que no se encuentra en la diagonal principal entre todas las variables aleatorias, es decir, mide como cambian las variables aleatorias en conjunto. La distribución Gaussiana multivariable tiene la siguiente densidad de probabilidad conjunta:

$$p(x|\mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^d|\Sigma|}} exp(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu))$$

dónde x es un vector random de tamaño d y $|\Sigma|$ es el determinante de la matriz dxd Σ.

La gaussiana multivariada es tan útil debido a sus propiedades algebraicas, que garantizan que también obtengamos distribuciones gaussianas al marginar y condicionar. La marginación significa integrar las variables del conjunto original de variables. El resultado es la distribución de un subconjunto de variables sin referencia a las que hemos integrado. Así, si sólo nos interesa la densidad de probabilidad de X:

$$P(x) = \int_y P(x, y)dy = \int_y P(x|y)p(y)dy$$

Esto también resulta en una distribución Gaussiana:

$$X \sim N(\mu_x, \Sigma_{xx})$$

De igual manera, esto pasa con el condicionamiento, que es la probabilidad de que una variable dependa de otra, como se puede observar a continuación:

$$X|Y \sim N(\mu_x + \Sigma_{xy}\Sigma_{yy}^{-1}(y-\mu_y), \Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx})$$

$$X|Y \sim N(\mu_y + \Sigma_{yx}\Sigma_{xx}^{-1}(y-\mu_x), \Sigma_{yy} - \Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy})$$

Una vez que se entienden las propiedades de la Gaussiana multivariada, es hora de combinarlas para resolver problemas de regresión. La idea básica cuando realizamos un proceso de regresión Gaussiano es que nuestro proceso Gaussiano asignará una probabilidad a las funciones potencialmente infinitas que podrían ajustarse a los datos que estamos manejando. Esta probabilidad expresa incertidumbre del modelo, lo que nos da una fuerte indicación de cuánto debemos confiar en la predicción puntual, es decir, la media de la distribución de probabilidad resultante.

Los kernels definen el tipo de funciones que podremos muestrear de nuestra distribución de funciones. A pesar de que los procesos GAussianos son métodos no paramétricos, los kernels nos permitirán controlar la forma de ese tipo específico de funciones. Uno de los kernels más usados es el kernel exponencial cuadrado, el cuál esta definido de la siguiente manera:

$$k(x_i, x_i) = n^2 exp(-\frac{1}{2l^2}(x_i - x_j)^T(x_i - x_j))$$

dónde la longitud l controla la suavidad y n la amplitud de la función.

Lo que se define anteriormente es la matriz de covarianza de la distribución multivariante, es decir, ahora tenemos acceso a las realizaciones de las funciones, con esto podemos muestrear la distribución multivariante Gaussiana, y esto se puede realizar sin observar ningún punto del entrenamiento.

Para finalizar con la introducción teórica, se introducen los puntos de prueba, es decir, los puntos en lo que estimaremos nuevas funciones, y para esto se utiliza la propiedad condicional Gaussiana multivariante. Si no se considera ningún término de error en el primer caso, el conjunto de funciones resultantes del condicionamiento pasará por cada punto del entrenamiento. En el segundo caso, la inclusión de ruido puede considerarse como la adición de un núcleo de ruido y esto ayudará a modelar pequeñas irregularidades en las predicciones. Una manera de estimar los parametros es maximizar la probabilidad marginal logarítmica.

II. **Trabajos previos**

Xiaolei Zhang y colegas (2020) [1] emplean un modelo de Poisson segmentado para analizar los datos disponibles de nuevos casos de los brotes de COVID-19 en diferentes países occidentales, incorporando las intervenciones de los gobiernos como consejos, órdenes de permanecer en casa, cierres patronales, cuarentenas, etc. Este análisis permitió hacer una predicción estadística sobre el punto de inflexión, la duración y la tasa de ataque para los países estudiados.

A.M.Mishra y colegas (2020) [2] desarrollaron un modelo matemático que considera las clases suceptible, expuesta, infectada, asintomática, en cuarentena y recuperada como en el caso de la enfermedad COVID-19. Tanto las clases expuestas como las infectadas tienen la posibilidad de estar en cuarentena/aislamiento. Los individuos asintomáticos se recuperaron sin someterse a tratamiento o pasaron a la clase infectada después de cierto tiempo. Formularon el número de reproducción para el modelo propuesto. El análisis de elasticidad y sensibilidad indica que el modelo es más sensible a la tasa de transmisión de las clases expuestas a las infectadas que a la tasa de transmisión de la clase susceptible a la expuesta. El análisis de la estabilidad global del modelo propuesto se estudia mediante la función de Lyapunov.

Gabriele Martelloni y Gianluca Martelloni (2020) [3] definen un modelo con 4 poblaciones: total de infectados, positivos actuales, recuperados y muertos. Y proponen un método alternativo a un modelo SIRD clásico para la evaluación de la epidemia de Sars-Cov-2 mediante la introducción de una ley simple de conservación. Dicho modelo es de igual maner aplicable para otras enfermedades.

Amit Singhal y colegas (2020) [4] desarrollaron dos modelos diferentes para captar la tendencia de un número de casos y también para predecir los casos en los próximos días, de modo que se puedan hacer los preparativos adecuados para luchar contra esta enfermedad. El primero es un modelo matemático que tiene en cuenta varios parámetros relacionados con la propagación del virus, mientras que el segundo es un modelo no paramétrico basado en el método de descomposición de Fourier (FDM), ajustado a los datos disponibles.

III. **Conjunto de datos**

Para llevar a cabo la creación del modelo de proceso de regresión Gaussiana se utilizó el conjunto de datos de *Our World in Data* [5] el cual corresponde a los datos históricos mundiales

desde que comenzaron a darse los casos de COVID-19 en cada país hasta el mes de Junio de 2021. Como se menciona anteriormente, para ejemplificar este trabajo se utilizaron tres países: México, Nueva Zelanda y Taiwan, de los cuales se utilizaron las siguientes características:

- **País:** definida en el conjunto de datos como *country*, se utilizó para extraer únicamente los países necesarios.
- **Continente:** definido en el conjunto de datos como *continent* , se utiliza como referencia de la característica país.
- **Fecha:** definido en el conjunto de datos como *date*, es utilizada como referencia sobre en que día se registraron los casos.
- **Casos nuevos:** definido en el conjunto de datos como *new cases*, es donde se encuentran todos los datos de los nuevos casos referentes a la enfermedad COVID-19.

IV. Modelo de proceso de regresión Gaussiana

En las figuras 1, 2 y 3 podemos observar la evolución de los nuevo casos de COVID-19 en México, Nueva Zelanda y Taiwan respectivamente. La propagación de una pandemia tiene una dinámica conocida. Por lo tanto la primera ola y la segunda tienden a ser muy diferentes ya que se implementan diferentes tipos de medidas de sanidad.

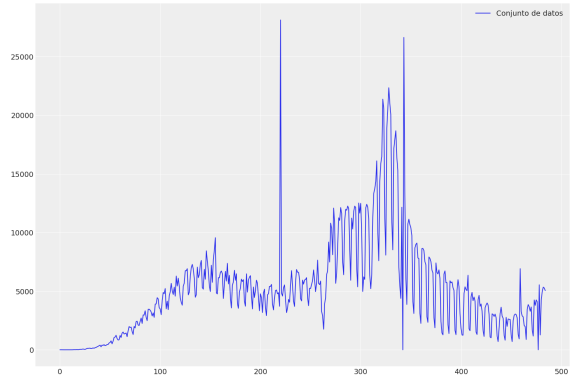

Figura 1: Nuevos casos de COVID-19 registrados en México

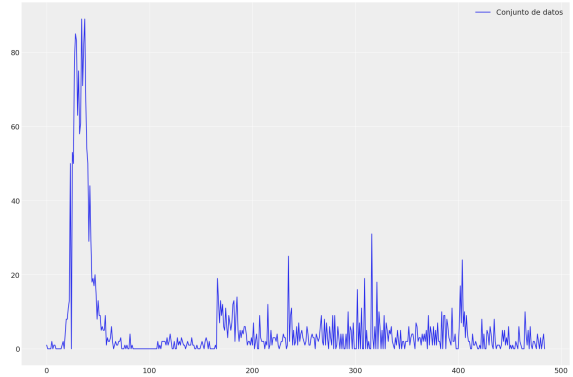

Figura 2: Nuevos casos de COVID-19 registrados en Nueva Zelanda

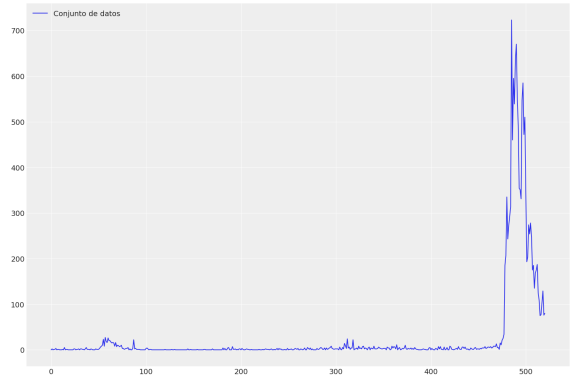

Figura 3: Nuevos casos de COVID-19 registrados en Taiwan

IV.A. **Implementación del modelo**

Los kernels pueden clasificarse en estacionarios y no estacionarios. Los kernels estacionarios siempre volverán a la media del proceso Gaussiano en las regiones más alejadas de los puntos observados. Esto es debido a que la covarianza que se está modelando entre los puntos depende solo de su posición relativa y no de su distancia absoluta. Por otro lado, un kernel lineal es un kernel no estacionario.

Este modelo es una suma de tres procesos Gaussianos para la señal y un proceso Gaussiano para el ruido:

- Un kernel cuadrático exponencial para modelar las irregularidades a medio plazo.
- Un término periódico para la estacionalidad semanal.
- Un kernel lineal para dar cuenta de la tendencia.
- El ruido se modela como un kernel de ruido blanco.

La prioridad de y en función del tiempo es la siguiente:

$$f(t) \sim GP_t(0, k_1(x, x')) + GP_{per}(0, k_2(x, x')) + GP_{lin}(0, k_3(x, x')) + GP_{ruido}(0, k_4(x, x'))$$

Si no se especifica alguna función media en el modelo, se supone que nuestro proceso Gaussiano tiene una media de cero. Si sólo utilizáramos núcleos estacionarios, esto significaría que la función acabaría volviendo a cero a medida que realizamos la previsión en el futuro. Las medidas adoptadas tienen como objetivo principal controlar la tasa de propagación, no exactamente hacerla desaparecer, es por eso que se crea una función distinta de cero.

En este caso tratamos con datos de conteo, es por esto que se utiliza uno o más procesos Gaussianos como procesos latentes que estimen una media de Poisson, los cuales se puede ver a continuación:

$$\theta \sim g(\phi)$$
$$f \sim MvNormal(0, K_{theta}(x))$$
$$y_i \sim Poisson(exp(f_i))\forall i \epsilon 1, ..., n$$

Al realizar comprobaciones predictivas a priori obtenemos los siguientes resultados mostrados en las figuras 4, 5 y 6 para México, Nueva Zelanda y Taiwan respectivamente. De la misma manera, podemos ver que las muestras están algo contenidas en el rango de nuestros datos.

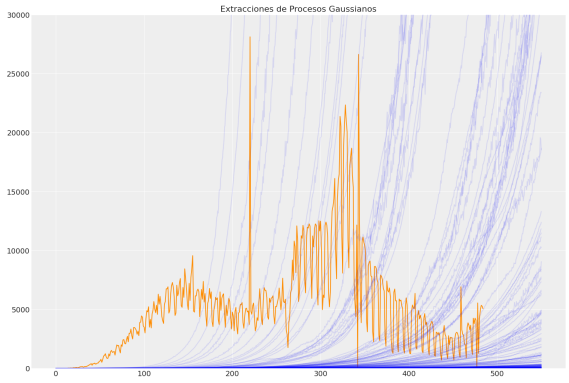

Figura 4: Comprobaciones predictivas a priori para México

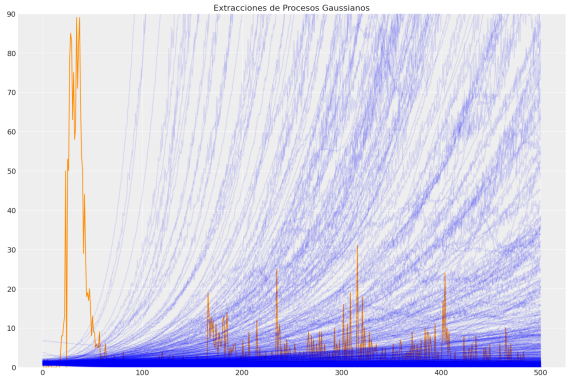

Figura 5: Comprobaciones predictivas a priori para Nueva Zelanda

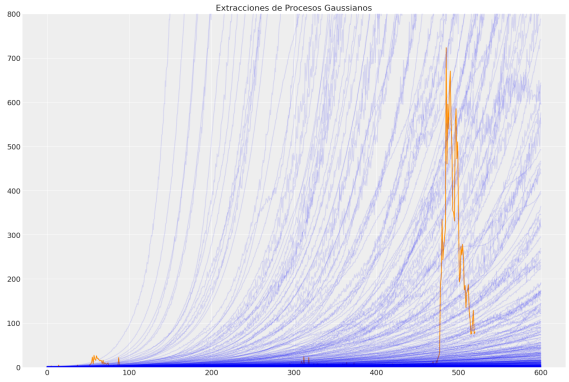

Figura 6: Comprobaciones predictivas a priori para Taiwan

Por último se complementa todo para obtener el modelo final, el cual utilizaremos para ajustar los nuevo casos de COVID-19 en México, Nueva Zelanda y Taiwan desde que la pandemia llego a cada uno de los países y se condicionará a los nuevo datos, es decir, a los 90 nuevos puntos temporales que pronosticaremos.

V. **Discusión y Resultados**

Como podemos observar en la figura 7 y 10, el escenario para México es malo, ya que el modelo estima una media de casos positivos de cerca de 50,000 para el día 90 después de la fecha marcada como fecha de inicio de predicción. Estas figuras también nos muestran que en el caso de que las medidas como el encierro resulten muy eficaces, el pico de 50,000 podría ser tan bajo como 25,000 (la parte inferior del intervalo del 50 %) o tan alto como 62,000 (la parte superior del intervalo del 50 %). También se pueden ver los casos más extremos con respecto al intervalo de credibilidad del 90 %.

Por otro lado, como podemos observar en la figura 8 y 11, el escenario para Nueva Zelanda es muy bueno, ya que el modelo estima una media de casos positivos de cerca de 10 para el día 90 después de la fecha marcada como fecha de inicio de predicción. Estas figuras también nos muestran que en el caso de que las medidas como el encierro resulten muy eficaces, el pico de 10 podría ser tan bajo como 0 (la parte inferior del intervalo del 50 %) o tan alto como 6 (la parte superior del intervalo del 50 %). También se pueden ver los casos más extremos con respecto al intervalo de credibilidad del 90 %.

Por último, como podemos observar en la figura 9 y 12, el escenario para Taiwan bueno, ya que el modelo estima una media de casos positivos de cerca de 60 para el día 90 después de la fecha marcada como fecha de inicio de predicción. Estas figuras también nos muestran que en el caso de que las medidas como el encierro resulten muy eficaces, el pico de 60 podría ser tan bajo como 6 (la parte inferior del intervalo del 50 %) o tan alto como 60 (la parte superior del intervalo del 50 %). También se pueden ver los casos más extremos con respecto al intervalo de credibilidad del 90 %.

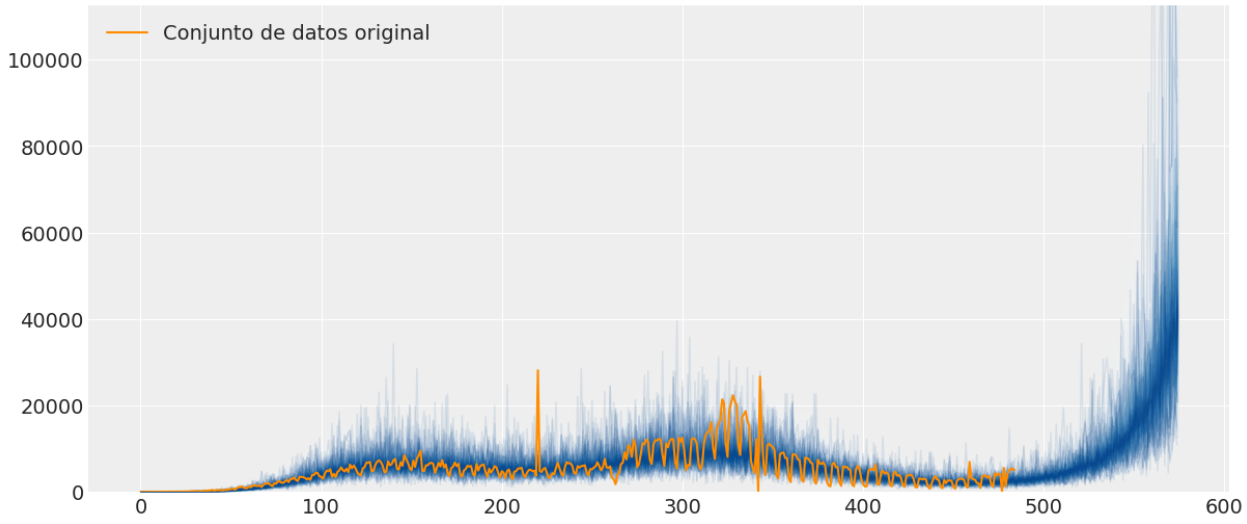

Figura 7: Predicción de los próximos 90 días para México

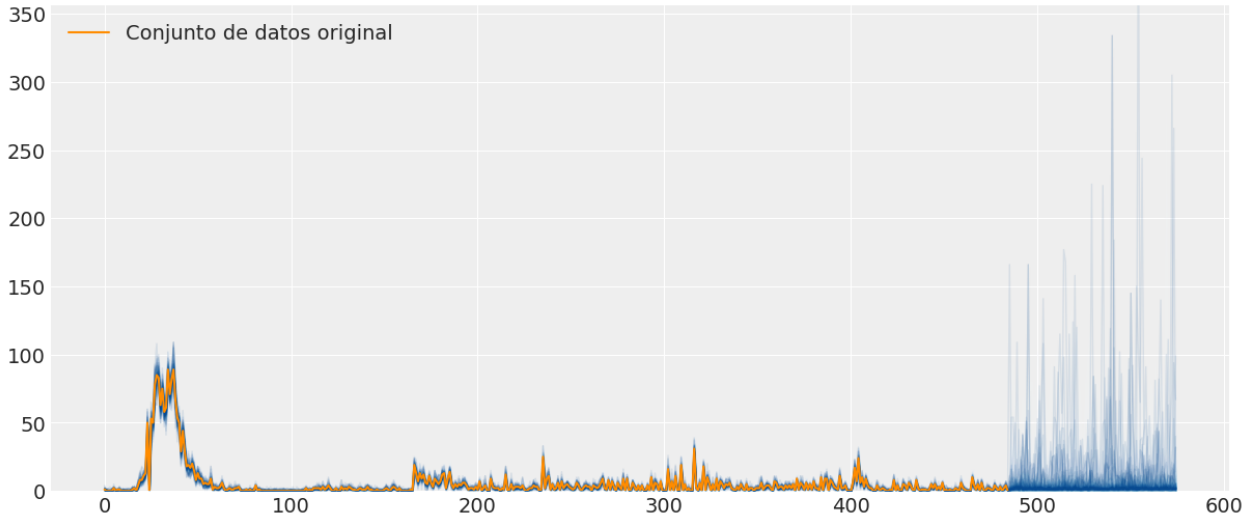

Figura 8: Predicción de los próximos 90 días para Nueva Zelanda

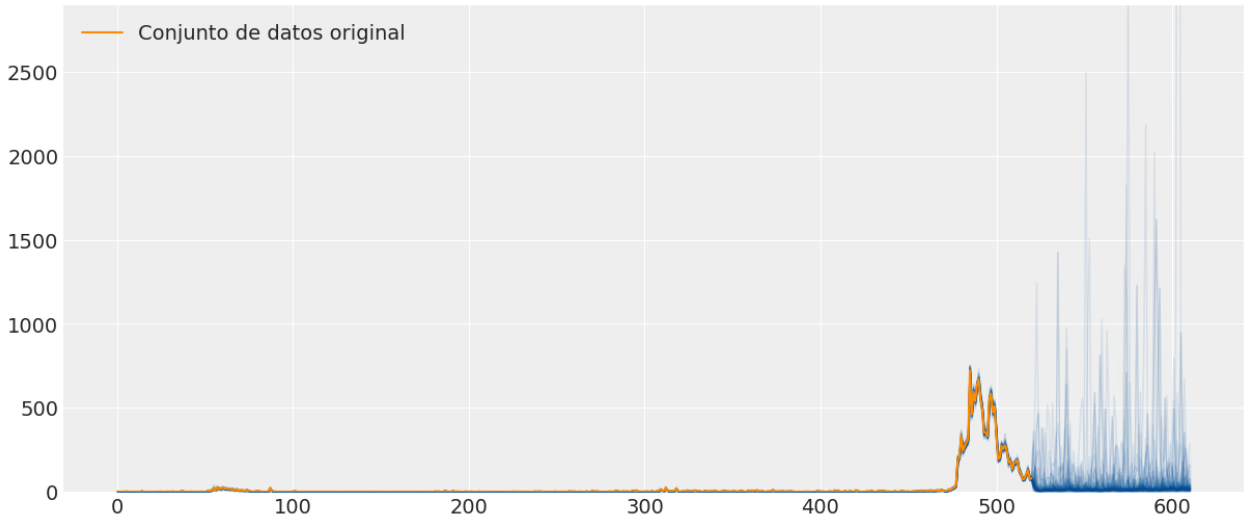

Figura 9: Predicción de los próximos 90 días para Taiwan

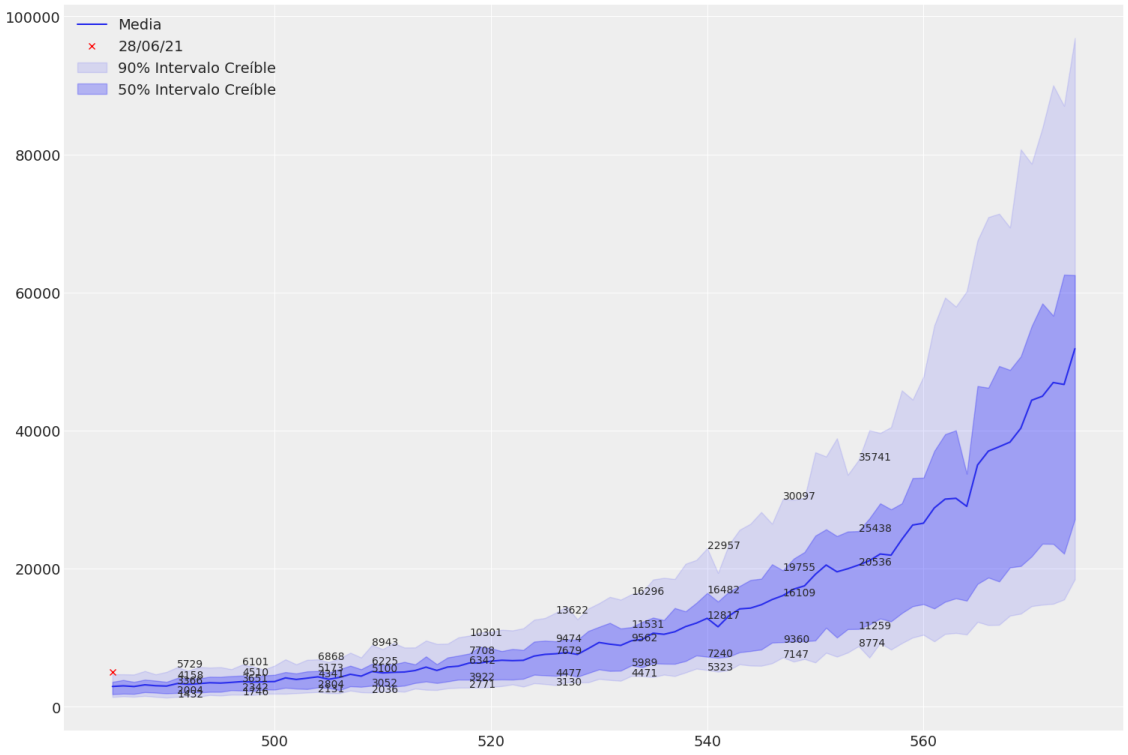

Figura 10: Predicción de los próximos 90 días para México con intervalos creíbles

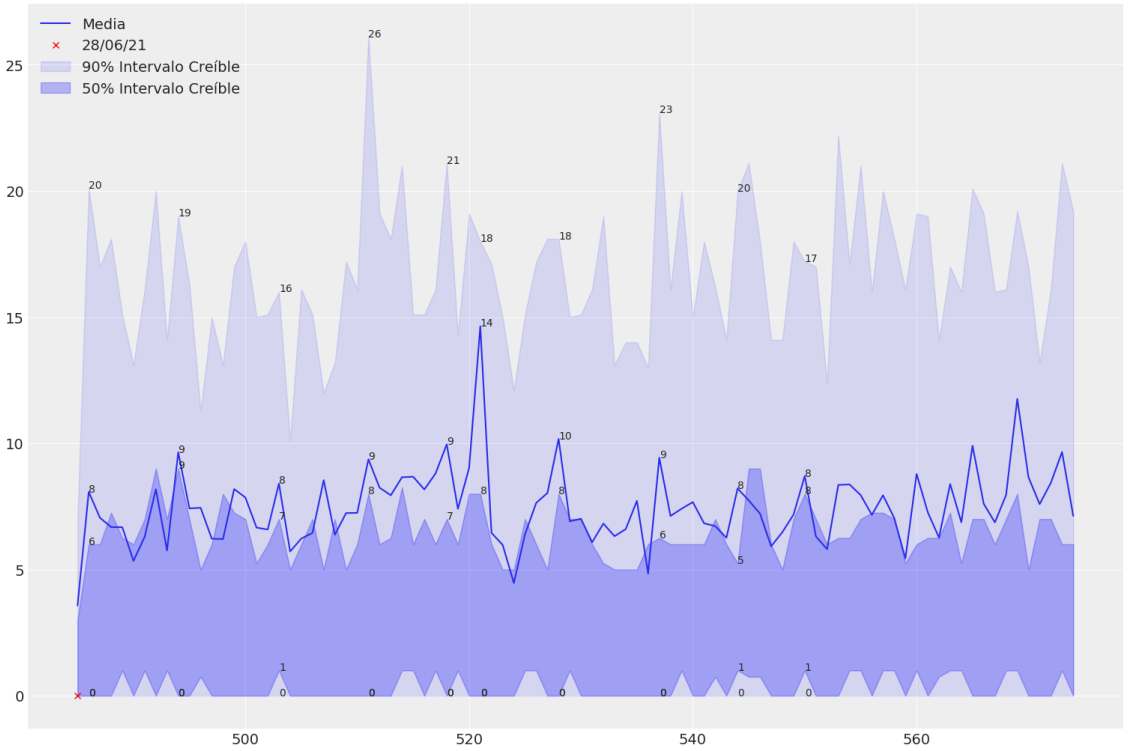

Figura 11: Predicción de los próximos 90 días para Nueva Zelanda con intervalos creíbles

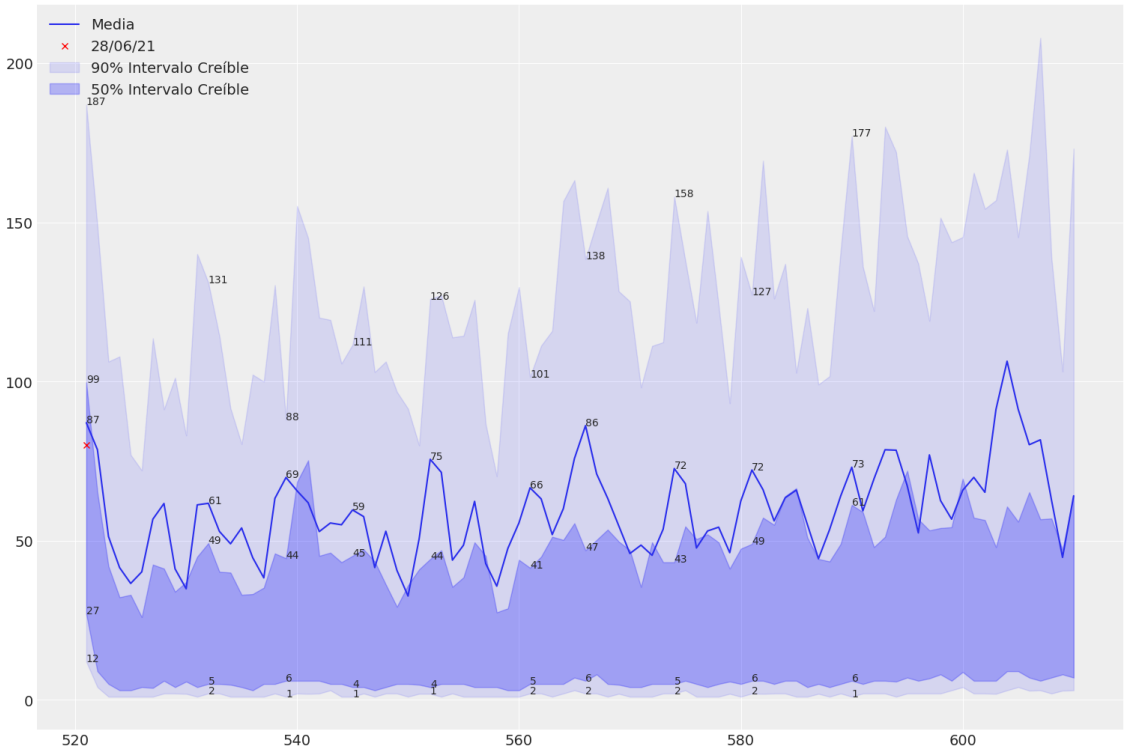

Figura 12: Predicción de los próximos 90 días para Taiwan con intervalos creíbles

VI. Conclusiones

Este artículo muestra el gran uso que se le puede dar a los procesos de regresión Gaussiana cuando los patrones son dificiles de ajustar. Los resultados del modelo diseñado en este trabajo son muy similares a los resultados de otros modelos basado en otros modelos estadísticos.

Por otra parte para México pudimos ver un patrón muy alarmante, ya que la siguiente ola de contagios de COVID-19 en este país será muy fuerte e incluso vendrá con mayor fuerza que las anteriores. En cuanto a Nueva Zelanda y Taiwan pudimos observar patrones muy congruentes con lo datos históricos, ya que se mantienen estables e inclusive llegan a bajar en algún punto. Esto se puede deber a las medidas de sanidad que han llevado hasta ahora.

Referencias

[1] Xiaolei Zhang, Renjun Ma y Lin Wang. "Predicting turning point, duration and attack rate of COVID-19 outbreaks in major Western countries". En: *Chaos, Solitons and Fractals* 135 (2020), pág. 109829. ISSN: 0960-0779. DOI: https://doi.org/10.1016/j.chaos.2020.109829. URL: https://www.sciencedirect.com/science/article/pii/S0960077920302290.

[2] A.M. Mishra, S.D. Purohit, K.M. Owolabi e Y.D. Sharma. "A nonlinear epidemiological model considering asymptotic and quarantine classes for SARS CoV-2 virus". En: *Chaos, Solitons and Fractals* 138 (2020), pág. 109953. ISSN: 0960-0779. DOI: https://doi.org/10.1016/j.chaos.2020.109953. URL: https://www.sciencedirect.com/science/article/pii/S0960077920303520.

[3] Gabriele Martelloni y Gianluca Martelloni. "Modelling the downhill of the Sars-Cov-2 in Italy and a universal forecast of the epidemic in the world". En: *Chaos, Solitons and Fractals* 139 (2020), pág. 110064. ISSN: 0960-0779. DOI: https://doi.org/10.1016/j.chaos.

2020.110064. URL: https://www.sciencedirect.com/science/article/pii/S0960077920304616.

[4] Amit Singhal, Pushpendra Singh, Brejesh Lall y Shiv Dutt Joshi. "Modeling and prediction of COVID-19 pandemic using Gaussian mixture model". En: *Chaos, Solitons and Fractals* 138 (2020), pág. 110023. ISSN: 0960-0779. DOI: https://doi.org/10.1016/j.chaos.2020.110023. URL: https://www.sciencedirect.com/science/article/pii/S0960077920304215.

[5] Our World in Data. *Rate of Daily New Confirmed Cases of Covid-19*. 2021. URL: https://ourworldindata.org/grapher/rate-of-daily-new-confirmed-cases-of-covid-19-positive-rate?yScale=linear.