Modelación de la incidencia de covid en México

Martin Alonso Flores Gonzalez

15 de febrero de 2021

Introducción

A consecuencia del brote epidémico ocurrido en diciembre del 2019 en la ciudad de Wuhan y a causa de la rápida propagación de esta nueva enfermedad hacia otros países, el 30 de enero de 2020 el virus SARS-Cov-2 fue declarado emergencia sanitaria de nivel internacional por la Organización Mundial de la Salud (OMS). El 11 de marzo del 2020 , ya con casos confirmados en más de cien piases, la OMS cataloga a la enfermedad como pandemia.

Dada la nueva categoría de la enfermedad y con el fin de poder anticipar acciones que disminuyeran la tasa de contagios así como alistar los recursos necesarios para el combate contra la enfermedad, han surgido varios trabajos de investigación científica que simulan la dinámica de la enfermedad desde diferentes acercamientos científicos.

Como una de las medidas de prevención, la mayoría de los Mandatarios en diversos países optaron por cerrar sus fronteras evitando la entrada de cualquier persona a sus países. Esto permite estudiar la dinámica de la enfermedad en cada país como un sistema cerrado, donde modelos epidemiologicos compartimentales como Susceptible -Infectado – recuperado (SIR) o Susceptible – Expuesto – Infectado – Recuperado (SEIR) pueden ser muy buen acercamiento a la evolución de la emergencia sanitaria.

En este proyecto se estudiará la dinámica de los casos nuevos de enfermos por Covid en México al emplear el modelo compartimental SEIR, para ello se utilizará la incidencia diaria obtenida de datos reales otorgados en clase, además se estimarán los parámetros del modelo antes mencionados por medio de la estadística Bayesiana así como del uso de MCMC.

SEIR vs SIR.

Antes de comenzar el análisis de los datos, resulta conveniente el explicar el porque de la elección del modelo a utilizar, por ello se da a continuación una breve introducción a los modelos SIR y SEIR.

En el Modelo SIR se cataloga a toda la población en tres compartimentos;

- S: El número de individuos susceptibles. Cuando un individuo susceptible y uno infeccioso entran en çontacto infeccioso", el individuo susceptible contrae la enfermedad y pasa al compartimento infeccioso.
- I: El número de individuos infecciosos. Estos son individuos que han sido infectados y son capaces de infectar a individuos susceptibles.
- R: El número de personas removidas (e inmunes) o fallecidas. Se trata de personas que han sido infectadas y se han recuperado de la enfermedad y han entrado en el compartimento extraído o han muerto. Se asume que el número de muertes es insignificante con respecto a la población total. Este compartimento también puede denominarse recuperado.º resistente".

Como se puede ver en las especificaciones de los anteriores compartimentos, una vez que las personas entran al compartimento R dejan de participar en la dinámica de la enfermedad, es por ello que pareciera que se erradica la enfermedad o se mueren todas las personas. Más aun el modelo clásico en realidad no toma

en cuenta la dinámica de vida, es decir, no hay nuevas incorporaciones ni decesos en la población, por lo que se trabaja con un sistema completamente aislado. Habiendo explicado lo anterior, es momento de presentar el sistema de ecuaciones diferenciales que modela la dinámica de la enfermedad en la población

$$\frac{dS}{dt} = -\frac{\beta SI}{N}$$
$$\frac{dI}{dt} = \frac{\beta SI}{N} - \gamma I$$
$$\frac{dR}{dt} = \gamma I$$

en donde las unidades de N, S, I, R son personas y las unidades de β, γ son 1/tiempo.

En el modelo SEIR se le agrega el compartimento E (expuestos) al modelo SIR, que representa el número de personas que han sido expuestas al virus. Con este nuevo compartimento la dinámica del modelo se ve de la siguiente forma.

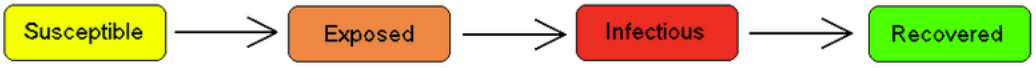


Figura 1: Dinámica de las personas entre compartimentos

El sistema de ecuaciones diferenciales que genran la dinamica de la poblacion en el modelo SEIR se muestra acontinucion.

$$\frac{dS}{dt} = -\frac{\beta SI}{N}$$
$$\frac{dE}{dt} = \frac{\beta SI}{N} - \kappa E$$
$$\frac{dI}{dt} = \kappa E - \gamma I$$
$$\frac{dR}{dt} = \gamma I$$

EL modelo SEIR clásico tampoco tomo en consideración la dinámica vital y por ello también es aplicable a poblaciones cerradas. Es claro que el modelo SEIR es una extensión del modelo SIR por lo que tienen muchas similitudes.

Notemos que en el modelo SIR se toma en cuenta que todas las personas que son susceptibles están expuestas a entornos con agentes infecciosos desde el inicio, por lo que es bastante bueno modelar micro sistemas con este modelo. Un ejemplo sería modelar la dinámica de la enfermedad en un edificio que este aislado de la demás comunidad, en este ejemplo se tendrían personas infectadas en el mismo entorno que el resto de la población susceptible, por lo que seria sencillo aceptar el supuesto de que todos los susceptibles están expuestos desde el comienzo de la dinámica.

Por otro lado, en el modelo SEIR se piensa que no todas las personas susceptibles están expuestas a los agentes infecciosos en el comienzo y que tarda cierto tiempo a que las personas sean expuestas a estos entornos. Un ejemplo claro de esto seria modelar la dinámica de covid en México. Supongamos que el comienzo de la epidemia se da en la C.D México, entonces las personas que están en Jalisco no están expuestas a los agentes infecciosos en el comienzo y de hecho tarda cierto tiempo en que la gente de Jalisco comience a estar expuesta, ya que para esto se tiene que tener cierta evolución en la dinámica de la enfermedad en la ciudad de origen y después haber ciertas migraciones para llevar el virus de un estado a otro. AL haber cierto tiempo hasta que ciertos sectores de la población sean expuestos a agentes infecciosos, el modelo SEIR es mejor para modelar la contingencia sanitaria en todo México y por ello se ha elegido como modelo a utilizar en este proyecto.

Elección de los datos de entrenmiento

En el comienzo del estudio se pre seleccionó la información a utilizar de la base de datos otorgada, selección que se realizó al tomar en cuenta la fecha del inicio de síntomas[4] así como la columna de respuesta en donde se verificaba si la causa de los malestares eran por el covid. De esta manera se iban registrando el número de casos positivos por día (incidencia diaria), obteniendo lo que se muestra a continuación.

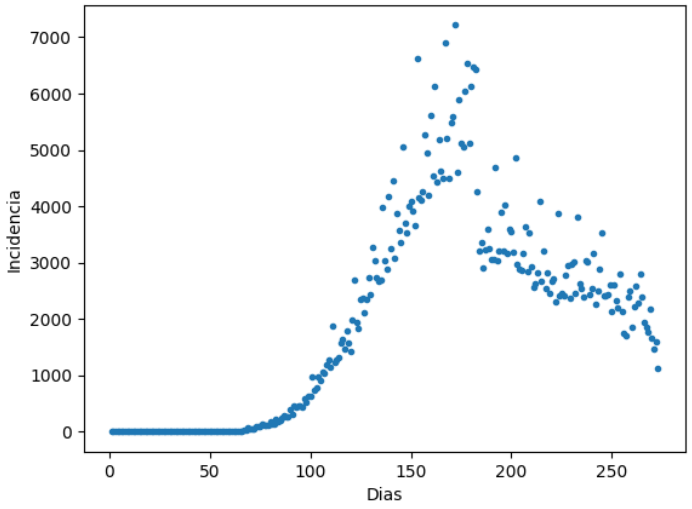


Figura 2: Incidencia diaria del 01/1/2020 al 29/9/2020

Para la selección de los datos de entrenamiento se tomo en consideración que la dinámica de la enfermedad comienza cuando una persono está infectada y como el primer infectado en México aparece el 22 de febrero (según la base de datos), se tomó como muestra de entrenamiento los datos de incidencia diaria desde el 22 de febrero hasta el 1 de septiembre, obteniendo la siguiente grafica.

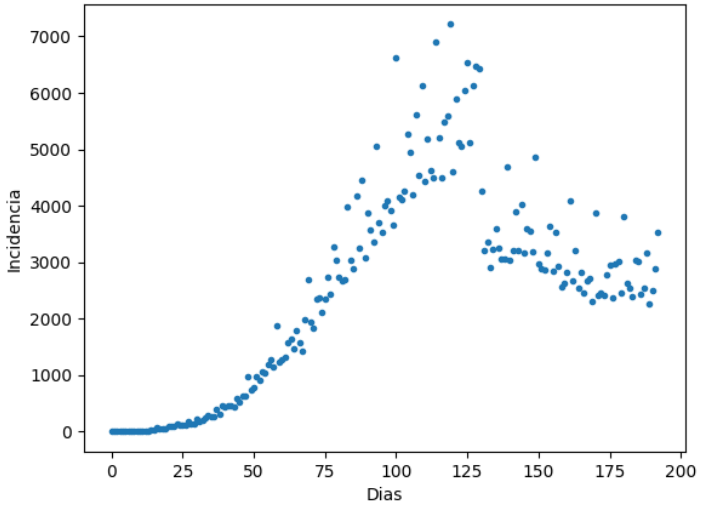


Figura 3: Incidencia diaria del 22/2/2020 al 29/9/2020

Si bien es cierto que se tienen datos desde el primero de enero, los datos previos al primer infectado son inútiles dado que no se puede modelar nada cuando una enfermedad no está presente en el sistema. Tomemos en cuenta que nos interesa la evolución de la incidencia en México y aunque para enero en otros países ya se tenia cierta evolución en la dinámica de la enfermedad, en México todavía no comenzaba la dinámica.

Al estudiar de manera detallada la figura (3), se puede observar como los datos parecieran ser parte de dos dinámicas diferentes, es decir, los datos previos al día 125 parecieran tener la misma dinámica, si se puede notar cierto ruido en los datos pero a pesar de ello la nube de puntos sigue una clara tendencia. Por otro lado, los datos que se encuentran después del día 125 parecen tener una dinámica completamente diferente a los primeros datos, se puede observar también cierto ruido y los datos parecen agruparse en dos cúmulos de puntos diferentes.

Cabe señalar que la reducción en la incidencia diaria que se aprecia después del día 125 se puede deber a las medidas de prevención tomadas por el gobierno, tales medidas como la cuarentena y el uso obligatorio de cubre bocas, aunque no se ha podido confirmar dicha suposición.

SEIR utilizado

Ya se ha hablado del modelo SEIR, pero ahora se tiene que hacer notar que en el modelo descrito anteriormente no se habla de la incidencia en ningún momento, incluso no aparece en el sistema de ecuaciones que modela la dinámica. Aunque no aparezca explícitamente, la información de la incidencia se encuentra en el sistema de ecuaciones y es justamente el número de personas que van entrando al compartimiento de infectados por cada paso de tiempo y con el fin de obtener los datos de la incidencia, se trabaja con el siguiente sistema de ecuaciones diferenciales.

$$\frac{dS}{dt} = -\frac{\beta SI}{N} \tag{1}$$

$$\frac{dE}{dt} = \frac{\beta SI}{N} - \kappa E \tag{2}$$

$$\frac{dI}{dt} = \kappa E - \gamma I \tag{3}$$

$$\frac{dC}{dt} = \kappa E \tag{4}$$

$$\frac{dR}{dt} = \gamma I \tag{5}$$

Como se puede apreciar en el anterior sistema, la variable C no interactuá de manera explicita con el resto de las ecuaciones diferenciales, por lo que no afecta la dinámica. En este caso C representa la incidencia acumulada con forme avanza el tiempo, de manera que las incidencias diarias estarán dadas por lo siguiente

$$\begin{aligned} Y_0 &= C_0 \\ Y_t &= C_t - C_{t-1}, \end{aligned} \tag{6}$$

recordemos que la unidad de tiempo base es el día, por lo tanto $t = 0, 1, 2, ...$ Recordemos también que las unidades de las variables S, E, I, C, R estan en personas ($[p]$) y las unidades de los parámetros β, κ, γ están en 1/dias ($[1/d]$).

Teniendo una metodología clara para la recuperacion de la informacion deseada a partir del modelo $SEIR$, ya se puede comenzar a realizar la inferencia sobre los párametros de β, κ, γ, pero antes es necesario mencionar que las condiciones iniciales para la solucion del sistema son ($S_0 = 1.2 \times 10^8 - 5 - 1, E_0 = 5, I_0 = 1, C_0 = 1, R_0 = 0$).

Inferencia Bayesiana: elección de prioris y del modelo

La información relevante del coronavirus que se conoce, es el tiempo de incubación del virus en las personas (de 2 a 14 días) y el tiempo medio de recuperación a partir del inicio de los síntomas que es de 14 días. En cuanto al tiempo de exposición, se tiene que mencionar que no se ha encontrado datos reales que describan el tiempo que transcurre hasta que una persona se expone a entornos infecciosos, sin embargo muchos de los artículos proponen esta cantidad [**?**].

Después de leer varios artículos, se decidió seguir lo realizado en [**?**, **?**] en donde se eligen distribuciones Log Normales como distribuciones a priori de los parámetros, que en combinación con la información antes mencionada, se tiene que las distribuciones a priori utilizadas en este proyecto son las siguientes.

$$\begin{aligned} \beta &\sim LogN(\mu = 0, \sigma = 1) \\ \kappa &\sim LogN(\mu = \log(1/12), \sigma = 0{,}5) \\ \gamma &\sim LogN(\mu = \log(1/14), \sigma = 0{,}5) \end{aligned} \tag{7}$$

Los parámetros para la priori de β son los mismo que los propuestos es [**?**]. Los parámetros de las distribuciones a priori de κ y γ se hicieron al tomar en cuenta que κ^{-1} = días de incubación y γ^{-1} = días de recuperación.

Para la elección del modelo se siguió lo realizado en [**?**, **?**] en donde se propone a la distribución Binomial Negativa para el análisis de datos con mucho ruido, tal es el caso de los datos mostrados en la figura(3). En dicha propuesta se utilizan dos parámetros de dispersión θ y ω, tales que la varianza de la distribución es una función cuadrática de la media, es decir, si m y v^2 son la media y la varianza de la distribución Binomial Negativa respectivamente, entonces se cumple lo siguiente [**?**]

$$v^2 = \omega m + \theta m^2,$$

haciendo que los parámetros de la distribución sean $p = \frac{m}{v^2}$ y $r = \frac{m^2}{v^2 - m}$.

De manera que si x_i son los datos observacionales de incidencia diaria e Y_i es el dato de la incidencia obtenido por medio de las ecuaciones (4) y (6), el modelo de las x‘s está dado como

$$x_i \sim BN(Y_i, \theta, \omega),$$

donde $\theta = 0.5$ y $\omega = 2$, siguindo lo hecho en [**?**].

Habiendo explicado la elección de las distribuciones a priori y la elección del modelo para los datos de entrenamiento, fue posible conocer numéricamente el valor para la distribución posterior de $(\beta, \kappa, \gamma)|\vec{X}$, con la que se definió la función de energía y se procedió a realizar el twalk.

Simulación de MCMC

Para la parte de la simulación estocástica, se utilizo el algoritmo twalk desarrollado por el DR. Christen [**?**], con la función soporte como $\mathbf{1}_{\beta>\mathbf{0},\kappa>\mathbf{0},\gamma>\mathbf{0}}$ y la función de energía antes mencionada. Al simular un millón de iteraciones de la cadena, se obtiene la siguiente evolución.

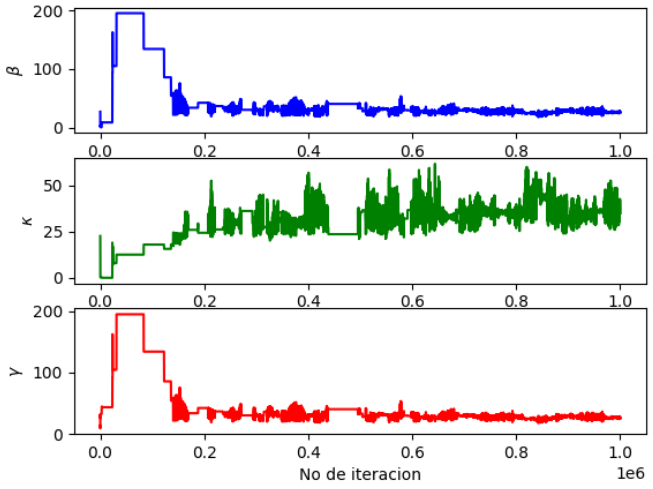


Figura 4: Evolución de la cadena pre burn in

En la anterior imagen se puede observar que desde la iteración $200,000$ la cadena comienza a tener cierta regularidad en los patrones, por ello se ha elegido este tiempo como tiempo de burn in. Al graficar los datos pos quema se tiene lo siguiente.

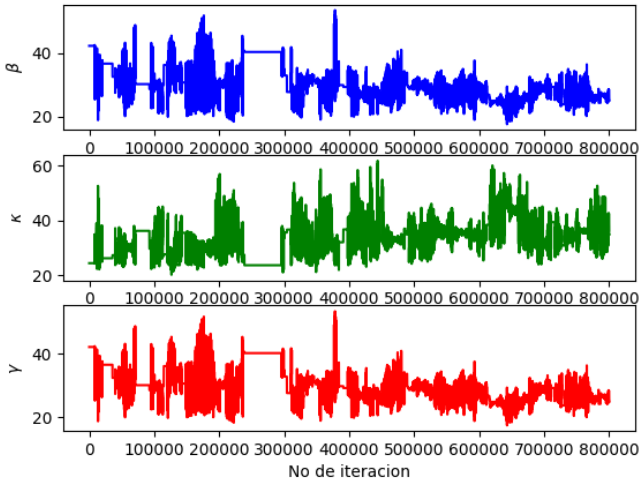
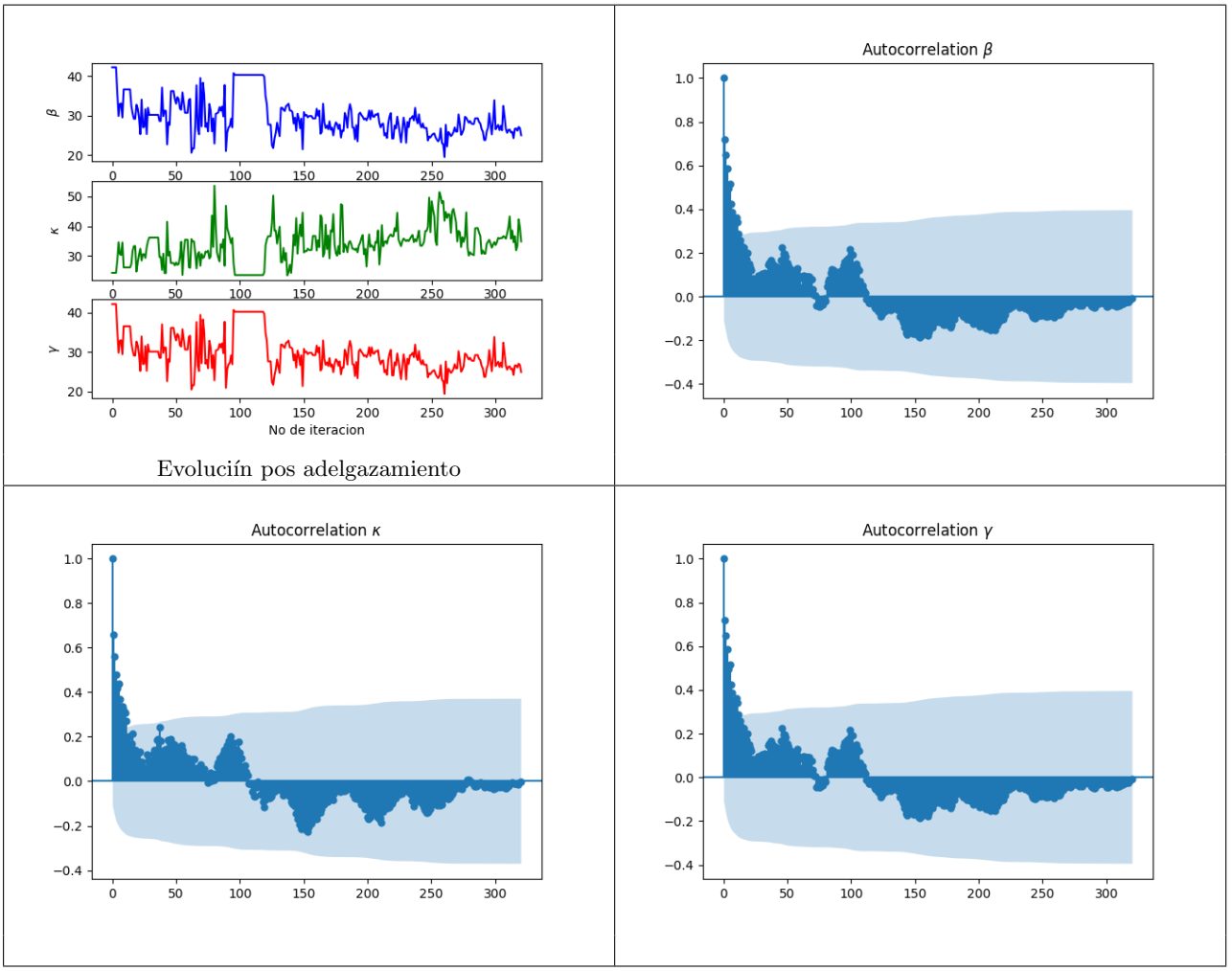


Figura 5: Evolución de la cadena pos burn in

Al analizar la anterior imagen se pueden apreciar pocos intervalos con lag, aunque se aprecia un intervalo bastante amplio. Si se toma en cuanta que el twalk pronosticó que el tiempo de muestra efectiva era de 2500, entonces es posible adelgazar la cadena para conseguir datos relativamente independientes tal y como se muestra en la siguiente a continuación.



Evoluciín pos adelgazamiento

Al ver la evolución de la cadena pos quema podemos seguir apreciando que la cadena se mantiene constante en un intervalo considerablemente grande, esto ocasionará cierto sesgo en los datos. Por otro lado, en las graficas de auto correlación se puede apreciar que los primeros datos están altamente correlacionados, pero esta correlación decrece para permanecer dentro de los limites permitidos.

Resultados

Al tener ya procesados los datos de la cadena, es posible hacer el siguiente analisis stedistico de ellos.

	β	κ	γ	R_0
conteo	321	321	321	321.
media	29.773406	33.703337	29.691522	1.002763
mediana	29.217913	33.847994	29.140261	1.002764
moda	27.34471215	35.7621316	27.27052642	–
std	4.783390	6.044023	4.770834	0.000205
min	19.532051	23.514574	19.467814	1.002299
max	42.284679	53.566151	42.166104	1.003388
$Q_{2,5\,\%}$	22.503982	23.628595	22.436230	1.002408
$Q_{25\,\%}$	26.614997	30.150399	26.536330	1.002621
$Q_{75\,\%}$	31.511268	36.642938	31.433408	1.002880
$Q_{97,5\,\%}$	40.353192	47.376264	40.238948	1.003173

Cuadro 1: Estadísticas descriptivas

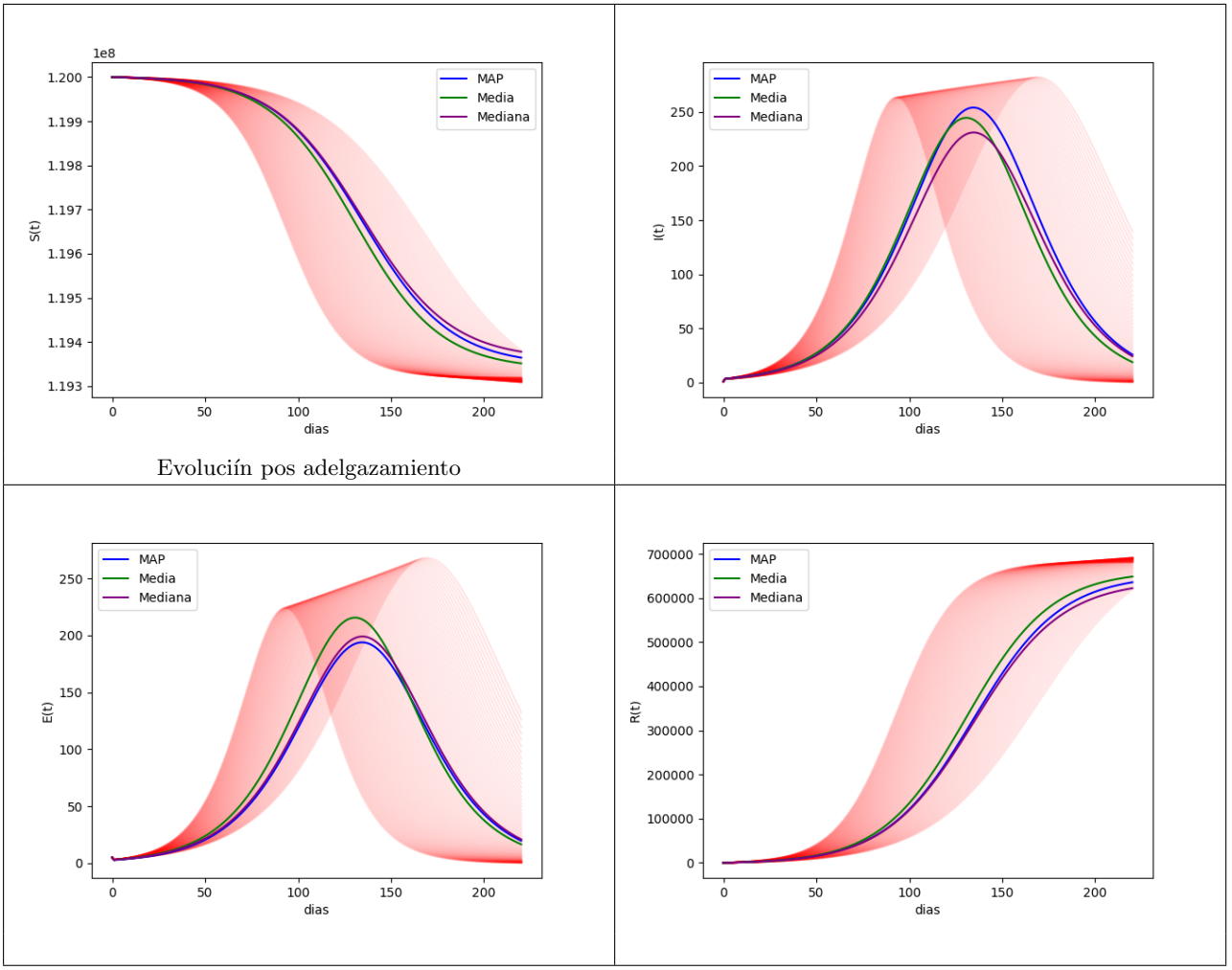
También es posible obtener los intervalos de máxima densidad posterior, los cuales se muestran a continuación.

β	κ	γ	R_0
(23.261, 40.353)	(23.514, 44.086)	(23.189, 40.2389)	(1.0024, 1.0031) .

Cuadro 2: IBC al 95 %

Como es posible observar, los IBC's no contienen a los datos que se proponen en la tarea 2, además es posible notar que los intervalos de los cuantíles del 95 % contienen a los IBC's, por ello se utilizaran los cuantíles para realizar las bandas de confianza.

A continuación se muestran las graficas de las soluciones al sistema de ecuaciones SEIR con los datos obtenidos en el análisis estadístico anterior.



Evoluciín pos adelgazamiento

En las anteriores graficas se muestran los resultados para S, E, I, R al utilizar los estimadores puntuales originados de las funciones de pérdida error cuadrático, valor absoluto y 0-1, así como las respectivas bandas de confianza del 95 % en color rojo.

A continuación se mostrará el ajuste de los datos restantes utilizados en la figura (3), datos que van desde el 22 de febrero hasta el 29 de septiembre.

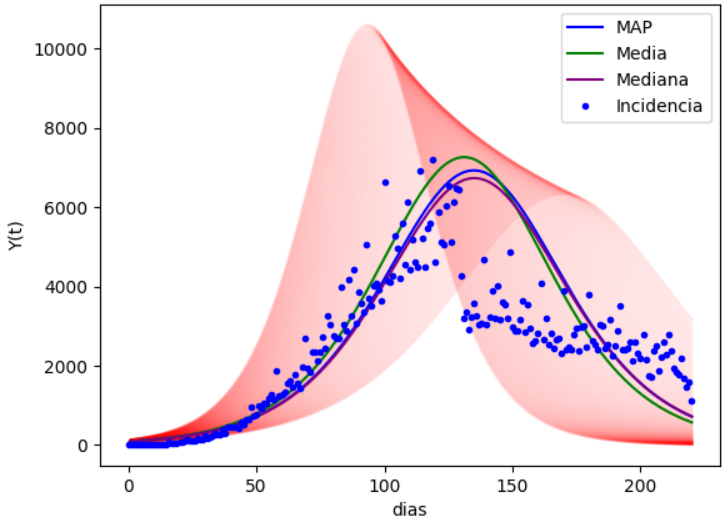


Figura 6: Ajustes para la incidencia

De igual manera que con las graficas anteriores, se muestran los ajustes obtenidos al utilizar los estimadores puntuales originados de las funciones de pérdida error cuadrático, valor absoluto y 0-1, así como las respectivas bandas de confianza del 95 % en color rojo.

Para el análisis de residuales del ajuste realizado a la incidencia diaria, se tiene que $e_i = y_i - \hat{y}_i$, donde y_i es la incidencia observada en el i-ésimo día, mientras que $\hat{y}_i$ es la i-ésima incidencia ajustada para el mismo día. Teniendo lo anterior, es posible graficar los residuales obteniendo lo siguiente.

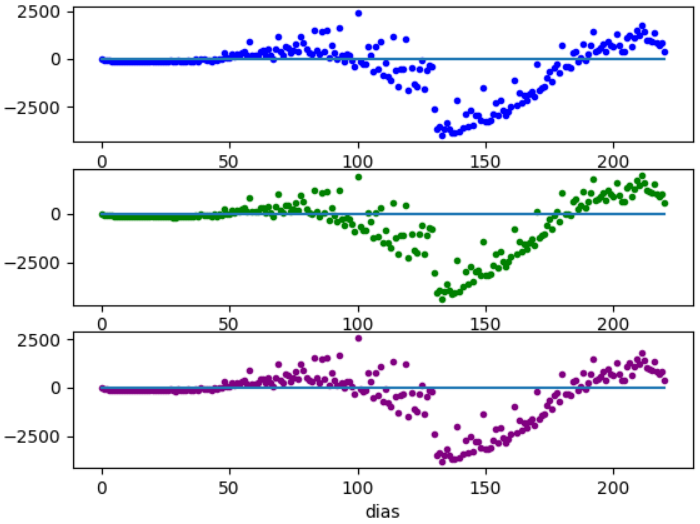


Figura 7: Grafica de residuales

En la anterior imagen se grafican los residuales con los ajustes hechos por los estimadores puntuales de moda (color azul), media (color verde) y mediana (color purpura).

Conclusión

Después de haber realizado el anterior estudio, es posible obtener varias conclusiones sobre el proceso y sobre los resultados obtenidos.

Elección de prioris y resultados de los parámetros: Como se menciono anteriormente, la elección de los parámetros de las distribuciones a priori se hicieron al conocer la relación que existe entre los parámetros κ y γ con los tiempos de incubación (12 días) y de recuperación (14 días). Sin embargo al analizar los intervalos de máxima densidad posterior nos encontramos que el periodo de incubación en realidad debería de ser de entre 0.02268 a 0.4252 días, mientras que el tiempo de recuperación debería de ser de entre 0.0248 a 0.04312 días, tiempos que no se ven en la vida real.

Número básico de reproducción R_0: En base al valor mínimo valor que se registra en el cuadro(1), R_0 es mayor que uno por lo que la enfermedad se puede expandir de manera masiva en la población si no se interviene y se toman las medidas de prevención adecuadas.

Figura (6): Al mirar esta grafica, es posible apreciar varios puntos que salen de la zona de confianza, señal de que el ajuste realizado no es lo suficientemente bueno para los datos con los que se cuenta. No obstante, se puede observar que la totalidad de los datos de septiembre están dentro de la zona de confianza.

Figura (7): Al ver la grafica de los residuales se puede confirmar las sospechas de una ajuste deficiente. Se puede ver que los puntos no se distribuyen de manera aleatoria al rededor del eje, de hecho se pueden ver claramente que los residuales tienen cierto patrón.

Pensamientos finales: En vista de los resultados obtenido, es necesario pensar en utilizar otros modelos que se puedan ajustar de mejor manera a los datos de campo que se obtienen y con ello tener mejor capacidad de predicción. Tras leer varios artículos, fue revelador el saber que en ninguno de ellos se emplea el modelo SEIR clásico, de hecho siempre se utilizan modificaciones que utilizan más compartimentos y más parámetros. También fue curioso notar que en varios artículos se tomaban en cuenta las acciones que los gobiernos habían tomado durante el brote para modelar los datos con diferentes sistemas SEIR, como si se tratara de una mezcla de modelos. Pensando en el Modelo Clásico, quizá una manera de mejorar el ajuste sería estimando las condiciones iniciales del sistema de ecuaciones diferenciales, así como la elección de otras distribuciones a prioris.

Referencias

[1] Mehmet E. Aktas, Esra Akbas and Ahmed El Fatmaoui , *Persistence homology of networks: methods and applications*, Springer open, Applied Network Science, 2019

[2] Slobodan Maletić, Milan Rajkovic , *Simplicial Complexes of Networks and Their Statistical Properties*, Researchgate, 2008

[3] Alexander P. Kartun-Giles, Ginestra Bianconi, *Beyond the clustering coefficient: A topological analysis of node neighbourhoods in complex networks*, Chaos, Solitons and Fractals: X, Volumen 1, 2019

[4] RAÚL RABADÁN , *TOPOLOGICAL DATA ANALYSIS FOR GENOMICS AND EVOLUTION*, Cambridge University Press, 2020