Mezcla de Gompertz con covariables para modelar la pandemia por COVID-19

XIX Escuela de Probabilidad y Estadística
CIMAT,Guanajuato

Graciela González Farías

CIMAT
farias@cimat.mx

Abril 2021

Trabajo conjunto con: Roberto Vásquez Martínez
(Universidad de Guanajuato-Ayudante de Investigador Nacional SNI III
roberto.vasquez @cimat.mx)

CONACYT

CIMAT

Colaboradores

- Dr. José Ulises Márquez Urbina (CIMAT)
- Dr. Rogelio Ramos Quiroga (CIMAT)
- Dr. Iván Rodríguez González (CIMAT)
- Dr. Norberto Alejandro Hernández Leandro (CIMAT)
- CentroGEO

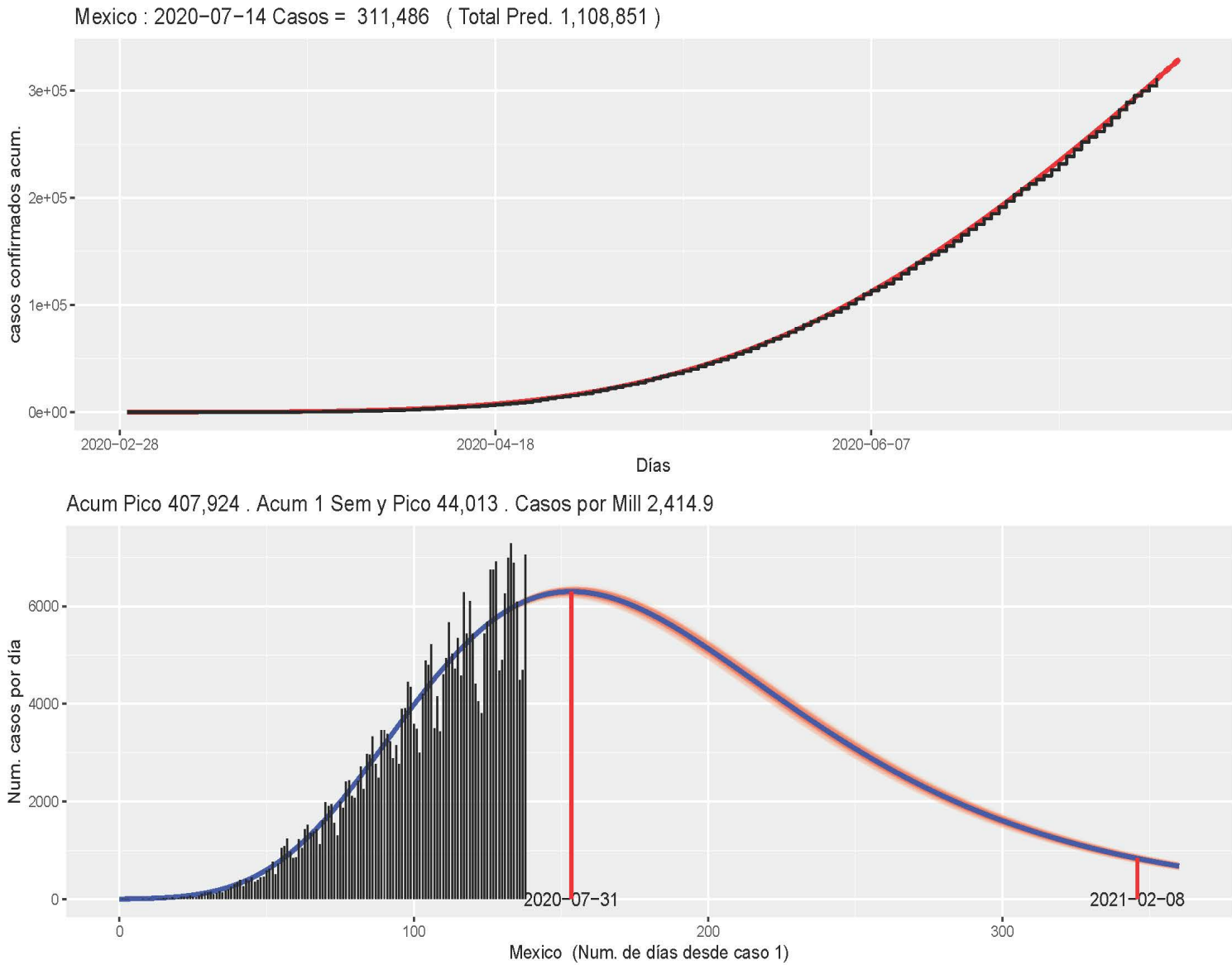
Outline

Ideas iniciales de modelación

- Modelo Gompertz
  - **Prats, C. *et al.*** (2020) Analysis and prediction of COVID-19 for different regions and countries. Daily report 27-03-2020. UPC, BioComSC, CMCiB, IGTP.
  - **IHME COVID-19 forecasting team** (2020). Forecasting COVID-19 impact on hospital bed-days, ICU-days, ventilator days and deaths by US state in the next 4 months. Report (Modelo Normal).

Etapa temprana

En un etapa temprana de la pandemia fue buena opción



Modelo Gompertz

- Se eligió un modelo de tres parámetros:

$$N(t) = \alpha \exp\left(-\beta\, e^{-\kappa t}\right)$$

donde $N(t)$ son los casos confirmados acumulados.

- La asíntota α corresponde al total de infectados al final de la pandemia. Los estimadores son muy sensibles a la cantidad y tipo de datos y las tasas de crecimiento
- La derivada del modelo es un "*proxy*" para el número de casos nuevos

$$C(t) = N'(t) = \beta\, \kappa\, N(t)\, e^{-\kappa t}$$

- El tiempo en el que se alcanza el número máximo de infectados diarios se obtiene derivando la función de casos nuevos $C(t)$

$$t_{\text{máx}} = \frac{\log(\beta)}{\kappa}.$$

Modelo Gompertz

- El número de casos en el tiempo máximo $t_{\text{máx}}$ está dado por

$$C(t_{\text{máx}}) = \kappa\,\beta\,N(t_{\text{máx}})\,e^{-\kappa\ t_{\text{máx}}}$$

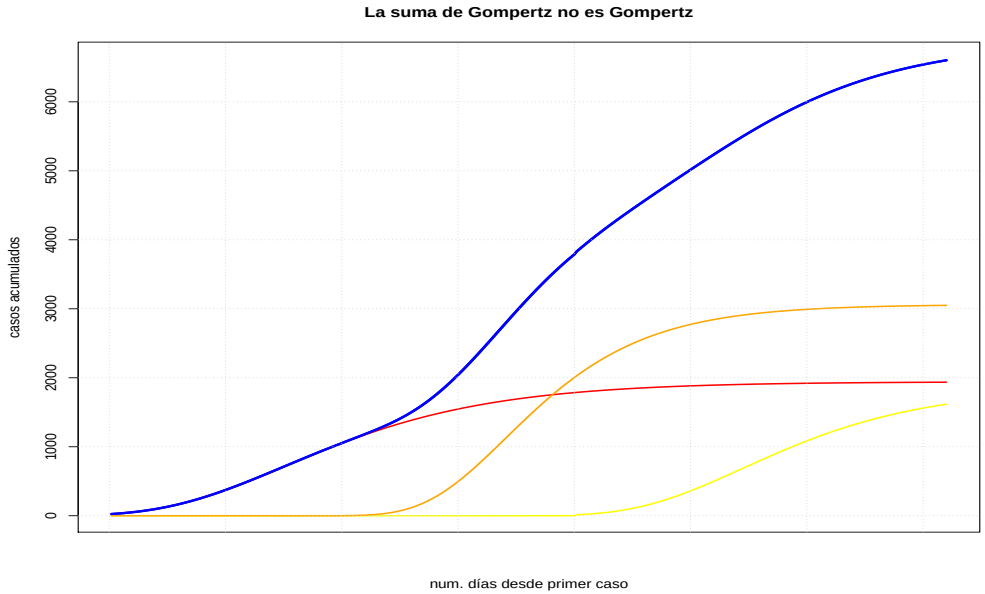
- El número de casos acumulados al momento de máxima incidencia es $N(t_{\text{máx}})$.
- El *fin de la epidemia* puede aproximarse por una fracción del valor de α.
- El modelo puede estimarse a través de mínimos cuadrados no-lineales, el cual, bajo normalidad, es equivalente a MC. Intervalos de confianza predictivos para la incidencia acumulada al tiempo t^* pueden encontrarse mediante

$$y(t^*;\widehat{\theta}) \pm z_{1-\nu/2}\left(\widehat{\sigma}^2 + \widehat{S}^2\right)^{\frac{1}{2}}$$

$$\widehat{S}^2 = \widehat{\sigma}^2\left(\frac{\partial}{\partial\theta}y(t^*;\widehat{\theta})\right)^T\left[\left(\frac{\partial}{\partial\theta}y(\mathrm{t};\widehat{\theta})\right)^T\left(\frac{\partial}{\partial\theta}y(\mathrm{t};\widehat{\theta})\right)\right]^{-1}\left(\frac{\partial}{\partial\theta}y(t^*;\widehat{\theta})\right)$$

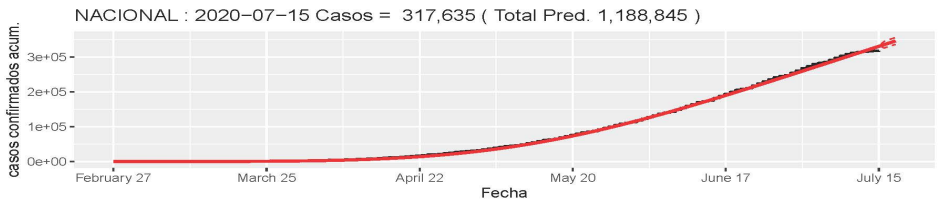
(Huet, S. *et al.* (2004). Statistical Tools for Nonlinear Regression. Springer).

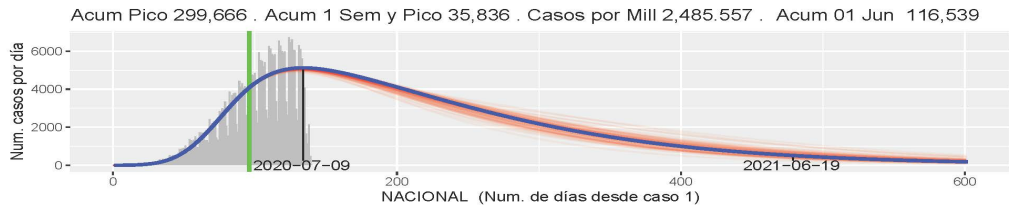
Modelo Gompterz por estado



CONACYT CIMAT

Estimaciones obtenidas a partir de los datos oficiales del Gobierno Federal al día 15/07/2020.





Se aplicó a los datos un promedio móvil, considerando para cada día el número de casos totales de ese día y los dos días anteriores.
Se descartaron, además, los últimos días para los que no hubo registro de nuevos casos.

Modelo Gompertz por estado: Modelos mixtos no lineales

- Elegimos un modelo con efectos aleatorios para datos activos a un nivel regional (estados y áreas metropolitanas), suponiendo que en una región específica los parametros puede considerarse como variables aleatorias siguiente una distribución conocida.
- El supuesto anterior es razonable para México. Aunque existen instrucciones federales, cada estado sigue sus propias políticas, lo que crea heterogeneidad entre estados diferentes.
- El modelo clásico de efectos aleatorios queda estructurado de la siguiente forma

$$y_{ij} = g(t_{ij}; \theta_i) + \epsilon_{ij}, \qquad i = 1, \ldots, n, \quad j = 1, \ldots, n_{ij}$$

donde g es el modelo Gompertz y

$$\epsilon_{ij} \sim N(0, \sigma^2)$$
$$\theta_i \sim N(\theta, \Omega)$$

son variables aleatorias independientes.

Modelo Gompertz por estado: Modelos mixtos no lineales

- El supuesto de normalidad en $\theta_i = (\alpha_1, \beta_i, \kappa_i)$ es flexible y en nuestro caso, dada la naturaleza no negativa de los parámetros, usamos una distribución log-normal.
- La inferencia en regiones específicas fue hecha considerando las distribuciones condicionales $p(\theta_i|y_i, \widehat{\theta}, \widehat{\Omega})$, que es equivalente a un método empírico bayesiano.
- Se han producido reportes con tablas y resúmenes gráficos por estados y áreas metropolitanas

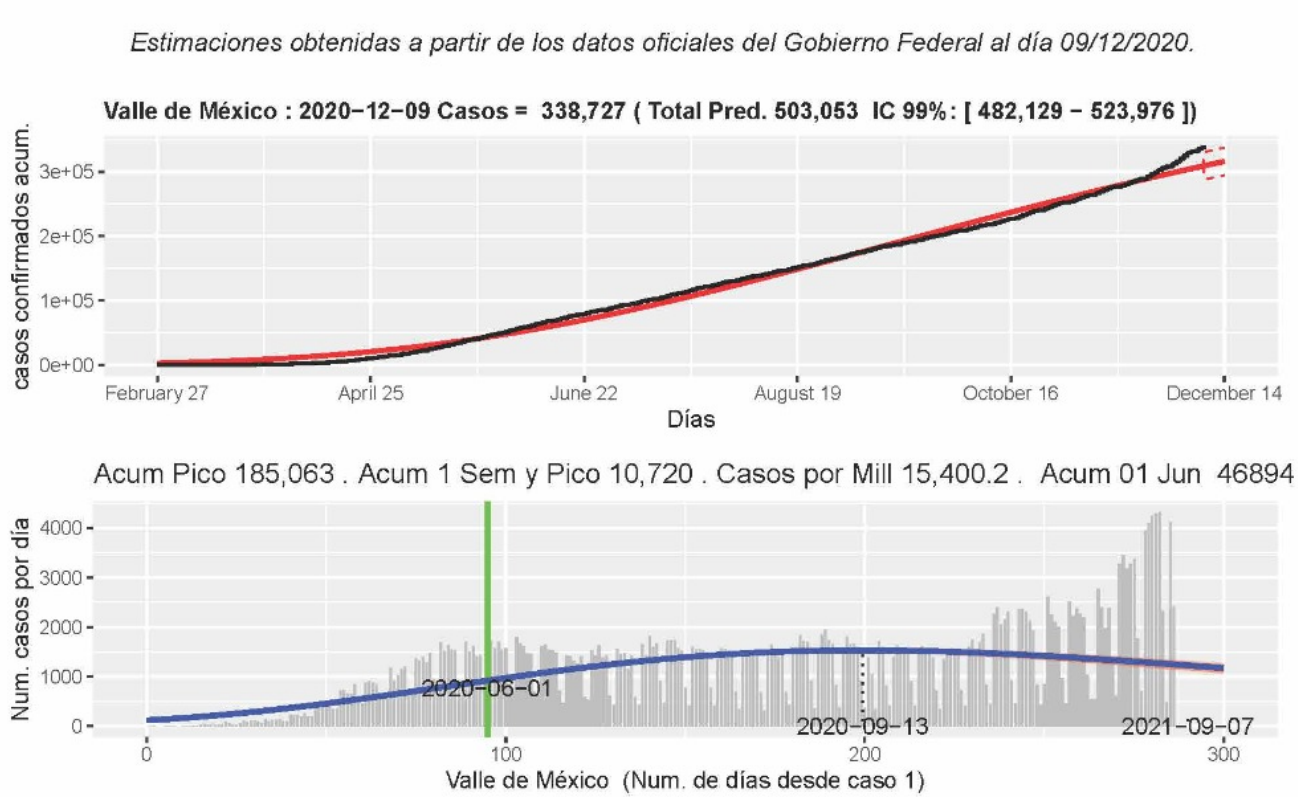
(https://coronavirus.conacyt.mx/).

Gompertz: Datos

- Datos de diferentes países no son comparables; algunos países hicieron pruebas masivas mientras que otros, como México, sólo consideraron casos positivos reportados en hospitales públicos.
- El modelo Gompertz no introduce en forma directa valores atípicos, cambios de dinámica producidos por efecto de las distintas políticas públicas, que generan el equivalente a mezclas de posibles modelos Gompertz y que se han observado en estos tiempos más recientes a todos los niveles: país, estado, zonas metropolitantes, etc.

Outline

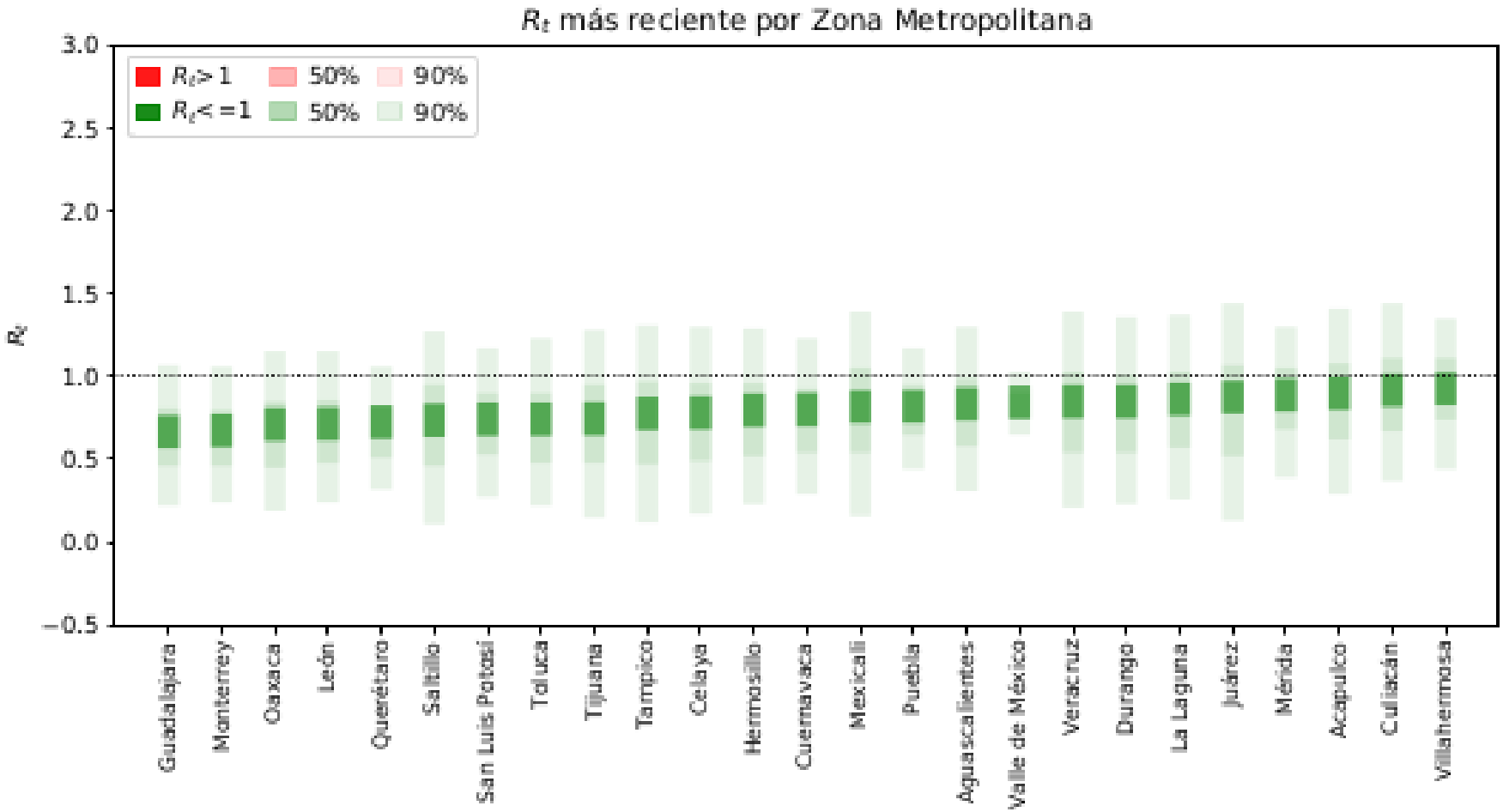
Modelo de Mezclas Gompertz



- El modelo de mezclas se desarrolló para usar una **covariable** que refleje la dinámica de la pandemia y trate de capturar el efecto bimodal de la misma.

Covariable R_t

- Es **la tasa reproducción efectiva al tiempo** t, es decir, el número promedio de infectados por una persona al tiempo t.



R_t Valle de México

Modelo de mezclas Gompertz

- R_t resulta entonces una variable directamente relacionada con el riesgo de contagio.
- Por ende se incorpora al modelo en el riesgo de contagio que representaremos como: **la función de riesgo de una mezcla de distribuciones Gompertz**.
- Para darle mayor flexibilidad a la mezcla consideraremos una **proporción logística relacionada con el R_t.**

Modelo de mezclas Gompertz

- Nos basaremos en la acumulada y la densidad de una mezcla de Gompertz, la densidad queda definida

$$f(t; a, \Psi) = \sum_{i=1}^{2} \pi_i(a; \alpha) f_i(t; a, \theta_i),$$

con $\theta_i = (\lambda_i, \kappa_i, \gamma_i)^T$, t el tiempo, a la covariable, $\psi = (\alpha, \theta_1, \theta_2)$ y

$$\begin{aligned} \pi_1(a; \alpha) &= 1 - \pi_2(a; \alpha) \\ &= \frac{e^{\alpha_1 + \alpha_2 a}}{1 + e^{\alpha_1 + \alpha_2 a}}, \end{aligned}$$

la proporción logística.

Modelo de mezclas Gompertz

Las parametrizaciones que utilizamos de la densidad y la función de riesgo son

$$
\begin{aligned}
f_i(t; a, \theta_i) &= h_i(t; a, \theta_i) \exp\{-\Lambda_i(t; a, \theta_i), \\
h_i(t; a, \theta_i) &= \exp(\gamma_i a + \lambda_i + \kappa_i t) = \exp(\gamma_i a) h_{i,0}(t|a, \theta_i), \\
\Lambda_i(t; a, \theta_i) &= \frac{1}{\kappa_i} [\exp(\gamma_i a + \lambda_i + \kappa_i t) - \exp(\gamma_i a + \lambda_i)]
\end{aligned}
$$

Modelo de mezclas Gompertz

- A través de una transformación podemos recuperar la acumulada de infectados ($N(t)$) a partir de la función de distribución.
- La curva de infectados diarios ($C(t)$) se recupera a partir de la densidad.
- El cálculo de los parámetros se hace vía un algoritmo ECM multiciclo (Expectation/Conditional Maximization)[1].
- Debemos hacer un muestro de los datos de infectados diarios observados para obtener una muestra de tiempos de fallo (el tiempo al que una persona se contagia).

[1] Ver [3]

Predicción de la cola

- Al momento de hacer el muestreo y aplicar el algoritmo ECM suponemos que la muestra proviene de una mezcla de distribuciones Gompertz, sin embargo proviene de una distribución truncada hasta donde tenemos datos observados.

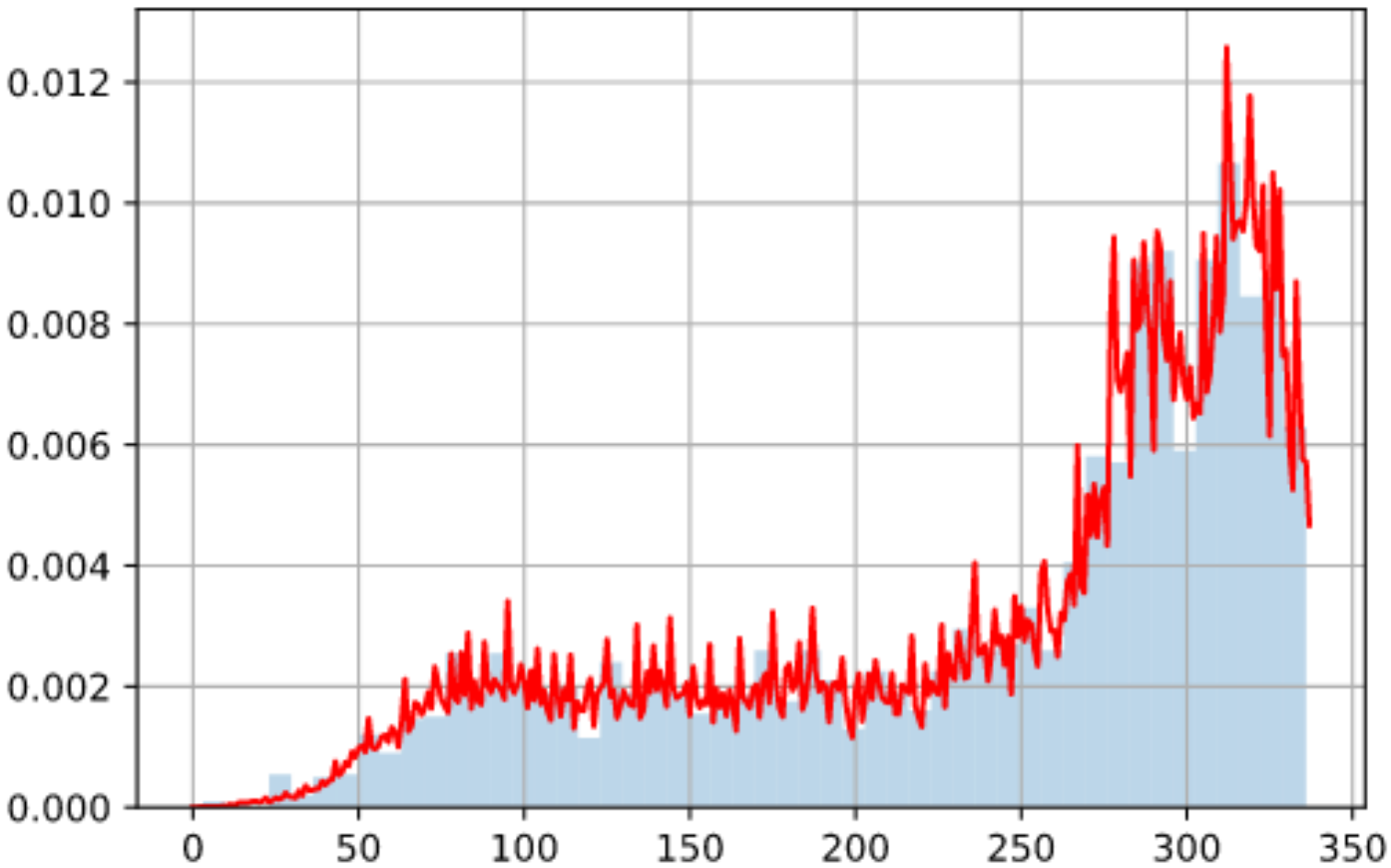
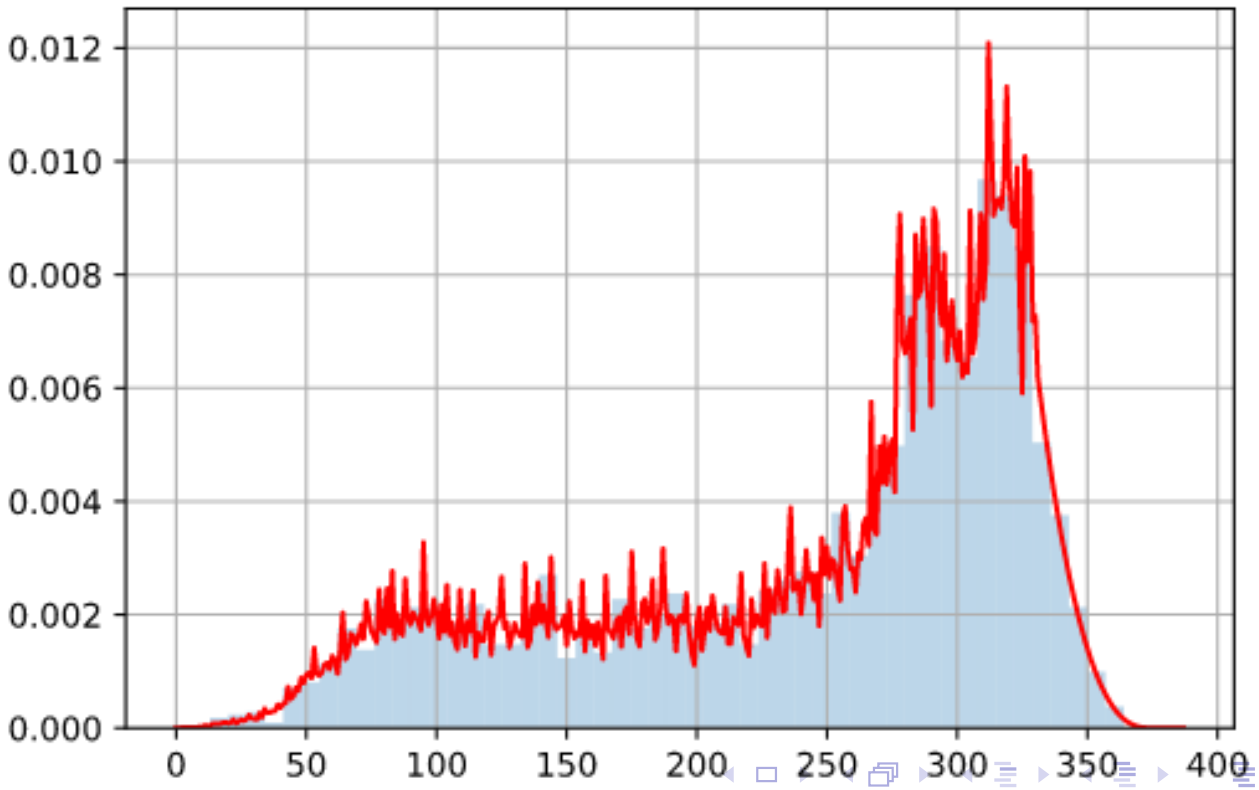


Figura: Datos observados del Valle de México (escalados)

Predicción de la cola

- Predecimos la parte faltante de los datos (cola de la distribución) usando técnicas de Valores Extremos.
- Se puede probar que la distribución Gompertz pertenece al dominio de atracción Gumbel.
- Usando el Método de Exceso sobre un Umbral podemos predecir los valores de la cola bajo el supuesto de que los datos siguen una mezcla de distribuciones Gompertz.



Predicción del R_t

- El R_t observado lo tenemos a disposición en el repositorio del CONACYT (`https://coronavirus.conacyt.mx`).
- No tenemos el valor del R_t para la cola de la distribución.
- Se calcula el R_0, y con ello, la proyección del R_t para la cola de la distribución que quedaría estimada por:

$$R_t = R_0 \cdot \frac{S_0 - N(t-1)}{S_0},$$

donde S_0 es el número inicial de susceptibles.

- NOTA: Este número inicial S_0 puede o no, dependiendo del tipo de epidemia en cuestión, resultar complejo de estimar per se. Aquí no es el caso.

Predicción del R_t

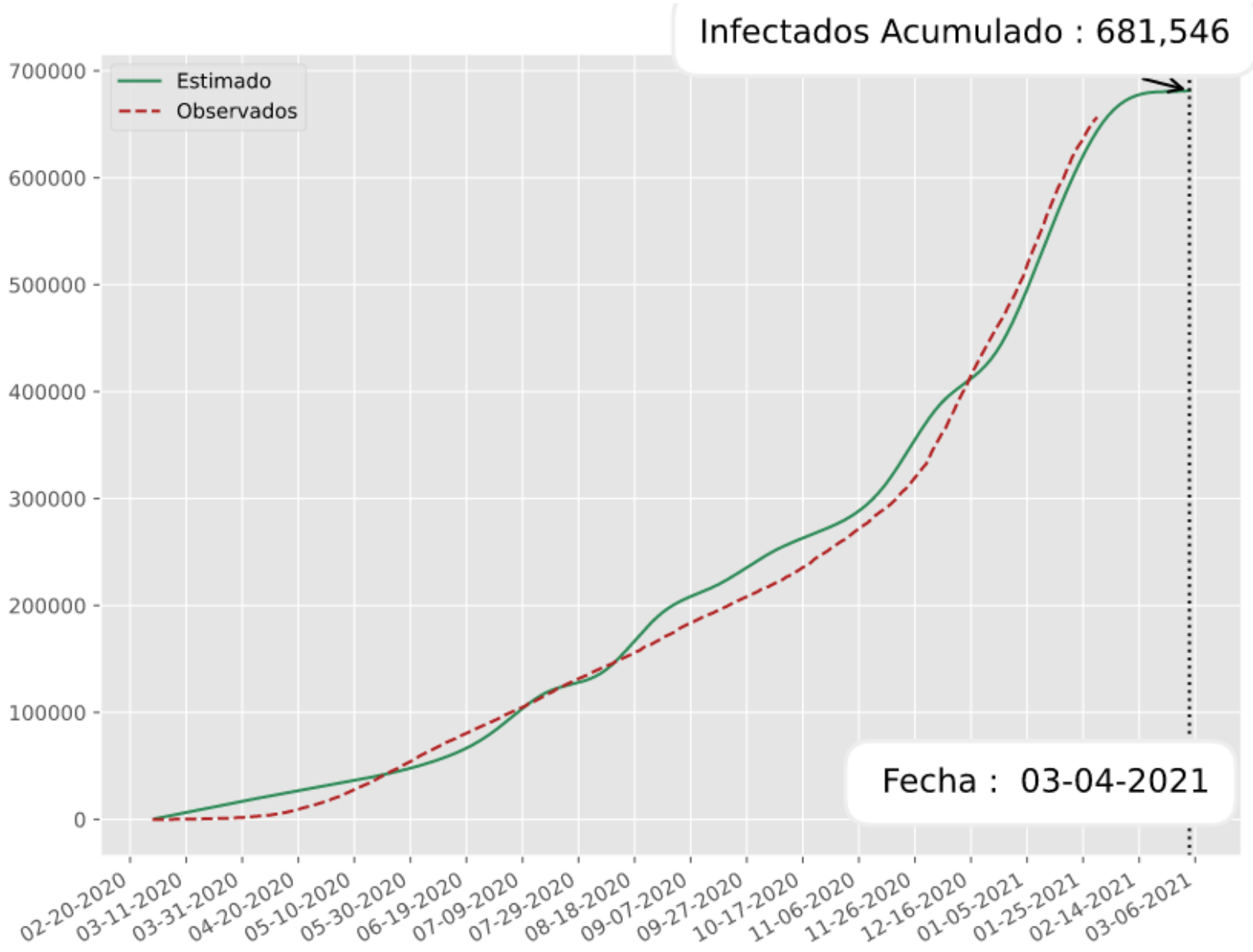
- El R_t observado lo tenemos a disposición en el repositorio del CONACYT (`https://coronavirus.conacyt.mx`).
- Se puede calcular por otros medios, como por ejemplo, a través de modelos compartimentales, y la relación con movilidad y exposición al riesgo. Es un trabajo que tenemos enviado en este momento para su posible publicación, trabajo conjunto con el Dr. Ulises Márquez, Dra. Leticia Ramírez, M.C. Iván Rodríguez y yo.

Outline

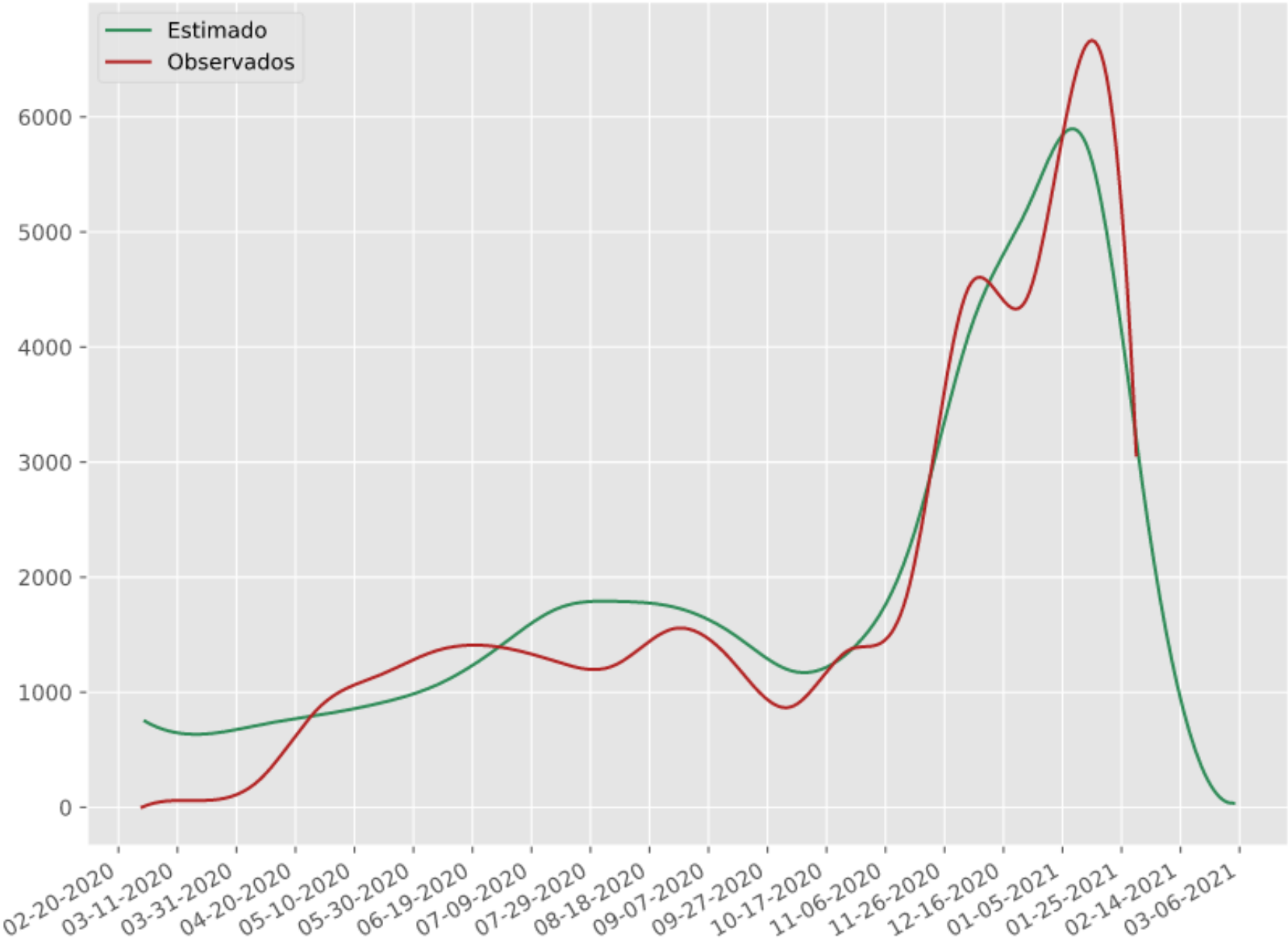
Valle de México

Presentamos el ajuste de este modelo para los datos del Valle de México al hasta el día 29/Enero/2021 en el que al momento había un total de $656,636$ infectados acumulados.

Infectados Acumulados



Infectados Diarios



Comentarios

- El modelo de mezclas con covariables nos proporciona mayor flexibilidad ya que permite un comportamiento bimodal además de tener visión sobre los cambios de política sanitaria.
- A través de este modelo podemos tener predicciones en periodos cortos de tiempo.
- Como trabajo posterior podría generalizarse este modelo para modelar n olas de contagios.
- La implementación de este modelo permite agregar covariables adicionales del R_t y con estas covariables también se podría introducir el concepto de censura.
- Este ejercicio se puede repetir en aquellas zonas en las que se observe un comportamiento bimodal en los casos de infectados diarios.

¡GRACIAS!

Referencias

G.J. McLachlan et al.
An Algorithm for fitting mixtures of Gompertz distributions to censored survival data.
s.f.

N.H. Gordon.
Maximum likelihood estimation for mixture of two Gompertz distributions when censoring occurs.
Comunications in Statistics-Simulation and Computations.
2(19):733–747, 1990.

H. Rinne.
The Hazard Rate: Theory and Inference.
Department of Economics and Management Science.
Justus–Liebig–University, s.f..

Referencias


A.W Marshall & I. Olkin
Life Distributions.
Springer, 2007.


L. Martino et al.
Independent Random Sampling Methods.
Springer, 2018.


J. Kalbfleish & R.L. Prentice
The Statistical Analysis of Failure Time Data.
Wiley & Sons,Inc. 2002.


G.J. McLachlan & T. Krishnan
The EM Algorithm and Extensions.
Wiley & Sons, Inc. 2008.

Referencias


Información referente a casos de COVID-19 en México. [En línea].
México: Información del Sistema de Vigilancia Epidemiológica de Enfermedades Respiratorias Viral Accessed: 2020-10-30
https://datos.gob.mx/busca/dataset/informacion-referente-a-casos-covid-19-en-mexico


Ecosistema Nacional Informático.
https://coronavirus.conacyt.mx/


Xiao-Li Meng & Donald B. Rubin
Maximum likelihood estimation via the ECM algorithm: A general framework.
1993


IHME COVID-19 forecasting team.
Forecasting COVID-19 impact on hospital bed-days, ICU-days, ventilator days and deaths by US state in the next 4 months.
Report
2020