© Revista de Matemática: Teoría y Aplicaciones 2023 **30**(1) : 141–172
cimpa – ucr  issn: 1409-2433 (Print), 2215-3373 (Online)
doi: https://doi.org/10.15517/rmta.v30i1.50927

# A GOMPERTZ MIXTURE APPROACH FOR MODELING THE EVOLUTION OF THE COVID-19 DYNAMICS

# MEZCLA DE GOMPERTZ PARA MODELAR LA EVOLUCIÓN DE LA DINÁMICA DEL COVID-19

Roberto Vásquez Martínez[*]
Graciela González Farías[†]
José Ulises Márquez Urbina[‡]  Rogelio Ramos Quiroga[§]

*Received: 11/May/2022; Revised: 14/Set/2022;*
*Accepted: 18/Oct/2022*

[*] University of Guanajuato, Department of Mathematics, Guanajuato, Mexico. E-Mail: roberto.vasquez@cimat.mx

[†] CIMAT, Probability and Statistics, Research Center in Mathematics, Guanajuato, Mexico. E-Mail: farias@cimat.mx

[‡] CIMAT, Probability and Statistics, Research Center in Mathematics, Monterrey and National Council on Science and Technology (CONACYT), Monterrey, Mexico. E-Mail: ulises@cimat.mx

[§] CIMAT, Probability and Statistics, Research Center in Mathematics, Guanajuato, Mexico. E-Mail: rramosq@cimat.mx

**Abstract**

Different countries used the growth Gompertz function at the beginning of the COVID-19 pandemic to model the number of cumulative infected cases since it provides reasonable results. Such a model allows only one mode, but the pandemic evolution has exhibited a multimodal behavior due to the different waves and variants of the COVID-19 virus. Thus, Gompertz's classical growth model is not well suited to describe a long pandemic with different virus variants. This work presents generalizations of the Gompertz model that can reproduce a multimodal behavior to model the dynamics of infected cases. The models are applied to COVID-19 data from Nuevo León, Mexico.

**Keywords:** COVID-19; Gompertz mixture; Poisson process; Cox process.

**Resumen**

Diferentes países usaron la función de crecimiento Gompertz al principio de la pandemia por COVID-19 para modelar el número acumulado de infectados dado que proporcionaba un ajuste razonable. Este modelo permite una única moda, pero la pandemia evolucionó exhibiendo un comportamiento multimodal debido a las diferentes olas y variantes del COVID-19. Por tanto, el modelo Gompertz clásico de crecimiento no ajusta bien para describir una pandemia larga con diferentes variantes del virus. Este trabajo presenta generalizaciones del modelo Gompertz donde se pueda capturar un comportamiento multimodal para modelar la dinámica de los casos infectados. Este modelo es aplicado a datos de COVID-19 de Nuevo León, México.

**Palabras clave:** COVID-19; Mezcla de Gompertz; Proceso de Poisson; Proceso de Cox.

**Mathematics Subject Classification:** 62F10; 60G55.

# Introduction

The Gompertz function provides a framework to model growth data thanks to its sigmoid form. Among its multiple applications, it has been used to model the growth of the population of different organisms, the growth of cancer tumors, and the cumulative cases in epidemics. In particular, it has been used by different groups [2] to understand the evolution of the COVID-19 pandemic in various regions around the globe. The Gompertz growth model permits assessing the impact of the control measures in those regions and obtaining short-term trends predictions.

The Gompertz growth model has valuable properties for modeling epidemics / pandemics:

1. it is flexible enough to model regions at different pandemic stages and various dynamics;
2. it allows to include uncertainty levels for the progress of the epidemic;
3. it provides an estimate for the total infected population, the date of maximum incidence, and the size of such maximum;
4. it allows determining confidence prediction intervals for the short term;
5. it is adaptable since it allows to include interventions that can reflect, for example, lockdown and social distancing policies;
6. it requires less computational resources than other epidemic models.

Besides, the Gompertz model can be further adapted to be used in conjunction with a random coefficient model, which provides greater flexibility and supports better decisions on public health. However, like most models, it has some issues. For instance, the model does not fit all types of outliers well; besides, sometimes, especially at the beginning of an epidemic/pandemic, interventions are insufficient to capture the effects of the outliers in the long term. Another problem is that the model cannot capture a multiple mode behavior. This aspect is especially relevant for the COVID-19 pandemic, in which we have had multiple variants that have induced more than one mode in the evolution of the pandemic. Such variants even can infect the vaccinated and recovered population.

In the present work, we introduce an extension of the classical Gompertz growth model to include multiple modes, thus capturing a more extensive family of dynamics for the evolution of an epidemic/pandemic. The extension is based on a distribution with a hazard function analogous to the Gompertz growth function, which we refer to as the Gompertz distribution. The daily infected cases are modeled via a Cox model with an intensity function based on the hazard function of Gompertz mixtures. A version of the Cox model including the effective reproduction number as a covariable is also discussed. This covariable is introduced to accelerate the Cox model. It is worth noting that there are other growth approaches for modeling cumulative cases which are not based on the Gompertz framework, for instance, based on Gaussian or logistic models (e.g. [17]). However, the Gompertz model exhibits can reproduce a more extensive dynamics class than those models.

To exemplify the use of the proposed models, they are applied to the cumulative COVID-19 cases in the state of Nuevo León, México. The state of Nuevo León is selected because it has a highly industrialized economy, has a critical trade exchange with the United States, and its COVID-19 pandemic dynamic includes different modes. In México, as in other countries, there have been many approaches for modeling the evolution of the COVID-19 pandemic, such as compartmental models (see, e.g. [9, 1, 23]). An advantage of the Gompertz growth model over many of these models is that it requires fewer resources while still providing a reasonable depiction of the pandemic's evolution. In Mexico, the pandemics dynamic has some features that resemble those found in other countries, but at the same time, it has its particularities. In Nuevo León, the affection of the three main COVID-19 variants- alpha, delta, and omicron- induced the most critical changes in the dynamic of the pandemic, which impacted its long-term behavior.

The article is organized as follows. Section 1 introduces the classical Gompertz growth model and explains how it can be extended via hazard functions. Section 2 describes how to fit the Gompertz-Mixture distribution via the Expectation - Maximization (EM) algorithm, and the distribution is fitted to Nuevo León's COVID-19 data. Section 3 discusses how to model the daily infected cases using a Cox model, with and without covariables, with intensity given as a mixture of Gompertz hazard functions. Section 3 also includes fitting the models to Nuevo León's infected daily cases. Section 4 augments the discussion on the results of adjusting the models to Nuevo León. Section 5 discusses how to proceed under the presence of truncated data; such a case corresponds to an epidemic/pandemic whose progression is still in progress. Section 6 is a conclusions section. The article also includes three appendices. Appendix A presents a methodology to find appropriate values to initialize the EM algorithm employed to estimate the mixture of Gompertz distributions. Appendix B discusses some aspects of how to approximately determine the time when a new COVID-19 variant becomes more prevalent. Appendix C briefly discusses a GitHub repository that includes the R and Python codes used to fit the models and produce the plots and results reported in the article.

# 1 A generalization of the Gompertz growth model via hazard functions

This section reviews some fundamental elements of the Gompertz growth model. For more details, see, for example, [26].

**Definition 1 (Gompertz growth model)** *The Gompertz-growth function, denoted as $N_G$, is given by*

$$N_G(t) = \alpha e^{-\beta e^{-\kappa t}}, \qquad \textit{with } t, \beta, \kappa, \alpha > 0. \tag{1}$$

Function $N_G$ is an increasing sigmoid function; thanks to this property, it can be used to model growth phenomena. At the beginning of the COVID-19 pandemic, this model was employed to describe the evolution of the cumulative infected cases. One important advantage of the Gompertz model is that it is interpretable. Namely, the parameter $\alpha$ represents the model asymptote, while $\kappa$, the epidemic growth coefficient, can be interpreted as the infectious coefficient.

As seen in Figure 1, the COVID-19 pandemic has more than one mode; therefore, we need to set a model that can allow several modes. Thus, it is necessary to extend such a model to accommodate this behavior. The present work describes methodologies to achieve this objective.

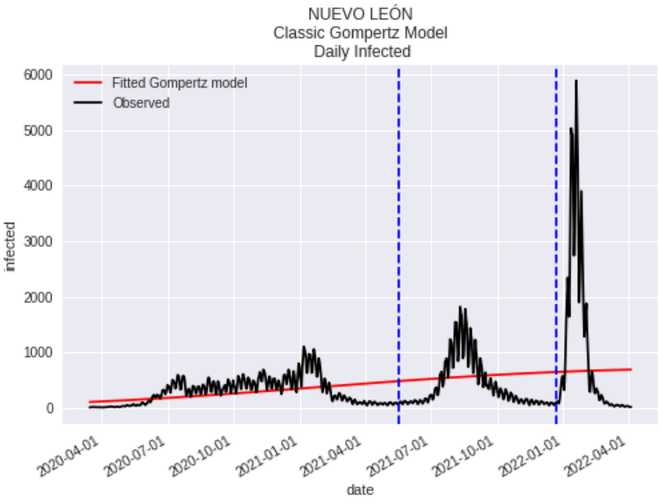


**Figure 1:** The figure shows the daily active cases (black), a fitted classical Gompertz model for the whole period (red), and divisions (dashed) that frame the periods during which each of the three main COVID-19 variants dominated the dynamic of the pandemic. The period spans from 2020-03-02 to 2022-04-04. In the figure, we can observe the multimodal behavior of the active daily cases and the classical model's insufficiency to capture the pandemic's whole dynamic.

The Gompertz model can be estimated through non-linear least-squares with restrictions, more precisely, by solving the optimization problem

$$(\hat{\alpha}, \hat{\beta}, \hat{\kappa}) = \text{argmin}_{(\alpha,\beta,\kappa)} \left\{ \sum_{j=1}^{n} (Y_j - N_G(j))^2 \right\}, \text{ for } \alpha, \beta, \kappa > 0,$$

where $Y$ represents the number of cumulative infected cases at day $j$, and $n$ is the last day registered. In this article, we estimate the parameters using lmfit Python's library. For computational purposes, we assume that the epidemic / pandemic starts the first day that there are 20 or more accumulated infected cases, i.e. $Y_0 \equiv N_G(0) \geq 20$.

Observe that the relative growth rate [27] of the model (1) is given by

$$\frac{N_G'(t)}{N_G(t)} = \beta\kappa e^{-\kappa t}, \qquad \text{with } \beta, \kappa > 0, \tag{2}$$

which corresponds to the hazard function of the Negative-Gompertz distribution [10]. A possible approach to generalizing the Gompertz growth model to capture multiple modes is modifying the hazard function to produce a Gompertz-mixture distribution. Thus, we assume that the infection rate (or failure rate) is an exponentially increasing function analogous to (2). Namely, we assume that the failure rate will be

$$h(t) = \lambda\xi e^{\xi t}, \qquad \text{with } \lambda, \xi > 0. \tag{3}$$

The hazard function (3) has been employed in [10]. A suitable reparametrization of this function is given by

$$h(t) = \exp\{\gamma + \xi t\}, \quad \text{with} \quad \xi > 0 \qquad \left(\text{i.e. } \lambda = \frac{e^{\gamma}}{\xi}\right), \tag{4}$$

which simplifies the restrictions needed in the estimation of the parameters. This hazard function has been considered in [14], page 171. Let $F$ denote the distribution with hazard function (4). We propose to extend the Gompertz growth model as a mixture of $F$ distributions. In this work, we refer to $F$ as the Gompertz distribution and to its mixture extension as the Gompertz-Mixture model. We do not propose a direct generalization of the classical Gompertz growth model since our proposal is based on a distribution with a hazard function in the form of an increasing Gompertz function. We did it because it produced accurate adjustments to growth curves.

In the following, when we refer to the Gompertz distribution, we mean the distribution with hazard function (4). As we discuss later, we assume that the infection times (failure times) follow a mixture of Gompertz distributions. Besides, if $N(t)$ is the number of accumulated infected at time $t$, we model $N(t+1) - N(t)$ with a Cox model whose intensity function is given in terms of the hazard function of the Gompertz mixture distribution of the failure times.

# 2 Estimating the Gompertz-Mixture distribution via the Expectation-Maximization (EM) algorithm

This section contains two subsections. In the first subsection, we discuss how to fit the Gompertz-Mixture distribution; in the second subsection, we fit such distribution from Nuevo León's COVID-19 cases.

## 2.1 Estimation

Consider the Gompertz-Mixture model with $G$ components, with mixture proportion

$$\pi = (\pi_1, \ldots, \pi_G)^T, \text{ such that } \sum_{i=1}^{G} \pi_i = 1,$$

and parameter vector

$$\mathbf{\Psi} = (\theta_1^T, \ldots, \theta_G^T)^T,$$

where $\theta_i^T = (\gamma_i, \xi_i)$ is Gompertz's paramter vector for $i$-th component. We denote this model in the following as MGomp$(G, \pi, \mathbf{\Psi})$ [13]. For estimating the Gompertz-Mixture model, we need to transform the growth data (cumulative cases) $\{Y_j : j = 1, 2, \ldots, n\}$ to failure time or individual's infected time data $\{T_j : j = 1, 2, \ldots, m\}$, where $m$ is the desired sample size. We assume that the failure times $T_j$ are independent and identically distributed random variables with distribution MGomp$(G, \pi, \mathbf{\Psi})$ for some unknown parameters. The independence assumption is necessary to have a finite mixture of distributions.

The previous distributional assumption can be better explained by looking at Figure 1. We observe that Nuevo León has had three upsurges in the daily infected curve. Thus, we can divide the COVID-19 timeline into three subperiods corresponding to these three spreads. The previous assumption means that the failure times during each subperiod follow a Gompertz-Mixture distribution, including the case with only one mode. A similar observation is made in [2]. Later in this manuscript, we describe how approximately determine when these subperiods start and end. It is worth noting that each of these subperiods can be identified with a COVID-19 variant: the subperiods were mainly driven by the alpha ($\alpha$), delta ($\Delta$), and omicron ($O$) COVID-19 variants, respectively.

For transforming the growth data into failure time data, we generate a discrete uniform sample on the interval $[Y_0, Y_n]$; that is, we sample $m$ random variables $I_1, \ldots, I_m$ from this distribution. The $k$-th sampled value represents the $I_k$ cumulative case: through a binary search, for each $k = 1, \ldots, m$, we can find $j \in \{0, 1, \ldots, n-1\}$ such that

$$Y_j < I_k \leq Y_{j+1}. \tag{5}$$

Therefore, we assign the $k$-th individual the $(j+1)$-th day of infection. Employing this algorithm, we build a sample $\{T_j : j = 1, 2, \ldots, m\}$ of failure times.

Sampling failure times through the relationship (5) is the first step to fit a Gompertz-Mixtures model to the cumulative cases. Given the failure times, the next step is to get Gompertz-Mixture parameters. It is possible to estimate the parameters through the EM algorithm as proposed in [12]. Such a methodology consists of an iterative process that requires initializing the vector parameters $\pi, \Psi$ and the number of components $G$. We discuss how to correctly choose the initial conditions and the number of components in Appendix A. We make available programs for estimating Gompertz-Mixture parameters and determining the initial conditions on the GitHub repository specified in Appendix C. Namely, the Python scripts opt_baseline.py and initial_conditions.py contain code to estimate the Gompertz-Mixture parameters and determine appropriate initial conditions. By appropriate, we mean that they could be considered near optimal Gompertz-Mixture parameters, thus producing a good performance in the Quasi-Newton method used for estimation. Table 1 summarizes the estimated parameters for the Gompertz-Mixtures obtained for each variant in Nuevo León.

**Table 1:** Gompertz-Mixture failure distribution for each strain

| | $G$ | $\pi$ | $\Psi$ |
|---|---|---|---|
| $\alpha$ | 4 | (0.11,0.41,0.31,0.17) | (-12,0.07;-9.07,0.02;-19.03,0.05;-10.64,0.02) |
| $\Delta$ | 4 | (0.24,0.29,0.43,0.04) | (-10.11,0.1;-9.6,0.07;-6.6, 0.02;-16.38,0.07) |
| $O$ | 3 | (0.44,0.41,0.15) | (-6.8,0.22;-7.1,0.15;-4.99,0.04) |

**Remark 1** *If the last wave is unfinished, the last Gompertz component of the mixture associated with such wave is a truncated distribution. In this case, it is necessary to estimate the right tail of such a Gompertz component. In Section 5, we present a methodology to fit the right tail in a Gompertz distribution. The GitHub repository mentioned in Appendix C includes the R script* tail_Gompertz.R *to perform such estimation.*

## 2.2 Estimating Gompertz-Mixture distribution for Nuevo León

The available COVID-19 data for Nuevo León spans from March 2nd, 2020, to April 4th, 2022. The first date corresponds to the first record of a COVID-19 case in Nuevo León, while the second date corresponds to the last available date when preparing the present manuscript.

As discussed briefly in the previous section, for practical purposes, in Nuevo León, we recognize a time domain of influence for the most relevant COVID-19 strains: $\alpha$, $\Delta$, and $O$. We refer to each subperiod by the variant that drives it, and we consider the whole dynamic of the pandemic for the state of Nuevo León as the concatenation of these subperiods. Besides, we use strain and variant as synonyms.

It is worth noting that the nature of the data in each subperiod is different. Data for the $\alpha$ subperiod, which corresponds to the first part of the pandemic, only comes from hospital reports. For the $\Delta$ subperiod, most of the employed data also comes from hospital reports, but it also includes data from other sources like testing. Data for the $O$ variant still considers hospital data, but it also includes data from massive testing. Thus, these subperiods mark the evolution of different susceptible populations. The changes in the susceptible populations are caused, among other things, because reinfection was possible and by the different interventions to control the growth of active cases. These interventions were, for example, the social distancing policies and the massive vaccination campaigns. During the $\Delta$ and $O$ subperiods, many reinfected individuals did not go to a hospital but could have had a positive COVID-19 test. Besides, some individuals had several positive tests for the same infection since a negative test was required for some services and workplaces. Then, for practical purposes, it is convenient to consider that different processes drive the dynamics in the subperiod corresponding to each variant, particularly that each strain has different Gompertz-Mixture distributions for the failure times and different inherent counting processes.

Table 2 presents the (approximate) starting and final dates for each subperiod in Nuevo León. By final date, we mean the point where a new variant replaces the most prevalent variant; it does not imply a 100% finish of the old

**Table 2:** Timeline of each strain in Nuevo León

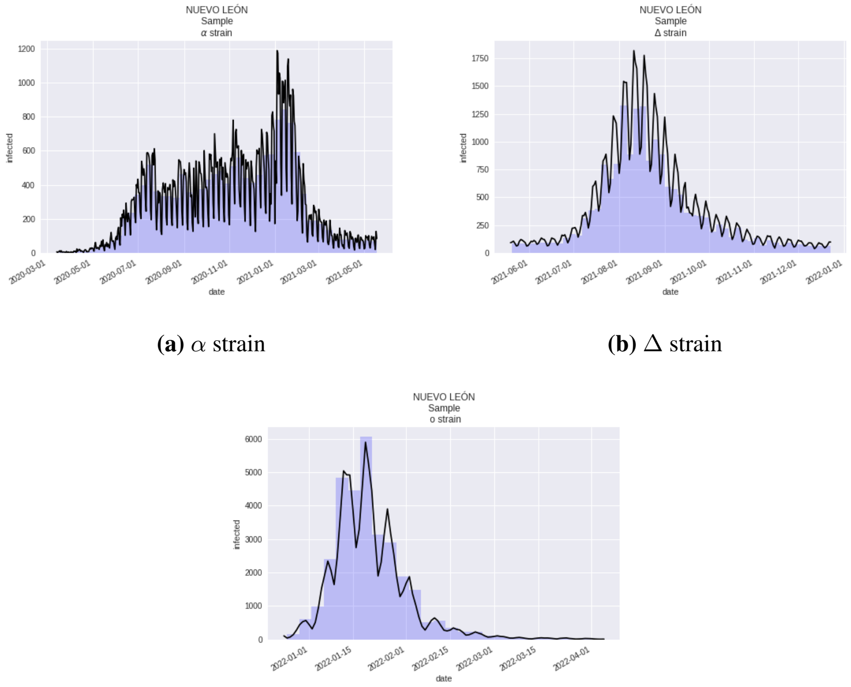
| Strain of COVID-19 | Start | Final |
|---|---|---|
| $\alpha$ | 2020-03-02 | 2021-05-18 |
| $\Delta$ | 2021-05-19 | 2021-12-23 |
| $O$ | 2021-12-24 | continue |

dominant variant but that it has been controlled. To determine the beginning and end of a subperiod, we pair the information of the effective reproduction number $R_t$ (see Appendix B) and the new cases by COVID-19 variant type. Intuitively, when $R_t$ has been below 1 for a long time, this might indicate the end of a variant subperiod. After a period with $R_t < 1$, if the effective reproduction number $R_t$ increases above 1 and cases produced by a new variant surpass the cases of the previous dominant, this signifies that a new subperiod of a new variant has started. The information regarding the first cases can be found at the site [5], where the Mexican authorities report scientific details on the COVID-19 pandemic.

Sampling failure times through the relationship (5) is the first step to fit a Gompertz-Mixtures model to the cumulative COVID-19 cases of Nuevo León. We sample at each subperiod, assuming the time and the dynamic restart at the beginning of each new strain of COVID-19. Besides, we consider that the failure times follow a different Gompertz-Mixture distribution during each subperiod. Figure 2 presents the results for this sampling, split by variants. Given the failure times, the next step is to get Gompertz-Mixture parameters for each strain. Let $\{T_{1j} : j = 1, 2, \ldots, m_1\}$, $\{T_{2j} : j = 1, 2, \ldots, m_3\}$, and $\{T_{3j} : j = 1, 2, \ldots, m_3\}$ the failure time data for the $\alpha$, $\Delta$, and $O$ subperiods, respectively. We assume that $T_{rj} \sim$ MGomp$(G_r, \pi_r, \Psi_r)$, for $r = 1, 2, 3$.

Figure 2 suggests that the $O$ subperiod is in its terminal phase with very few new infected cases. Thus, we can assume that this subperiod has already completed its full dynamics on April 4th, 2022, the last day available on the data. Then, we can perform the same type of analysis on the three subperiods under the assumption that the three of them have already completed their infectious cycle.

**Remark 2** *The methodologies explained here are based on Nuevo León's case. However, they can be applied to other cases, including those with more and fewer subperiods or with an unfinished final subperiod.*



**(a)** $\alpha$ strain

**(b)** $\Delta$ strain

**(c)** $O$ strain

**Figure 2:** The histogram (blue) is the sample of each strain's failure time data. The solid line represents COVID-19's daily cases with a three days mobile mean

# 3 Modeling the number of infections with the Gompertz-Mixture model

This subsection presents two subsections. In the first subsection, we present a Cox model for the number of failures (infections). The second subsection introduces a covariable in the Cox model from the first section. For both subsections, we employ the framework of subperiods/variants discussed in Section 2.2. In addition, in both subsections, we adjust the corresponding models to the data from Nuevo León.

## 3.1 Baseline Cox model

Let $h_{ri}$ and $S_{ri}$ be the hazard and survival functions, respectively, for the $i$-th component of the $r$-th Gompertz-Mixture. If we denote by $h_{0r}$ and $S_{0r}$

the hazard and survival functions, respectively, for the $r$-th Gompertz-Mixture (see [22]), then

$$S_{0r}(t) = \sum_{i=1}^{G_r} \pi_i S_{ri}(t), \qquad \text{and} \qquad h_{0r}(t) = \frac{\sum_{i=1}^{G_r} \pi_i S_{ri}(t) h_{ri}(t)}{S_{0r}(t)}.$$

Function $h_{0r}$ is the *failure rate* for the $r$ variant, and we call it the *baseline failure rate*. In the following sections, we obtain a baseline counting measure with failure rate $h_{0r}$ to estimate the number of infected cases in a certain period. In this way, we build a framework that allows obtaining the cumulative infected cases as a function of time. The parameters for the previous hazard and survival functions are estimated employing the EM algorithm as discussed in Section 2.

For $t \in \mathbb{N}$, assume that $X_{rt}$ is a random variable that counts the number of failures in the interval $(t-1, t]$ during the subperiod $r$. In other words, if we consider $t$ as the number of days after the beginning of the pandemic associated with variant $r$, $X_{rt}$ counts the infected, in the day, at $t$ days after that start. It is important to recall that we restart the time in every strain.

If $X_{r0}$ is the initial number of infected at the beginning of the subperiod $r$, then the cumulative number of infected cases at day $t$ for the variant $r$ is

$$N_r(t) = X_{r0} + \sum_{j=1}^{t} X_{rj}, \qquad \text{for } t \in \mathbb{N}.$$

A natural approach is to assume that $N_r(t)$ is a nonhomogeneous Poisson process with rate function $m_{0r}$ given by

$$m_{0r}(t) \propto \Lambda_{0r}(t) = \int_0^t h_{0r}(s) ds.$$

This kind of process is known as the *Cox process* or the doubly stochastic Poisson process [20]. The process increments over disjoint intervals are, in general, statistically dependent, and such dependency is transferred to the failure times distribution for each strain/subperiod. The rate function $m_{0r}(t)$ is stochastic in nature since it depends on the estimated parameters, which in general are given by a transformation of the sample.

Under the Poisson assumption, $X_{rt} \sim \text{Pois}(m_{0r}(t) - m_{0r}(t-1))$. We will assume that the mean of $X_{rt}$ is directly proportional to the cumulative risk and the proportion of susceptible individuals at that moment $t$. Let $P_r$ be the susceptible population at the beginning of strain/subperiod $r$ and $\alpha_{k_r} = P_r/k_r$ be the carrying capacity for some constant $k_r \in \mathbb{R}^+$ depending on the strain $r$. More precisely, we assume that

$$m_{0r}(t) - m_{0r}(t-1) = (\alpha_{k_r} - N_r(t-1)) \cdot (\Lambda_{0r}(t) - \Lambda_{0r}(t-1)), \tag{6}$$

for $t \in \mathbb{N}$, with the restriction

$$m_{0r}(0) = X_{r0}. \tag{7}$$

Doing something similar to Fisher's scoring method, after taking mathematical expectation in (6) and rearranging terms, we obtain

$$m_{0r}(t) + (a_{rt} - 1)m_{0r}(t-1) = \alpha_{k_r} \cdot a_{rt}, \quad \text{for } t \in \mathbb{N}, \tag{8}$$

where $a_{rt} = \Lambda_{0r}(t) - \Lambda_{0r}(t-1)$. In the previous equation, we use that $\mathbb{E}[N_r(t-1)] = m_{0r}(t-1)$.

Solving the first-order difference equation (8) with initial condition (7), we obtain that

$$m_{0r}(t) = \alpha_{k_r} + (X_{r0} - \alpha_{k_r}) \prod_{i=0}^{t} (1 - a_{ri}), \tag{9}$$

with $a_{r0} = 0$ and $a_{ri}$ as before.

The only unknown quantity to estimate $m_{0r}$ is $\alpha_{k_r}$; more precisely, we do not know $k_r$, the fraction of susceptible population for the $r$ strain. If we have observed $n_r$ days of the subperiod $r$, we can obtain an optimal $\widehat{k}_r$ by solving the optimization problem

$$\widehat{k}_r = \text{argmin}_{k_r \in \mathbb{R}^+} \left\{ \sum_{j=1}^{n_r} \left[ (X_{r0} - \alpha_{k_r})(b_{r,j} - b_{r,j-1}) - \widehat{X}_{rj} \right]^2 \right\}, \tag{10}$$

where $b_{r,j} = \prod_{i=0}^{j}(1 - a_{ri})$, and $\widehat{X}_{rt}$ is the number of infections at day $j$. We solve this optimization problem through the Levenberg-Marquardt algorithm [18] taking the objective function as a function of $k_r$. Thus, we obtain the optimal fraction of susceptible population $\widehat{k}_r$ for each strain.

For each strain, we restart the susceptible population. Considering conditions such as vaccination and the characteristics of each COVID-19 variant, if $P$ denotes the total population of the region under study, we take

$$P_r = \begin{cases} P & \text{if} \quad r = 1, \\ 0.51P & \text{if} \quad r = 2, \\ P & \text{if} \quad r = 3. \end{cases}$$

In the previous relation, it can be seen that for the $\alpha$ and $O$ variants, we consider the whole population as susceptible. For the $\alpha$ variant, the value of $P_r$ indicates that the entire population is deemed susceptible; for the $O$ variant, the value of $P_r$ means that we are assuming that reinfections were possible and that vaccination only protected against severe infections. We take these assumptions as a methodological approach as we have no more information to determine a more accurate form for the susceptible population. On the other hand, evidence suggests that vaccination prevented developing symptoms and that reinfections were rare for the $\Delta$ variant. At the time, for Nuevo León, the proportion of the vaccinated population was $0.7$. If we assume an overall efficiency of $0.7$ of the vaccines against the delta variant, we can estimate in $0.49P$ the protected population against diagnosis. Since the available data for Nuevo León during the delta variant consists mainly of hospital reports (i.e., we do not see those who do not develop symptoms), $0.51P$ of the population can be considered susceptible.

We compare the daily infected observed data with mathematical expectation of each $X_{rt}$, for $t = 1, 2, \dots, n_r$, which is given by

$$\begin{aligned}\mathbb{E}[X_{rt}] =& m_{0r}(t) - m_{0r}(t-1)\\ =& (X_{r0} - \alpha_{\widehat{k_r}}) \cdot (b_{r,t} - b_{r,t-1}), \qquad \text{for } t = 1, 2, \dots, n_r.\end{aligned} \tag{11}$$

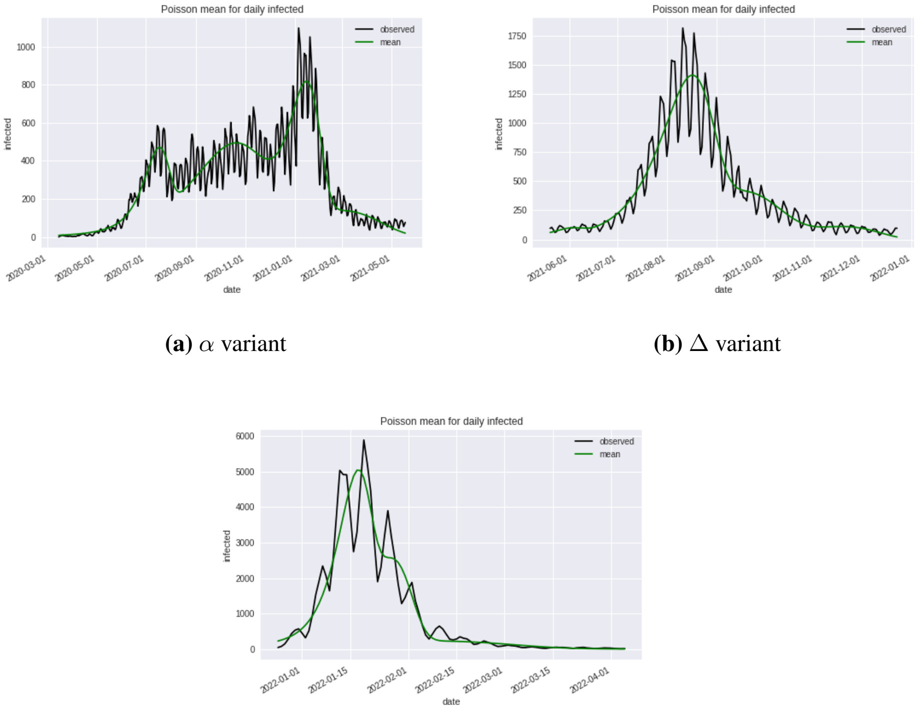
Figure 3 shows the result of this contrast for each strain.

**Table 3:** Estimated carrying capacity and initial susceptible population of each strain for baseline process

| | $\alpha$ | $\Delta$ | $O$ |
|---|---|---|---|
| $P_r$ | 4,653,458 | 2,373,263 | 4,635,458 |
| $k_r$ | 37.93 | 27.52 | 45.76 |
| $\alpha_{k_r}$ | 122,670 | 86,246 | 101,685 |
| observed accumulated infected cases | 123,385 | 86,671 | 104,258 |

Table 3 shows, for each strain, the carrying capacity, the initial susceptible population, and the fraction $k$. It is important to note that $\alpha_{k_r}$ provides the value of the asymptote in each subperiod. In addition, it presents the cumulative number of observed infected cases at the end of each strain in the following table. In Table 3, we can see that our carrying capacity estimation of each strain is plausible.

If we do not have seen the pandemic subperiod entirely, we need to estimate the Gompertz-tail to project observed data and then perform the same algorithm as before. That is, we can calculate a plausible asymptote despite incompleted or truncated data.



**(a)** $\alpha$ variant

**(b)** $\Delta$ variant

**(c)** $O$ variant

**Figure 3:** The black line is the daily infected cases curve smoothed through a three-day moving average. The green line represents the mean in (11) by solving (10) and using Gompertz-Mixture's parameters in Table 1

## 3.2 Accelerated Cox model

This section describes how to incorporate a covariable in the Cox process (5) as done in the proportional hazard models [8] to improve and *accelerate* the modeling framework. The covariable considered is the effective reproduction number $R_t$; the ideas presented work for other covariables. It is vital to notice that, due to the different nature of the available data during each variant, the dynamic of $R_t$ differs during each subperiod. Therefore, we accelerate the Cox process for each subperiod independently of the others.

We refer to the Cox model (6) as the baseline model. The approach is to accelerate the cumulative hazard function with a function of the covariable.

The covariable functional is given by

$$\varphi_r(R_{tr}) = \exp(\beta_r(R_{tr} - \overline{R}_{tr})),$$

where $R_{tr}$ is the effective reproduction number on the respective subperiod, $\overline{R}_{tr}$ is the mean of $R_{tr}$, and $\beta_r$ is a parameter to be estimated that accelerates the baseline Cox model (6).

Let $h_r$ be the accelerated hazard function. We use the relative risk or Cox model [8] approach for accelerating the baseline model (6); that is, we assume

$$h_r(t) = \varphi_r(R_{tr})h_{0r}(t) = \exp(\beta_r(R_{tr} - \overline{R}_{tr}))h_{0r}(t), \quad \text{for } r = 1, 2.$$

Table 4 contains the regression coefficient for the $\alpha$ and $\Delta$ variants. We omit the $O$ variant because the data for $R_t$ is only available from February 3rd, 2020, to January 23rd, 2022. The description and analysis of the effective reproduction number $R_t$ are available in [4, 16][1]. Figure 4 depicts the behavior of the mean function of the baseline and the accelerated Cox process against observed daily data. These plots are built by simulating Poisson variables. Contrary to Figure 3, where the baseline curves are created employing the expression for the mean value, the curves in Figure 4 are built from simulations.

**Table 4:** Coefficients of relative risk model

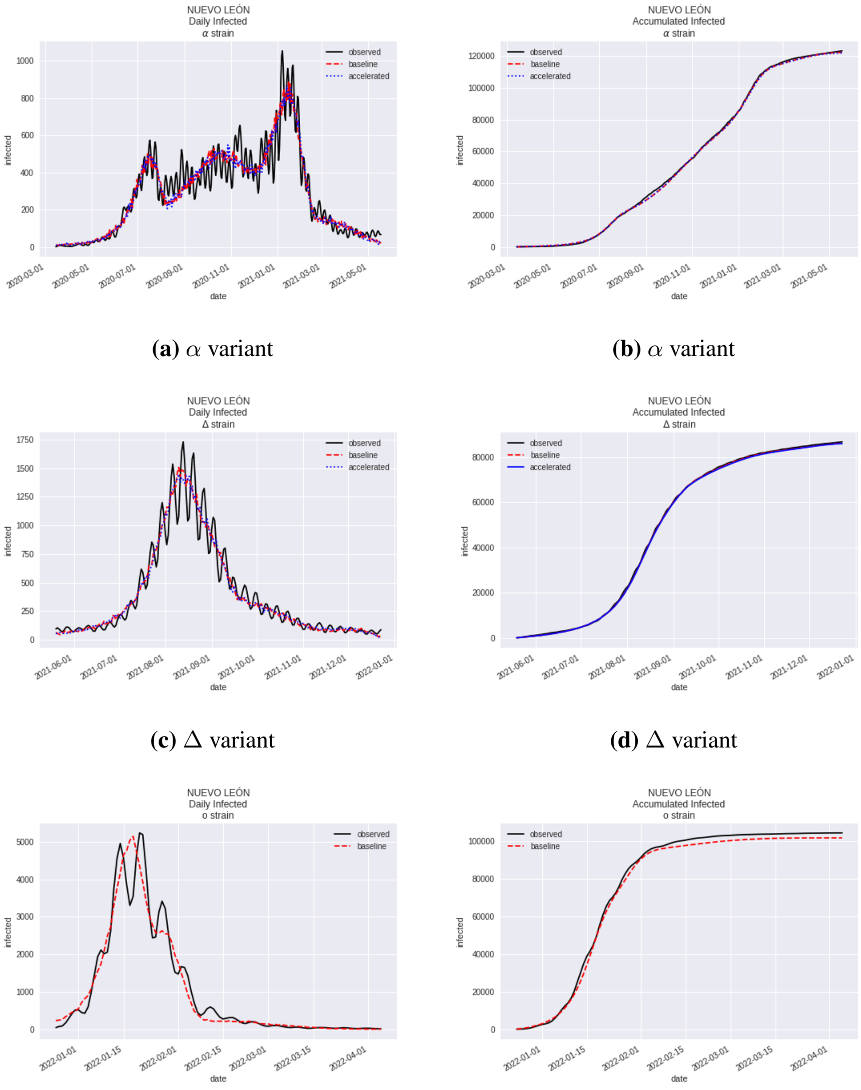
| | $\alpha$ | $\Delta$ |
|---|---|---|
| $\beta_r$ | -0.047 | -0.014 |

Figure 4 shows the fit of the growth models to the cumulative cases for each subperiod/variant. We can see that the Cox Gompertz-Mixture model exhibits good fits for the three variants.

The accelerated version of the mean function can be obtained by solving the equation

$$m_r(t) = \alpha'_{k_r} + (X_{r0} - \alpha'_{k_r}) \prod_{i=0}^{t} (1 - a'_{ri}), \quad \text{for } t \in \mathbb{N},$$

where $\Lambda_r(t) = \int_0^t h_r(s)ds$, $a'_{ri} = \Lambda_r(i) - \Lambda_r(i-1)$, for $i = 1, 2, \ldots$ and $a'_{r0} = 0$. The last equation is analogous to the difference equation in (6). As done with the Cox baseline model, we can estimate the carrying capacities of the accelerated process by solving the analogous version of the optimization problem in (10).



**(a)** $\alpha$ variant

**(b)** $\alpha$ variant

**(c)** $\Delta$ variant

**(d)** $\Delta$ variant

**(e)** $O$ variant

**(f)** $O$ variant

**Figure 4:** The figure presents the Gompertz-Mixture model (baseline and accelerated Cox model) fitted for each subperiod. On the left, we have the observed daily cases and the fitted model derivative; on the right, we have the cumulative cases and their corresponding fitting.

Table 5 presents the Poisson deviances of both Cox models estimated: baseline and accelerated. We have similar performances between the baseline and the accelerated models. Besides, we can see that the Poisson process model does not explain enough variability in data. Table 6 presents, for each subperiod/variant, the carrying capacity, the initial susceptible population, and the fraction $k$ for the accelerated model; it is analogous to Table 3. We can see that both tables show very similar values providing more evidence that both models show similar performances. Still, as shown in Figure 3, the models can at least describe the expected behavior of the process.

**Table 5:** Mean Poisson deviance for each model

| | $\alpha$ | $\Delta$ | $O$ |
|---|---|---|---|
| baseline | 23.47 | 28.6 | 107.66 |
| accelerated | 22.9 | 30.44 | |

**Table 6:** Estimated carrying capacity and initial susceptible population of each strain for accelerated process

| | $\alpha$ | $\Delta$ |
|---|---|---|
| $P_r$ | 4,653,458 | 2,373,263 |
| $k_r$ | 38.04 | 27.6 |
| $\alpha_{k_r}$ | 122,346 | 85,979 |
| observed accumulated infected cases | 123,385 | 86,671 |

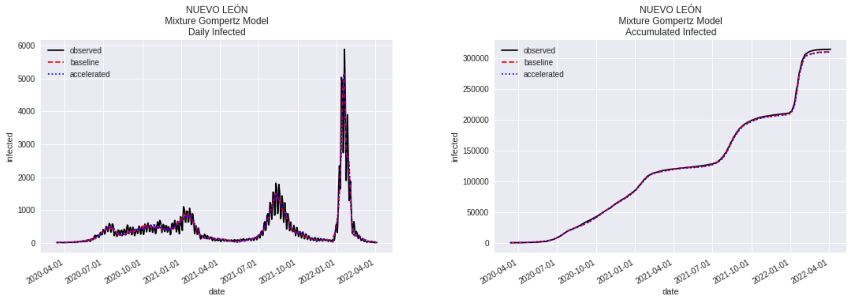
The previous results indicate no significant difference between the baseline and the accelerated Cox models. However, the accelerated could likely outperform the baseline model if additional information and suitable covariables are employed; in this work, we use the effective reproduction number since it is the only available information we have. We include the accelerated model for completeness to include an approach extensively used in the context of a Gompertz model [12].

[1] The data needs to be required from the Mexican authorities

# 4 Concatenating the mixed models

This section further discusses the results of applying the Gompertz-Mixture model to Nuevo León's cumulative COVID-19 cases data.

Figure 5 shows the result of concatenating the models for the three subperiods. We cannot see discrepancies due to the scale resolution. However, in Figures 3 and 4, we can see that the models for the $\alpha$ and $\Delta$ subperiods exhibit better performance than for the $O$ case. This visual confirms what the deviances reported in Table 5 indicate. It is important to remember that the asymptotes are estimated for each subperiod independently since it depends on the susceptible population and the database nature.



**(a)** Daily infected curve

**(b)** Growth curve

**Figure 5:** Density and growth model for the concatenated COVID-19 dynamic in Nuevo León.

With complete data, i.e., a curve that decreases enough at the end of each subperiod, it is plausible to transform growth data into failure times, then assume a Gompertz-Mixture model and see the whole COVID-19 dynamic as a mixture of mixtures. For incomplete data, Section 5 describes how to complete the tail. In that section, we can see in an example that it is possible to build an asymptote through the right tail estimation.

The Gompertz-Mixture model allows for comparing each subperiod/variant through the value of their parameters, such as growth coefficients. For instance, the $O$ subperiod has the most significant growth coefficient, which explains that the number of cumulative infected cases during this subperiod is similar to those reported on the $\alpha$ subperiod but in a shorter period. In addition, we can see at the beginning a more significant growth in the $\Delta$ variant with a $0.1$ growth coefficient compared to $0.0686$ for the $\alpha$ growth coefficient. In general, we can see more significant growth coefficients in the $\Delta$ strain than in the $\alpha$ strain; analogously, we see larger coefficients for the $O$ variant than for the $\Delta$ variant. This phenomenon, perhaps, responds to the increasing information sources during the COVID-19 timeline: from only hospital reports during the $\alpha$ subperiod to massive testing information during the $O$ variant.

Focusing on one subperiod, we can perform a clustering analysis for its timeline and see how the growth coefficients of each of its Gompertz components change in time. For example, for the $\alpha$ subperiod, we have that the best Gompertz-Mixture model has four components with corresponding growth coefficients $(0.068, 0.023, 0.05, 0.018)$. Thus, we can see that the $\alpha$ subperiod exhibited its most significant growth at its beginning.

We have assumed that the subperiods are technically complete with what we have done until this point. This assumption is used because it is the condition of the data used in the example. The following section discusses how to proceed when we consider a subperiod that is still in progress, that is, if the curve of daily infected cases is right-censored.

## 5 Gompertz tail estimation

This section describes how to estimate the right tail of a Gompertz distribution. The method is based on the Peaks-Over-Threshold method [3]. Thus, we can complete the observed density of the last Gompertz component of the failure times. The procedure can also be employed to complete the current subperiod of the daily infected cases (that is, to estimate the right tail of the subperiod) under the assumption that the conditions remain unchanged. By the tail, we mean the last right part of the distribution.

In the following, we include some theoretical results necessary to explain the methodology to present a complete description. These results are taken from [21].

**Definition 2 (Von Mises distribution)** *Let $F$ be a distribution function with right extreme $\omega_F \leq \infty$. Function $F$ is a* Von Mises *distribution if there exists $z < \omega_F$ such that*

$$\overline{F}(x) = c \exp\left\{ -\int_z^x \frac{dt}{a(t)} \right\}, \quad z < x < \omega_F,$$

*where $c > 0$ and $a$ is an absolute continuous positive function with density $a'$ satisfying* $\lim_{x \to \omega_F} a'(x) = 0$.

**Definition 3 (Domain of attraction)** *We say $F$ is in the domain of attraction of a distribution $G$, written $F \in D(G)$, if there exist sequences $a_n > 0$, $b_n$, $n \geq 1$, such that*

$$F^n(a_n x + b_n) \to G(x) \qquad \textit{weakly.}$$

**Theorem 1** *Let $F$ be a Von Mises distribution with auxiliary function $a$ and let $\overline{F}^{\leftarrow}$ denote the generalized inverse function of $F$. Then, $F$ belongs to the Gumbel domain of attraction $D(\phi_G)$, and the normalized constants for the maximum are $b_n = \overline{F}^{\leftarrow}(1/n)$ and $a_n = a(b_n)$.*

We need the following result to be able to apply the Peaks-Over-Threshold method.

**Theorem 2** *If $F$ is the Gompertz distribution function, then $F \in D(\phi_G)$.*

**Proof.** Recall that the hazard function associated with the Gompertz distribution is given by $h(t) = \exp(\lambda + \xi t)$. Taking $a(\xi) = [h(\xi)]^{-1} = \exp(-\lambda - \xi\xi)$, we have that

$$\overline{F}(t) = \exp\left(-\int_0^t \frac{d\xi}{a(\xi)}\right), \qquad t > 0.$$

Besides, since $\xi > 0$,

$$\lim_{\xi \to \infty} a'(\xi) = 0. \tag{12}$$

Therefore, the Gompertz distribution is a Von Mises distribution; besides, thanks to Theorem 2, it belongs to the Gumbel domain of attraction. ■

Since the Gompertz distribution belongs to the Gumbel domain, we can apply the following result [19].

**Theorem 3 (Pickands-Balkema-de Haan)** *Let $F$ be a distribution function. Then, $F$ belongs to a domain of attraction $D(\phi)$, i.e. $F \in D(\phi)$ where $\phi$ is a extreme value distribution, if and only if*

$$\lim_{u \to \omega_F} \sup_{0 < x < \omega_F - u} \left| \frac{\overline{F}(x+u)}{\overline{F}(u)} - \overline{P}_{\psi, a(u)}(x) \right| = 0,$$

*for some measurable and positive function $a(u)$ and where $\overline{P}_{\psi,a(u)}$ represents a Generalized-Pareto tail with shape parameter $\psi$ and scale $a(u)$.*

The previous result permits using the Peaks-Over-Threshold method to estimate the unobserved tail of a Gompertz distribution. It is practical, among other things, because it is possible to complete the right tail of truncated data with this procedure.

Thus, to estimate the right tail, we should choose a threshold $u$ large enough to obtain a good approximation

$$\overline{F}(x+u) \approx \overline{F}(u)\overline{P}_{\psi,a(u)}(x). \tag{13}$$

If the sample size allows it, we can select $u$ such that the proportion of failure time data greater than $u$ is roughly $5\%$.

Before applying the method, it is recommended to smooth the failure time data; Kernel Density Estimation [7] can be used for such a task. Then, we recommend resampling from the estimated density and then applying the described methodology for the tail estimation to this new sample [7, 11].

Using the package ismev, we implemented the methodology described in this section in R. The corresponding code is in the GitHub repository mentioned in Appendix C under the title tail_Gompertz.R.

In the following, we present an application of the method described above to estimate the right tail of the cumulative infected cases. Let's assume we only observed the pandemic until February 1st, 2022; Figure 6 shows this graphically.

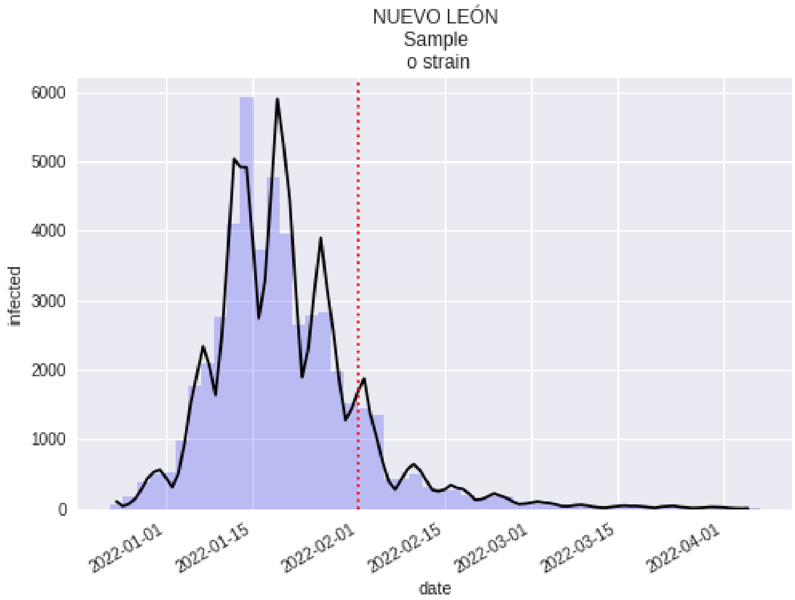


**Figure 6:** Right truncated data

As recommended, we start by smoothing the failure data; then, we estimate the right tail by the described procedure. The result is shown graphically in Figure 7.

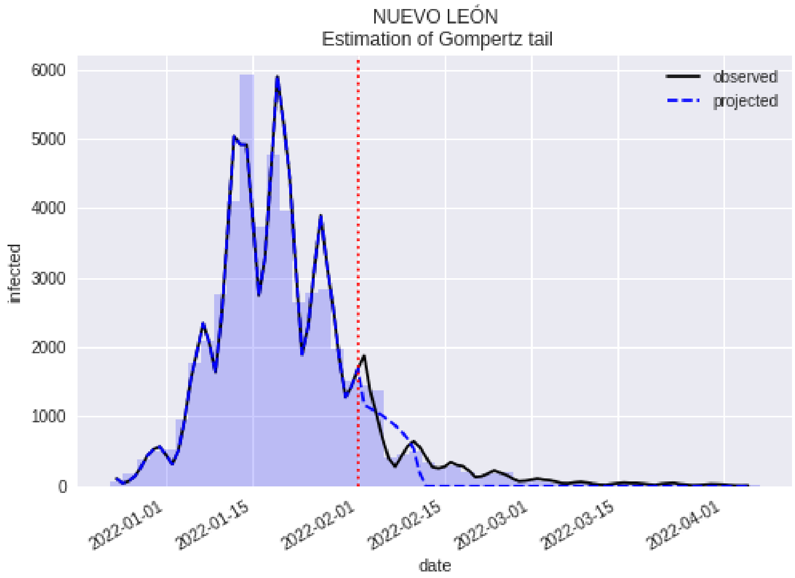


**Figure 7:** Gompertz tail estimation

We estimate the right tail of the daily cumulative infected cases by rescaling the tail estimation (13). More precisely, we employ the approximation

$$\widehat{D}_t = Y_u \cdot \widehat{\overline{F}}(u) \cdot p_{\widehat{\psi},\widehat{a}(u)}(t-u),$$

where $\widehat{D}_t$ is the estimation of the daily cumulative cases at day $t$; $Y_u$ is the cumulative infected cases at day $u$; $\widehat{\overline{F}}(u)$ is an empiric estimation of the susceptible population proportion at day $u$; and, $p_{\widehat{\Psi},\widehat{a(u)}}$ is a generalized-Pareto density with the estimated parameters. The estimator $\widehat{\overline{F}}(u)$ must be taken as empirical values according to the available information. For instance, it is possible to employ the susceptible population of another subperiod/variant at a similar time if the dynamic of both subperiods is similar or use the information of what has happened in other regions if the dynamics are similar again. Using this empirical value, it is possible to reestimate it using the Levenberg-Marquardt algorithm from the initial susceptible population.

Estimating the right tail of the cumulative cases is important for the present modeling framework. In particular, it allows to estimate a plausible asymptote for the cumulative cases and do the same type of analysis that the classical Gompertz model allows despite truncated data.

In the case of unobserved peak (when we are observing the subperiod beginning), we suggest getting the Cox process' rate function $m$ analytically as (9), i.e., we obtain

$$m(t) = \alpha_k + (X_0 - \alpha_k)\prod_{i=0}^{t}(1 - a_i),$$

where $a_i$ is as before considering the cumulative hazard function of a Gompertz distribution. We fix $\alpha_k$ to a reasonable value which can come from considering a similar susceptible population of another subperiod/variant or additional information. Then, with a closed form of the rate function $m$, we obtain preliminary Gompertz parameters solving the following optimization problem

$$(\hat{\gamma}, \hat{\xi}) = \text{argmin}_{(\gamma,\xi)}\left\{\sum_{i=1}^{r}(m(i) - \hat{Y}_i)^2\right\}, \tag{14}$$

and the solution can be found with a similar procedure as the used in the optimization problem in (10). These parameters could be used to provide a projection of the cumulative infected cases. This last part is not implemented in the repository.

## 6 Conclusions

This work proposes an interpretable parametric model to adequate the classical Gompertz growth model for epidemic modeling. The model allows multiple modes and heterogeneity. In addition, the model accommodates a framework that works even when the subperiods are still in progress which allows for building an asymptote to predict the epidemic. The particularities of the COVID-19 pandemic caused the epidemic curve for the infected cases to exhibit a multimodal structure. The primary motivation to develop the model discussed in this work was to have a growth model able to capture this dynamic.

The model treats the COVID-19 dynamic as the concatenation of three subperiods corresponding to the COVID-19 strain's dominance. That is assumed because each epidemic subperiod has its particularities. This has been observed for different regions around the globe; Mexico is not an exception. For instance, the starting dates for the pandemic and the applied public policies have varied, causing different types of dynamics in the reported infected cases. During a significant part of the pandemic in Mexico, the main concern was the management of the health infrastructure available to attend to the most critical infected individuals. However, currently, the most critical changes in the dynamic have been induced by the different variants of the virus.

The main contribution of this work is to propose a model that allows growth models with a multimodal structure throughout a mixture of Gompertz distributions. The model initially was constructed assuming that all, or almost all, the infections have already been seen. The model is then extended to the case where the epidemic is still in process, i.e., when the process is censored to the right (end of Section 5).

The Accelerated Cox Model did not significantly differ from the Baseline Cox Model. We could search for other covariables for improving the Accelerated Cox Model adjustment, this will be addressed in a future work. For incomplete subperiods, employing a good projection into the future for $R_t$ as a covariable, the accelerated Cox model could provide more assertive information on the asymptote for the Gompertz model. We are thinking of this as further work. In fact, we already have some methodologies for projecting $R_t$ [9].

For a more accurate depiction of reality, it is necessary to have information regarding all cases. A possible way to do this is through models that permit estimating the size of the undetected cases [6, 9]. In the present article, we studied a subset of an infected population since there were individuals infected with COVID-19 that never went to the hospital or did a COVID-19 test.

Finally, it is relevant to note that the proposed models can be used to model other regions and incomplete subperiods, provided there is enough data.

## Acknowledgements

The author R. Vásquez Martínez is grateful to Iván Rodríguez González (CIMAT) for the multiple discussions and data sharing. In addition, the four authors are very thankful to Dr. Javier Trejos for the invitation to submit this article.

## Financial support

The author R. Vásquez Martínez is grateful with National Council on Science and Technology (CONACYT) for the financial support of the Research Assistant's Scholarship directed by G. González Farías.

# Appendix A Initial conditions

This appendix discusses how to determine initial conditions for the Gompertz-Mixture model estimation. The proposed method is heuristic.

The first requirement for applying the EM Algorithm to the Gompertz-Mixture model is to determine the number of components. As in clustering analysis under a mixture model, we assume that each Gompertz component corresponds to one of several waves/peaks observed. The Gaussian Mixture Model (GMM) can relatively well capture the number of waves in each strain, even for underlying Gompertz distributions.

To estimate the number of components, we estimate a GMM model with 1, 2, 3, and 4 components and select the one with the lowest Bayesian information criterion (BIC). The number of components of the selected GMM model corresponds to the number of components of the Gompertz-Mixture model. The next step is to initialize the vectors $\pi_r$ and $\Psi_r$. We select $\pi_r^{(0)}$ as the mixture proportion of the GMM model with the lowest BIC.

Determining $\Psi_r^{(0)}$ is a little bit harder. For each subperiod/variant $r$, let $G_r$ be the number of components that comprise the corresponding subperiod. We employ the selected GMM model for each subperiod to cluster the failure times data $\{T_{rj} : j = 1, 2, \ldots, m_r\}$. For the $r$-th subperiod, let $D_{ir}$, $i = 1, 2, \ldots, G_r$, denote each corresponding cluster. We assume that the hazard function for each cluster satisfies

$$h_{ri}(t_j) = \exp(\gamma_{ri} + \xi_{ri} t_j), \qquad \text{for} \qquad t_j \in D_{ir}.$$

Employing the Nelson-Aalen estimator [25], we obtain a nonparametric estimator $\widehat{h}_{ri}$ for the hazard function $h_{ri}$ at each cluster. Then, we obtain initial parameters $\gamma_{ri}^{(0)}$ and $\xi_{ri}^{(0)}$ by solving the regression

$$\log \widehat{h}_{ri}(t_j) = \gamma_{ri} + \xi_{ri} t_j + \varepsilon, \qquad \text{for} \qquad t_j \in D_{ir}.$$

These values can be used to initialize EM Algorithm [14] used to estimate the Gompertz-Mixture parameters as presented in Section 2.2.

The GitHub repository mentioned in Appendix C includes the Python routine estimate_initial_parameters_risk, located at module initial_conditions.py, that implements the heuristic methodology described above.

**Remark 3** *Some elements in the heuristic procedure described in this appendix can be adapted if necessary. For instance, a higher number of components for the GMM model or a different nonparametric method to estimate the hazard function can be considered. However, we report those with the best performance in our tests for many different regions.*

# Appendix B $\mathrm{R_t}$ and COVID-19's variants

After a severe infectious period in Mexico in the summer of 2020, the epidemic curve of infected cases followed a slow increasing behavior until the end of 2020. This slow-growing period is likely a consequence of the public policies implemented to reduce virus transmission to avoid the saturation of health facilities. In February 2021, after Christmas, the most infectious period associated with the alpha variant was registered. This is a complete description of the dynamic of the alpha variant in Mexico.

In Mexico, during the alpha subperiod, most of the data for infected cases came from hospital reports, principally from public hospitals. More information was available from public testing sites and private laboratories for the delta and omicron variants. This increment in the available data resulted from the fact that individuals infected with these variants generally required fewer hospitalizations, although not few. These characteristics apply to Nuevo León, the selected region to exemplify the proposed models, primarily because it has an extensive health services infrastructure, including hospitals and laboratories. The type of information used to compose the data is fundamental to understanding the results provided for the models presented. In particular, it means that the results reflect just the dynamic of the observed cases, leaving out the undetected cases. For a more accurate depiction of reality, it is necessary to have information regarding all cases.

For the case of Nuevo León, the observed data cover until the beginning of April, when all the variants/subperiods are in remission. This means that we did not have to estimate the right tail, and predicting the asymptotes was direct. However, as mentioned earlier, the present work also discusses a way to estimate the right tail and compute an asymptote.

In this Appendix, we discussed some aspects of determining the start and final of the subperiods mentioned in Section 2. As we mentioned, this is partly based on the effective reproduction number $R_t$. The quantity $R_t$ describes the infection rate of an epidemic during the time. It is worth recalling that for a value $R_t < 1$, an epidemic/pandemic is controlled, while for $R_t > 1$, the epidemic/pandemic is in a growth phase. Figure 8 shows $R_t$ for the COVID-19 pandemic in Nuevo León, where we can observe that their values fluctuate between above and below 1.

The $R_t$ curve is obtained with a methodology similar to the one shown in [24] but taking 12 days as the epidemiologic week and different variances for each region studied in Mexico [9]. The $R_t$ calculation uses an initial value $R_0$ based on epidemical properties, for instance, the infectious disease period. Thus, the proposed start date in Table 2 is the moment when, based on a new $R_0$,

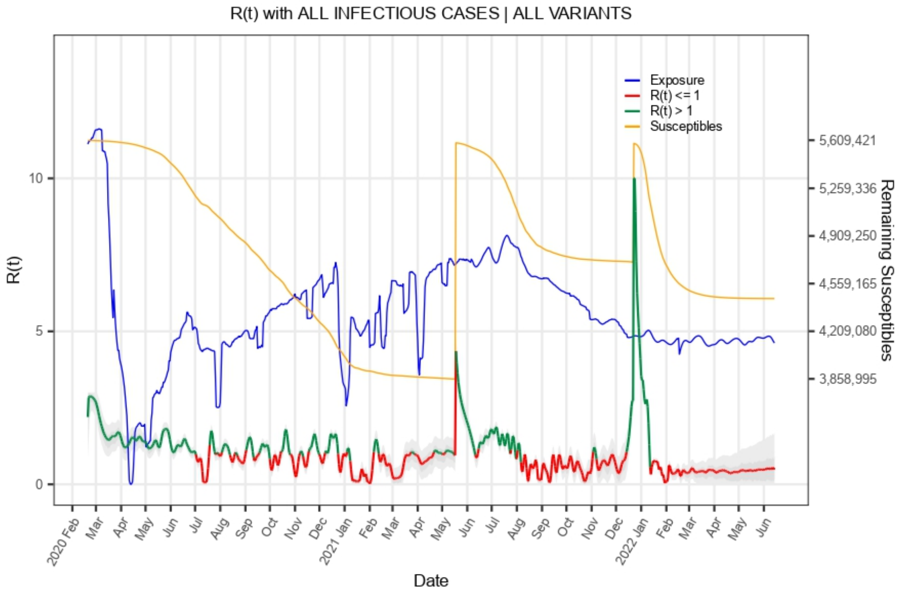
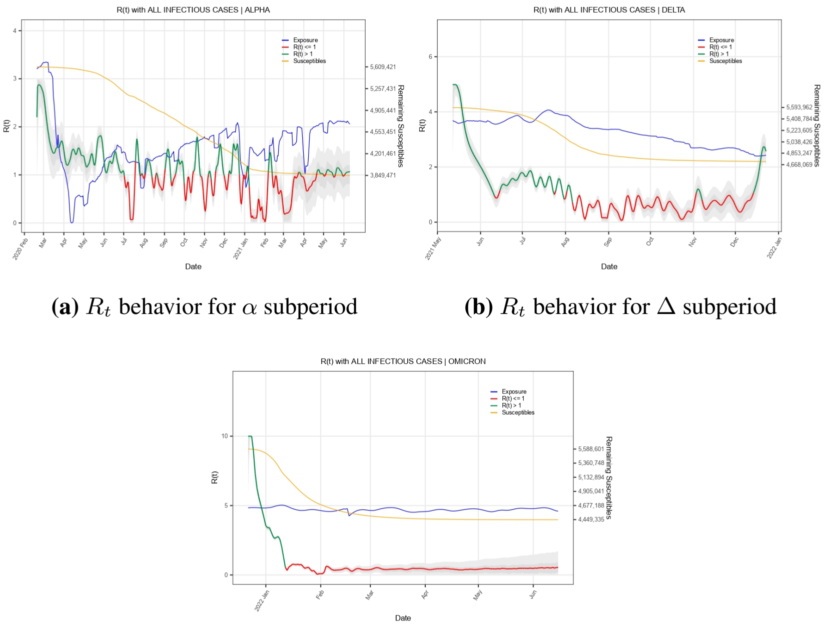


**Figure 8:** The figure presents the $R_t$ curve (red/green) from March 2020 to June 2022. The curve was generated using only the information up to March 2020. After this date, the values for the curve were produced by employing a projection (see [9]) with the available information. In the figure, we can also see how the susceptible populations evolve (yellow) for each subperiod under consideration described before.

we obtain a new $R_t$ that attempts to capture the epidemical features at that moment. We could see it as a restart of $R_t$ computation. The final of a subperiod and the beginning of a new one is, for practical purposes, determined when $R_t$ has been below 1 for a long time, and then it increases above, and the cases of a new variant surpass the cases of the previous dominant strain. The beginning of a new subperiod match approximately with the new variants' first cases; the prevalence and knowledge about the new variant are used, particularly in determining the values of $R_0$. The final date of a subperiod is one day less than the start date because we are treating the global dynamic as a concatenation of the subperiods' dynamic. For better visualization, Figure 9 shows the $R_t$ curve disaggregated by subperiods. Notice the difference in scales for each subperiod.

It is worthy to point out that the influence of each strain does not disappear totally, but we see how its dominance decreased through $R_t$, and because of $R_t$ nature, we can conclude that strain enters a controlled phase. When we include knowledge of a new strain in $R_t$ computation, we take the strain that shows dominance in the current epidemic dynamics.



**(a)** $R_t$ behavior for $\alpha$ subperiod

**(b)** $R_t$ behavior for $\Delta$ subperiod

**(c)** $R_t$ behavior for $O$ subperiod

**Figure 9:** The figure presents the curve of the effective reproduction number $R_t$ (green/red) for each strain. Notice that the scales are different; the $O$ had an explosive increment of cases, while during the $\alpha$ subperiod, the increment was somehow controlled. As explained before, this is caused because the observed cases are not the same for each subperiod. It is important to recall that the $O$ subperiod includes a projection of $R_t$, as we described in Figure 8.

# Appendix C Software and data

We prepared a GitHub repository with Python and R codes to reproduce the methodologies discussed in this work. Such a repository is located at https://github.com/robervz22/code-A-Gompertz-mixture-approach-for-modeling-the-evolution-of-the-COVID-19-dynamics, and the code also produces the figures included in this article.

On the other hand, the daily and accumulated infected data is open-access and available at the official site in [15]. A copy of this data is in the repository. The data of the effective reproduction number $R_t$ need special permissions, and therefore we omit it; however, the Jupyter Notebook GMM.ipynb indicates how that data can be used for the Gompertz-Mixture model.

# References

[1] I. Ahmed, G. U. Modu, A. Yusuf, P. Kumam, I. Yusuf, *A mathematical model of Coronavirus Disease (COVID-19) containing asymptomatic and symptomatic classes*. Results in physics **21**(2021), 103776.

[2] M. Català, D. Pino, M. Marchena, P. Palacios, T. Urdiales, P.-J. Cardona, S. Alonso, D. López-Codina, C. Prats, E. Alvarez-Lacalle, *Robust estimation of diagnostic rate and real incidence of COVID-19 for European policymakers*. PLoS One **16**(2021), no. 1, e0243701. DOI: 10.1371/journal.pone.0243701

[3] S. Coles, *An introduction to statistical modeling of extreme values*. Springer Series in Statistics. Springer-Verlag London, Ltd., London, 2001, xiv+208. DOI: 10.1007/978-1-4471-3675-0

[4] CONACYT, *Estimaciones de la Tasa de Reproducción Efectiva Rt de COVID-19 para los Estados y Zonas Metropolitanas de México*. 2022. URL: https://salud.conacyt.mx/coronavirus/investigacion/productos/. Accessed: 29-Apr-2022, 10:06 a.m.

[5] CONACYT, *Vigilancia de variantes del virus SARS-CoV-2*. 2022. URL: https://salud.conacyt.mx/coronavirus/variantes/. Accessed: 30-Apr-2022, 11:03 p.m.

[6] C. Gourieroux, J. Jasiak, *Time varying Markov process with partially observed aggregate data: An application to coronavirus*. J. Econometrics **232**(2023), no. 1, 35–51. DOI: 10.1016/j.jeconom.2020.09.007

[7] T. Hastie, R. Tibshirani, J. Friedman, *The elements of statistical learning*. Springer Series in Statistics. Springer, New York, 2009, xxii+745. DOI: 10.1007/978-0-387-84858-7

[8] J. D. Kalbfleisch, R. L. Prentice, *The statistical analysis of failure time data*. Wiley Series in Probability and Statistics. Wiley-Interscience, Hoboken, NJ, 2002, xiv+439. DOI: 10.1002/9781118032985

[9] J. U. Márquez Urbina, G. González Farías, L. L. Ramírez Ramírez, D. I. Rodríguez González, *A multi-source global-local model for epidemic management*. PLoS One **17**(2022), no. 1, e0261650. DOI: 10.1371/journal.pone.0261650

[10] A. W. Marshall, I. Olkin, *Life distributions*. Springer Series in Statistics. Structure of nonparametric, semiparametric, and parametric families. Springer, New York, 2007, xx+782.

[11] L. Martino, D. Luengo, J. Míguez, *Independent random sampling methods*. Statistics and Computing. Springer, Cham, 2018, xii+280. DOI: 10.1007/978-3-319-72634-2

[12] G. McLachlan, A. Ng, P. Adams, D. C. McGiffin, A. Gailbraith, *An algorithm for fitting mixtures of Gompertz distributions to censored survival data*. Journal of Statistical Software **2**(1997), 1–23. DOI: 10.18637/jss.v002.i07

[13] G. McLachlan, D. Peel, *Finite mixture models*. Wiley Series in Probability and Statistics: Applied Probability and Statistics. Wiley-Interscience, New York, 2000, xxii+419. DOI: 10.1002/0471721182

[14] G. J. McLachlan, T. Krishnan, *The EM algorithm and extensions*. Wiley Series in Probability and Statistics. Wiley-Interscience, Hoboken, NJ, 2008, xxviii+359. DOI: \10.1002/9780470191613

[15] Mexican Government, *Datos Abiertos Dirección General de Epidemiología*. 2022. URL: https://www.gob.mx/salud/documentos/datos-abiertos-152127. Consulted on 30-Apr-2022, 11:06 p.m.

[16] Mexico's Secretariat of Health, *Sistemas de Información de la Red IRAG*. 2021. URL: https://www.gits.igg.unam.mx/red-irag-dashboard/reviewHome. Consulted on 07-Sept-2022, 11:00 p.

[17] C. Murray et al., *Forecasting COVID-19 impact on hospital bed-days, ICU-days, ventilator-days and deaths by US state in the next 4 months*. MedRxiv. 2020. DOI: 10.1101/2020.03.27.20043752

[18] J. Nocedal, S. J. Wright, *Numerical optimization*. Second. Springer Series in Operations Research and Financial Engineering. Springer, New York, 2006, xxii+664. URL: https://www.math.uci.edu/~qnie/Publications/NumericalOptimization.pdf.

[19] J. Pickands, *Statistical inference using extreme order statistics*. Ann. Statist. **3**(1975), 119–131. URL: http://links.jstor.org/sici?sici=0090-5364(197501)3:1%3C119:SIUEOS%3E2.0.CO;2-O&origin=MSN.

[20] M. A. Pinsky, S. Karlin, *An introduction to stochastic modeling*. Elsevier/Academic Press, Amsterdam, 2011, x+563. DOI: 10.1016/B978-0-12-381416-6.00001-0

[21] S. I. Resnick, *Extreme values, regular variation, and point processes*. Vol. 4. Springer Science & Business Media, 2008.

[22] H. Rinne, *The Hazard Rate: Theory and Inference (with Supplementary MATLAB-Programs)*. 2014. URL: https://books.google.co.cr/books?id=welNuwEACAAJ.

[23] G. R. Shinde, A. B. Kalamkar, P. N. Mahalle, N. Dey, J. Chaki, A. E. Hassanien, *Forecasting models for coronavirus disease (COVID-19): a survey of the state-of-the-art*. SN Computer Science **1**(2020), no. 4, 1–15.

[24] K. Systrom, *Estimating COVID-19's in Real-Time*. 2022. URL: https://github.com/k-sys/covid-19/blob/master/Realtime%20R0.ipynb.

[25] T. M. Therneau, P. M. Grambsch, *Modeling survival data: extending the Cox model*. Statistics for Biology and Health. Springer-Verlag, New York, 2000, xiv+350. DOI: 10.1007/978-1-4757-3294-8

[26] K. M. Tjørve, E. Tjørve, *The use of Gompertz models in growth analyses, and new Gompertz-model approach: An addition to the Unified-Richards family*. PloS one **12**(2017), no. 6, e0178691. DOI: 10.1371/journal.pone.0178691

[27] C. P. Winsor, *The Gompertz curve as a growth curve*. Proceedings of the national academy of sciences **18**(1932), no. 1, 1–8.