# Report on the forecast of cumulative COVID-19 cases in Mexico

Mario Santana-Cibrian, Cátedra CONACyT
and Jorge X. Velasco-Hernandez
Nodo Multidisciplinario de Matemáticas Aplicadas,
Instituto de Matemáticas UNAM-Juriquilla

April 5, 2020

## Disclaimer

This is a technical report elaborated on behalf of the Nodo Multidisciplinario de Matemáticas Aplicadas del Instituto de Matemáticas UNAM-Juriquilla. The analysis it contains is based on the publicly available information as released by the Secretaría de Salud. Its aim is to contribute to the knowledge necessary to fight the sars-cov-2 epidemic. Any questions or comments, please contact Dr. Jorge X. Velasco Hernández, jx.velasco@im.unam.mx

## 1 Background

The National trajectory of cumulative COVID-19 cases in Mexico was reanalyzed in order to create short-term real-time forecasts of the outbreak. The data consists of daily cumulative cases from February 28, 2020 to April 5, 2020. The data was obtained daily at 19:00 p.m. (GMT-6) from the Secretariat of Health press conferences.

To create the forecasts, we use Richards model where cumulative cases $C(t)$ are given by

$$C'(t) = rC(t)\left[1 - \left(\frac{C(t)}{K}\right)^a\right], \tag{1}$$

where $r$ represents the growth rate of infection, $K$ is the carrying capacity (or final epidemic size) and $a$ is a scaling parameter. This model is an extension of the simple logistic growth model that has been recently used to predict cumulative COVID19 cases in China [5].

Parameters $a$, $r$ and $K$ must be estimated in order to properly describe the observed data and to provide accurate forecasts. We use an statistical approach through Bayesian inference.

## 2 Previous Estimates: March 24th

In our previous report [1]:

- The growth rate of confirmed cases $r$ was close to 0.2 (95% credible interval 1.9 to 2.1).
- With data up to March 24th we estimated the basic reproduction number for the whole country $R_0$= 2.7 (2.5, 3.2) using the exponential growth method, and $R_0 = 2.3$ (2.0, 2.6) with the maximum likelihood (//CRAN.R-project.org/package=R0) [1]. Given that, for the Richards curve, the growth rate for Mexico was $r = 0.2$ we determined that the more likely estimate for $R_0$ came from the maximum likelihood method. Fit was performed for data up to March 24 and used the next available days to test the performance of the algorithm.
- The total size of the outbreak $K$ is difficult to estimate since Mexico is at a relative early stage of the outbreak. We estimate that the final size of the outbreak will be around 710 392 (95% credible interval 65 163 -1 258 296). At this stage of the outbreak the long-term predictions have large uncertainity. We performed a 14 days prediction from March 25, 2020 to April 7, 2020 shown in Figure 1. We see that on March 31, 2020, the Richards curve predicts 1,103 cases (95% credible 401-2 353); the reported confirmed cases on that day was 1 215, which is within the range.

## 3 Key Findings: April 5th

- For April 7, 2020, the model predicts a median number of 4 202 cases ( 95% credible interval 1 440- 11 012). On April 5, 2020 the number of confirmed cases reported by Secretaria de Salud were 2 143, that indicates, in our view, that by April 7th the observed number of cases will be below the median number of cases predicted by the model. From the beginning of April there seems to be a reduction in the incidence compared to what was expected using the cases up to March 24. This could be caused for several reasons. Among the most importanta are a) the effect of the social distancing measures implemented by the government 15 days ago, b) a significative number of cases are under-reported.
- On, April 5th, we use the most recent data available to estimate the parameters of the Richards model. This time we start the analysis at March 10 since from this day the number of cumulative cases begins a sustained growth trend. The updated growth rate has a median value of 0.16 (95% 0.15-0.17), which is lower than the March 24th estimation. The final size of the epidemic is now estimated as 718 394 cases (95% 89 757-1 264 073). The prediction for the next 15 days can be seen in Figure 2. On April 12 we expect 1 842 to 14 143 cases with a median of 6 287. By April 19, from 5 076 to 41 173 cases are expected with a median of 17 477.

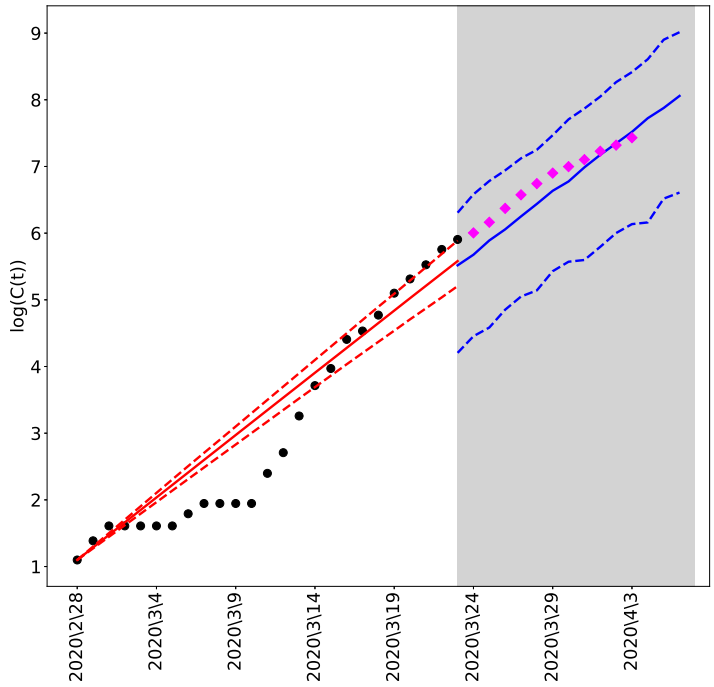


Figure 1: Logarithm of the cumulative COVID19 cases in Mexico from February 28, 2020 to April 5, 2020. Only data up to March 24 was used to fit the model. Black dots show the observed cases. Pink diamonds are the most recent observations that were not used to fit the model in order to test its short term accuracy. The red solid line shows the mean trajectory of the outbreak estimated from the data, red dashed lines are 95% probability intervals. The blue solid lines represent the mean posterior predictive trajectory of the epidemic and the blue dashed lines are 95% predictive probability intervals.

- The new estimation for $R_0$ is 1.91 (1.84, 1.98) with the exponential growth method and 1.76 (1.65, 1.86) with the maximum likelihood method. This is consistent with the reduction observed in the growth rate $r$ for the Richards model.
- It is important underline that Richards model and $R_0$ estimates are both based on the number of confirmed cases which can be significantly underestimating the true number of COVID-19 cases.

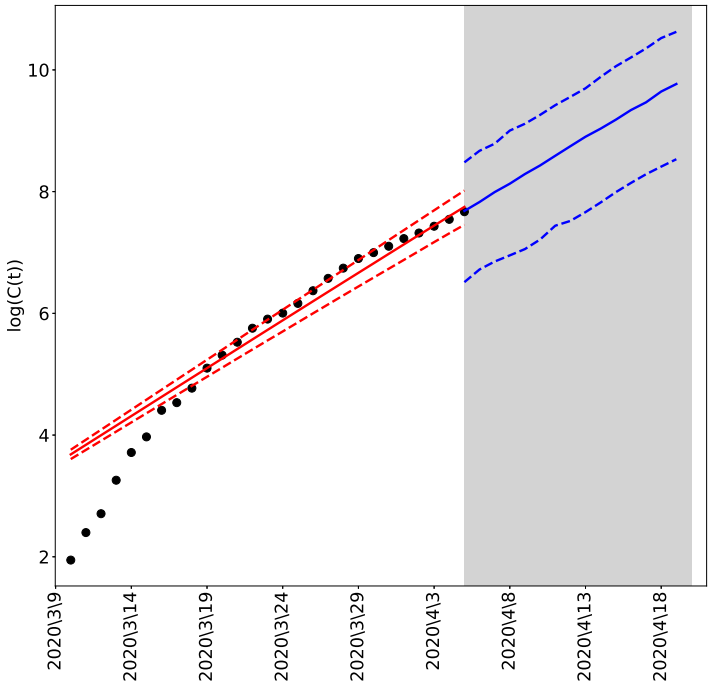


Figure 2: Logarithm of the cumulative COVID19 cases in Mexico from March 10, 2020 to April 5, 2020. All available data was used to fit the model. Black dots show the observed cases. The red solid line shows the mean trajectory of the outbreak estimated from the data, and red dashed lines are 95% probability intervals. The blue solid lines represent the mean posterior predictive trajectory of the epidemic and the blue dashed lines are 95% predictive probability intervals.

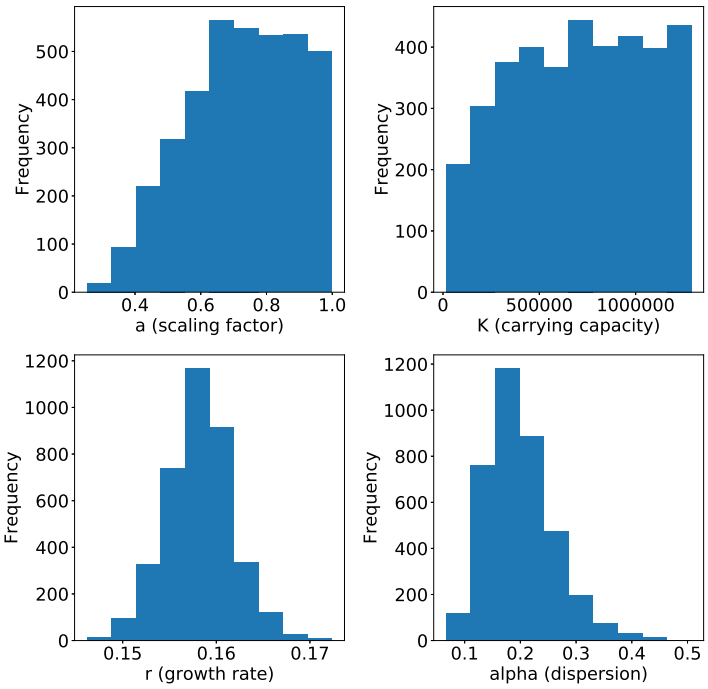


Figure 3: Posterior distribution estimates of the parameters involved in Richards model using data from March 10, 2020 to April 5, 2020.

# Technical Appendix

## A Parameter estimation

Let $Y_j$, for $j = 1, 2, ...n$, be the number of observed cumulative cases at time $t_j$, with $t_j$ given in days. We assume that $Y_j$ follows a Negative Binomial distribution with mean value $C(t_j|a, r, K)$ and dispersion parameter $\alpha$. Here, $C(t_j|a, r, K)$[1] is the solution of Richards model presented in (1). Assuming that, given the parameters, the observations $Y_1, Y_2, \ldots, Y_n$ are conditionally independent, then

$$E[Y_j|a, r, K, \alpha] = C(t_j|a, r, K) \tag{2}$$

$$Var[Y_j|a, r, K, \alpha] = C(t_j|a, r, K) + \alpha C(t_j|a, r, K)^2 \tag{3}$$

The Negative Binomial distribution allows to control the variability of the data by considering over-dispersion which is common for epidemiological data. If $\alpha = 0$, then we return to the Poisson model which is often used in this context.

Let $\boldsymbol{\theta} = (a, r, K, \alpha)$ be the vector of parameters to estimate. The inclusion of the parameter $\alpha$, which is related to the variability of the data, not to the Richards model, is necessary since in practice this variability is unknown. Then, the likelihood function, which represent how likely is to observe the data under the Negative Binomial assumption and Richards model if we knew the parameters, is given by

$$\pi(y_1, \ldots, y_n|\boldsymbol{\theta}) = \prod_{j=1}^{n} \frac{\Gamma(y_j + \tau)}{\Gamma(y_j)\Gamma(\tau)} \left(\frac{\tau}{\tau + C(t_j|a, r, K)}\right)^{\tau} \left(\frac{C(t_j|a, r, K)}{\tau + C(t_j|a, r, K)}\right)^{y_j}. \tag{4}$$

where $\tau = 1/\alpha$. Consider that parameters $a$, $r$, $K$ and $\alpha$ as random variables. Assuming prior independence, the joint prior distribution for vector $\boldsymbol{\theta}$ is

$$\pi(\boldsymbol{\theta}) = \pi(a)\pi(r)\pi(K)\pi(\alpha),$$

where $\pi(a)$ is the probability density function (pdf) of a Uniform(0,1) distribution, $\pi(r)$ is the pdf of a Gamma(shape=2, scale=1), $\pi(K)$ is the pdf of a Uniform($K_{\min}$, $K_{\max}$), and $\pi(\alpha)$ is the pdf of a Gamma(shape=2, scale=0.1). To select the prior for parameter $r$, we consider that previous estimation of $r$ are close to 0.3 [5], and a Gamma(2,1) represent a weekly informative prior as it allows for a wide range of values of $r$. Also, there is no available prior information regarding the final size of the outbreak $K$. This is a critical parameter in the model and, in order to avoid bias, we assume a uniform prior over $K_{\min}$ and $K_{\max}$. To set these last to values, we consider that the minimum number of confirmed cases is the current number of observed cases $Y(t_n)$ times ten, i.e.

[1] We changed the notation with respect to equation 1 to emphasize that the solution of the ODE depends on the parameters of interest.

| | Lower bound | Median | Upper bound |
|---|---|---|---|
| $a$ | 0.39 | 0.73 | 0.99 |
| $r$ | 0.15 | 0.16 | 0.17 |
| $K$ | 8.98 x $10^4$ | 7.18 x $10^5$ | 1.26 x $10^6$ |
| $\alpha$ | 0.10 | 0.19 | 0.35 |

Table 1: Posterior estimates for each parameter and 95% probability intervals. Data from March 10, 2020 to April 5, 2020 was used to produce these estimates.

$K_{\min} = y_n * 10$. The reasoning behind this number is that it has been suggested that the observed cases are under reported by a factor of 10 as reported by the Secretaria de Salud. To set the upper bound for $K$, we consider a fraction of the total population $K_{\max} = N * 0.01$, where $N$ is the population size of Mexico. This fraction was determined base on the observations of other countries such as Italy where the proportion of infected represents one of the worst case scenarios considering its population size.

Then, the posterior distribution of the parameters of interest is

$$\pi(\boldsymbol{\theta}|y_1, \ldots, y_n) \propto \pi(y_1, \ldots, y_n|\boldsymbol{\theta})\pi(\boldsymbol{\theta}),$$

and it does not have an analytical form since the likelihood function depends on the solution of the Richards model, which must be approximated numerically. We analyze the posterior distribution using an MCMC algorithm that does not require tuning called *t-walk* [2]. This algorithm generates samples from the posterior distribution that can be used to estimate marginal posterior densities, mean, variance, quantiles, etc. We refer the reader to [4] for more details on MCMC methods and to [3] for an introduction to Bayesian inference with differential equations. We run the t-walk for 200,000 iterations, discard the first 50,000 and use 2,000 samples to generate estimates of the parameters and short-term predictions for the next 14 days. Table 1 shows the median posterior estimates for each parameter and 95% probability intervals. Figure 3 shows the marginal posterior distributions estimates for each parameter.

# References


[1] Manuel Adrian Acuna-Zegarra, Andreu Comas-Garcia, Esteban Hernandez-Vargas, Mario Santana-Cibrian, and Jorge X. Velasco-Hernandez. The sars-cov-2 epidemic outbreak : a review of plausible scenarios of containment and mitigation for mexico. *medRxiv*, https://doi.org/10.1101/2020.03.28.20046276, 2020.

[2] J. Andrés Christen and Colin Fox. A general purpose sampling algorithm for continuous distributions (the t -walk). *Bayesian Anal.*, 5(2):263–281, 06 2010.

[3] J. Kaipio and E. Somersalo. *Statistical and computational inverse problems.* Springer-Verlag New York, 2005.

[4] Christian P. Robert and George Casella. *Monte Carlo Statistical Methods*. Springer Verlag, 2nd edition, 2004.

[5] K. Roosa, Y. Lee, R. Luo, A. Kirpich, R. Rothenberg, J.M. Hyman, P. Yan, and G. Chowell. Real-time forecasts of the covid-19 epidemic in china from february 5th to february 24th, 2020. *Infectious Disease Modelling*, 5:256 – 263, 2020.