Analysis of the early COVID-19 epidemic in Mexico through back-projection and a segmented regression model MT5020, 2025

Nora Ghenciulescu, Regina Crespo Lopez Oliver, Lena Weyer

21. May 2025

1 AI use

ChatGPT was used to improve the formulation of less than 10 single sentences. It was also used to debug code and to assist with LaTex issues.

2 Introduction

Mexico was one of the Latin American countries that was hit the hardest during the COVID-19 pandemic [4]. As of April 13, 2024 (last update date), Mexico had 7,702,809 reported cases, out of which there were 334,958 reported deaths [7]. During the pandemic, the local government was greatly criticized due to the lack of restrictions put into place [4]. On 15 March 2020, 8 states decided to close schools, but it was not until 31 March that all non-essential activities were suspended [8] [9]. Politicians and important figures pushed an early reopening to the lockdown in June 1st 2020, under a "traffic light" system, where red indicated a great danger for all people and focused on letting essential workers go to work, while green lifted restrictions in open spaces. This led to an overwhelmed medical system, longer lockdown for schools and affected vulnerable groups of people. Under such management, Mexico had the second highest excess deaths per thousand people in Latin America, after Peru [4].

In this report, we wish to analyse the early Coronavirus Disease 2019 (COVID-19) pandemic in Mexico by making use of backprojection to estimate infection times and employing a segmented regression model to estimate the 'change points' of the epidemic curve. The goal of our report is to relate critical moments of the COVID-19 crisis with the identified change points, placing focus on the first year of the pandemic.

3 Theoretical Background and Related Work

3.1 Regression with change points for Covid-19 epidemic curve in Germany

This work is inspired by a paper of Kuechenhoff et al.[5] analyzing the Covid-19 pandemic in Germany and its federal state Bavaria, in the time between March and the end of April 2020. The analysis succeeds in correlating the change points in a regression with interventions related to the Covid-19 pandemic. For the analysis, a segmented regression model with unknown change points was used. Models with different numbers of change points were compared and the model with the lowest Bayesian information criterion (BIC) was chosen. For the Bavarian data, the best performing model has five change points; for Germany, the model with four change points is chosen.

To have gain knowledge about the course of the pandemic a age stratified analysis was conducted. The researchers connected the change points with political and social events. With some change points a specific event could be assigned, like a political appeal for self-enforced social distancing with a decrease in numbers. Other change points were related to a variety of measures taken multiple days earlier. However, the authors stress that their analysis shows associations rather than causal connections.

3.2 Back-projection

Most of the time, the time of infection is unknown; we only know the time when a case was reported. If this infection time is needed in research, it can be estimated with the help of back-projection. Smoothed non-parametric back-projection was established in 1991 and has few assumptions compared to the parametrial back-projection methods used before. Non-parametrical back-projection is often unstable because of that it gets smoothed.

In this method, it is assumed that time can be divided into discrete intervals, resulting in an approximate representation. Another assumption is that the length of the incubation period does not depend on the infection time. f_d, defined as the probability that the duration of incubation is d days, is also assumed to be known and it is assumed that N_t, the number of individuals that got infected at day i, are independent Poisson variates. The expectation of N_t is defined as $E[N_t] = \lambda_t$. Under these assumptions, the likelihood of the number of corona cases diagnosed at day T is

$$L(y = (y_1, ...y_T)) = \prod_{t=1}^{T}(\sum_{i=1}^{t} \lambda_i f_{t-i}) exp(-\sum_{i=1}^{t} \lambda_i f_{t-i})$$

The goal is to get non-parametric and stable maximum likelihood estimates of the λ_t the expected number of individuals infected at day t. In smoothed non-parametric back-projection, the maximum likelihood is optimized numerically with an EM algorithm that includes a smoothing step. [1]

4 Data

The data used comes from the Secretary of Health branch of government in Mexico [2]. This data is available publicly, and contains several .csv files. Each one of these reports a broad number of specific data, such as co-morbidities, state of hospitalization, patient data, lab results etc. It is important to clarify that the dataset contains the date of symptoms, admission date to the hospital, and the time of death (if it is such a case). For our purposes, we refer to onset date as the date of symptoms, and reported date as the time the patient entered the clinic/hospital. We also find it worthy to mention that we only worked with one of these .csv files (corresponding to only the year 2020) due to the size and complexity of the data.

5 Methods

5.1 Data preparation and description

The original dataset contains individual cases from February 2020 to January 2021. For this specific report, we focused on age, time of symptoms, time of ingress (clinic or hospital), and type of patient (ambulatory or hospitalized). We also filtered only for positive COVID-19 cases, regardless of if the individuals were hospitalized or not.

For easier implementation with the code from the original paper, we summed up the cases grouping by age group, state, and date. In Figure 1 we can appreciate the cumulative number of cases through individuals' age. In Figure 2, we can see the distribution of the frequency of cases per age group. It is important to note how much this data appear to be skewed to the right. The data was cleaned, filtered and plotted in Python.

5.2 Back-projection

In order to correlate the timing of the change points with social events, not the time of disease onset but the time of infection should be used in the analysis. This is done by taking the incubation period into account with the help of a smoothed non-parametric back-projection.

As discussed in Section 3.2, the incubation-time must be known to apply this method. The incubation time was chosen based on the literature. In other acute respiratory viral infections, the days of infection seem to be log-normally distributed; this was also done by Lauer et al. for Covid-2019 [6]. They found that the median incubation period is 5.1 days, and the 97.5th percentile is 11.5 days.

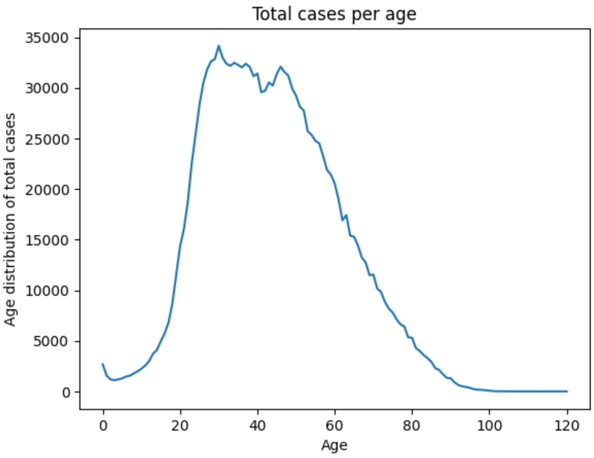
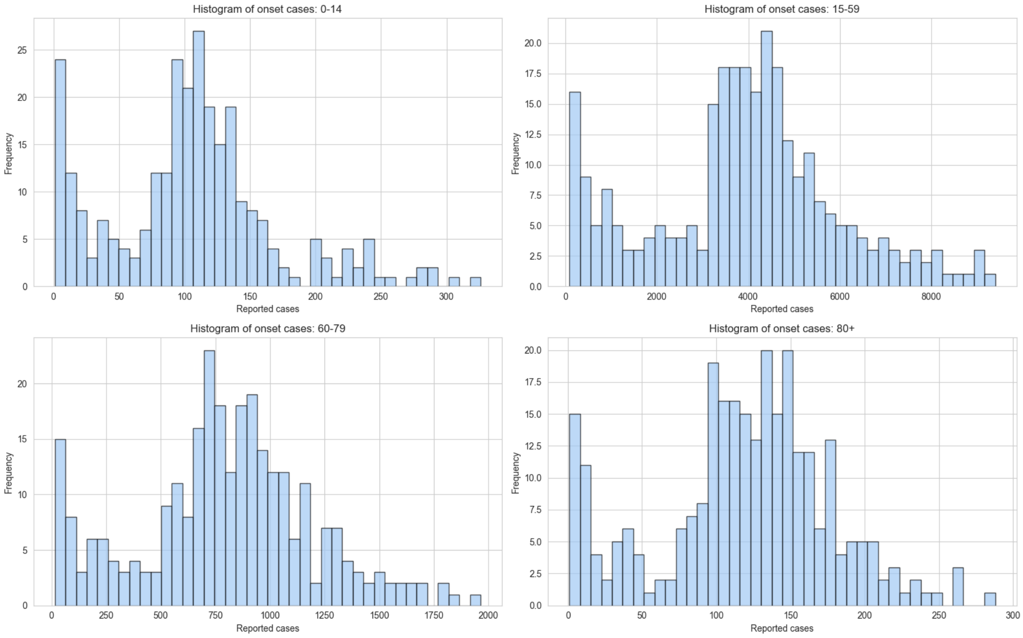

Total cases per age

Figure 1: Distribution of age in onset cases.


Histogram of onset cases: 0-14

Histogram of onset cases: 15-59

Histogram of onset cases: 60-79

Histogram of onset cases: 80+

Figure 2: Frequency of onset cases reported per age group.

With this knowledge the backprojNP function of the surveillance package in R [3] can be chosen to estimate the time of infection. We implemented this using adapted code from the original paper [5].

5.3 Segmented regression model

We use a segmented regression model with to fit the epidemic curve. This is a regression model where the data is partitioned into intervals, divided by 'change points', and a separate regression equation is fit for each interval. As in the original paper [5], the model is a generalised linear model given the change points; the model equation is:

$$E[log(Y_t)] = \beta_0 + \beta_1 t + \sum_{k=1}^{K} \gamma_k (t - CP_k)_+$$

Here, again as in the original paper, Y_t is the number of detected cases by time t of infection, β_0, β_1 and γ_k are the regression parameters, K is the number of change points and $x_+ = max(x, 0)$.

We run the model for various number of change points ($K \in \{2, 3, 4, 5, 6\}$, as in [5]). Subsequently, we use the Bayesian information criterion (BIC) to select the optimal K. Optimisation is done using the R package 'segmented', coupled with discrete optimisation. In our analysis, we directly use the functions provided in [5] to run the model on our data, with minor modifications to the parallelization procedure. We do not run their analysis files, but instead use our own code for calling the model functions. All analyses were conducted using R version 4.4.1.

6 Results

6.1 Dataset irregularities

Following the first case of Covid-19 in Mexico, reported on February 27th 2020, the pandemic had a slow beginning, as seen in Figure 3. Lockdown was implemented in Mexico on March 31st and lifted on June 1st (resumption of non-essential job activities); surprisingly, there are no dips in the infection curve around the period of the lockdown (Figure 3).

As seen in Figure 2 and Figure 3, the data is heavily skewed and noisy for a model to fit properly. In the case of Figure 2, if we pay attention to the frequency of small accumulated cases, we can understand why it is a problem for the model to numerically compute the models. Furthermore, looking at Figure 3, we can see that, for two of the most important dates (the lockdown start on March 31st and the restriction lift on June 1st), the data doesn't seem to follow any specific trend that the segmentation model might be able to pick up. As a result, the subsequent analyses discussed in this report, particularly Section 6.3, do not yield conclusions that can be tied to real-life data as neatly as in the original paper [5].

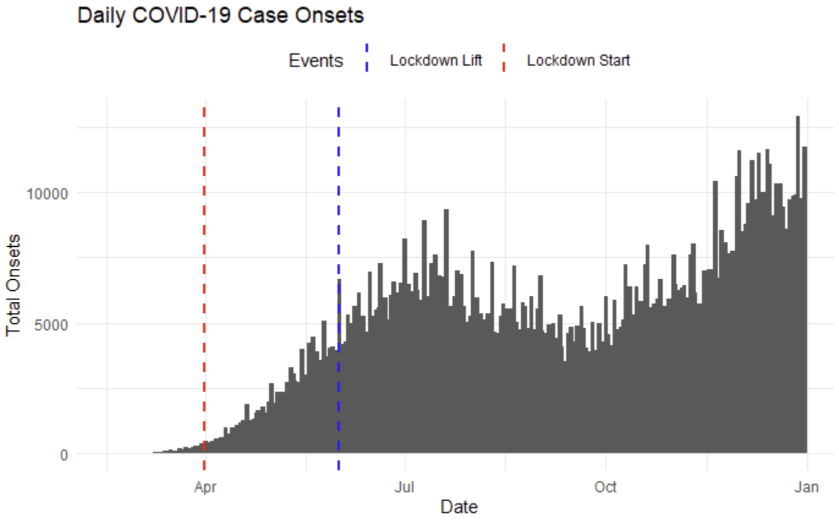


Figure 3: Barplot of disease onsets across time. The red and blue dashed lines denote the lockdown start (March 31st 2020) and lift (June 1st 2020), respectively.

6.2 Backprojection

We successfully manage to infer the back-projected data from the onset data using the smoothed non-parametric back-projection algorithm. In Figures 4 and 5, the curves of reported cases and disease onset can be compared with the back-projected data. It can be seen that the curves are shifted. The curve representing infection times is smoother than the onset curve due to the smoothing applied by the algorithm.

6.3 Attempts to fit regression model

In the original paper [5], the authors use data between February 19th 2020 and May 15th 2020. After performing back-projection, they use a complicated optimisation procedure in order to estimate the

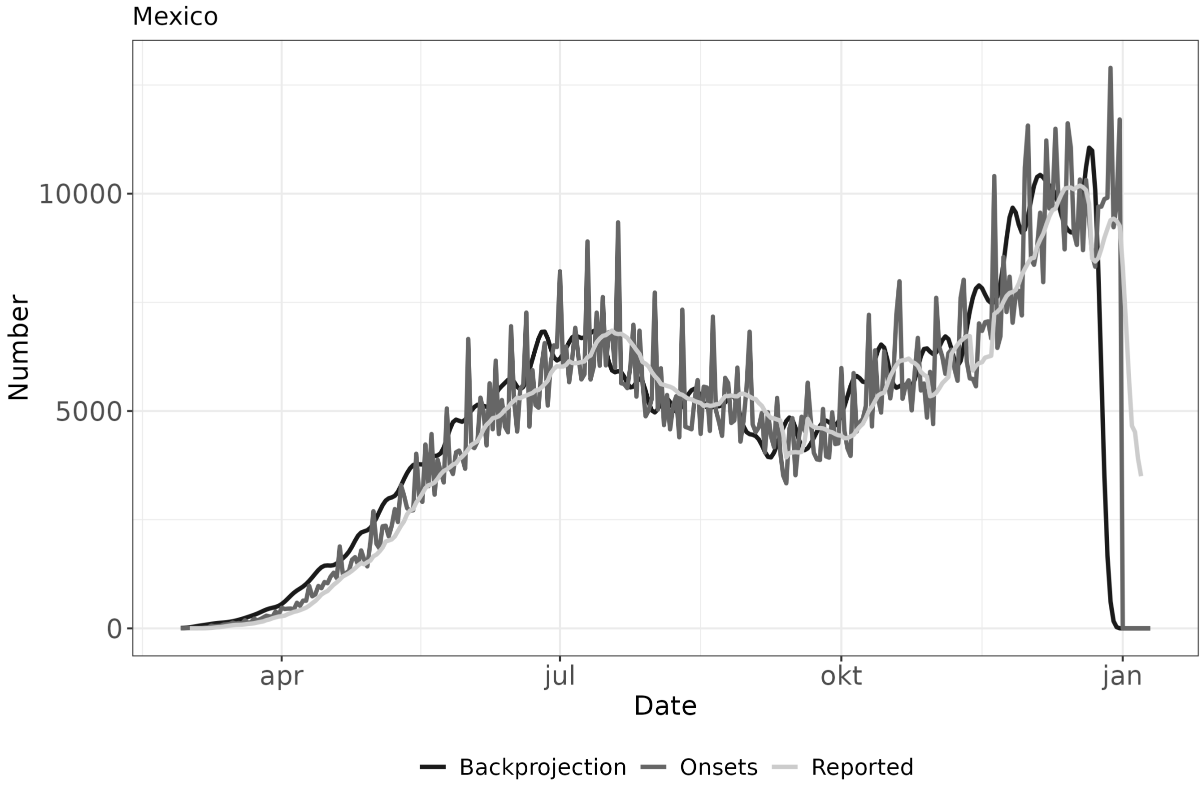


Figure 4: Comparison of time series of daily reported cases, disease onsets, and back-projection for the full data.

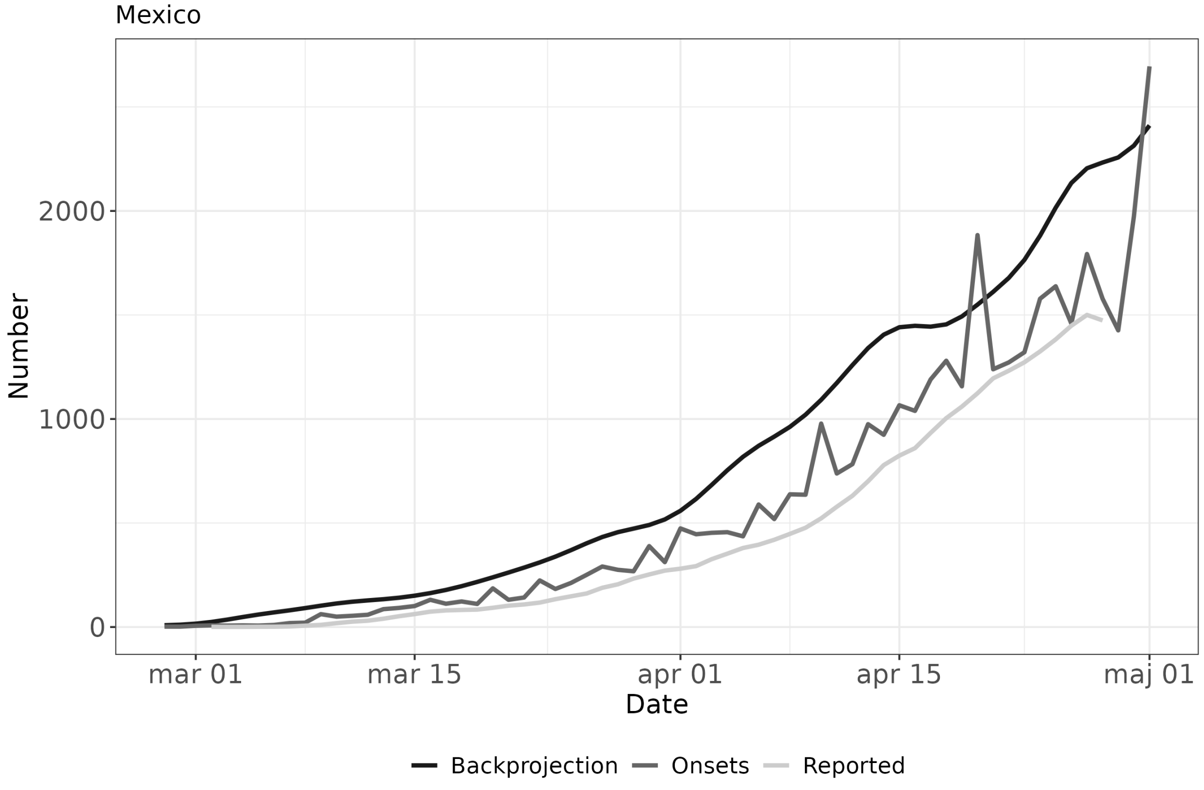


Figure 5: Comparison of time series of daily reported cases, disease onsets, and back-projection for March and April.

change points - this includes a discrete optimisation of the starting values, followed by fitting the segmented regression model. Parts of the optimisation code are hard-coded to fit the data used in [5] (so a time series consisting of only 61 data points); additionally, updates to certain R packages led to us being unable to extract the model deviance for the full data. Because of that, the code was run for the three-month period from the original paper (February 19th 2020 - May 15th 2020).

The fitted model of the epidemic curve for this 3-month period is shown in Figure 6; the optimal number of breakpoints is 3 and these are shown in Table 1. These change points could not be correlated to specific COVID-19-related events in the Mexican timeline [9]. The first two breakpoints, on March 4th and March 9th 2020, occur after the first COVID-19 case was confirmed in Mexico (February 27th 2020), but before COVID-19 was declared a global pandemic by the World Health Organization (March 11th 2020). The third breakpoint, on April 12th 2020, occurs some time after a lockdown was imposed (March 31st 2020) - interestingly, the epidemic appears to grow even more rapidly after this change point; this could indicate a failure of the lockdown measures to prevent the viral spread.

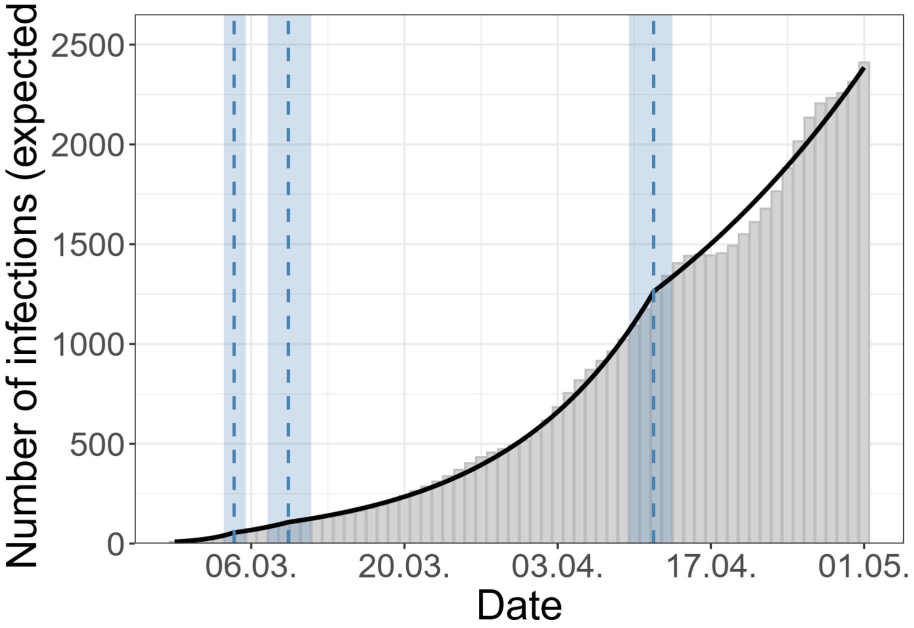


Figure 6: Results of the segmented regression model for reported cases in the period February 19th 2020 - May 15th 2020. The bars show the backprojected data (time of infection for each case), the solid line is the fitted curve according to the model. The number of change points is 3, as selected using BIC. The dashed lines and surrounding shaded areas indicate estimated change points and their respective 95% confidence intervals.

Break Point	**Lower CI Bound**	**Upper CI Bound**
6.4 (2020-03-04)	6.2 (2020-03-04)	6.7 (2020-03-05)
11.4 (2020-03-09)	10.5 (2020-03-08)	12.3 (2020-03-11)
44.7 (2020-04-12)	43.5 (2020-04-10)	46 (2020-04-13)

Table 1: Break point values for the February 19th 2020 - May 15th 2020 period, with confidence intervals. The first value is the index of the break point date in our data, the actual date is given in parentheses afterwards.

Next, we also stratify the data by age, keeping the same age groups as in the original paper: 0-14, 15-59, 60-79, 80+. The segmented regression model is subsequently run on each age group separately. This was again done only for the three-month period February 19th 2020 - May 15th 2020. We then standardize the number of infections for each age group using the overall age distribution over time in our dataset. Note that, since the model was fit separately for each age group, the number of change points varies per age group; also, for the 80+ segment of the population. Due to computational issues the model was fitted for not more than 2 change points. The results are shown in Figure 7. The 15-59 and 60-79 age groups behave similarly, exhibiting a change in slope around the previously identified April 12th change point. The 0-14 population exhibits a slower rise in the number of infections, perhaps indicating that the closing of schools was successful for slowing down the spread of COVID-19 among children in Mexico. The 80+ age group also exhibits distinct behaviour - the infection curve appears exponential, with no shift in slope around the previously identified breakpoints; this indicates that the mechanisms of infection were different among the older population compared to the middle-aged population.

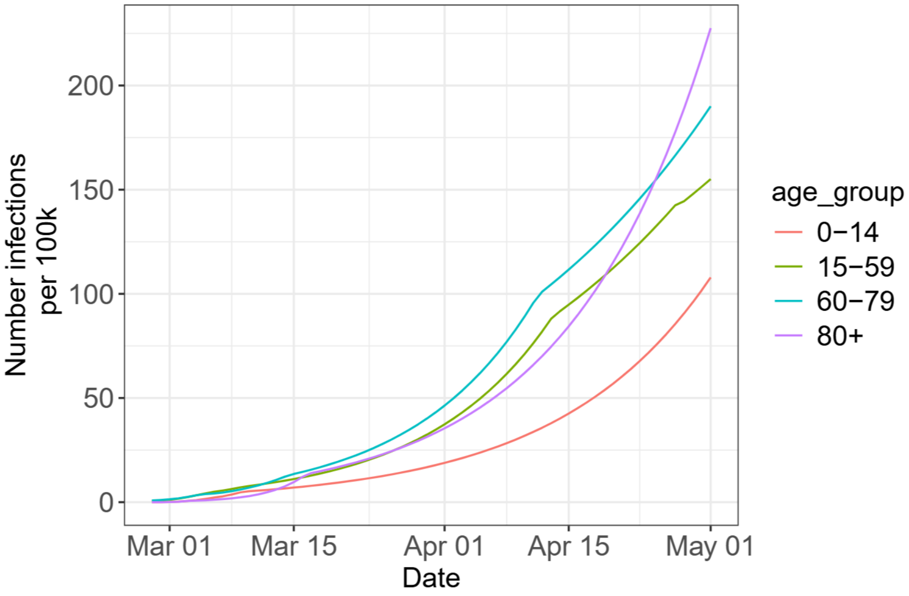


Figure 7: Results of the segmented regression model in the four age groups 0-14, 15-59, 60-79 and 80+. The lines show the number of infections per 100 000 individuals.

7 Discussion and Limitations

Compared to the German data of the original paper [5], the Mexican pandemic data appears to be much more erratic and lacking a clear peak through time. This made it difficult for the segmentation model to break the epidemic curve into distinct intervals and fit the regression lines; often, the models would simply not converge. While these problems may have been fixed by increasing the time span of the data, increasing the number of data points, or altering the presented model to be more robust since the original paper works with just 50 bootstrap samples and 1000 iterations at most on different functions), these solutions were not implemented due to a lack of time and the complexity of the original code (e.g. parallelization).

In terms of the quality of the data, the project would benefit from estimating the unreported cases. Given the distribution of the data shown in Figure 3, we believe that irregularities in the case reporting negatively impacted the data analysis process and made interpretation difficult. However, in a country where restrictions were lacking or implemented late [4], these irregularities make sense; even when policies were eventually put into place to prevent the spread of the virus, failure to follow them led to escalation of the pandemic. More than a case for segmentation, this shows that without strict restrictions, a pandemic can break out of control.

8 Conclusion

In this report, we successfully adapted part of the methodology in [5] to a Mexican dataset of the early COVID-19 pandemic. We employ smoothed non-parametric back-projection to infer the infection times from the case onset times for data spanning a full year. Furthermore, we fit a segmented regression model on a three-month period to estimate the change points in the epidemic curve. The optimal number of change points in this case was found to be 3. Interestingly, epidemic growth seemed to increase after each change point and these dates could not be correlated to established events in the Mexican pandemic timeline such as the start of lockdown. This could indicate that failure to follow preventive measures or implement such measures earlier led to uncontrollable growth of the epidemic, in contrast to the pandemic development in Germany [5]. Overall, our report highlights the difficulties of adapting infectious disease models across countries. This showcases the need to look for ways to properly model ”noisy” epidemic curves, corresponding to countries with a less reliable case reporting system and a less developed prevention apparatus.

9 Code and Data Availability

All code and data that were used in these project are available at `https://github.com/ReggieScript/InfectiousDisease-COVIDMX`. The relevant code is also uploaded on Moodle.

References

[1] Niels G. Becker, Lyndsey F. Watson, and John B. Carlin. “A method of non-parametric back-projection and its application to aids data”. en. In: *Statistics in Medicine* 10.10 (1991). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.4780101005, pp. 1527–1542. ISSN: 1097-0258. DOI: 10.1002/sim.4780101005. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.4780101005 (visited on 05/19/2025).

[2] *Datos Abiertos Dirección General de Epidemiología — Secretaría de Salud — Gobierno — gob.mx.* URL: https://www.gob.mx/salud/documentos/datos-abiertos-152127 (visited on 05/21/2025).

[3] Michael Hoehle et al. *surveillance: Temporal and Spatio-Temporal Modeling and Monitoring of Epidemic Phenomena.* Nov. 2024. URL: https://cran.r-project.org/web/packages/surveillance/index.html (visited on 05/20/2025).

[4] Felicia Marie Knaul et al. “Punt Politics as Failure of Health System Stewardship: Evidence from the COVID-19 Pandemic Response in Brazil and Mexico”. English. In: *The Lancet Regional Health – Americas* 4 (Dec. 2021). Publisher: Elsevier. ISSN: 2667-193X. DOI: 10.1016/j.lana.2021.100086. URL: https://www.thelancet.com/journals/lanam/article/PIIS2667-193X(21)00082-X/abstract (visited on 05/21/2025).

[5] Helmut Küchenhoff et al. “Analysis of the early COVID-19 epidemic curve in Germany by regression models with change points”. en. In: *Epidemiology and Infection* 149 (2021), e68. ISSN: 0950-2688, 1469-4409. DOI: 10.1017/S0950268821000558. URL: https://www.cambridge.org/core/product/identifier/S0950268821000558/type/journal_article (visited on 05/20/2025).

[6] Stephen A. Lauer et al. “The Incubation Period of Coronavirus Disease 2019 (COVID-19) From Publicly Reported Confirmed Cases: Estimation and Application”. en. In: *Annals of Internal Medicine* 172.9 (May 2020), pp. 577–582. ISSN: 0003-4819, 1539-3704. DOI: 10.7326/M20-0504. URL: https://www.acpjournals.org/doi/10.7326/M20-0504 (visited on 05/19/2025).

[7] *Mexico COVID - Coronavirus Statistics - Worldometer.* en. URL: https://www.worldometers.info/coronavirus/country/mexico/ (visited on 05/21/2025).

[8] V. Suárez et al. “Epidemiología de COVID-19 en México: del 27 de febrero al 30 de abril de 2020”. en. In: *Revista Clinica Espanola* 220.8 (May 2020), p. 463. DOI: 10.1016/j.rce.2020.05.007. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC7250750/ (visited on 05/21/2025).

[9] *Timeline.* URL: https://www.comisioncovid.mx/linea-del-tiempo.html (visited on 05/21/2025).