Nowcasting COVID-19 deaths with Reporting Delay corrections

Universitat Autònoma de Barcelona

Master in Modelling for science and engineering

Saioa Etxebarria[1], Jorge Piqueras[2], Sergio Villar Mesurado[3],
Arturo del Cerro Vergara[4], Sharika Vijayakumar[5], supervised by Xavier Mora[6]
[1] *saioa.etxebarria@e-campus.uab.cat* [2] *jorge.piqueras@e-campus.uab.cat* [3] *sergio.villar@uab.cat*
[4] *arturo.delcerro@ e-campus.uab.cat* [5] *sharika.vijayakumar@e-campus.uab.cat* [6] *xmora@mat.uab.cat*

Introduction

- When an event takes place in an epidemic (a new infection, death, recovery, etc), it is not always reported on the same day that it occurred. **The difference in time between the appearance of an event and the entry of the event into a dataset, is the so-called reporting delay**. Reporting delays can be caused by multiple factors. For instance, the diagnostic tests (such as the RT-PCR or the antigen test) take time to produce the results, and once they are produced, the event must be collected in a dataset.
- For this reason, when a bar graph of the new or total events is created, the **counts attributed to the most recent days are lower than they should be, as some of the events that actually occurred for these days have not yet been reported**. This is called time-length bias, and it can lead to a missinterpretation and underestimation of the current development of the epidemic, as it might make it seem like the events curve is getting smoother or more downward trending, when in reality it is not.
- The ability to correct the data and give an estimation of the actual events that took place on a certain day, building a **statistical model that takes account of the reporting delays**, is key to a correct interpretation of the day-to-day evolution of the epidemic, and consequently, to have better grounds for the policy-making against it.

The statistical method

■ Let $\mathbf{n_{td}}$ denote the number of events that happened on day t, and were reported on day $t+d$, where d is the reporting delay.

■ If the total number of days is C, then for a day $t = 0, 1, .., C$, the reporting delay can only take values $d = 0, 1, ..., C - t$. By means of the counts n_{td}, we can perform a Poisson regression with formula:

$$\log(n_{td}) = \alpha_t + \beta_d + \sum_x I(t + d = x)\gamma_x,$$

where $x = Monday, Tuesday, ..., Sunday$, is a day where a certain special reporting effect might occur, such that we can include the indicator function I in order to take account of it in the model.

■ Then, the estimated proportion of events on day t being reported with delay d, among all of the reported with delay smaller than d is:

$$g_{td} = \frac{exp\{\beta_d + \sum_x I(t + d = x)\gamma_x\}}{\sum_{i=0}^{d} exp\{\beta_i + \sum_x I(t + i = x)\gamma_x\}}$$

■ Given the total number of events on day t, as reported by time C, $N(t, C)$; that is, the uncorrected events for day t; we can calculate the estimated corrected events on day t, $\hat{N}(t)$ (that would be reported in time $t >> C$) as follows:

$$\hat{N}(t) = \frac{N(t, C)}{\prod_{j=C-t+1}^{C}(1 - g_{td})} \qquad (1)$$

■ The model's performance is ideal for big values of C and considers that the reporting system is stationary and does not evolve in time and among the individuals.

The problem

■ After a rough Christmas, the curve of new events and deaths in **Canarias, Spain** would start decreasing in mid January, due to the restrictions applied by the government. We use daily reported data from a public database for the death incidence in Canarias, in the time span **from 10/01/2021 to 01/02/2021**, so $C = 23$.

■ The goal of the study is to apply the statistical model to the new deaths in this time span, and compare it to the "real" counts that would be reported weeks later for each day.

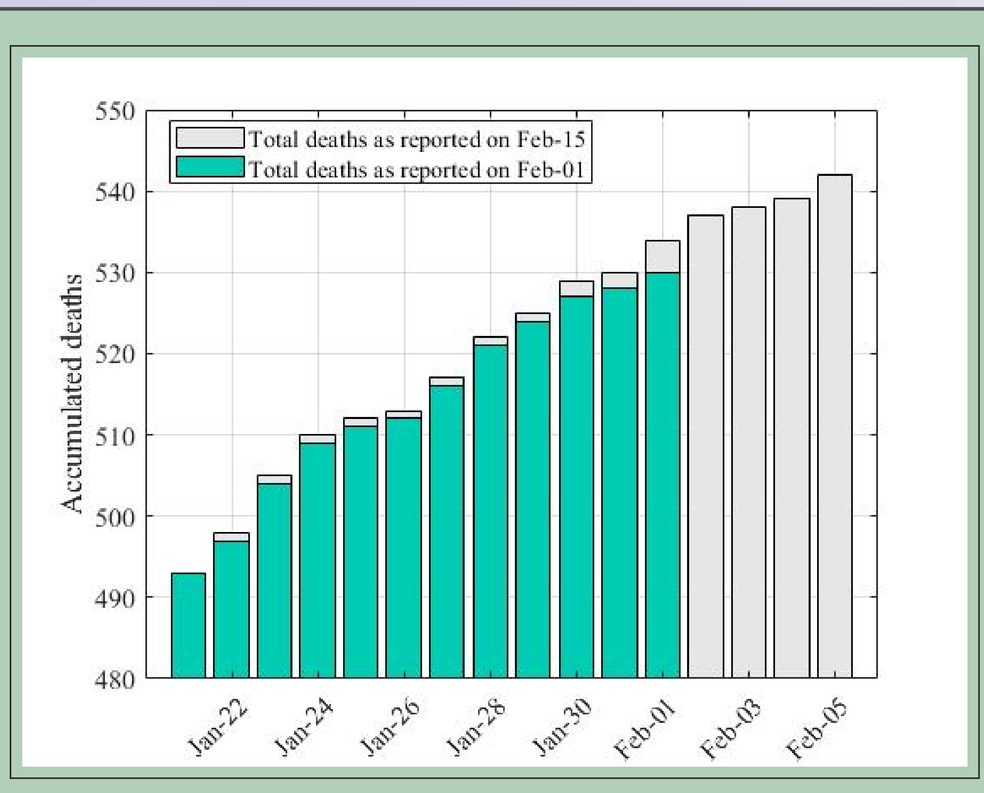

Figure 1: Comparison between the total deaths in a time span as reported by two different days. The difference of counts is due to reporting delays.

Results

■ We perform the Poisson regression adding two Indicator functions, one for Mondays, and one for Sundays. In Figure 2 we show the following results:

- Uncorrected new deaths $N(t, C)$, as reported on 01/02/21
- Corrected counts $\hat{N}(t)$, obtained from the model (1)
- Real counts as reported weeks later

The corrected counts from the model are a solid estimation of the effect of the reporting delays.

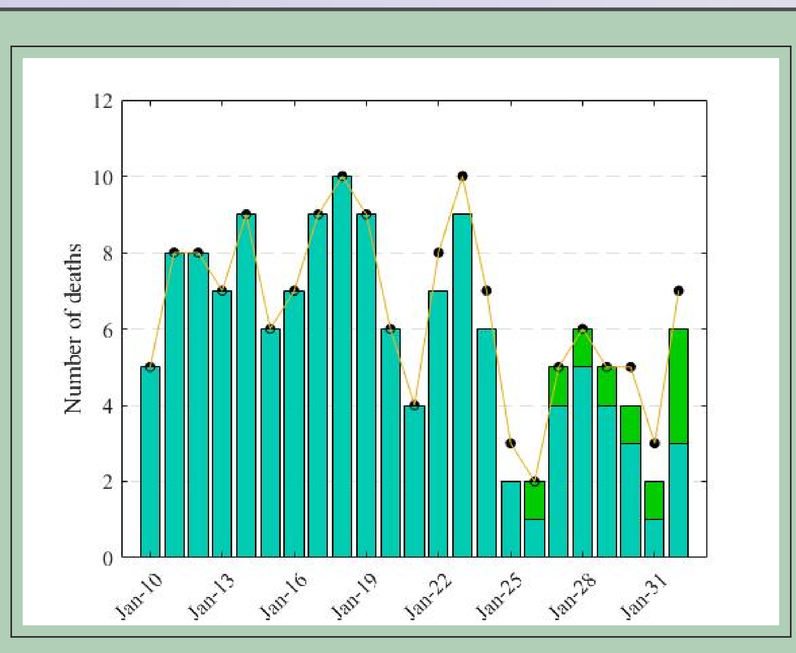

Figure 2: Statistical model prediction compared with the real data

Conclusion

- As Figure 2 shows, the data as reported on 01/02/21, $N(t, C)$, can be misleading, as it displays a rather downward sloping curve for the new deaths. The statistical model is successful at correcting the tendencies of the curve so that it resembles reality more precisely.
- This type of correction can be done daily for live current data, in a nowcasting exercise which rectifies the data in a way that might be counter-intuitive just by looking at the raw counts.
- Reporting delay statistics are a key tool for data processing in the current COVID-19 epidemic, just as they were for the SARS (2003, Canada, Hong Kong and Singapore) or the AIDS (1993, Canada) outbreaks.

Acknowledgements

The data used in this study can be found online on the `Escovid19data` project: `https://github.com/montera34/escovid19data`. A more comprehensive analysis of the reporting delay effect and time-length bias can be found on the chapter 7.3 of book by P.Yan and G. Chowell, *Quantitative Methods for Investigating Infectious Disease Outbreaks* (2019).