COVID-19: Using an Automated Chain Ladder technique to predict ultimate Daily Deaths in a live environment

You are here

10 June 2020

Richard Kelsey, Director of RTP Analytical Services Limited, discusses how an Automated Chain Ladder technique can be used to predict ultimate Daily Deaths in a live environment.

The COVID-19 pandemic has focused the world's attention on models of illness and death. The key input to these models, and hence the decision-making of governments is data. This blog argues that relying on daily reported deaths may not tell the full story due to the delay in reporting deaths, sometimes up to 10 days. This slows down the identification of peaks in the pandemic. I propose a solution that uses a standard general insurance actuarial technique to predict the number of deaths happening each day, allowing for delays in reporting.

The model was built in such that it could have run live from 7 April, in an automated manner estimating ultimate deaths by date of death from 1 April up to the as at date of the model run. Used in this way, it would have identified the peak in deaths in the UK two days earlier than was actually identified. Furthermore, since the peak (8 April), the number of deaths per day can be seen to be reducing by 5% per day, or equivalently halving every 14-15 days since then.

This approach could be used if there is a second wave, and ideally should be applied to data in different regions. It could also be used on total deaths to measure excess deaths in the period, i.e. the number of deaths in excess of what would be experienced in normal times.

Using the time to report as well as the date of death to predict total deaths

Many of the statistics on COVID-19 deaths are based on a date of reporting. Unfortunately, using data by reporting date only shows a misleading picture. The under-reporting at weekends, with subsequent catch-up during the following week, is well known. There are other reporting delays and, following the Coronavirus 2020 Act in late March, there is a confounding effect of speeding up in the reporting of deaths making it look like more deaths have occurred.

The Office for National Statistics (for England and Wales, or NRS in Scotland and NISRA in Northern Ireland) present deaths by date of reporting and a one-week lagged presentation of deaths by date of death. Unfortunately, perhaps due to the one week lag, this data gets reduced coverage. Maybe "new" data such as yesterdays or last weeks reported count is considered more useful than say two week old, better presented data, despite the risk of misleading

conclusions. A two dimensional analysis of the data by date of death and reporting delay identified these features, but is a little cumbersome to present and not always easy to interpret. The Chain Ladder technique [1] is a simple technique used by general insurance actuaries to use the extra dimension of reporting delay to project the date of deaths count to an estimated ultimate amount. Christofides, Oka and Heneghan used the Chain Ladder method on COVID-19 deaths data from daily NHS England data for the month of April [2]. In their paper, they identified a number of limitations. In particular, (i) assuming the data is stable over time, and (ii) assuming lower week-end reporting does not materially affect the results. The aim of their paper is to produce accurate daily predictions of ultimate death counts by date of death by using an automated Chain Ladder.

An automated Chain Ladder with adjustments for the two factors above (stability and weekend reporting) could have been deployed live to produce predictions for all prior dates of death from 7 April; some 7 days after the start of the release of daily data. This paper described such an automated chain ladder.

Key findings from this adjusted analysis are:

- An automated Chain Ladder would have predicted the peak for deaths was 8 April and this could have been identified by 11 April i.e. after running and updating the model daily, the run on 11 April would have identified 8 April was the peak and that both 9 and 10 April would be lower. On the more widely circulated daily reported data which is more volatile, the same call would have been made cautiously on the 14 of April, but with trepidation as reported deaths rose and fell without obvious trend up to 24 April.

- The reduced weekend reporting is material and needs to be allowed for. Since the model was designed for a live environment, it was developed using data from 5 April to adjust for this. The adjustment is a simple "grossing up of cumulative deaths reported by the Sunday to the value they would have been had it been a normal working day and then derive the chain ladder projection factor. Likewise, the count of deaths was also grossed up to a non-Sunday amount before projecting to 'ultimate'. Lower volumes were not reported during Easter weekend – perhaps because of increased working. Knowing the process and understanding staffing levels and 'how it works in practice is the key to accurate early days prediction. It is certainly not "24/7".

- There was a speeding up of reporting and reduction of reporting lag identified in two waves. To identify reporting speed, the distribution of deaths reported in the 10 days following death was investigated. A higher proportion reported in the immediate days following death would be indicative of a speeding up, particularly if the distribution then followed this new pattern.

  1. 6-14 April (The proportion of deaths in the first two days compared to deaths reported in the first 10 days increased from 50% to 60%, and the first 4 days increased from 80% to over 85%)
  2. From 15 April (The proportion of deaths in the first two days compared to deaths reported in the first 10 days increased from 60% to 70%, and the first 4 days increased from over 85% to over 90%)
  3. The timing and quantum of these were not expected, so their effects had to run through the prediction model.

- There are still significant reporting delays on recording deaths. For example 1 April, a further 3% of deaths were recorded between 30 April to 12 May.

- Since the end of April, a new Friday effect has been observed where there seems to be a speeding up of reporting before the weekend.

Data

The NHS England data records daily reported deaths from COVID-19 who had previously tested positive by date of death daily. The first report is at 2 April. The most recent data used for this analysis is the 16 May update. A Consolidating summary of by date of death for deaths prior to 2 April was also available. There is to the authors knowledge no other dataset available capable of showing development of deaths by date of death. Thus the analysis specifically excludes Scotland and Northern Ireland and deaths occurring outside of hospital (e.g. Care homes or hospices).

Methodology

Weighted Chain Ladder using last 7 observations, for first 10 days.

6% (10% before 15 April) tail factor at 10 days, with exponential decay to 0% after 30 days (40 days before 15 April) to accommodate extended reporting delay.

Initially no adjustment was assumed. However due to the magnitude (see Chart 1) the affected days (Friday, Saturday and Sunday deaths) were grossed up to derive a "weekday" daily development factor. If the "as at" date of the projection was a Sunday, Friday, Saturday and Sunday deaths were further grossed up to a "expected weekday amount" by an adjustment development factor.

Since no prior information was available for speeding up of reporting, no a priori adjustment could be made.

Results

The format of the results is a two dimensional array of predictions of ultimate deaths by date of prediction. Date of death ranges from 1 April 2020 to 12 May 2020. Date of prediction ranges from 7 April 2020 to 12 May 2020. To ensure no "benefit of hindsight", an automated Chain Ladder was used.

Chart 1: This shows the predictive modelling results. This shows that the prediction improves with the aging of the model. The actual reported deaths reported on that day is overlaid with the red line.
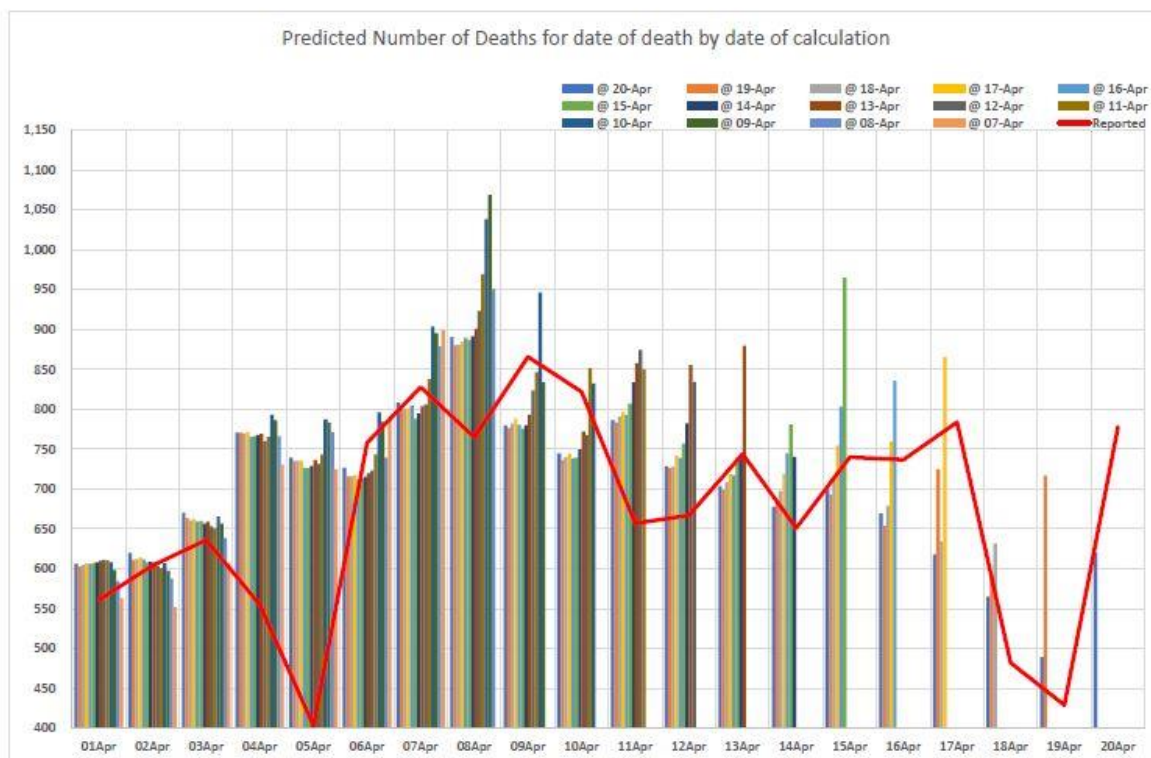
Predicted Number of Deaths for date of death by date of calculation

[3] Source: The data used in the above model was taken from NHS England Covid-19 Daily Deaths.

Most of the volatility has reduced by "date of death+2", this can be seen in chart 2 below. Chart 2: Convergence of Predictor. This shows convergence after 2 days therefore as a predictor the 8 April fall was predicted from 11 April in this robust model. Daily Deaths are now predicted to be down from the peak of just under 900 to below 150 from 16 May, and these lower numbers are inducing more volatility in the estimates. The "decay curve after the 8 April peak" reduces by 5% per day compound.
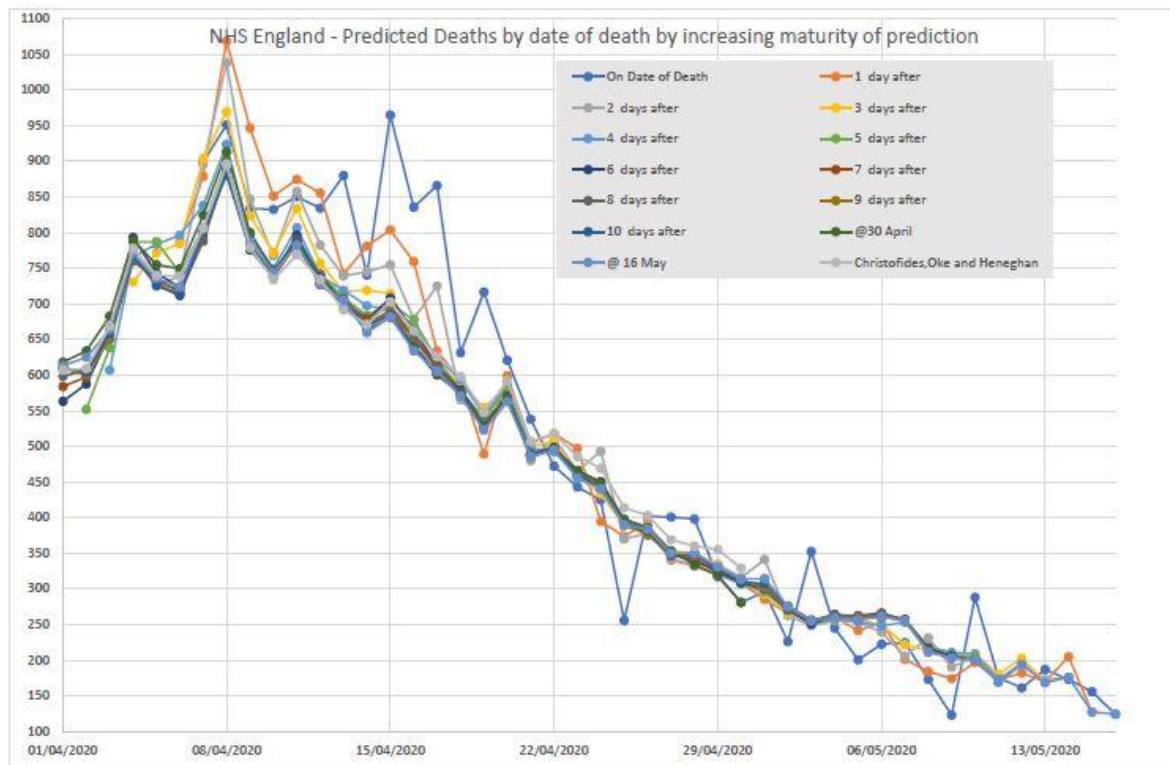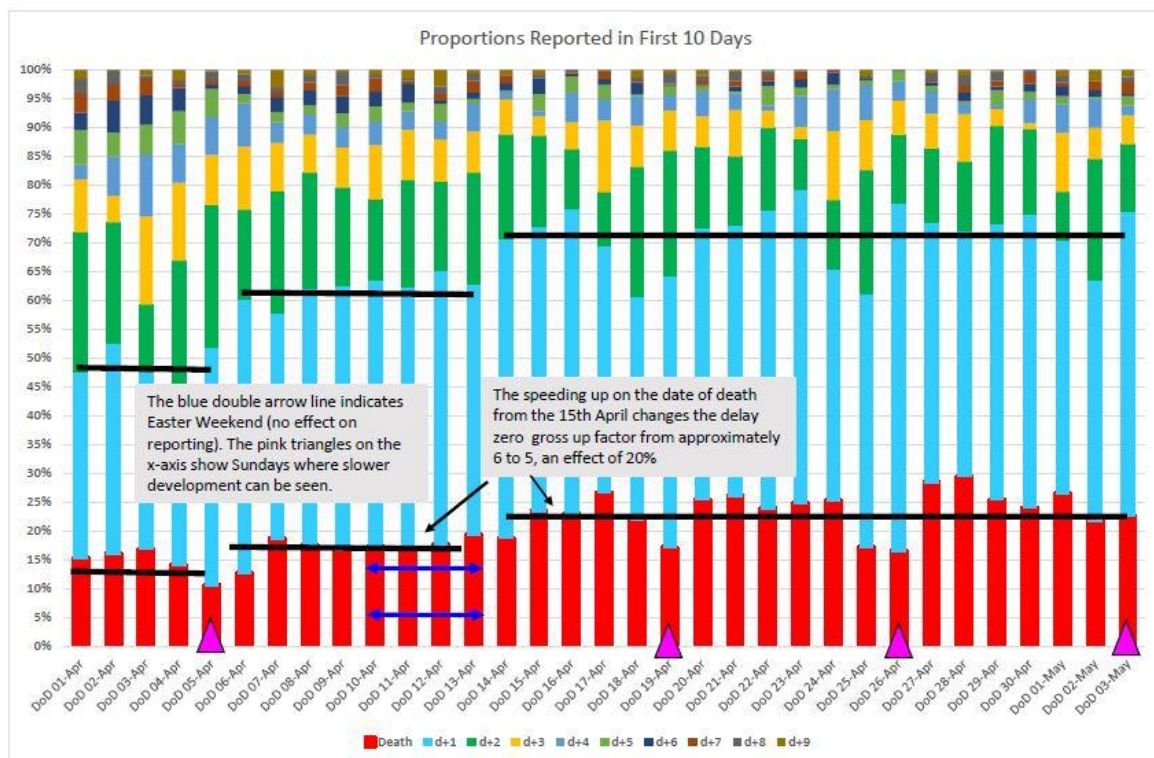
[3] Source: The data used in the above model was taken from NHS England Covid-19 Daily Deaths.

There are two clear stages of speeding up of reporting. The first from 6-14 April and the second after 14 April. This can be seen in chart 3.

Chart 3: Proportion reported by delay. Here the distribution of deaths reported in the first 10 days following death is pictured on a day by day basis. 100% represents the first 10 days. There is later reporting of deaths after 10 days – between a further 5% and 10%,but with limited data the first 10 days best shows speeding up effects. This highlights the significance of the weekend effect where death reporting lagged behind actual date of death, except for Easter which was staffed as a normal weekday. This also shows the speeding up of reporting from the 15 April which gross is a 20% effect.

Proportions Reported in First 10 Days

[3] Source: The data used in the above model was taken from NHS England Covid-19 Daily Deaths.

Results for the month of April are summarised in the table below along-side the reported numbers and the results of the earlier Christofides prediction. In both cases, a prediction could be provided by expert judgement or smoothing by eye outside of an automated process. The Christofides approach acknowledges limitations regarding weekend working patterns and speeding up of data recording. The automated approach in this paper adjusts for known features *after* they have been identified in the data, *or, there is a priori* information which could suggest an estimated adjustment (such as normal working during the Easter Weekend), since it tries to mimic a live environment.

Table 1: Deaths in April. Christofides et al. prediction was lower than actual for 1-2 April, but adequate overall for the period 1 April to 14 April. The automated approach assumes a longer development for deaths prior to 15 April (a death on 4 April due to COVID-19 was recorded on 16 May) and because of this, overestimated deaths as at 30 April. As the tail factor recedes the prediction reduces. The predictions for 15 April to 26 April (after the speeding up of notification) are now looking a little high. This margin carried forward to 27-30 April where both this prediction and the automated approach as at 30 April had not allowed for a "Friday catch up" which occurred on 1 May. This 'catch up' now seems to be a repeating feature. Overall there is approximately a 1% difference throughout the whole of April for total number of deaths predicted by both models as at 30 April.

Total Death Occurance Predictions	Full Month	1-14 April	15-26 April	27-30 April
Total reported 30 April	16,595	9,831	6,193	571
Christofides, Oke and Heneghan	18,162	10,233	6,895	1,044
Total reported 16 May	17,608	10,156	6,494	958
Automated Chain Ladder @ 30 April	17,957	10,395	6,629	934
Automated Chain Ladder @ 16 May	17,811	10,293	6,577	995

[4] Source: The data used in the above table was taken from NHS England Covid-19 Daily Deaths.

Conclusion

An automated Chain Ladder works and would predict peak deaths 2 days earlier than by date of reporting and with more certainty (due to volatility in daily reporting data).

The peak is now over, but such a tool would be useful for early signposting a second wave by region and/or picking up trends in lower death counts and changes of cause of deaths being registered.

It would be better to use all death data and not just COVID-19 deaths, then project separately by location of death (e.g. hospital, care-home, home). This could project deaths by major 'cause of death' thereby identifying trends in excess deaths earlier. Excess deaths is a better measure, because there was initially under-recording of COVID-19 deaths. Now there is greater awareness of COVID-19, there may well be reporting of COVID-19, particularly where there are a number of genuine COVID-19 deaths.

Furthermore, Individual COVID-19 only deaths are becoming more volatile at present due to lower volumes. A more robust count of all deaths would reduce parameter error. This, however, would require daily ONS data which is not currently available. This approach may also be useful to identify the extent to which COVID-19 accelerated deaths rather than caused death in its own-right. Lower deaths later in 2020 and in 2021 from causes such as dementia for example could be part of that measure.

[1] https://en.wikipedia.org/wiki/Chain-ladder_method

[2] https://www.cebm.net/covid-19/covid-19-the-chain-ladder-method-to-estimate-deaths-not-yet-reported/

[3] NHS England Covid-19 Daily Deaths: https://www.england.nhs.uk/statistics/statistical-work-areas/covid-19-daily-deaths/