Stockholms universitet

Evaluating Nowcasting Methods for COVID-19 Related Fatalities in Sweden

Markus Olofsson Lindroos

Masteruppsats i matematisk statistik
Master Thesis in Mathematical Statistics

Masteruppsats 2021:13
Matematisk statistik
Juni 2021

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Stockholm University

Mathematical Statistics
Stockholm University
Master Thesis **2021:13**
http://www.math.su.se

Evaluating Nowcasting Methods for COVID-19 Related Fatalities in Sweden

Markus Olofsson Lindroos*

June 2021

Abstract

When a fatality occurs due to COVID-19 there is a delay before the authorities receive a report of it in their database. The authorities then report the aggregated number of COVID-19 related fatalities to the public. The number of daily COVID-19 related fatalities for the most recent days close to today is hence only partially observed. For this reason, it is difficult to determine the current trend in daily fatality counts from the reported cases alone. Nowcasting is the task of inferring total counts based on partially observed data by extrapolating cases based on previous knowledge about the reporting delays. Nowcasting methods are applicable to predict the number of daily COVID-19 related fatalities in Sweden, and the results of one nowcasting method is published on the web site Altmejd et al. (2021a), which is often referred to. In this thesis, we aim to comparatively evaluate this nowcasting method with another recent nowcasting method of Günther et al. (2020a). To do so, we define and use proper scoring rules and other evaluative metrics, such as the coverage of 95% prediction intervals. The nowcasting methods can be evaluated in retrospect, since fatalities are rarely reported with a delay in excess of 1-2 months, and hence the true number of daily fatalities can be assumed to be known after 2 months. When applying the nowcasting methods for a "now" varying over 17 instances from February 2 until March 2, 2021, we find that our implementation of of Günther et al. (2020a) performs better than that published at Altmejd et al. (2021a) by almost all metrics.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: Markus Olofsson Lindroos. Supervisor: Michael Höhle.

Acknowledgements

I would like to thank my supervisor Michael Höhle for his helpful and insightful remarks and comments, curious encouragement and suggestions, and general support during the process of writing this thesis. I would also like to thank the nowcasting research group at Stockholm University consisting of Fanny Bergström, Tom Britton, Felix Günther and said supervisor for the opportunity to sit in on some of their meetings.

Contents

1 Introduction **4**
- 1.1 Outline of thesis . . . 6

2 Data **7**
- 2.1 Overview of data . . . 9
- 2.2 Delay distribution . . . 12

3 Nowcasting **15**
- 3.1 General nowcasting problem for count data . . . 15
  - 3.1.1 Structure of data: The reporting triangle . . . 17
- 3.2 Model of Günther et al. (2020a) . . . 19
  - 3.2.1 Discrete hazard regression model . . . 19
  - 3.2.2 Fatality counts model . . . 22
- 3.3 Method of Altmejd et al. (2021b) . . . 23
  - 3.3.1 Sampling step . . . 24
  - 3.3.2 Smoothing step . . . 25
  - 3.3.3 Deduction of predictive distribution . . . 26
- 3.4 Other points of view . . . 26
  - 3.4.1 Chain ladder approach . . . 27
  - 3.4.2 Removal method and multinomial mixture approach . . . 28

4 Evaluation **31**
- 4.1 Scoring rules . . . 32
  - 4.1.1 Squared and absolute error score . . . 33
  - 4.1.2 Log-score . . . 33
  - 4.1.3 Ranked probability score . . . 34
- 4.2 Other evaluative metrics . . . 35
  - 4.2.1 Error and relative error . . . 35
  - 4.2.2 Width and coverage of prediciton intervals . . . 36
- 4.3 Practical implementation of evaluation . . . 36

5 Results **38**
- 5.1 Implementation . . . 39
- 5.2 Main comparative results . . . 39
  - 5.2.1 Overview . . . 39
  - 5.2.2 Summary results of evaluation . . . 40
  - 5.2.3 Detailed results of evaluation . . . 44
- 5.3 Conclusion . . . 48

6 Discussion **50**
- 6.1 Possible improvements . . . 50
- 6.2 Closing remarks . . . 51

1 Introduction

The COVID-19 pandemic has had an immense impact across the world. In Sweden, at the time of writing this thesis, more than one million people have tested positive for the SARS-CoV-2 virus [FHM (2020-21)], which causes COVID-19; The true number of cases being greater, since not all afflicted by the virus and the disease it causes are tested. Furthermore, more than fourteen thousand individuals have perished with COVID-19 being listed as a contributing factor in their passing [Soc. (2021a)].

When a COVID-19 related fatality occurs, a death certificate is issued by a physician and sent to the Swedish Tax Agency (in Swedish: Skatteverket) at latest on the next business day after the fatality. The Tax Agency subsequently informs other government agencies about the fatality, such as the Public Health Agency (in Swedish: Folkhälsomyndigheten) and the National Board of Health and Welfare (in Swedish: Socialstyrelsen). At this point, these agencies do not know that the fatality is COVID-19 related. The Public Health Agency deduce whether or not it is COVID-19 related by checking if the deceased individual has tested positive for COVID-19 at most 30 days before their passing. The National Board of Health and Welfare waits until a physician has amended the death certificate with a cause of death, and check whether it includes COVID-19 as a contributing factor. This amendation should be done at most three weeks after the passing [Soc. (2021b)]. When the fatality is confirmed to be COVID-19 related, by each agency's own definition of confirmation, it is included in the time series for the number of daily fatalities by the date of their occurrence. The time series are published by each of these agencies and available to the public [FHM (2020-21), Soc. (2021a)]. The steps between the occurrence of the fatality and its subsequent reporting take time. We refer to this time as *reporting delay*.

Indicators relating to the current state of the pandemic, such as the time series for the daily number of fatalities, are of interest for policy makers, as to determine whether interventions should be implemented or relaxed. But because of the reporting delay, the number of COVID-19 related fatalities that occurred in the recent past is–partially–unknown, since a substantial proportion of them have not yet been reported. The task of *nowcasting* is to use the partially observed fatality counts as to estimate the true number of daily fatalities.

Nowcasting has a long history in actuarial sciences in the form of the chain-ladder method for estimating necessary claims reserves; Insurance claims are also subject to a reporting delay. Mack (1993) provided the first distribution free method for computing the variance of the necessary claims reserves estimate of the chain-ladder method. Lawless (1994) mention the long history of epidemiological nowcasting methods for estimating the number of cases of AIDS in countries, with Morgan and Curran (1986) being his earliest reference to such an application. Höhle and an der Heiden (2014) generalized the method of Lawless (1994) in a Bayesian nowcasting model, which, together with the model of McGough et al. (2020) served as the basis for the model of Günther et al. (2020a), an implementation of which will be presented in this thesis.

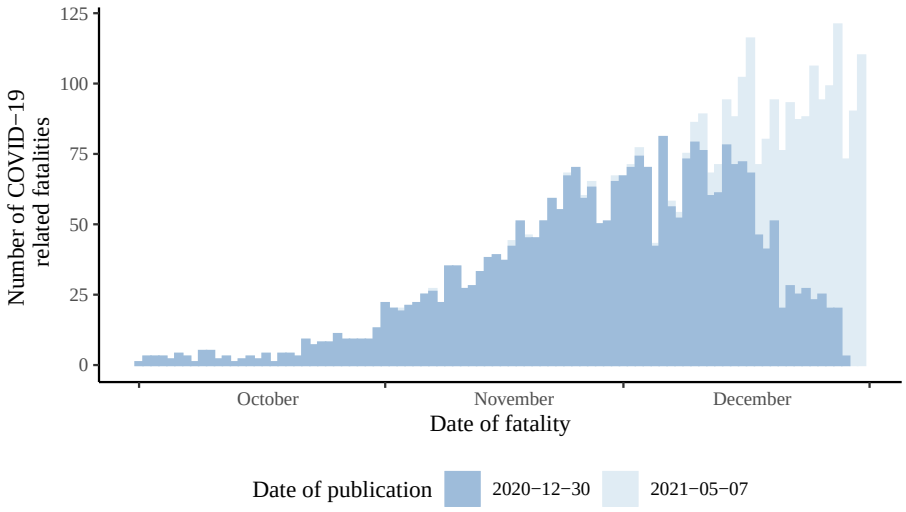

Figure 1: Daily number of COVID-19 related fatalities in Sweden during the last three months of 2020.

As an example of delayed reporting, we look at Figure 1. In this figure, we see the number of COVID-19 related fatalities that occurred in Sweden per day during the last three months of 2020, as known to us at two different dates: December 30, which was the last day of 2020 on which new data were published by the Public Health Agency of Sweden, and on May 7, 2021, which includes the latest data used for the completion of this thesis. On December 30, one is obviously only privy to the data reported until then, which corresponds to the dark blue columns of Figure 1. But these columns do not convey what is the most recent trend, as of December 30, in daily fatality counts, since the more recent counts are subject to delayed reporting. In particular, if one were completely unaware of the existence of a reporting delay, it would appear as though the number of daily fatalities are decreasing during the latter half of December. In retrospect, on May 7, 2021, we see that the number of daily fatalities were actually *increasing* during the latter half of December, which is the complete opposite conclusion.

Certainly, on December 30, the policy maker would benefit from an accurate *nowcast* of the more recent daily fatality counts for the latter half of December.

A web page, which has gained some attention, that provides informative illustrations of the fatality counts in conjunction with their reporting delay is that of Altmejd et al. (2021a). In one of their figures, they include the results of a nowcasting method, which is defined in terms of `R` computer code, and made publicly and readily available at Adam Altmejd's GitHub covid repository, Altmejd et al. (2021b).

Another recent nowcasting method is that of Günther et al. (2020a), which,

in their paper, is applied to nowcast the number of daily COVID-19 disease onsets in Bavaria during the first pandemic wave in the spring of 2020. We apply an appropriately modified version of this model as to nowcast the daily fatality counts in Sweden. We note that the translation from disease onsets to fatality counts demand no modification of the model, since both are of the count nature, and are subject to a reporting delay.

In retrospect, the true number of fatalities can be assumed to be known, since fatalities reported with delays in excess of 1-2 months are rare. This enables us to retrospectively determine how well a nowcasting method performs. In this thesis, we aim to comparatively evaluate the method of Altmejd et al. (2021b) with that of Günther et al. (2020a), using different evaluative metrics. We shall find that, when applied to the period between February 2 and March 2, 2021, the method of Günther et al. (2020a) outperforms that of Altmejd et al. (2021b) by almost every metric. The Altmejd et al. (2021b) method, however, seems to perform better during the last week-and-a-half of the considered period.

1.1 Outline of thesis

The outline of this thesis is as follows: In section 2 we detail how the COVID-19 fatality data in Sweden are obtained, and how the data structure is. We also provide an exploratory analysis, which focuses heavily on the distribution of the reporting delay; In section 3 we provide the reader with an introduction to the general subject of discrete time nowcasting for count data, and define the model of Günther et al. (2020a) and how we modify it to adapt to the structure of the Swedish data. Also, we attempt to describe the method of Altmejd et al. (2021b), but since this is defined in terms of `R` computer code only, the definition given is rather algorithmic. Finally, we give a very brief overview of other methods for nowcasting count data in discrete time; In section 4, we define the evaluative metrics we use, and briefly motivate their use. Importantly, we also define what constitutes proper scoring rules; In section 5, we present our results, and, in section 6, we provide a brief discussion.

2 Data

The data used in this thesis originate from the Public Health Agency of Sweden [FHM (2020-21)]. However, for reasons we will get to shortly, we obtain the data via a secondary source, Altmejd et al. (2021b). Since the advent of the COVID-19 pandemic presence in Sweden, the Public Health Agency has, on a near daily basis (see Tab. 2 below for details), published several indicators as to describe the dynamics of the pandemic and its impact in the population. They include time series for the daily number of newly confirmed cases, intensive care unit (ICU) admissions and fatalities by the date of their occurrence[1]. Cumulative equivalents of these indicators are also published stratified with respect to (either of) age, sex, region and municipality. The contents of the publications have varied throughout the course of the pandemic. More lately, data on the number of partially and fully vaccinated individuals have also been included. In this thesis, however, we are solely concerned with the number of fatalities. In Figure 2, we see the fatality time series illustrated, and compared to that of another source, National Board of Health and Welfare [Soc. (2021a)], which count COVID-19 related fatalities slightly differently (the specifics of which we will get to shortly).

The Public Health Agency does not, however, publish the dates on which fatalities were reported as such – only the dates of their occurrence. Consequently, the delay between a fatality and its reporting is not directly available from published data. This yields nowcasting models mostly inapplicable, since their inferences rely heavily on the reporting delay distribution. In its lieu we follow Altmejd et al. (2020) in considering the delay between the fatality and its inclusion in the published time series. This *publishing delay* – as it were – is obtainable from data, if one keeps records of the Public Health Agency's past publications, which are not readily available from FHM (2020-21). Altmejd et al. (2020) have providently kept such records, and made them available as a part of the GitHub repository Altmejd et al. (2021b). The data used in this thesis are thus collected from this repository.

In sections 2.1 and 2.2 we present an exploratory analysis of this data. The former section contains an overview of the fatality time series, and provides details on the Public Health Agency's publishing pattern. The latter section is devoted to the delay distribution. For the remainder of this section preamble, we define which fatalities are considered to be COVID-19 related according to the data, and explain how one may deduce the publishing delay from the act of keeping records of past publications.

According to the National Board of Health and Welfare Soc. (2021b), the Public Health Agency considers a fatality to be COVID-19 related if the deceased was confirmed to be infected with the novel coronavirus (SARS-CoV-2) by laboratory test, despite their actual cause of death. This may be reported either by the treating physician, or by comparing the national death registry with

[1] Newly confirmed cases are tallied by "statistical date" (in Swedish: statistikdatum), a term which appears limited to the jargon of the Public Health Agency (and related authorities). A more informative indicator, if tractable, would be the dates of e.g. disease onset.

the registry of confirmed coronavirus infections, counting fatalities occurred at most 30 days after infection confirmation to be COVID-19 related. In both cases, but particularly the latter, fatalities may retrospectively be de-classified as COVID-19 related, when the cause of death is determined to be another, e.g. a traffic accident. Once such information becomes available, they are removed from the fatality time series. The Public Health Agency, however, notes that this occurs only in a limited number of cases. The proportion of fatalities reported by physicians, and the frequency at which the death registry is compared to the registry of confirmed infections are not readily available from the data documentation. We should also note that the Public Health Agency, the National Board of Health and Welfare, and the regions of Sweden may publish slightly differing fatality counts [FHM (2020-21), Soc. (2021b)]. The reasons for this are somewhat ambiguously put, however: FHM (2020-21) refer to Soc. (2021b) relating to the definition of what constitutes a COVID-19 related fatality, yet they use different definition. Generally, the Public Health Agency relies on laboratory test cross-referencing more, while the National Board of Health and Welfare use the cause of death as per the death certificate [Soc. (2021b)]. A third, possible, measure of the number of COVID-19 related fatalities is the excess mortality, computed as the difference between the actual number of fatalities and a base-line corresponding to the average number of fatalities during years with relatively low numbers influenza related fatalities. In a study conducted by the Public Health Agency, these three measures are compared, yielding the conclusion that they follow the same trend, and produce fatality counts close to one another [FHM (2020a)]. They also point out that the testing capacity for COVID-19 was limited during the first wave of the pandemic during the spring of 2020, whence the National Board of Health and Welfare, which go by the cause of death as per the death certificate, counted a greater number of fatalities during this period, while later in the course of the pandemic, they count fewer. This is seen in Figure 2.

In Figure 2, we see a comparison of the Public Health Agency's and National Board of Health and Welfare's counts of daily COVID-19 related fatalities.

One may deduce the reporting delay by keeping track of past publications, simply by computing the difference in fatality counts of subsequent publications on corresponding dates. For instance, consider the number of fatalities occurred on April 15, 2020: On April 15, the Public Health Agency counted 6 fatalities on this date, which hence were reported on the day of their occurrence, having a delay of zero days; On April 16, the Public Health Agency counted 41 fatalities for April 15, whence $41 - 6 = 35$ fatalities were reported with a delay of one day; On April 17, 45 fatalities were reported yielding $45 - 41 = 4$ fatalities with a delay of two days; and so forth. In Table 1, we illustrate this calculations for fatalities occurred on, and reported before, April 14-17, 2020.

However, due to the fact that fatalities previously classified as COVID-19 related may be de-classified as such, this technique will not yield an exact count of the number of fatalities reported with a certain delay. On April 16, for instance, the Public Health Agency might have added 40 fatalities to the time series, while removing 5, yielding us to falsely assume that 35 fatalities were re-

ported with a delay of one day. Also, there are instances where the difference in fatality counts between subsequent publications on corresponding dates is *negative*, due to the de-classifications, inducing *observable* de-classifications. These instances are however of low order relative to the fatality counts, and rather rare for later fatality dates: As is seen in Figure 3, the greatest number of observable de-classifications is 5 on April 6, 2020, a day on which 90 COVID-19 related fatalities occurred (as known by May 7, 2021). Note also that observable de-classification are reported with a particularly long delay.

2.1 Overview of data

The Public Health Agency has published the time series for the daily number of fatalities since April 2, 2020. All past publications, except that for July 30, 2020, are available at Adam Altmejd's GitHub, whence they may be downloaded by cloning his `covid` repository [Altmejd et al. (2021b)]. Each publication consists of an Excel file (i.e. a file with file extension `.xlsx`), in which each Excel-sheet contains data corresponding to either of the indicators mentioned in the preamble of this section. Additionally, they contain a sheet stating when data are tallied, and when they are published. From this we find that the publishing pattern of the Public Health Agency has varied throughout the course of the pandemic, becoming increasingly sparse. We distinguish 4 distinct publishing patterns, outlined in Table 2.

Consequently, the difference between the internal reporting delay observed by the Public Health Agency and the *publishing delay* observed by the public was relatively minute during period I, with fatalities being included in the time

		Reporting day			
		April 14	April 15	April 16	April 17
Fatality day	April 14	5	31	49	60
	new	5	$31-5=26$	$49-31=18$	$60-49=11$
	April 15		6	41	45
	new		6	$41-6=35$	$45-41=4$
	April 16			10	38
	new			10	$38-10=28$
	April 17				4
	new				4

Table 1: Number of daily fatalities occurred on April 14-17, 2020, as reported on April 14-17, 2020. The black numbers indicate the cumulative number of fatalities for each calendar day, the form at which the Public Health Agency present their data, while the grey numbers indicate the number of *newly* reported fatalities, i.e. those that were reported "today". Gray entries on the main diagonal correspond to fatalities reported with zero days delay, on the first superdiagonal one day delay, on the second two days delay, and the top right grey entry was reported with a delay of three days.

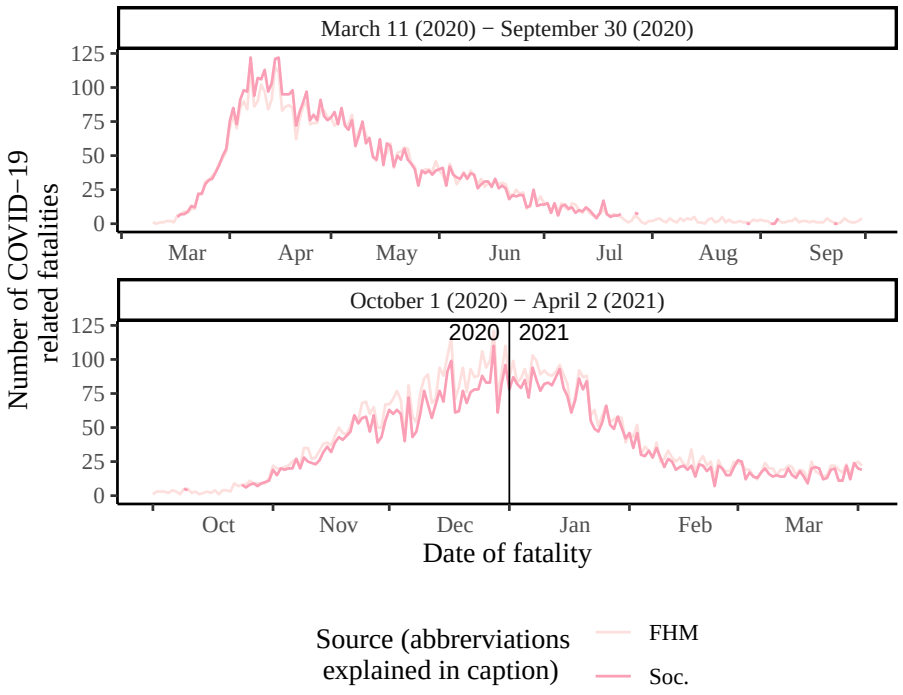

Figure 2: Time series of daily COVID-19 related fatalities as counted by the Public Health Agency (Folkhälsomyndigheten, FHM) [FHM (2020-21)], and by the National Board of Health and Welfare (Socialstyrelsen, Soc.) [Soc. (2021a)] as reported by 7 May, 2021.

series at most one day after their internal reporting to the Public Health Agency. During periods II, III and IV, the publication may occur at most three or four days, respectively, after the internal report. That is, when the internal report occurs on a Friday, it will not be published until Monday or Tuesday, inducing an additional delay of three or four days. We shall see in section 2.2 that this induces a peculiar publishing delay distribution.

The reason for the sparsification of publishing occations is that laboratory SARS CoV-2 test results are reported to the Public Health Agency particularly slowly during the weekend [FHM (2020b)]. This effects the fatality time series through the additions due to comparing laboratory tests with the death registry. One could imagine that fatalities reported by physicians are also incurred with a longer delay during weekends, but this is not explicitly given as a reason for the publication sparsificatoin. At any case – the number of fatalities added to the time series were fewer during weekends and Mondays during time periods I, II and III, as is illustrated in Figure 4.

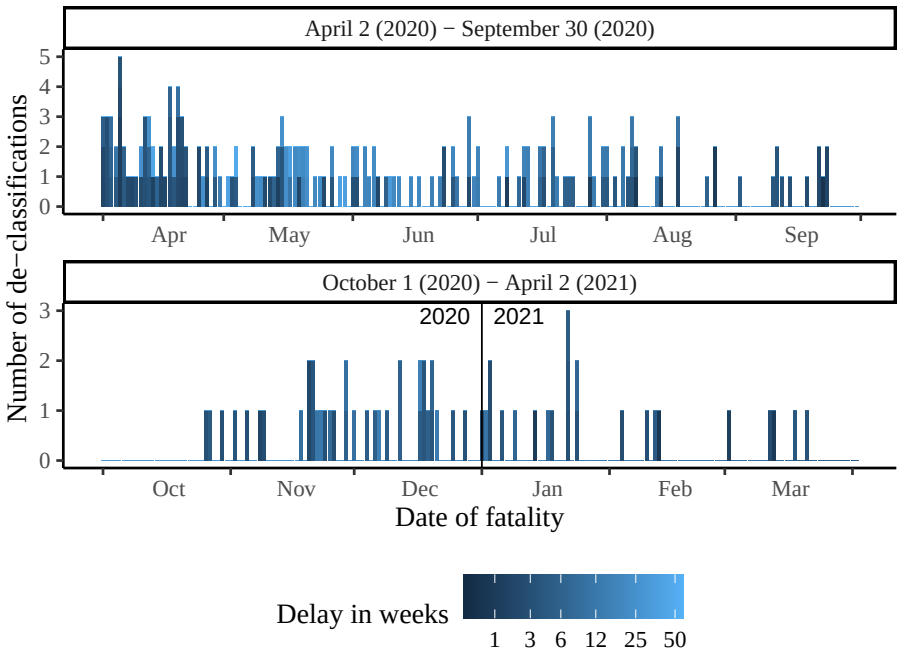

Figure 3: Number of observed instances of fatalities being de-classified as being COVID-19 related.

I **April 2 - June 14 (2020)**: Daily publications using data as of 11:30 the same day.

II **June 15 - August 10 (2020)**: Publications on business days (Monday - Friday) using data as of 11:30 the same day. No publication was made on June 19, as it is a public holiday in Sweden (Midsummer's Eve). Also, July 30 is missing.

III **August 11 - September 13 (2020)**: Publications on business days using data as of the previous day. (Supposedly as of 24:00 the previous day, but no explicit tallying time is stated.)

IV **September 14 (2020) - present (May 7, 2021)**: Publications on business days except Mondays using data as of the previous day. No publications were made on the various public holidays around Christmas and New Year.

Table 2: Periods of different publication patterns of the Public Health Agency. The date breaks correspond to Monday - Sunday when applicable.

2.2 Delay distribution

While the main quantity of nowcasting is the most recent fatality counts, the type of nowcasting considered in this thesis relies heavily on the distribution of the *reporting delay*, yielding the latter an ubiquity of nowcasting. In this section, we make two main points about the delay distribution, as observed retrospectively: That it is not stationary with respect to time, and that it is highly dependent on the weekday on which the fatality occurred, owing to the uneven reporting distribution with respect to weekday (cf. Fig. 4). Both of these effects relate to the fact that the publication schedule has changed over time, but neither is solely due to this fact. As an example of the latter phenomenon, we look at Figure 5, where the empirical probability mass functions of the delay distribution is illustrated stratified with respect to the weekday of the fatality, but shifted such that they are aligned with respect to the weekday of the report. Two time periods are compared. Clearly, the probability masses align rather neatly for both time periods, but only the latter is effected by the sparser publication schedule. This plot is also illustrative of the fact that the delay distribution changes more generally: During the earlier period most reports occurred during the first and second week after the fatality, but during the later period, the third and fourth weeks had rather many reports, too.

Generally, the reporting delay seems to have become progressively longer during the course of the pandemic. In Figure 6, we see the number of fatalities reported for each combination of date and delay. Here, too, one sees that fatalities occurred during the first half of 2020 were generally reported with a shorter delay than those during the second half of 2020 and the first half of 2021. The behaviour of the median and 95th percentile (computed over a rolling 21-day period) is also indicative of this. Note that fewer fatalities occurred during July until September, 2020, whence the 95th percentile is more "wobbly" during this period (being computed from less data).

The diagonal high-frequency report streaks apparent in Figure 6 are due to reporting being less frequent during the weekend and on Mondays, and more frequent on other weekdays. This should be compared to the publication periods of Table 2, and the consequent frequency of reporting on certain weekdays illustrated in Figure 4. In particular, since June 15, 2020, no reports occur on the weekend, yielding while diagonal streaks in Figure 6.

A common practise in nowcasting is that one assumes a *maximum delay*. We will give some reasons for this in section 3.1.1; An immediately visible reason is the sparsity of data for longer delays, as is seen in Figure 6, where we also see delay for which at least 95% of fatalities were reported. In Figure 7, we illustrate the proportion of fatalities not yet reported 1, 2, 3, 4, 5 and 6 weeks after the date of their occurrence, for fatalities reported before May 7, 2021 (which is 5 weeks after April 2, 2021), computed for a 21-day rolling period. Clearly, there is some variation, but for the entire period, approximately (and frequently less than) 2% of fatalities were reported with a delay longer than 5 weeks.

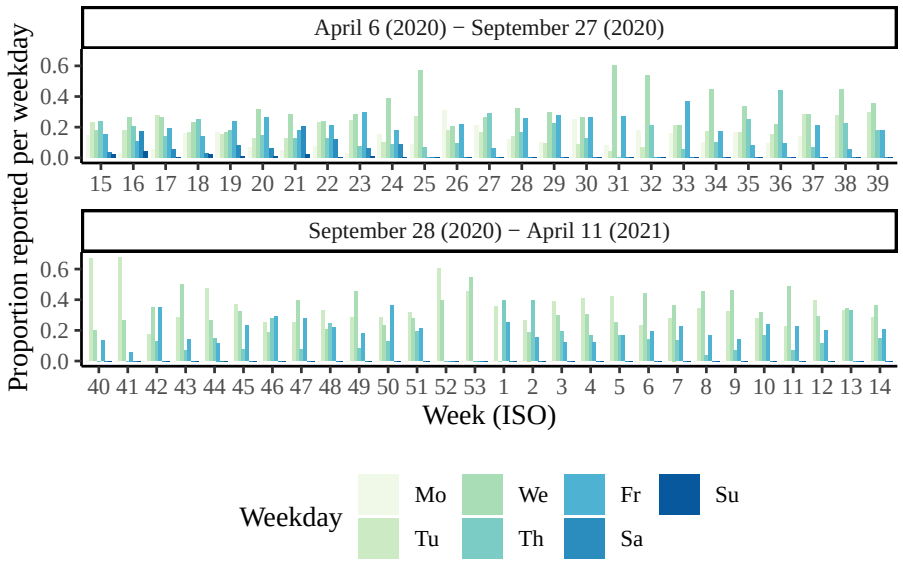

Figure 4: Frequency histograms for the number of fatalities reported on each weekday for different weeks.

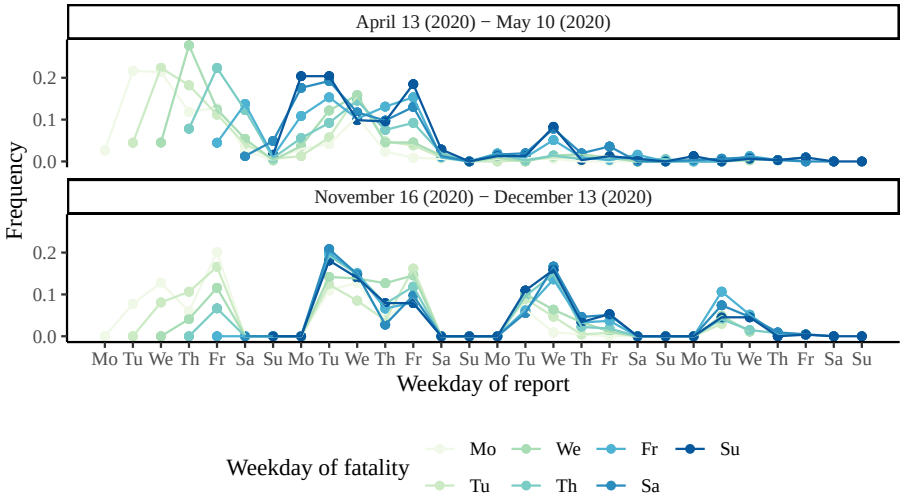

Figure 5: Empirical probability mass functions for the delay of two four week periods during the first and second pandemic waves. The distributions have been shifted in the x-axis direction such that the different distributions are aligned with respect to the weekday of the report. In particular, the probability mass function for the ith day of the week is shifted $i - 1$ steps to the right.

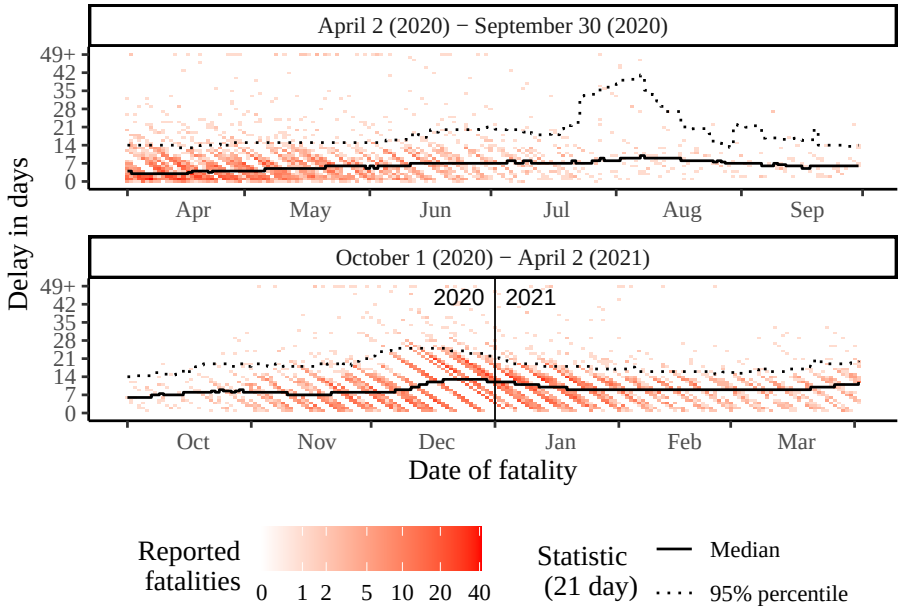

Figure 6: Number of fatality reports for each combination of fatality date and reporting delay, where 49+ signifies a delay of 49 days or more, and median and 95th percentiles for a rolling 21-day period around each fatality date.

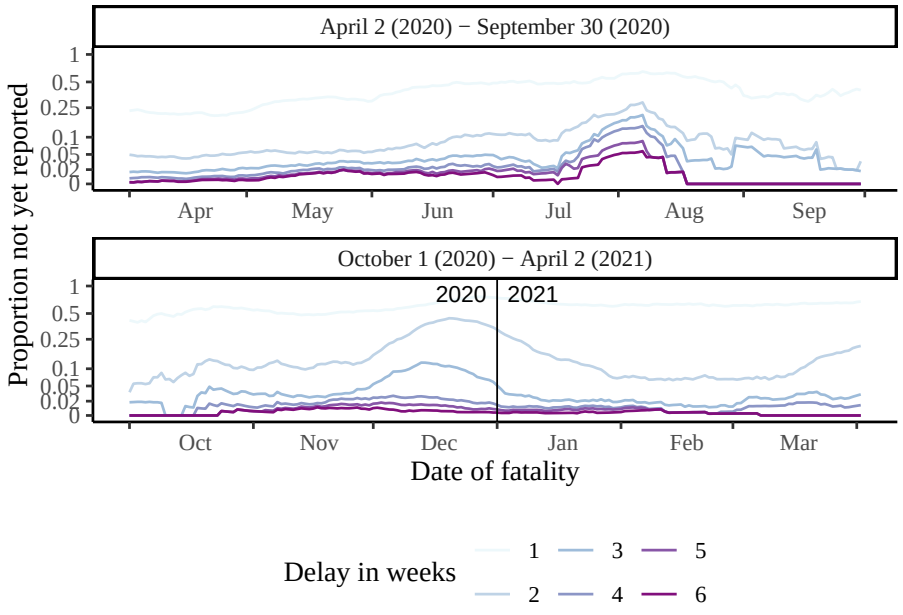

Figure 7: Proportion of fatalities that had not yet been reported after a delay of x weeks, for fatalities reported before May 7, 2021, for a rolling 21-day period.

3 Nowcasting

While the titular topic of this thesis is the *evaluation* of nowcasting methods, the underlying methods themselves and the further underlying framework of nowcasting on which they stand require an equally–if not even more–thorough treatise. In this section, we do that. First, in section 3.1, we provide an introduction to the "general" subject of nowcasting, noting that we confine generality to count data reported on a discrete time scale. In particular, we introduce useful notation that will be used throughout the thesis, define the structure of our data, and clarify how this structure poses a mathematical problem well suited for statistical model solutions. The notation used large follows Lawless (1994) with some minor modifications. We reiterate the nowcasting model of Günther et al. (2020a) in section 3.2, noting some minor modifications as to make it (more precisely) applicable to the Swedish fatality data. Next, we attempt to define the method of Altmejd et al. (2021b), noting that this is an approximate description extrapolated from code rather than from a scientific article; The paper, Altmejd et al. (2020), describes a model different from the one implemented in the code, Altmejd et al. (2021b). Lastly, we provide a brief overview of other points of view of the nowcasting problem in section 3.4. In particular, we reiterate the novel link to the removal method from ecology posed by Altmejd et al. (2020), note its similarities with the more classical approach for nowcasting count data of Lawless (1994), and briefly present the chain ladder method stemmed from actuarial sciences.

But what is nowcasting? Until now, we have largely relied on the reader to extrapolate its meaning and purpose from the etymological root of the word *nowcasting*, which is a portmanteau of the words *now* and *forecasting*. Clearly, this refers to predicting (or "forecasting") the present state ("now"). The set of situations for which nowcasting proves useful and necessary is thereof also implied: When the present state is not fully observed, due to some reporting delay. It has been pointed out [Höhle and an der Heiden (2014), McGough et al. (2020), Günther et al. (2020a), Altmejd et al. (2020)], and we have seen (in Fig. 1), that the incidence of e.g. disease onsets and–in our case–fatalities may be subject to reporting delays and thus require nowcasting. The more general subject of nowcasting–however–is far more general, including applications in meteorology, economics and actuarial science. Altmejd et al. (2020) further provides a link to statistical methods from ecology, but the use of the word *nowcasting* seems to be confined to the other mentioned fields.

3.1 General nowcasting problem for count data

In this section, we will describe the general discrete-time nowcasting problem for count data, but we choose to use a phrasing only consistent with the nowcasting of fatality counts subject to daily reporting delays, i.e. we refer to the discrete-time increments as *days* and the counts that are reported as *fatalities*, noting that these could be replaced with e.g. weeks and disease onsets, respectively, without inducing any changes in the mathematical description of the problem.

We are chiefly concerned with two closely related quantities, and their "parametric counterparts", as it were:

- The daily number of fatalities, which we denote by

$$N_t = \#\text{ fatalities occurred on day } t.$$

  This is the quantity we would ultimately like to estimate for $t = 1, 2, \ldots, T$, where T represents the latest day for which data are available, i.e. T corresponds to "now" or "today". Obviously, T is known.

- The parametric counterpart of N_t is its mean, which we define by

$$\lambda_t = \mathbb{E}[N_t]\,.$$

  This parameter is particularly important in epidemiological nowcasting, since one generally assumes there to be a dependence between the number of fatalities of subsequent days, i.e. $\ldots, N_{t-1}, N_t, N_{t+1}, \ldots$. This is particularly true of infectious diseases, since fatalities thereof depend on the number of cases in the population, which in turn influence the number of new infections, which constitutes the number of future cases, some of which cause fatalities, and so forth. In section 3.2.2, we define a model for λ_t such that the dependence of the sequence $N_1, N_2, \ldots, N_T$ is taken into account in the modelling. That is, we let the dependence of $N_1, \ldots, N_T$ be carried by the sequence of its means $\lambda_1, \ldots, \lambda_T$. We note that, at least, Höhle and an der Heiden (2014) also include a dependency structure, while not modelling for an *infectious* disease.

- The number of fatalities occurred with certain delays, which we denote by

$$n_{td} = \#\text{ fatalities occurred on day } t \text{ and reported on day } t + d,$$

  i.e. n_{td} fatalities were reported with a delay of d days out of the N_t that occurred on day t. The relationship between N_t and the set $\{n_{td}, d \geq 0\}$ is, of course, that the former is the sum of the latter over $d \geq 0$, as defined in equation (3) in the next subsection. The attentive reader may relate n_{td} to the number of "newly reported" fatalities for day t reported on day $t + d$, corresponding to the grey entries of Table 1 in section 2. But recall that, due to the possibility of fatalities being de-classified as such, the "newly reported" fatality count may actually be negative. The modelling framework we use in this thesis does not allow for this, whence we opt to define $n_{td} \geq 0$, mapping negative "newly reported" fatality counts to zero.

- The parametric counterpart governing n_{td} is the distribution of the reporting delay. We will quantify this by the probability mass function, which we denote as

$$p_{td} = \mathbb{P}(\text{delay} = d \,|\, \text{fatality occurred on day } t).$$

  As such, we allow for the possibility of the reporting delay to vary with respect to the time of the fatality.

This representation immediately conveys a mathematical problem suitably solved by statistical modelling, since we assume the delays to stem from a probability distribution determined by p_{td}. However, N_t may actually be assumed to be a scalar integer-valued parameter (e.g. as in sec. 3.4.2).

Another notational tool that we use in sections 3.3 and 4.2, but not in our main treatise, is the number of fatalities occurred on day t and reported on or before day $s \geq t$. We write this as

$$N_t(s) = \sum_{d=0}^{s-t} n_{td}. \tag{1}$$

In particular, $N_t = N_t(\infty) = N_t(t+D)$, where the latter equality comes from the assumption that there is a maximum delay D. We will introduce and motivate this assumption in the next subsection.

3.1.1 Structure of data: The reporting triangle

Recall that T denotes the latest day for which data are available, i.e. "now" or "today". The essential problem of nowcasting is that fatalities have *occurred* for $t = 1, 2, \ldots, T$ (where $t = 1$ denotes the first day at which a fatality was occurred due to the disease and outbreak in question), but we do not know their number, since their reporting to us is subject to delays. Consequently, at time T, we do not observe N_t directly, but rather, we observe reports of it at times $t, t+1, t+2, \ldots, T-1, T$, i.e. relative to time t, with delays $d = 0, 1, 2, \ldots, T-1, T-t$. As previously mentioned, we let n_{td} denote the number of fatalities occurred on day t and reported with a delay of d days. At time T, we observe the set of n_{td} such that

$$\{n_{td} : t + d \leq T\}, \tag{2}$$

with which the task of nowcasting is to infer

$$N_t = \sum_{d=0}^{\infty} n_{td} = \underbrace{\sum_{d=0}^{T-t} n_{td}}_{\text{Observed.}} + \underbrace{\sum_{d=T-t+1}^{\infty} n_{td}}_{\text{Unobserved.}}. \tag{3}$$

An illustration of the structure of data is present in Table 3. This should be compared to the example representation of the Swedish fatality data in Table 1 in section 2. The grey "new" entries in Table 1 correspond to n_{td}, but they are ordered column-wise by the reporting day $t + d$ instead of the delay (as in Table 3), whence the columns of Table 3 corresponds to the diagonals of Table 1. In the chain ladder approach to nowcasting, one usually uses the partial sums $N_t(t+d) = \sum_{i=0}^{d} n_{ti}$ in stead of n_{td} to represent the data [Mack (1993)], but otherwise the structure is identical.

A practice used by all epidemiological nowcasting methods encountered by the author of this thesis in its preparation [Lawless (1994), Höhle and an der

		Reporting delay d					
		$d=0$	$d=1$	$\dots$	$\dots$	$d=T-1$	$d=T$
Fatality day t	$t=1$	n_{10}	n_{11}	$\dots$	$\dots$	$n_{1,T-1}$	$n_{1,T}$
	$t=2$	n_{20}	n_{21}	$\dots$	$\dots$	$n_{2,T-1}$	
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\iddots$		
	$\vdots$	$\vdots$	$\vdots$	$\iddots$			
	$t=T-1$	$n_{T-1,0}$	$n_{T-1,1}$				
	$t=T$	n_{T0}					

Table 3: Structure of data as defined by equation (2). The empty lower right triangle of this table correspond to indices of t and d which are yet to be observed, i.e. such that $t+d>T$.

Heiden (2014), McGough et al. (2020), Günther et al. (2020a), Altmejd et al. (2020)] is to introduce a *maximum delay*. The basic reason for doing so is that fatalities with longer delays are generally more sparsely distributed, thus yielding it impossible for them to be "estimated reliably". [Höhle and an der Heiden (2014)]. We let D denote the fixated maximum delay. Obviously, one chooses D such that fatalities with a delay greater than d make up a negligible proportion of the total number of fatalities, i.e. $\mathbb{P}(\text{delay} > D) \approx 0$. Subsequently, the number of fatalities at day t becomes definable by a finite sum: $N_t = \sum_{d=0}^{D} n_{td}$, which itself simplifies the modelling.

The reporting triangle will also get a different structure once a maximum delay has been introduced. This is illustrated in Table 4. There is an open question in how to define n_{tD}, however: Should index D signify delays of D days or longer, or simply delays of D days? In the former case, one might include fatalities with observed delays greater than D days in the fatality counts n_{tD}, i.e. $n_{tD} = \sum_{d=D}^{T-t+1} n'_{td}$, where n'_{td} is the observed count without a maximum delay, i.e. the "raw" counts observed by the nowcaster. But this would yield n_{tD} to have a different meaning for different t, whence we choose to define it as $n_{td} = n'_{td}$ for $d \leq D$, and $n_{td} = 0$ for $d > D$, thus assuming n'_{td} to be irrelevant for $d > D$.

A second modification of data included in several nowcasting models (e.g. Höhle and an der Heiden (2014), McGough et al. (2020)) is to introduce a "moving window" over the temporal index t. This means that one only includes data occurred at most m days before now. Thus, the upper rectangle of Table 4 would be excluded from the data for $t < T-m$. A reason for including a moving window is that the delay distribution changes over time, and the reporting delay of very distant fatalities might not be considered informative about the current delay distribution. Introducing a moving window assumes the information contained by $\{n_{td}, t < T-m\}$ to be completely depreciated. Another reason is that the practical implementation of some Bayesian nowcasting models (e.g. the one of Günther et al. (2020a) presented in section 3.2) demands computing power proportional to the dimension of data. These two reasons in conjunction motivate the inclusion of a moving window.

		Reporting delay d					
		$d=0$	$d=1$	$\dots$	$\dots$	$d=D-1$	$d=D$
Fatality day t	$t=1$	n_{10}	n_{11}	$\dots$	$\dots$	$n_{1,D-1}$	$n_{1,D}$
	$t=2$	n_{20}	n_{21}	$\dots$	$\dots$	$n_{2,D-1}$	$n_{1,D}$
	$\vdots$	$\vdots$	$\vdots$	$\ddots$		$\vdots$	$\vdots$
	$t=T-D$	$n_{T-D,0}$	$n_{T-D,1}$			$n_{T-D,D-1}$	$n_{T-D,D}$
	$t=T-D+1$	$n_{T-D+1,0}$	$n_{T-D+1,1}$			$n_{T-D+1,D-1}$	
	$\vdots$	$\vdots$	$\vdots$		$\iddots$		
	$\vdots$	$\vdots$	$\vdots$	$\iddots$			
	$t=T-1$	$n_{T-1,0}$	$n_{T-1,1}$				
	$t=T$	n_{T0}					

Table 4: Structure of data when modified as to include a maximum delay. The upper rectangle of this table corresponding to $t \leq T-D$ is assumed to be fully observed, while the lower rectangle (corresponding $t \geq T-D$) has the same structure as in Table 3.

In the application of Günther et al. (2020a) (see sec. 3.2), we choose $D = 7 \cdot 5 = 35$ and $m = 7 \cdot 10 = 70$, while Altmejd et al. (2020) choose $D = 25$; It being unclear whether or not they use a moving window. During the period from April 2 until December 1, 2020, 1.00% of fatalities were reported with a delay longer than 35 days, while 1.91% were reported with a delay longer than 25 days (as observed by February 2, 2021, which is the earliest "now" with which we perform nowcasting in section 5, cf. also Fig. 7).

3.2 Model of Günther et al. (2020a)

In this section, we present the model of Günther et al. (2020a), which constitutes a hierarchical Bayesian model. As was mentioned in section 1, this model was initially defined as to nowcast the number of daily COVID-19 disease onsets in Bavaria. In section 3.2.1 we introduce the discrete hazard regression model, which will be our only tool for testing different variations of the model on the Swedish data. This model is used to estimate the probabilities p_{td} for $t = 1, 2, \dots T$ and $d = 0, 1 \dots, D$. We will also introduce a small modification in considering days with zero reporting (cf. Tab. 2), but also note that this could be handled approximately but arbitrarily well by the original model. In section 3.2.2, we define the model for the fatality counts.

3.2.1 Discrete hazard regression model

The discrete hazard function is defined by

$$h_{td} = \mathbb{P}(\text{delay} = d \,|\, \text{delay} \geq d).$$

One may equivalently define the delay distribution in terms of the discrete hazard function, since the former is uniquely determined by the latter: Noting that, for $d \geq 1$,

$$\begin{aligned} p_{td} &= \mathbb{P}(\text{delay} = d \,|\, \text{delay} \geq d, t)\, \mathbb{P}(\text{delay} \geq d \,|\, t) \\ &= h_{td} \sum_{d'=d}^{\infty} p_{td'} = h_{td} \left(1 - \sum_{d'=0}^{d-1} p_{td'}\right), \end{aligned} \tag{4}$$

and that $p_{t0} = h_{t0}$ we see that the delay distribution in terms of p_{td} can be obtained from the set of discrete hazard functions h_{td} in an iterative fashion.

Note that $h_{td} \in (0, 1)$ for $d = 0, 1, \ldots, D-1$. In other words, h_{td} is not constrained by $h_{td'}$ for $d' \neq d$, as opposed to p_{td} which necessarily sum to one over d. Hence, the discrete hazard function is well suited to be modelled by linear regression using the log-odds (or logit) link. The use of discrete hazard modelling in epidemiological nowcasting originates from Höhle and an der Heiden (2014). We let the linear predictor element of the model be denoted by

$$g_{td} = \gamma_d + \boldsymbol{W}_{td}^{\mathrm{T}} \boldsymbol{\eta}, \tag{5}$$

where γ_d is a delay-specific intercept, $\boldsymbol{W}_{td}$ is a vector of k covariate effects depending on time and delay, and $\boldsymbol{\eta}$ is the corresponding k-length vector of coefficients. In equation (8), we elect to use a different representation of (5) consistent with the covariate effects we actually use; This representation is less general, though, and equation (5) suffices to clearly define the regression element of the model.

Günther et al. (2020a) define $h_{td} = \text{logit}^{-1}(g_{td})$ for $d = 0, 1, \ldots, D-1$ and $h_{tD} = 1$, the latter being due to the assumption of a maximum delay D inducing all "remaining" fatalities to be reported on day $t + D$. But a particularity of the Swedish fatality data is that, for more recent periods (cf. Tab 2 periods II-IV), fatality counts are not reported every day. For non-reporting days, the probability of reporting should be unequivocally *zero*, but since g_{td} is a real number, $h_{td} = \text{logit}^{-1}(g_{td})$ maps to the open interval $(0, 1)$, i.e. $h_{td} > 0$ which implies that $p_{td} > 0$ for all $d \in \{0, 1, \ldots, D-1\}$. To remedy this, we simply introduce an indicator representing whether or not day $t + d$ is a reporting day:

$$R_{td} = \mathbb{I}_{\{t+d \text{ is a reporting day}\}}. \tag{6}$$

We include this in the model for h_{td} such that it fixates certain discrete hazard probabilities to zero:

$$h_{td} = \begin{cases} R_{td} \cdot \text{logit}^{-1}(g_{td}) & \text{for } d \in \{0, 1, \ldots, D-1\} \\ 1 & \text{for } d = D \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

We note that the definition $h_{td} = 1$ for $d = D$ ignores whether or not $t + d$ is a reporting day, but in order for a maximum delay D to exist, the discrete

hazard must evaluate to 1 for some $d \leq D$. A more careful implementation would be to define $h_{td} = 1$ for the greatest $d \leq D$ such that $t + d$ is a reporting day. However, since the maximum delay is supposed to be chosen such that fatalities are sparsely reported with a delay of D, we deem the influence of a more carefully chosen time-dependent maximum delay to be negligible, if one chooses D large enough.

In applying this model to the Swedish data, we use one of the suggested covariate effects used by Günther et al. (2020a)–namely–an effect of the weekday of the reporting day. Considering only period IV of Table 2, this regression element takes the form

$$\begin{aligned} g_{td} = \gamma_d &+ \eta_1 \mathbb{I}_{\{t+d \text{ is a Wednesday}\}} \\ &+ \eta_2 \mathbb{I}_{\{t+d \text{ is a Thursday}\}} \\ &+ \eta_3 \mathbb{I}_{\{t+d \text{ is a Friday}\}}. \end{aligned} \tag{8}$$

Using this representation, it is clear to see how one could approximately include the domain knowledge of there existing days with no reporting that we chose to quantify using an indicator R_{td}. Suppose equation (8) included a term $\eta_0 \mathbb{I}_{\{t+d \text{ is not a reporting day}\}}$. Choosing a suitable prior on η_0, e.g. $\eta_0 \sim \mathcal{N}(-C, 0.01)$, where $C \to \infty$ would induce $p_{td} = \text{logit}^{-1}(g_t d) \to 0$ for $t + d$ being a non-reporting day without the necessity for an indicator R_{td}. In practise, one would simply choose a "large" C, since software used for sampling from the posterior (and predictive) distribution does not generally allow for infinite parameters (such is the case for e.g. `Stan` [Stan Development Team (2019a)]).

We use an informative prior on γ_d, such that

$$\gamma_d \sim \mathcal{N}\Big(\text{logit}(\tilde{h}_d), 1\Big), \tag{9}$$

where $\tilde{h}_d$ is obtained by $\tilde{h}_d = \tilde{p}_d / (\sum_{d'=d}^{D} \tilde{p}_{d'})$ (equivalent to the discrete hazard function defined by $\tilde{p}_d$), and

$$\tilde{p}_d = \frac{1 + \sum_{t=T-m}^{T-D} n_{td}}{D + \sum_{t=T-m}^{T-D} N_t}, \tag{10}$$

i.e. $\tilde{p}_d$ is approximately the empirical delay as obtained by from the upper fully observed rectangle spanned by $t = T - m, T - D$ and $d = 0, \ldots, D$ (cf. Tab. 4), but where each delay has been given an additional fatality in order to insure that $\tilde{p}_d > 0$. This is because $\tilde{p}_d = 0$ would imply that $\tilde{h}_d = 0$ which induces $\text{logit}(\tilde{h}_d)$ to evaluate to negative infinity, which is (as just mentioned) practically impossible in implementation.

The use of an informative prior for γ_d is not consistent with Günther et al. (2020a), but it has been previously suggested in epidemiological nowcasting literature (McGough et al. (2020) *suggest* it, but do not use it in their main analysis); Our choice of doing so is however mainly motivated by the fact that we would like to–somehow–include the domain knowledge that fatalities are

generally reported "sooner rather than later", as it were, consistently with our fixation of maximum delay D. The easiest way to do this is to simply use an informative prior based on observed data.

For the regression coefficients, we let $\eta_i \sim \mathcal{N}(0, 0.25)$ consistently with the parametrization of Günther et al. (2020a) (formally retrieved from their code at Günther (2020b)).

We should note that the choice of prior for γ_d in equation (9) is not entirely consistent with the regression model used in equation 8; Since we assume the base level with respect to weekday to be Tuesday, i.e. we assume that $g_{td} = \gamma_d$ on Tuesdays, it would be more consistent to estimate (10) using only fatality counts reported on Tuesdays. It is however consistent with the prior assumption put on the regression coefficients that the reporting weekday has no effect, since the prior mean of η_i is zero. Also, we will implement a version of the model where we do not include reporting weekday effect, and in this case the prior on γ_d as defined by equations (9) and (10) makes complete sense.

3.2.2 Fatality counts model

In this brief subsection, we define the model for n_{td}, from which N_t are defined as the sum of n_{td} over $d = 0, 1, \ldots, D$. With p_{td} defined in terms of h_{td} by equation (4), and λ_t denoting the expected value of N_t, the individual fatality counts are distributed as

$$n_{td} \,|\, p_{td}, \lambda_t \sim \mathcal{NegBin}(p_{td} \cdot \lambda_t, \phi)\,. \tag{11}$$

This thus constitutes our likelihood. We also define $\mathcal{NegBin}(0, \phi) = 0$ deterministically, as to incorporate the zero probability of reporting induced by their being days on which no reporting is done. That is, when $R_{td} = 0$ in equation (6), then $p_{td} = 0$ whence $n_{td} = 0$ non-stochastically.

We parametrize the negative binomial distribution as per the build-in parametrization of the programming language `Stan` [Stan Development Team (2019a)] (see sec. 5.1), where

$$Y \sim \mathcal{NegBin}(\mu, \theta) \Rightarrow \mathbb{E}[Y] = \mu, \mathbb{V}\mathrm{ar}(Y) = \mu + \frac{\mu^2}{\theta}. \tag{12}$$

The dispersion parameter ϕ is put with an improper uniform prior on $\mathbb{R}_+$ (which is possible in `Stan`).

We put a prior on $\{\lambda_t, t \geq 0\}$ originating with McGough et al. (2020), which invented the "Bayesian smoothing" element for the mean process of N_t. The mean process for $\{\lambda_t, t \geq 0\}$ is defined by letting $\{\log(\lambda_t), t \geq 0\}$ constitute a Gaussian process, whereby

$$\begin{aligned} \log(\lambda_0) &\sim \mathcal{N}(0, 1) \\ \log(\lambda_t) \,|\, \lambda_{t-1} &\sim \mathcal{N}\big(\log(\lambda_{t-1}), \sigma^2\big) \;\text{ for } t \geq 1 \end{aligned}$$

But since we use a moving window, indices $t < T - m$ are not included. We hence define the origin of the random walk by $\log(\lambda_{T-m}) \sim \mathcal{N}(\log N_{T-m}, 1)$.

Also, we put a prior on σ proportional to $\mathcal{N}(0, 3^2)$ on the positive real axis (i.e. σ is half-normal), consistent with the prior used by Günther et al. (2020a) (again, formally retrieved from Günther (2020b)).

3.3 Method of Altmejd et al. (2021b)

In this section, we *attempt* to recapitulate the method of Altmejd et al. (2021b). We refer to it simply as a *method* rather than a model, since we are not privy to a written-down version of it using mathematical notation with text of explanations–Only the source code is available. We mention immediately that this method is a work-in-progress, and ask the reader to take note of the date of code retrieval in the references. While the source code, by the nature of programming languages, provide an exact description of the models algorithmic proceedings, the work-in-progress nature of its implementation makes it difficult to decipher, because lines of code used in a previous version might (and do) remain. For this reason, we ingress this section with the disclaimer that the description of the method of Altmejd et al. (2021b) given here may contain inaccuracies; Therefore this section is to be seen as an overview of their methods than a complete description.

The method of Altmejd et al. (2021b) has two main steps. First, they sample from the predictive distribution of N_t. We give an overview of what this step constitutes in section 3.3.1. It turns out, however, that this distribution is to variable in relation to data, when looking at the data from day to day, whence a smoothing mechanism is implemented. This is done by a Gaussian process separately from the initial sampling step. We give some more detail on how this is done in section 3.3.2. Finally, they present a point estimate and prediction intervals from this Gaussian process, which are made readily available at Altmejd et al. (2021b). In section 3.3.3, we mention how this enables us to extrapolate the predictive distribution, which in turn enables us to comparatively evaluate it against our application of the model of Günther et al. (2020a). In short, the point estimate gives us an approximate value for the mean of the Gaussian process at time points $t = T - 25, T - 24, \ldots, T$, and the prediction intervals gives us approximate values for the diagonal elements of the covariance matrix.

The fact that we *can* evaluate the method of Altmejd et al. (2021b) is sufficient motivation that we *do*, since their results are made public at Altmejd et al. (2021a), and deserve comparison to other methods. The intellectual critique that should accompany such an evaluative comparison is however made difficult by their not existing a written-down model description, however.

Having pointed out the difficulty of extrapolating a mathematical method from computer code, and that this section would have greatly benefited from a written-down version of Altmejd et al. (2021b), we should also point out that this section would have been impossible to produce were it not for the level of transparency kept by Altmejd: The source code of the method is made publicly available, and–as mentioned in section 2–the providence of keeping records of past publications made by the Public Health Agency of Sweden [FHM (2020-

21)], and making these records public, have made it tractable for others–the author of this thesis included–to perform nowcasting for Swedish COVID-19 indicators (e.g. fatalities).

3.3.1 Sampling step

The sampling step is, from our point of view, somewhat of a "black-box", albeit slightly greyshly tinted, since we are able to extrapolate some important information from the code. In particular, it is apparent that Altmejd et al. (2021b) take into account the effect of the weekday of the reporting day, and whether or not this day is a public holiday in Sweden. Non-reporting days in the Swedish data consist of certain weekdays, e.g. Saturdays, Sundays and Mondays since September 14, 2020 (cf. Tab. 2), and public holidays. As such, the effect of certain days being non-reporting days are taken into account by the method.

It appears as though the method of Altmejd et al. (2021b) samples for the number of daily fatality counts N_{T-d} for each delay $d \in \{0, 1, \ldots, 25\}$ separately. The sampling step consists of two sub-steps. First, a generalized additive model is fitted to past, fully observed, data. Since a maximum delay of $D = 25$ is assumed, data are fully observed for $t < T - 25$. Let $N_t(t+d) = \sum_{d'=0}^{d} n_{td}$ (as per eq. (1)) denote the number of fatalities occurred on day t and reported with a delay less than or equal to d. The model assumes that

$$N_t(t+d) \sim \mathcal{B}\mathit{inom}\big(N_t, \text{logit}^{-1}(f_{td})\big)\,,$$

where

$$f_{td} = s(t) + \beta_0 \mathbb{I}_{\{t+d \text{ is a public holiday}\}} + \sum_{i=1}^{7} \beta_i \mathbb{I}_{\{t+d \text{ is the } i^{\text{th}} \text{ day of the week}\}}, \quad (13)$$

and $s(t)$ is a spline based on the thin plate regression basis. (See e.g. Wolf (2017) ch. 5.5.1 pgs. 215-221 for a treatise on thin plate regression splines.) Important to note in (13) is that covariate effects are put in place as to account for the reporting structure of Swedish data.

Let $N_t(T) = N_t(T - d + d)$ denoting the number of fatalities occurred on day t and reported before time T. The total fatality counts N_{T-d} are sampled from

$$\mathbb{P}(N_{T-d} = N \,|\, N_t(T), f_{td}) \propto \mathbb{P}(N_t(T) = N_t(T) \,|\, N, f_{td}) \text{ for } N \in \{0, 1, \ldots, 200\},$$

where the right hand side is the probability distribution of a binomial distribution with size and probability parameters N and $\text{logit}^{-1}(f_{td})$, respectively. Note that the probability of the event $\{N_{T-d} < N_t(T)\}$ is zero. Also, no motivation for the upper bound of 200 put on N_{T-d} is given, but it appears to be a compromise in order to obtain a tractable grid of values to consider for N_{T-d}. For comparison, the highest number of COVID-19 related fatalities recorded in Sweden, as of May 7, 2021, was 121 on December 28, 2020.

For each $N_{T-d}, d \in \{0, 1, \ldots, 25\}$, 2000 samples are produced.

3.3.2 Smoothing step

As mentioned in the preamble of this section, the predictive distribution produced by the samples is too variable in relation to the retrospectively observed true fatality counts. In particular, it produces too wide prediction intervals, and the day-to-day variation is also much larger than what is observed in data. For this reason, Altmejd et al. (2021b) apply a smoother; the basic assumption of which appears to be that the square root fatality counts $\sqrt{N_t}$ are realizations of a Gaussian process. The process is defined in terms of its covariance structure, which in turn is defined in terms of the Matérn covariance function $C_\nu(|t_i - t_j|; \sigma^2, \rho)$, where $|t_i - t_j|$ is the temporal distance between two fatality days. Since fatalities may occur on every day, this is simply the absolute difference of indices (see eq. (14) below). The Matérn covariance function is defined in terms of special functions–particularly the Gamma and modified Bessel function–whence we refrain from defining it as to avoid a lengthy definition for a function we use only once. We use a parametrization consistent with Wikipedia (2021).

The $(i, j)^{\text{th}}$ element of the covariance matrix is given by

$$\Sigma_{ij} = C_\nu(|i - j|; \sigma^2, \rho) + \tilde{\sigma}^2 \mathbb{I}_{\{i=j\}}, \tag{14}$$

where $\tilde{\sigma}^2 > 0$ is a noise term added on the diagonal of the matrix. The parameters ν, σ^2, ρ and $\tilde{\sigma}^2$ are estimated using maximum likelihood and fully observed past data since April 2, 2020 (indexed by $t = 1$ below). The likelihood is defined by

$$\left(\sqrt{N_1}, \sqrt{N_2}, \ldots, \sqrt{N_{T-26}}\right)^{\mathrm{T}} \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}^{(0)}\right),$$

where $\boldsymbol{\Sigma}^{(0)}$ is the $T - 26 \times T - 26$ covariance matrix corresponding to times $t = 1, 2, \ldots, T - 26$.

From the sampling step of section 3.3.1, the median and 95% equal-tailed prediction intervals are kept. Let $q_\alpha^{(t)}$ denote the α^{th} quantile of the samples obtained for N_t. The smoothing step takes as its input a location and a scale parameter defined in terms of the aforementioned quantiles by

$$\boldsymbol{y} \text{ such that } y_i = \sqrt{q_{0.5}^{(T-i+1)}}$$

and

$$\boldsymbol{s}^2 \text{ such that } s_i^2 = \left(\frac{\sqrt{q_{0.975}^{(T-i+1)} + 1} - \sqrt{q_{0.025}^{(T-i+1)} + 1}}{2 \cdot 1.96} \right)^2.$$

With $\boldsymbol{\Sigma}$ denoting the 26×26 covariance matrix corresponding to times $t = T - 25, T - 24, \ldots, T$, the smoothing step concludes with producing

$$\begin{aligned} \hat{\boldsymbol{\mu}} &= \boldsymbol{\Sigma} \left(\boldsymbol{\Sigma} + \operatorname{diag}(\boldsymbol{s}^2)\right)^{-1} \boldsymbol{y} \\ \hat{\boldsymbol{\Sigma}} &= \boldsymbol{\Sigma} - \boldsymbol{\Sigma} \left(\boldsymbol{\Sigma} + \operatorname{diag}(\boldsymbol{s}^2)\right)^{-1} \boldsymbol{\Sigma}. \end{aligned} \tag{15}$$

The parameters $\hat{\boldsymbol{\sigma}}$ and $\hat{\boldsymbol{\Sigma}}$ represent the Gaussian process on which the squared root of the mean fatality counts are defined.

3.3.3 Deduction of predictive distribution

In this subsection, we briefly explain how we are able to obtain an approximate version of the predictive distribution for the fatality counts N_t, consistent with the code of Altmejd et al. (2021b) and corresponding public version of Altmejd et al. (2021a).

For each "prediction day", Altmejd et al. (2021b) provide point-estimates and upper and lower prediction interval limits for their predictive distribution. For now, assume the prediction day T to be fixed. Let $N_t(T) = \sum_{d=0}^{\min\{T-t,D\}} n_{td}$ denote the number of fatalities occurred on day t and reported to us before now. The statistics provided are obtained from the parameters $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$ defined in equation (15) by

$$\mathrm{PE}_t = \left\lfloor \hat{\mu}_t^2 + 0.5 \right\rfloor \tag{16}$$

$$\mathrm{LPI}_t = \left\lfloor \left(\max\left\{ \hat{\mu}_t - 2 \cdot \hat{\Sigma}_{tt}, \sqrt{N_t(T)} \right\} \right)^2 \right\rfloor \tag{17}$$

$$\mathrm{UPI}_t = \left\lceil (\hat{\mu}_t + 2 \cdot \hat{\Sigma}_{tt})^2 \right\rceil \tag{18}$$

We take this to mean that the predictive distribution is such that the probability density at n, say, should be evaluated as the probability of the interval $\sqrt{n \pm 0.5}$ in the normal distribution parametrized by $\hat{\mu}_t, \hat{\Sigma}_{tt}$ if $n > N_t(T)$, and that, from (17), $n < N_t(T)$ should be assigned a probability mass of zero, and–we take it–that $N_t(T)$ should be assigned the probability in the interval $(-\infty, \sqrt{N_t(T) + 0.5})$.

An approximate value for $\hat{\mu}_t$ is obtained simply by ignoring the rounding of equation (16), i.e. such that $\hat{\mu}_t \approx \sqrt{\mathrm{PE}_t}$. Using this approximate value, we obtain an approximate value for $\hat{\Sigma}_{tt}$ by reversing equations (17) and (18)

$$\hat{\Sigma}_{tt} \approx \begin{cases} \min\{(\sqrt{\mathrm{UPI}_t} - \hat{\mu}_t)/2, (\hat{\mu}_t - \sqrt{\mathrm{LPI}_t})/2\} & \text{if } \mathrm{LPI}_t < N_t(T) \\ \sqrt{\mathrm{UPI}_t} - \hat{\mu}_t)/2 & \text{otherwise.} \end{cases} \tag{19}$$

As such, we obtain $\hat{\Sigma}_{tt}$ simply by considering the distance between the point estimate PE_t to the limits of the prediction interval, LPI_t and UPI_t. Since the floor and ceiling functions, respectively, are used in the definitions of these limits (eqs. (17), (18)), this integer distance is greater than the corresponding real distance, whence we take the minimum of the two distances in the first case of equation 19. Also, when $\mathrm{LPI}_t = N_t(T)$, the lower limit is due to the truncation induced by the observed fatality count, whence we use only the upper prediction interval limit in this case.

We note that it is possible to obtain a grid of values for $\hat{\mu}$ and $\hat{\Sigma}$ which satisfy equations (15)-(18), thus quantifying the approximate nature of our deduction, but since the error is rather small, we deem it unnecessary.

3.4 Other points of view

In this section, we give an overview of some other points of views on nowcasting, which also stem from the structure of data as defined in section 3.1.1. First,

in section 3.4.1, we present the chain ladder method, which is an ubiquity of actuarial sciences in estimating the monetary reserve necessary to cover the costs of occurred but not yet reported insurance claims. The *number* of insurance claims follow exactly the same data structure as the daily fatality counts, while the claims costs are generally assumed to have positive, real support. In section 3.4.2, we present the novel point of view posed by Altmejd et al. (2020), that nowcasting of count data subject to reporting delay in discrete time can be described as a special case of the removal method for estimating population sizes from ecological statistics. We note, however, that we arrive at a likelihood identical to that of Lawless (1994), whose treatise is epidemiological.

3.4.1 Chain ladder approach

The chain-ladder method is an approach to nowcasting that originates from actuarial sciences, where it is used to estimate the cost of insurance claims for accidents, or some other insurance claimable events, that have occurred but not yet been reported. The necessity for nowcasting in this case stems for the fact that insurance companies are obliged to keep reserves as to be able to make payouts once the claims are reported. We use the terminology of Mack (1993), whom we follow in this subsection, in referring to the time of an accident as its *accident year* and the time of its reporting as its *development year*. These are the analogues of the fatality day t and reporting delay $d-1$ in days. The "minus one" is due to one using ordinal numbers in referring to development years, e.g. an accident reported in the same year of its occurrence is reported during its *first* development year, while having a reporting delay of *zero* years. The essential temporal structure of the data is hence identical to that of the reporting triangle in Table 3, although its entries may be different; This data structure is referred to as the "chain ladder", supposedly since each columns has the appearance of a chain ladder.

Let $t = x, x+1, \ldots, T$ denote the accident year, where x refers to some origin year and T is the current year, and let $d = 0, 1, \ldots, D$ denote the $(d+1)^{\text{th}}$ development year. For instance, an accident occurred during 2014 and reported during 2016 has a reporting delay of $d = 2$ years, and hence it is reported during its third development year. The index $(D+1)$ corresponds to the the latest available development year, or, as previously, D fixates a maximum delay.

But how are the entries in the chain ladder different from those in the reporting triangle? First, in the basic chain ladder method each entry represents the cumulative claims *costs* of accidents occurred during year t and reported by the $(d+1)^{\text{th}}$ development year, which is a positive real value, i.e. it is not an integer count. Second, the cumulatives of accident year t and development year $d+1$ are entered for each instance of (t, d) rather than the claims costs specific for accident year t and development year t. The former is analogous to $N_t(t+d)$ from equation (1) with respect to t and d, while the latter is analogous to n_{td}.

Let $C_t(t+d)$ denote the total claims cost for accidents occurred during year t and reported before or during development year $d+1$. The defining assumption of the chain ladder method is that the mean cumulative claims cost for accident

year t as reported by development year $d+1$, conditioned on the past reports for accident year t, is given by

$$\mathbb{E}[C_t(t+d)\,|C_t(t+d-1), C_t(t+d-2), \ldots, C_t(t)] = f_{d-1}\mathbb{E}[C_t(t+d-1)]\,, \tag{20}$$

where $f_{d-1} > 1$ is referred to as the development factor. The terminology of *development* hence stem from the fact that the cumulative claims costs for a certain accident year "develop" as more accidents are reported each development year. From this assumption, point estimates for $f_0, f_1, \ldots, f_{D-1}$ are tractable, and given by Mack (1993).

The total claims costs for accidents occurred during year t would hence be written as $C_t(t+D)$, and the task of the actuary is to estimate these quantities for $t = x, x+1, \ldots, T$. Their tool for doing so are the estimated development factors. At time T, $C_t(s)$ is known for $t \leq s \leq T$, and–in particular–we know $C_t(T)$ for $t \leq T$. By using the law of total expectation, it can be shown that equation (20) implies that

$$\mathbb{E}[C_t(t+D)\,|C_t(T), C_t(T-1), \ldots, C_t(t)] = f_{d-1}f_{d-2}\ldots f_{d-(T-t)}\mathbb{E}[C_t(T)]\,, \tag{21}$$

which corresponds to a special case of Theorem 1 of Mack (1993). The conditioning in the expectation on the left hand side of equation (21) corresponds to what have been observed from accident year t until the current year T. Replacing f_i with estimates $\hat{f}_i$ for $i = d-1, \ldots, d-(T-t)$ yields an estimate for the total claims cost for accident year t, $\hat{C}_t(t+D)$, say.

But equations (20) and (21) only provide point estimates for the total claims costs. By including an assumption relating to the conditional variance similar to the assumption on the mean (eq. (20)), and assuming that the total claims cost from different accident years are independent of one another, Mack (1993) derive estimates for the variance of the estimated total claims cost $\hat{C}_t(t+D)$.

While Mack (1993) is directly concerned with the claims costs, Mikosch (2009) (ch. 11 sec. 2, pgs. 365-381) provide an approach wherein the *number* of claims are instead considered, in parallel to the their costs. The assumptions are however identical–replacing the letter C with the letter N–and the approach for finding estimates for the development factors and total number of claims is also identical, whence the consequent estimate $\hat{N}_t(t+D)$, say, would take the form of a real positive number, and strictly represent the mean number of claims eventually reported.

3.4.2 Removal method and multinomial mixture approach

In this section, we follow section 2 of Altmejd et al. (2020) in linking epidemiological nowcasting with the removal method, as it was defined by Pollock (1991). We note–however–that the link posed by Altmejd et al. (2020) is rather tenuous in that the traditional solutions to the removal method are not immediately translatable to the epidemiological nowcasting problem; The problem formulation will also need some modification. We deem, however, the novelty of the

link worthy of reiteration. In the end of this subsection, we make some comments on the more classical solutions to the removal method problem, as they are summarised by Pollock (1991).

Recall that the discrete hazard function is defined as

$$h_{td} = \mathbb{P}(\text{delay} = d \,|\text{delay} \geq d, \text{fatality on day } t)\,,$$

and that the reporting delay in terms of p_{td} can be obtained from h_{td} using the iterative scheme laid out in equation (4) for $t = 1, 2, \ldots$ and $d = 0, 1, 2, \ldots, D$.

Now let, for a moment, N_t represent the number of specimens of some species within a confined geographical space not subject to any migration, births or deaths, i.e. N_t, once realized, is a fixed count (which, obviously, is consistent with its interpretation as the number of fatalities at day t). The index t bears no meaning temporally in this setting; Let it denote e.g. the index of the geographical space in question. The removal method is a capture-retain procedure (as opposed to capture-recapture procedures), in which on day $d = 0$, one catches n_{t0} specimens out of the N_t available ones; on day $d = 1$, one catches n_{t1} specimens from the $N_t - n_{t0}$ remaining ones, and so fourth. In general, on day d one catches n_{td} specimens from a remaining population of $N_t - \sum_{d'=0}^{d-1} n_{td'}$. Consequently, if one denotes the probability of capture on day d by h_{td} this produces a binomial likelihood:

$$\begin{aligned}
n_{t0} \,|\, N_t, h_{t0} &\sim \mathcal{B}inom(N_t, h_{t0}) \\
n_{t1} \,|\, N_t, h_{t1}, n_{t0} &\sim \mathcal{B}inom(N_t - n_{t0}, h_{t1}) \\
&\vdots \\
n_{td} \,|\, N_t, h_{td}, n_{t0}, n_{t1}, \ldots, n_{t,d-1} &\sim \mathcal{B}inom\left(N_t - \sum_{d'=0}^{d-1} n_{td'}, h_{td}\right)
\end{aligned} \tag{22}$$

At this point, the basic removal method diverges from its basic epidemiological counterpart, for one basic reason: In epidemiology, one generally assumes that the true number of fatalities will be eventually (albeit approximately) observed in full, and the reason for nowcasting is that one needs timely estimates of the recent number of fatalities. But for an ecologist it might suffice with a handful of samples constituting a small proportion of the total population. That is, a "final day" D, when all remaining specimens are assumed to be caught, does not exist, thus yielding the maximum delay of epidemiological nowcasting not generally translatable. Thus we revert to the epidemiological jargon.

Assuming a maximum delay to be existent, we might continue (22) until its final entry (note that $h_{tD} = 1$):

$$\begin{aligned}
&\vdots \\
n_{tD} \,|\, N_t, n_{t0}, n_{t1}, \ldots, n_{t,D-1} &= N_t - \sum_{d'=0}^{D-1} n_{td'} \text{ w.p.1.}
\end{aligned} \tag{23}$$

Putting (22) and (23) together yields a multinomial joint likelihood

$$(n_{t0}, n_{t1}, \ldots, n_{t,D-1}, n_{t,D})^{\mathrm{T}} \,|\, N_t, \{h_{td}, d = 0, 1, \ldots, D-1\} \\ \sim \mathit{Multinom}\left(N_t, (p_{t0}, p_{t1}, \ldots, p_{t,D-1}, 1)^{\mathrm{T}}\right), \quad (24)$$

which is the basis for Höhle and an der Heiden (2014), and to which, together with a (e.g. Poisson or negative binomial) distribution assumption on N_t, Stoner and Economou (2019) refer to as the "multinomial mixture approach" to nowcasting. Lawless (1994) further lays out a frequentist framework to estimating p_{td}, and thence estimate N_t, under the censoring induced by the structure of data (see sec. 3.1.1), but we refrain from recapitulating his results, instead electing to mention ecological methods. Lawless (1994) also provides asymptotic results as to the variance of his estimators valid for large N_t.

Pollock (1991) mentions three methods from ecology used to solve the problem. His preference is the numerical estimation of maximum likelihood estimates, though, whence two methods remain. These are generally not (immediately) applicable to epidemiological nowcasting. The simplest one relies on only obtaining two samples, n_{t0} and n_{t1}, say, and assuming the probability of capture to be constant, i.e. the discrete hazard rate is constant $h = h_{t0} = h_{t1}$. Due to this assumption, a point estimate for N_t can be obtained by solving $h = n_{t0}/N_t = n_{t1}/(N_t - n_{t0})$, which yields $\hat{N}_t = n_{t0}^2/(n_{t0} - n_{t1})$. The glaring weakness of this method is that it fails whenever $n_{t1} \geq n_{t0}$, which–of course–is a possibility when n_{t0} and n_{t1} are stochastic. The third method also assumes a constant capture probability / discrete hazard rate. Let $N_t(t + d - 1) = \sum_{d'=0}^{d-1} n_{td}$ denote the number of specimens caught before day d. Noting that the mean of n_{td} is a linear function of $N_t(t+d-1)$ with an intercept depending on N_t, formally

$$\mathbb{E}[n_{td} \,|\, n_{t0}, n_{t1}, \ldots, n_{t,d-1}] = h\,(N_t - N_t(t + d - 1)),$$

with $N_t(t-1) = 0$, and N_t can be obtained by performing linear regression for n_{td} as a function of $N_t(t+d-1)$. This would however also only provide a point estimate for N_t. Pollock (1991) cite the simplicity of this method as it strength, but recommend a maximum likelihood approach.

4 Evaluation

The aim of this thesis is the evaluation of nowcasting methods, or–more specifically–comparatively evaluating the *predictive performance* of different methods when applied to Swedish fatality data. To do so, we use several metrics, which are defined and briefly motivated in this section. The particular choice of metrics were heavily influenced by the equivalent choices of Günther et al. (2020a), as we use largely an identical set of evaluative metrics, but also note that e.g. Höhle and an der Heiden (2014) and McGough et al. (2020) also use similar sets of metrics in comparing the predictive performance of epidemiological nowcasting methods. In particular, they make use of *proper scoring rules*. We devote section 4.1 to introducing three such rules, and also define what propriety refers to with respect to a scoring rule. In section 4.2, we define the more self-explanatory evaluation metrics of prediction error and relative error, as well as coverage and width of $100 \cdot (1 - \alpha)\%$ prediction intervals. These do not, however, constitute "scoring rules" in the sense defined in section 4.1, and for this reason we simply refer to them as generic *evaluation metrics*.

Evaluation of any probabilistic nowcasting method, or of any probabilistic forecast, relies on two things: First, we need a probability distribution quantifying our uncertainty of the prediction. Second, we need the true value as it realizes. In order to avoid confusion with quantities from section 3, we opt to use a different notation in this section; This also allows us to drop the temporal subscript, and underlies the fact that the contents of this section are not stemmed from the field of nowcasting.

With some minor modifications and additions, we follow the notation of Czado et al. (2009), which is our primary source on proper scoring rules. Let X denote a non-negative integer valued stochastic variable, and let x denote its realization. Before knowing its realization, we produce a probabilistic forecast of it quantified by the infinite length vector $\boldsymbol{P}$, such that $\mathbb{P}(X \leq i) = P_i$ for $i = 0, 1, 2, \ldots$. It will prove convenient to also define the infinite length vector $\boldsymbol{p}$ by $\mathbb{P}(X = i) = p_i$ for $i = 0, 1, 2, \ldots$. Also, let $\hat{x}^{(\boldsymbol{P})}$ denote a point estimate for X based on $\boldsymbol{P}$, e.g. the mean or median, and let $q_z^{(\boldsymbol{P})}$ denote the z^{th} quantile of $\boldsymbol{P}$. Finally, let $\boldsymbol{Q}$ and $\boldsymbol{q}$ be the counterparts of $\boldsymbol{P}$ and $\boldsymbol{p}$, but for a probabilistic forecast that is (possibly) different from $\boldsymbol{P}$.

We should also point out that we define the scoring rules and other evaluation metrics in terms of a single instance of X, and correspondingly as argument the single tuple $(\boldsymbol{P}, x)$. In implementation, however, we report "averages over suitable sets of probabilistic forecasts" [Czado et al. (2009)], stemmed from the full set of predictions we produce. We specify more precisely how this is done in subsection 4.3, once we have introduced all the evaluation metrics. For the reader's convenience, we give a specific example here, though: In implementing the model of Günther et al. (2020a) and the method of Altmejd et al. (2021b), by their design we produce $D = 35$ and $D = 25$, respectively, probabilistic forecasts corresponding to the random variables $N_T, N_{T-1}, \ldots, N_{T-D}$. But we might be more interested in the latest, predictions, i.e. those for $N_T, \ldots, N_{T-7}$, since these are more informative with respect to the current trend. Hence, we

might compute average scores for our predictions over $N_T, \ldots, N_{T-7}$ in order to comparatively assess the ability of the methods to predict recent fatality counts.

4.1 Scoring rules

A *scoring rule* is, paraphrasing the first paragraph of section 3 of Czado et al. (2009), a function taking as its arguments a probabilistic forecast, i.e. a predictive distribution, and its subsequent realized value, and returning a numerical value, a "score", that makes the predictive distribution comparable to alternative predictive distributions. We follow Czado et al. (2009) in "take[ing] scoring rules to be negatively oriented penalties that a forecaster wishes to minimize", i.e. a lesser score is better.

Let $\mathrm{S}(\boldsymbol{P}, x)$ generically denote a score assigned to $\boldsymbol{P}$ when x is the realized value, and let $\mathrm{S}(\boldsymbol{P}, \boldsymbol{Q}) = \mathbb{E}_{\boldsymbol{Q}}[\mathrm{S}(\boldsymbol{P}, X)]$ denote its expected value when the distribution of X is determined by $\boldsymbol{Q}$. For the remainder of this section, we let $\boldsymbol{Q}$ represent "the forecaster's best judgement" [Czado et al. (2009)], or–heuristically–the true distribution of X, were the task to guess its distribution. A scoring rule is defined to be proper if

$$\mathrm{S}(\boldsymbol{Q}, \boldsymbol{Q}) \leq \mathrm{S}(\boldsymbol{P}, \boldsymbol{Q}), \tag{25}$$

which corresponds to equation (4) of Czado et al. (2009). This means that, on average, quoting $\boldsymbol{Q}$ will produce at least an as good score as quoting $\boldsymbol{P}$, if S is proper and the distribution of X is determined by $\boldsymbol{Q}$. In other words, the forecaster is incentivized to quote the truest distribution, to the best of his or her knowledge. A scoring rule is *strictly proper* if a strict inequality holds for equation (25), meaning that the best predictive distribution is unique, i.e. $\mathrm{S}(\boldsymbol{Q}, \boldsymbol{Q}) = \mathrm{S}(\boldsymbol{P}, \boldsymbol{Q})$ only if $P_i = Q_i$ for all $i \geq 0$.

A subtlety of probabilistic forecasting apparent from the careful wording of Czado et al. (2009) of $\boldsymbol{Q}$ being "the forecasters best judgement" is that the task of forecasting is not simply to guess the distribution of X, but to quantify one's own uncertainty with respect to the outcome of X, as indicated by past experiences (i.e. data). Certainly, a source of this uncertainty is the innate stochasticity of X, but it is not the only source, since data from which we infer $\boldsymbol{P}$ are subject to stochasticity as well. Consequently, when we use the heuristic of referring to $\boldsymbol{Q}$ as the "true" distribution X, we are referring to its distribution from the forecaster's point of view, as opposed to that of an entity that impossibly knows the true randomness of future events.

As an example of an extreme case making the distinction less subtle, suppose there are 101 different coins, which are flipped one after another. For the forecaster, the coins are indistinguishable from one another. The first 100 coins are unbiased, such that the probability of them turning up heads equals that of them turning up tails. But the 101^{st} coin is weighted such that it always turns up heads. Having observed the first 100 coin flips turn up 50 heads and 50 tails in no particular order, what should our probabilistic forecast be? Since the 101^{st} coin is indistinguishable from the first 100 coins, it is arguable that the

forecaster should quote $\mathbb{P}(\text{heads}) = 0.5$, this this is their best judgement based on data, even if $\mathbb{P}(\text{heads}) = 1$ would produce a better score. Better yet would, of course, be to realize that the first 100 coin flips contain no information as for the outcome of the 101^{st} coin flip, and this would be the realization of the entity that knowns the true randomness of future events, but such an entity does not generally exist.

4.1.1 Squared and absolute error score

We begin with a well-known and well-used scoring rule: the square error score. It is defined by

$$\text{SES}(\boldsymbol{P}, x) = \left(x - \hat{x}^{(\boldsymbol{P})}\right)^2,$$

for a point estimate $\hat{x}^{(\boldsymbol{P})}$ based on $\boldsymbol{P}$. The intuition behind this scoring rule is very straight-forward: a prediction inducing a point estimate close to the realized value is preferable. If the point estimate is taken to be the expected value of X induced by $\boldsymbol{P}$, i.e. $\hat{x}^{(\boldsymbol{P})} = \mathbb{E}_{\boldsymbol{P}}[X]$, the squared error score is proper. Suppose, as before, that $\boldsymbol{Q}$ is the best predictive distribution of X. Computing the difference between the expected squared error score of $\boldsymbol{P}$ and $\boldsymbol{Q}$ under probability law $\boldsymbol{Q}$ yields, after some algebra,

$$\begin{aligned}
\text{SES}(\boldsymbol{P}, \boldsymbol{Q}) - \text{SES}(\boldsymbol{Q}, \boldsymbol{Q}) &= \mathbb{E}_{\boldsymbol{Q}}[(X - \mathbb{E}_{\boldsymbol{P}}[X])^2] - \mathbb{E}_{\boldsymbol{Q}}[(X - \mathbb{E}_{\boldsymbol{Q}}[X])^2] \\
&= \mathbb{E}_{\boldsymbol{Q}}[X^2] - 2\mathbb{E}_{\boldsymbol{Q}}[X]\mathbb{E}_{\boldsymbol{P}}[X] + \mathbb{E}_{\boldsymbol{P}}[X]^2 \\
&\quad - \mathbb{E}_{\boldsymbol{Q}}[X^2] + 2\mathbb{E}_{\boldsymbol{Q}}[X]\mathbb{E}_{\boldsymbol{Q}}[X] - \mathbb{E}_{\boldsymbol{Q}}[X]^2 \\
&= \mathbb{E}_{\boldsymbol{Q}}[X]^2 + \mathbb{E}_{\boldsymbol{P}}[X]^2 - 2\mathbb{E}_{\boldsymbol{Q}}[X]\mathbb{E}_{\boldsymbol{P}}[X] \\
&= (\mathbb{E}_{\boldsymbol{Q}}[X] - \mathbb{E}_{\boldsymbol{P}}[X])^2 \geq 0, \qquad (26)
\end{aligned}$$

from which we see that the squared error score satisfies equation (25), when the point estimate is taken to be the mean. We note, however, that the difference in the last line of (26) does not imply a strict inequality, since there exist $\boldsymbol{P}$ and $\boldsymbol{Q}$ such that $\mathbb{E}_{\boldsymbol{Q}}[X] = \mathbb{E}_{\boldsymbol{P}}[X]$ where $P_i \neq Q_i$ for some $i \in \{0, 1, \dots\}$. For example, $\boldsymbol{p} = (1/3, 1/3, 1/3, 0, \dots)^{\mathrm{T}}$ produces a mean of 1, as does $\boldsymbol{q} = (0, 1, 0, 0, \dots)^{\mathrm{T}}$. Hence, the squared error score, while being proper, is not *strictly* proper.

Another well-known and well-used scoring rule is the *absolute* error score, defined as the square root of the squared error score

$$\text{AES}(\boldsymbol{P}, x) = \sqrt{\text{SES}(\boldsymbol{P}, x)} = \left|x - \hat{x}^{(\boldsymbol{P})}\right|.$$

The advantage of the absolute error score is that it conveys the order of the error.

4.1.2 Log-score

A score that is always *strictly* proper is the log-score, defined by

$$\text{logS}(\boldsymbol{P}, x) = \begin{cases} -\log p_x & \text{if } p_x > 0 \\ \infty & \text{if } p_x = 0. \end{cases}$$

As is pointed out by Höhle and an der Heiden (2014), this score is "very intuitive–the higher the probability of the [predictive] distribution for the [realized] value the better". The fact that $\text{logS}(\boldsymbol{P}, x) = \infty$ if $p_x = 0$ is somewhat problematic, though, since we will evaluate the average scores over several instances of X. If the predictive distribution assigns a probability of zero for either of these instances, the corresponding log-score will be evaluated to infinity, whence the average score will also be evaluated to infinity. This problem is exaggerated when our knowledge of the predictive distribution is obtained from sampling (see sec. 5.1), whereby no samples may be obtained for certain low-probability outcomes, yielding an erroneously exact zero probability for those events. This might be construed as a weakness of the log-score.

In Figure 8 we see an illustration of the log-score when applied to a predictive distribution corresponding to a negative binomial distribution with mean parameter $\mu = 12$ and dispersion parameter $\phi = 25$, and a realized value of $x = 13$. The simplicity of the log score is visually apparent, as well.

In ensuring the strict propriety of the log-score, we provide a link to the KullbackLeibler divergence $\mathcal{D}_{\text{KL}}(\boldsymbol{Q} \parallel \boldsymbol{P})$ between two probability distributions:

$$\begin{aligned} \text{logS}(\boldsymbol{P}, \boldsymbol{Q}) - \text{logS}(\boldsymbol{Q}, \boldsymbol{Q}) &= \mathbb{E}_{\boldsymbol{Q}}[-\log p_X] - \mathbb{E}_{\boldsymbol{Q}}[-\log q_X] \\ &= \mathbb{E}_{\boldsymbol{Q}}[\log (q_X / p_X)] \\ &= \mathcal{D}_{\text{KL}}(\boldsymbol{Q} \parallel \boldsymbol{P}) \geq 0. \end{aligned}$$

The Kullenback-Leibler divergence equals zero if and only if $\boldsymbol{P} = \boldsymbol{Q}$ (e.g. thm. 2.6.3 of Cover and Thomas (2006), *p. 28*), whence the log-score is indeed *strictly* proper.

4.1.3 Ranked probability score

Lastly, we introduce the ranked probability score. According to Czado et al. (2009), this score is also strictly proper. It is defined by

$$\text{RPS}(\boldsymbol{P}, x) = \sum_{i=0}^{\infty} \left(P_i - \mathbb{I}_{\{x \leq i\}} \right)^2 . \tag{27}$$

In contrast to the log-score, the ranked probability score takes into account the entire probability distribution $\boldsymbol{P}$, not just its value at its realized value. Also, it does not evaluate to infinity, whence averaging over several instances of predicted quantities does not suffer as much from one of them being particularly inaccurate. In the bottom right panel of Figure 8, we see an illustration of the contributions to the ranked probability score from each instance of $i \in \{0, 1, \ldots, 30\}$, transformed by the squared root, i.e. $\left| P_i - \mathbb{I}_{\{x \leq i\}} \right|$. Clearly, a sharper predictive distribution, yielding a steeper slope of the cumulative density function, would produce lesser individual contributions to the sum, and thus producing a better score; As would having the realized value fall closer to the median. Czado et al. (2009) note that the ranked probability score is a generalization of the absolute error: if the predictive distribution were a point

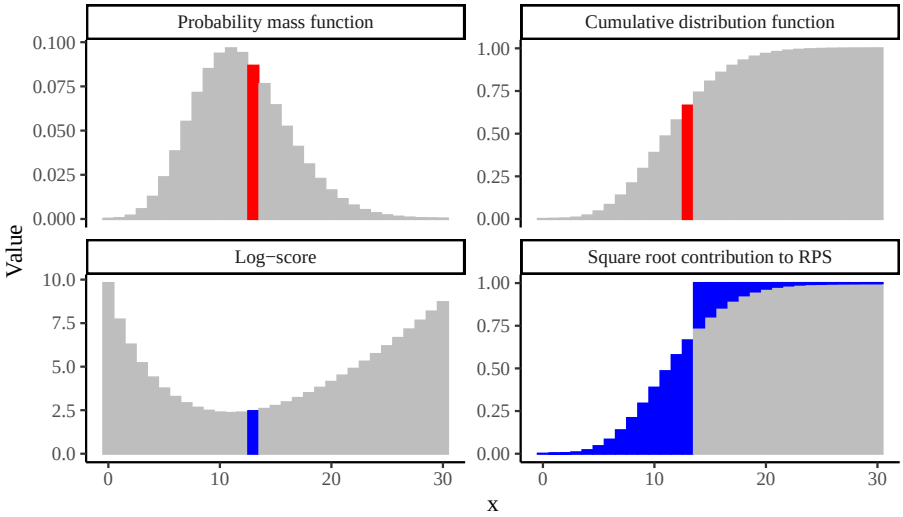

Figure 8: **Top:** Probability mass function and cumulative density function of a $\mathit{NegBin}(\mu = 12, \phi = 25)$-distributed random variable (parametrized as in eq. (12)), an indicated realization of $x = 13$ (red). **Bottom:** Log-score and square root contributions to the ranked probability score induced by such a predictive distribution and realized value (blue).

estimate, i.e. $P_k = 1$ for some $k \in \{0, 1, \dots\}$, then the contributions to the summand of equation (27) are 1 for the $|k - x|$ instances between k and x, and zero otherwise.

4.2 Other evaluative metrics

We use four other metrics, which do not constitute proper scoring rules, but are yet informative. We use a similar notational convention in that metric$(\boldsymbol{P}, x)$ takes as its arguments a predictive distribution $\boldsymbol{P}$ and a realized value x, except for the width of prediction intervals, which does not require a realized value.

4.2.1 Error and relative error

In this sections, we define two metrics measuring the divergence of a point estimate $\hat{x}^{(\boldsymbol{P})}$ based on the predictive distribution $\boldsymbol{P}$, e.g. its median. We define the *simple* error by

$$\mathrm{err}(\boldsymbol{P}, x) = \hat{x}^{(\boldsymbol{P})} - x,$$

which should not be confused with the absolute error, $\left|\hat{x}^{(\boldsymbol{P})} - x\right| = |\mathrm{err}(\boldsymbol{P}, x)|$. The latter is a scoring rule in that it has the property that lesser values are preferable (by our definition of scoring rules in section 4.1), while the former

is not. Rather, a simple error close to zero is preferable. The motivation for considering the simple error is that it may indicate whether the predictive distribution is positively or negatively biased. For instance, if the average simple error is greater that the realized value for several instances of its implementation, one might suspect a positive bias in the predictive distribution.

Another measure of the divergence of the point estimate is the *relative* error, defined by.

$$\text{relErr}(\boldsymbol{P}, x) = \hat{x}^{(\boldsymbol{P})}/x - 1.$$

The main advantage of the relative error as opposed to the simple error is that it is not dependent on the order of X, which may have an overly great influence on the simple error when averaging over several instances of X if either of these are of a significantly higher order than other instances.

4.2.2 Width and coverage of prediciton intervals

Another non-score metric is the coverage of $100 \cdot (1-\alpha)\%$ equal-tailed prediction intervals. Following the convention used throughout this section of defining evaluation metrics using a single instance of the predictive distribution $\boldsymbol{P}$ and its realized value x, we define the coverage for a single instance as well. Thus,

$$\text{coverage}_\alpha(\boldsymbol{P}, x) = \mathbb{I}_{\left\{q_{\alpha/2}^{(\boldsymbol{P})} \leq x \leq q_{1-\alpha/2}^{(\boldsymbol{P})}\right\}}.$$

is equal to 1 if x is contained by the $100 \cdot (1-\alpha)\%$ equal-tailed prediction interval induced by $\boldsymbol{P}$. This metric is however rather meaningless when applied to a single instance. When averaged over several instances of X, however, it carries meaning: The average coverage of a $100 \cdot (1-\alpha)\%$ should, obviously, be close to $(1-\alpha)$.

The last evaluation metric we use is only useful in conjunction with the average coverage, the width of equal-tailed prediction intervals, defined by

$$\text{width}_\alpha(\boldsymbol{P}) = q_{1-\alpha/2}^{(\boldsymbol{P})} - q_{\alpha/2}^{(\boldsymbol{P})}.$$

Clearly, one would prefer a narrow width, but only while keeping an average coverage close to $(1-\alpha)$. In particular, the width does not take into account the realized value.

4.3 Practical implementation of evaluation

As mentioned in the preamble of this section, in practice, we evaluate the probabilistic forecasts of the different nowcasting methods using *average* scores over several instances of predictive distributions and realized values. Let $\mathcal{P} = \{\boldsymbol{P}^{(1)}, \boldsymbol{P}^{(2)}, \ldots, \boldsymbol{P}^{(J)}\}$ denote a set of probabilistic forecasts for the random variables $X^{(1)}, X^{(2)}, \ldots, X^{(J)}$, and suppose these random variables are known to in retrospect have realized as $\mathcal{X} = \{x^{(1)}, x^{(2)}, \ldots, x^{(J)}\}$. The order of these sequences are not generally important, but it is important to distinguish $\boldsymbol{P}^{(i)}$ to be the forecast of $X^{(i)}$, which has realized as $x^{(i)}$ for all $i \in \{1, 2, \ldots, J\}$.

As such, we have J probabilistic forecasts for J realized values. We define the average score of a scoring rule, or other evaluative metric, S over a subset $\mathcal{J} \subseteq \{1, 2, \ldots, J\}$ of the available forecast-realized value pairs by

$$\text{avgS}_{\mathcal{J}}(\mathcal{P}, \mathcal{X}) = \frac{1}{|\mathcal{J}|} \sum_{i \in \mathcal{J}} \text{S}(\boldsymbol{P}^{(i)}, x^{(i)}) \tag{28}$$

where $|\mathcal{J}|$ is the size of the subset $\mathcal{J}$.

We may compute several average scores for $(\mathcal{P}, \mathcal{X})$ over different subsets $\mathcal{J}$. For instance, in section 5, we choose $\mathcal{J}$ such that it includes forecasts for N_{T-d} for specific delays $d \geq 1$, such that we may investigate whether certain nowcasting methods perform better than others in predicting the number of daily fatalities for recent days relative to now. For example, it is preferable to produce more accurate nowcasts for recent days than for more past days. If we were to only consider the average score over $\mathcal{J} = \{1, 2, \ldots, J\}$, such detailed advantages in performance would be obscured by the averaging. In fact, in section 5 we do not even present the average scores over all instances for which we produce nowcasts, since fatality counts N_{T-d} for days far in the past are, firstly, more completely reported and, second, less informative with respect to the current trend in fatality counts.

5 Results

In this section, we perform a comparative analysis of the model of Günther et al. (2020a) and the method of Altmejd et al. (2021b) from sections 3.2 and 3.3, respectively. We implement the nowcasting methods for a "now", denoted by T in section 3, consisting of reporting days between February 2 and March 2, 2021. We do not implement the nowcasting methods on non-reporting days, since, by definition, no new data are available on these days. This results in 17 instances of T.

Since the discrete hazard regression element of the model of Günther et al. (2020a) allows us to easily adapt some variations of it, we do so. In Table 5, we give names to, and briefly describe, the methods we implement; The latter three are variations of the model of Günther et al. (2020a). First, however, we make a slight modification to all of these variations, related to the Public Health Agency's publication pattern during February 2 until March 2, 2020.

By the definition of the Public Health Agency's current publication schedule outlined in Table 2, fatalities are not reported with a delay of zero since August 11. In other words, $p_{t0} = h_{t0} = 0$ for $t \geq$ 2020-08-11. We choose to incorporate this domain knowledge when implementing the model of Günther et al. (2020a). Formally, we may do so by modifying the definition of R_{td} given by equation (6) such that it induces a zero probability for zero delay:

$$R_{td} = \mathbb{I}_{\{d>0\}} \cdot \mathbb{I}_{\{t+d \text{ is a reporting day}\}}. \tag{29}$$

In practice, however, we simply removed the first column of the reporting triangle (cf. Tab. 4), as well as the corresponding discrete hazard regression intercept, γ_0. This is equivalent to defining R_{td} by equation (29).

- **Altmejd** refers to the method of Altmejd et al. (2021b), as extrapolated from their publicly available data (see sec. 3.3.3).
- **Günther main** refers to the model of Günther et al. (2020a) when *not* including a covariate effect corresponding to the weekday of the reporting day. This corresponds to simply letting $g_{td} = \gamma_d$ in equation (5).
- **Günther weekday** refers to the model of Günther et al. (2020a) when including a covariate effect for the weekday of the reporting day, wherein γ_d is defied by equation (8).
- **Günther naïve** refers to a *naïve* implementation of the model of Günther et al. (2020a), where we do not take into account that certain days are not weekdays. This corresponds to letting $R_{td} = 1$ for all $d > 0$. As such, we still include the domain knowledge of zero day delay reports being impossible.

Table 5: Names and descriptions of the implemented methods.

5.1 Implementation

The variations of the model of Günther et al. (2020a) were implemented using `Stan` [Stan Development Team (2019a)] using the `RStan` interface [Stan Development Team (2019b)] for the `R` programming language [R Core Team (2021)]. All computations, and production of figures and (some) tables were done using the `R` programming language.

`Stan` is a programming language that implements a Markov chain Monte Carlo posterior sampler for Bayesian inference. In implementing Günther et al. (2020a), for each instance of T, we run 4 chains for 3 500 iterations each, where the samples from the first 1000 iterations of each chain are discarded as burn-ins. Hence, we produce 10 000 samples for each instance of T. Each sample consists of a set of parameters, and a set of predicted fatality counts $\{n_{td}, t < T, 1 \leq d \leq D \text{ s.t. } t + d > T\}$, drawn form the predictive distribution defined by the parameter set and likelihood (eq. (11)). The syntax of `Stan` produces these fatality counts by pseudo-random number generation.

5.2 Main comparative results

In this section, we compare the method of Altmejd et al. (2021b) presented in section 3.3 with the variations of Günther et al. (2020a) presented in section 3.2; the variations being defined in Table 5. In particular, we present average scoring rules computed over suitable instances of probabilistic nowcasts, and determine which, if any, of the methods outperform the others. We will see that the Günther Main and Weekday methods outperform Altmejd, which in turn outperforms Günther Naïve.

5.2.1 Overview

Before delving into these main results, we presents some more general illustrative results with respect to the actual nowcasts. As was illustrated in Figure 1 in section 1, the necessity of nowcasting stems from the fact that the observed daily fatality counts at a time when the pandemic is ongoing understates the true daily fatality counts, due to a reporting delay, and the purpose of nowcasting is to provide a fatality curve that is closer to reality. In Figure 9, we illustrate the Altmejd and Günther Main nowcasting methods implemented for "now" being either February 2, 10, 18 or 26 in 2021. We see that, for all these dates, the nowcasts perform rather similarly with respect to the median estimate, but that the Altmejd method has a narrower 95% prediction intervals. We also see that the nowcasts seem to manage rather well since the true number of fatalities generally fall within the 95% confidence intervals. We note, however, that this is indistinguishable for longer delays, since the prediction intervals are very narrow there.

Figure 9 only provides glimpses of the nowcasts, however. To illustrate the outcome of all the 17 considered instances of T, we turn to Figure 10. In this figure, instead of illustrating the entire sequence of predicted fatality counts for

each T, i.e. $N_{T-D}, N_{T-D+1}, \ldots, N_{T-1}$, we instead consider only four instances of this sequence for each T, namely $N_{T-1}, N_{T-4}, N_{T-7}, N_{T-14}$. That is, for each of the 17 instances of T, we look at the predicted number of fatalities 1, 4, 7 and 14 days before T. Included are again 95% equal-tailed prediction intervals, indicated by error bars, the observed number of fatalities as reported by day T indicated by crosses $\times$, and the retrospective true number of fatalities indicated by a dotted line. Our conclusion from looking at Figure 9 that the Altmejd method produce narrower prediction intervals is again apparent in Figure 10, but for a delay of 14 days, or even 7 days, we see less of a difference in the width of the prediction intervals.

We excluded the Günther Weekday model from Figure 9 and Figure 10, since it produces an almost identical median of the nowcasting distribution, and almost identical prediction intervals. We shall see in section 5.2.2 that the Günther Weekday model produces almost identical values to Günther Main in all evaluative measures that we consider in this thesis. Also, the Günther Naïve model produces very wide prediction intervals, obscuring the difference between the Altmejd and Günther Main models, whence it was also excluded from Figure 9 and Figure 10.

5.2.2 Summary results of evaluation

Now, let us turn to our main results. We compute average scores, simple and relative errors, and coverage and width of 95%, 80% and 50% equal-tailed prediction intervals for *three* subsets of the predictive distributions for the fatality counts produced by the nowcasts. For each instance of T, these subsets can be written as $\{N_{T-1}, N_{T-2}, \ldots, N_{T-\delta}\}$ where $\delta \in \{7, 14, 21\}$. In other words, we compute the average when including predictions made one, two and three weeks into the past relative to T. In general, we are more interested in more recent fatality counts, since they serve as an indication of the current trend in the daily number of fatalities. Also, as is seen in Figure 9 and Figure 10, predictions made for dates further back in time relative to now are less uncertain, since a greater proportions of fatalities occurred then have already been reported.

The results relating to the average scores as well as simple and relative errors are presented in Table 6.

One may immediately be struck by the fact that the average log-scores evaluate to infinity for all models when including predictions from at least 14 days in the past. The reason for this is in fact rather straight-forward, once one recalls the peculiarity of the Swedish data presented in section 2 that fatalities may be de-classified as such. We made no attempts in section 3.2 in accounting for this, nor do Altmejd et al. (2021b), but Altmejd et al. (2020) do point out that it might be reasonable to do so. At any case, both methods yield a lower bound for the predicted fatality count on day t, corresponding to the observed fatality count on day T, since, according to the models, the eventually observed number of fatalities on day t can only be greater than the number observed on day T. Thus, if sufficiently many de-classifications occur such that the eventually observed number of fatalities on day t is less than the number observed on day

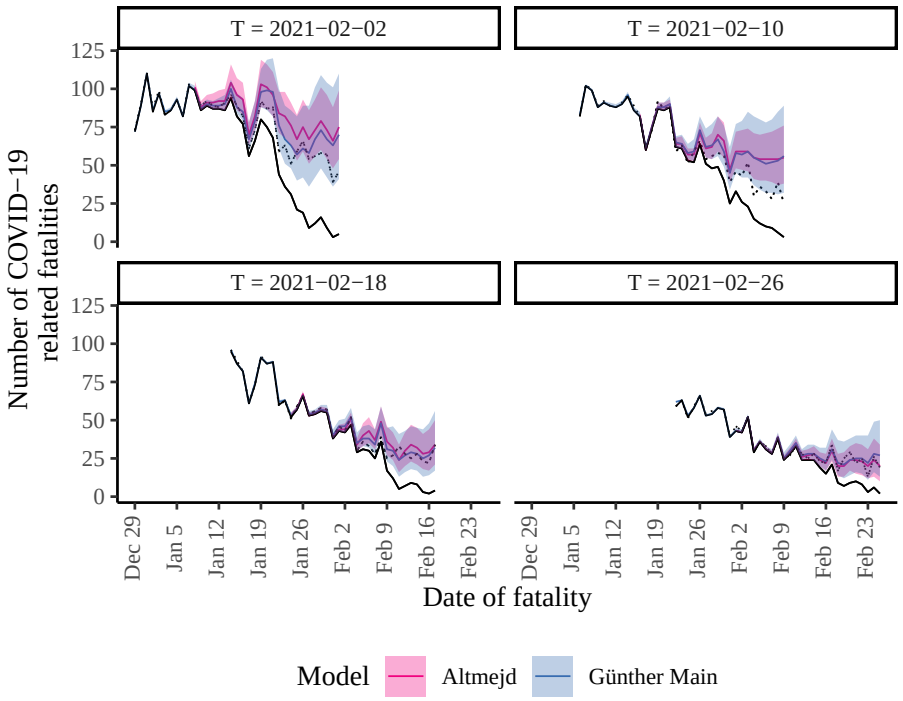

Figure 9: Nowcasts for the Altmejd and Günther Main for four different instances of T. The coloured lines correspond to the median of the predictive distribution, and the filled ribbons correspond to 95% equal-tail prediction intervals. The observed number of fatalities observed on day T is indicated by a black solid line, whereas the retrospective true number is indicated by a dotted line (using data as of May 7, 2021).

T, the probability of the realized value is, by definition, zero, whence the log-score evaluates to infinity. While the instances of observable de-classifications are rather rare (cf. Fig. 3), it suffices that only one of the instances of T include a de-classification as to induce the average score to evaluate to infinity. We mention finally that, since we defined $n_{td} \geq 0$ in section 3.1, even if we observe a de-classification in the "raw" data, we do not incorporate this in the observed fatality count.

The other evaluation metrics do not suffer from the scourge of de-classifications, producing comparable numbers for all instances of model and δ. The squared error score was computed using two different point estimates, while the absolute, simple and relative errors were computed using only the median point estimate. We see that the Günther Main and Weekday models produce very similar results, which are in turn better than those produced by Altmejd, expect for the relative error, where all three methods perform similarly. They perform

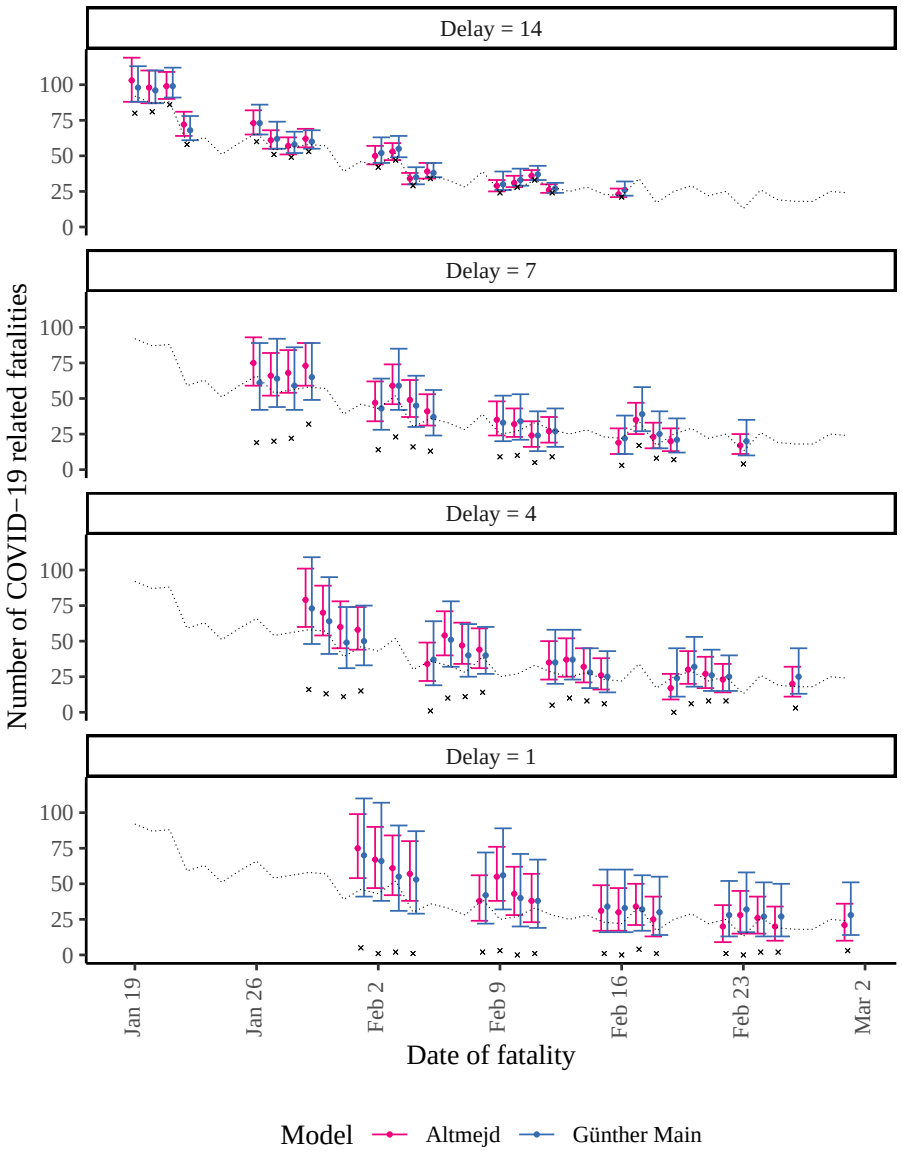

Figure 10: Nowcasts for N_{T-d}, where T is varied over reporting days between 2 February and 2 March, 2020, for $d \in \{1, 4, 7, 14\}$. The retrospective true number of fatalities is indicated by a dotted line (using data as of May 7, 2020), and the observed count for each T are indicated by $\times$.

Model	$d \leq$	logS	RPS	SES[1]	SES[2]	AES[2]	err[2]	relErr[2]
Altmejd	7	4.09	6.97	151.77	146.37	12.10	8.02	0.25
	14	∞	6.46	123.30	119.65	10.94	7.71	0.21
	21	∞	4.81	85.37	82.88	9.10	5.57	0.15
G. Main	7	3.61	5.67	124.46	107.34	10.36	6.88	0.25
	14	∞	5.07	93.32	80.74	8.99	6.17	0.19
	21	∞	3.89	65.32	56.62	7.52	4.62	0.14
G. Weekday	7	3.58	5.60	121.71	104.42	10.22	6.79	0.24
	14	∞	5.01	91.61	79.09	8.89	6.11	0.19
	21	∞	3.85	64.19	55.45	7.45	4.56	0.14
G. Naïve	7	4.23	13.98	914.18	567.96	23.83	20.07	0.66
	14	∞	11.76	661.81	406.97	20.17	16.55	0.49
	21	∞	8.94	461.59	283.10	16.83	12.52	0.35

[1] using the mean of the predictive distribution

[2] using the median of the prediction distribution

Table 6: Average scores, error and relative error for the nowcasting methods, applied for T varying over reporting days in the period from February 2 until March 2, 2021. For each instance of T, averages are computed over $t \in \{T-1, T-2, \ldots, T-\delta\}$, where $\delta \in \{7, 14, 21\}$, indicated by the second column above.

Model	$d \leq$	Coverage			Width		
		95%	80%	50%	95%	80%	50%
Altmejd	7	0.76	0.55	0.32	27.61	17.75	9.42
	14	0.69	0.46	0.25	22.93	14.83	7.85
	21	0.75	0.55	0.36	17.55	11.45	6.16
Günther Main	7	0.95	0.78	0.50	39.62	25.53	13.37
	14	0.87	0.66	0.42	31.22	20.15	10.56
	21	0.88	0.69	0.45	23.45	15.18	7.95
Günther Weekday	7	0.95	0.79	0.48	39.59	25.56	13.32
	14	0.87	0.67	0.40	31.22	20.15	10.53
	21	0.87	0.70	0.43	23.44	15.18	7.93
Günther Naïve	7	0.99	0.77	0.35	116.63	70.90	35.89
	14	0.91	0.66	0.30	94.37	57.02	28.71
	21	0.90	0.67	0.33	71.30	42.93	21.58

Table 7: Coverage and average width of 50%, 80% and 95% equal-tailed prediction intervals for the nowcasting methods, applied for T varying over reporting days in the period from February 2 until March 2, 2021. For each instance of T, averages are computed over $t \in \{T-1, T-2, \ldots, T-\delta\}$, where $\delta \in \{7, 14, 21\}$, indicated by the second column above.

similarly for the relative error, because the relative error of the Altmejd method is less for more recent instances of T, and greater for earlier instances, whereas the relative error of the Günther Main and Weekday methods is more constant for all instances of T. This can be seen in Figure 12, and we provide more detail relating to this in section 5.2.3.

Notably, all methods seem to have a bias toward greater fatality counts, since the simple and relative errors are greater than zero. This may be because the data used for the likelihood encompass the end-of-year holiday period, during which the reporting was generally later (cf. Fig. 6 and 7), and having not taken this into account yield the nowcasting models to expect late reporting also for the predicted fatality counts.

It is immensely apparent that one should take into account that certain days are not reporting days, since the Günther Naïve model performs worse by any metric. In particular, since it assumes the by-definition-zero instances of n_{td} for $t + d$ being a non-reporting day to be zero due to randomness, it overestimates the variation of the data, and thereby produce very wide predictive distributions for the fatality counts N_t. This is clearly seen in Table 7, where the average width of the 50% prediction intervals of the Günther Naïve model is greater than the average width of 95% prediction intervals of the Altmejd method, which produces the narrowest prediction intervals. These turn out to be too narrow, however, since their coverage is less than what is intended for all instances of δ. The Günther models that take into account that there exist non-reporting days perform remarkably well with respect to the coverage metric. In particular, for $\delta = 7$, they are almost spot-on. Averaging over longer delays, i.e. for $\delta \in \{14, 21\}$, produces slightly too narrow prediction intervals, but still they are rather close to their intended coverage. Note that the Günther Naïve model, while producing very wide 50% prediction intervals, does not manage to have these encompass the realized value, with a coverage only barely above 0.30.

5.2.3 Detailed results of evaluation

The results of Table 6 and Table 7 are rather summary. Let us have a look at the performance of the nowcasting methods when we average over more specific subsets, in order to evaluate the performances of the nowcasting methods for specific delays $d \in \{1, 2, \dots 35\}$ and specific instances of today T varying over reporting days from March 2 until February 2, 2021.

Let us begin with specific delays. In Figure 11 we average over the different instances of T for each instance of delay $d \in \{1, 2, \dots, 35\}$ separately. Doing so allows us to evaluate the performance of the methods specific to how far back in time the prediction is done. Note that the log-score evaluates to infinity for delays equal to or greater than 14 days, due to the aforementioned reasons. Unsurprisingly, all metrics, except the log-score, indicate that the models perform better for predictions of fatality counts further back in time, since a larger proportion of these fatalities have already been reported.

There is however a noticeable upward "bump" around delays of $d \in \{9, 10, 11\}$.

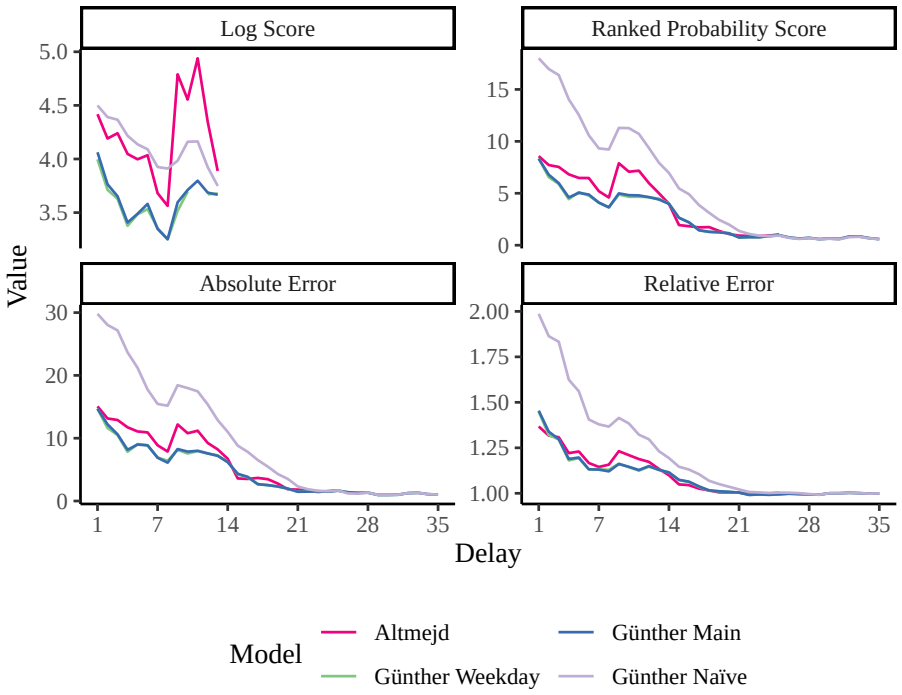

Figure 11: Log-score, ranked probability score, as well as absolute and relative error when using the median as point estimate for different models and delays. For each instance of delay $\in \{1, 2, \ldots, 35\}$, the score is computed for the 17 instances of "now" consisting of reporting days between 2 February and 2 March, 2021. Note that the "Günther Main" and "Günther Weekday" models are generally not distinguishable from one another in this figure, and that the "Altmejd" model has a maximum delay of $D = 25$, whereas the Günther models have a maximum delay of $D = 35$.

This is particularly noticeable for the log-score of the Altmejd method, which produces its worst scores for these instances of d. While higher scores and errors are generally correlated with earlier dates, when the daily fatality counts were greater relatively to more recent dates, we have not found any other systematic cause of the bump. Higher scores are generally correlated to higher fatality counts, since the predictive distribution is wider for high fatality counts (cf. Figs 9 and 10). We note, however, that in the Altmejd log-score case, there are a few outliers, consisting of non-averaged log-scores of approximately 10. In particular, for now being February 2, 2021, the Altmejd method produces log-scores slightly above 10 for the forecasts for January 22 and 24, 2021. This can be seen in the top left panel of Figure 9, where the Altmejd forecast greatly overestimates the number of fatalities occurred during these dates.

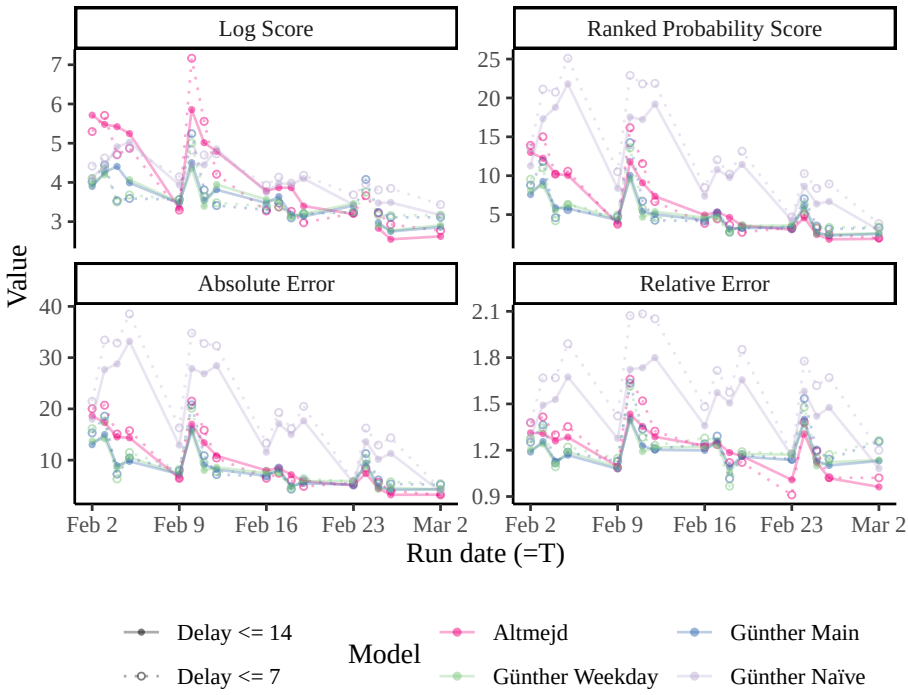

Figure 12: Log-score, ranked probability score, as well as absolute and relative error when using the median as point estimate for different models and delays. For each instance of T consisting of reporting days between February 2 and March 2, 2020, the average score is computed over delays of either less than 7 or less than 14.

We note that the "bump" is not an as distinguished feature for the average scores and errors of the Günther Main and Weekday methods, however.

In Figure 12 we instead average over $\{N_{T-1}, N_{T-2}, \ldots, N_{T-\delta}\}$ for $\delta \in \{7, 14\}$, similarly to Table 6 and 7, but for each instance of T separately. We see that the relative error is relatively constant for different T, while the other measures make seem the model to perform better for later instances of T. But if we compare this with Figure 9 and Figure 10, we may notice that later instances of T are subject to fewer fatality counts, which necessarily produce sharper predictive distributions, since we are concerned with count data. Furthermore, we see in Figure 12 that the performance of the Günther Naïve model seems to be dependent on the weekday of T, since Tuesdays systematically produces lower scores than the other weekdays. This may be because Tuesdays is the first reporting day of the week, following a three-day hiatus.

We also note that for $\delta = 14$, the average log-score evaluates to infinity for all models for "now" being February 24. This is due to the observed fatality count

for February 10, as of February 24, constituting a delay of 14 days, was 28, but the fatality count for February 10 as of May 7 is 27, due to a de-classification on March 2. Additionally, the Günter Weekday model fails to produce any sample with the true number of fatalities on January 24 for "now" being February 4, corresponding to a delay of 11 days, despite the true fatality count being higher that the observed fatality count, yielding an infinite log-score for this instance. For comparison, the Günther Main model produces a single sample with the true value for this instance. The reason for this is that very few fatalities that occurred on January 24 were reported with a delay of 12 days or longer, and since 2 fatalities were furthermore observably de-classified as being COVID-19 related. Thus, the nowcasting models produce particularly high over-estimations for the number of fatalities for this instance of T and delay. Specifically, in the fully observed upper rectangle of the reporting triangle used for this instance of T, 47.4% of fatalities were reported with a delay of 12 days or more, but the difference between the true fatality count (51) for January 24 and the observed count by February 2 (49) only constitutes 3.92% of the true count.

Notably, the Altmejd method outperforms the Günther methods during the last 5 instances of T by the metrics of Figure 12, corresponding to February 23 until March 2, 2021. For this reason, we look at this period a little closer. In Table 8 we present the evaluation metrics computed for said instances of T. Here we only compare Altmejd to Günther Main. The scoring rules of the top panel of Table 8 only confirm what we see in Figure 12, that the Altmejd method performs better. But the absolute error in conjunction to the simple error shows something more subtle: The average absolute error of the Altmejd method is rather close to the average absolute error of the Günther method, but the average simple error of the Altmejd method is, relatively, much less than the simple error of the Günther method. Hence, while both methods apparently still produce a positively biased median point estimate for the number of daily fatalities, the Altmejd method produces a noticeably less biased estimate. In our implementation of the Günther Main model, we used a moving window of $m = 70$ days. This includes the end-of-year holiday period for which we found a noticeable increase in the reporting delay. For the latest implemented instance of "now", March 2, 2021, at that time reported data for fatalities occurred on or after December 22, 2021, and later are used. It may be that the Altmejd method–somehow–puts less weight on so temporally distant fatalities, e.g. by a narrower moving window $m < 70$. As previously mentioned (in sec. 3.1.1), we did not find a moving window equivalent in the code from Altmejd et al. (2021b), however.

Since the daily fatality counts were generally decreasing during the considered instances of "now", e.g. on a weekly scale, these five instances roughly correspond to the instances of "now" for which the daily fatality counts were the least. This may also be a part of the reason for it performing better during this period of time. The Altmejd method generally produces sharper predictive distributions compared to the Günther methods, and the log-score and ranked probability score give better scores to shaper predictive distributions. In the lower panel of Table 8, we see that, again, the Altmejd method produces too

Model	$d \leq$	logS	RPS	SES[1]	SES[2]	AES[2]	err[2]	relErr[2]
Altmejd	7	3.16	3.20	33.19	32.31	5.68	1.11	0.11
	14	∞	2.76	24.83	24.24	4.92	1.30	0.08
	21	∞	2.06	17.13	16.72	4.09	1.01	0.06
G. Main	7	3.35	4.03	59.52	49.26	7.02	4.29	0.25
	14	∞	3.40	42.11	35.09	5.92	3.43	0.18
	21	∞	2.57	29.22	24.31	4.93	2.62	0.13

[1] using the mean of the predictive distribution
[2] using the median of the prediction distribution

Model	$d \leq$	Coverage			Width		
		95%	80%	50%	95%	80%	50%
Altmejd	7	0.91	0.77	0.54	21.74	13.94	7.40
	14	0.89	0.73	0.46	17.06	11.10	5.87
	21	0.90	0.76	0.55	12.64	8.30	4.51
G. Main	7	0.94	0.80	0.60	31.86	20.43	10.71
	14	0.87	0.74	0.54	24.30	15.63	8.24
	21	0.89	0.78	0.56	18.01	11.68	6.18

Table 8: Evaluation metrics for the Altmejd and Günther Main nowcasting methods, applied for T varying over reporting days in the period from February 23 until March 2, 2021. This corresponds to the last five instances for which the nowcasting methods are implemented. For each instance of T, averages are computed over $t \in \{T-1, T-2, \ldots, T-\delta\}$, where $\delta \in \{7, 14, 21\}$, indicated by the second column above.

narrow prediction intervals, and that the Günther Main prediction intervals have a more accurate coverage.

Lastly, we take a look at the performance of 95%, 80% and 50% equal-tailed prediction intervals for the different nowcasting methods, when considering specific delays as was done in Figure 11. This is illustrated in Figure 13. In the lower panel, we see the widths of the prediction intervals, which consistently behave in agreement with our conclusions about the prediction intervals in our discussion relating Table 7 in the last paragraph of section 5.2.2. Note that we only average over 17 instances for each delay, here. For all methods, there appears to be an downward "bump" somewhere between a delay of 7 and 14 days. This could be compared with the upward bump for similar delays in Figure 11, and in this case we have not found any particular systematic reason for the decline in coverage either.

5.3 Conclusion

In this section, we have come to three main conclusions. First, we have found that the model of Günther et al. (2020a), when appropriately modified for Swedish fatality data, performs better than the method of Altmejd et al. (2021b) by almost every measure. In particular, the Günther Main and Weekday models

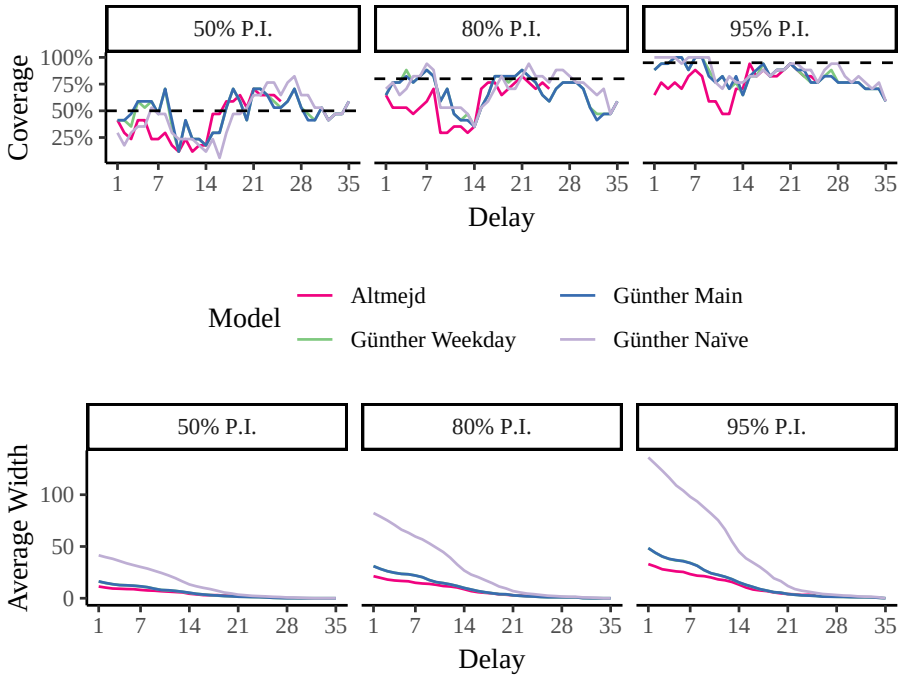

Figure 13: **Top:** Coverage of 50%, 80% and 95% equal-tailed prediction intervals with respect to delay for different models. **Bottom:** Average width of 50%, 80% and 95% prediction intervals with respect to delay for different models. For each instance of delay $\in \{1, 2, \ldots, 35\}$, the score is computed for the 17 instances of "now" consisting of reporting days between 2 February and 2 March, 2021. Note that the "Günther Main" and "Günther Weekday" models are generally not distinguishable from one another in this figure, and that the "Altmejd" model has a maximum delay of $D = 25$, whereas the Günther models have a maximum delay of $D = 35$.

produce surprisingly accurate prediction intervals.

Our second conclusion is, however, perhaps our most apparent conclusion: One should take into account that certain days are not reporting days. The Günther Naïve model performed worse than all other models.

The third conclusion relates to the realization that the Altmejd method performs better for the five most recent instances of "now" we considered. We found that the Günter methods consistently produced positively biased forecasts, due to the data for the likelihood including the end-of-year holiday period of 2020, for which reporting delays were generally longer, while the bias of the Altmejd method decreased as a function of "now". It may hence be prudent to take into account that the reporting distribution changes over time.

6 Discussion

In this thesis, we have seen that nowcasting methods may produce good estimates for the number of daily COVID-19 related fatality counts in Sweden. In particular, we found that the method of Günther et al. (2020a), when accounting for non-reporting days, produces probabilistic distributions with a surprisingly accurate coverage of the prediction intervals. An ensemble of proper scoring rules accompanied by other evaluative metrics have proved informative in comparing different nowcasting methods to one another for the period of "now" varying from February 2 until March 2, 2021.

The fact that certain days incur no new fatality reports, due to the publishing pattern of the Public Health Agency, proved important to take into account; However, additionally taking the reporting weekday into account offered barely noticeable improvements in terms of the accuracy of the nowcasts.

However, we considered a rather narrow slice of time, with our retrospective "now" only spanning over the course of one month. We noted that the method of Altmejd et al. (2021b) performed better in terms of scoring rules during the latter five "nows", out of 17 considered. This coincides with the period during which the daily fatality counts were the least. It might be the case that the method of Altmejd et al. (2021b) is better at nowcasting counts of low order compared to the method of Günther et al. (2020a). If we had evaluated the nowcasting methods for a broader slice of time, then we would perhaps be able to say something more definite about this. However, computer run-times of approximately 40 minutes per model per instance of "now" prompted us to consider only 17 instances of "now". We would also have benefited from a written-down version of the Altmejd method by its authors, Altmejd et al. (2021b), as to more properly investigate the reason for the better performance of the method for these instances of "now".

6.1 Possible improvements

We noted that all of the nowcasting methods produced positively biased nowcasts for all instances of T, and concluded that this is due to the data used for the inference included the end of December and beginning of January, during which several public holidays occur, and the reporting delay was observably longer during this period. It is possible to include a more general time-of-reporting element in the discrete hazard regression model of Günther et al. (2020a), which might have had remedied this bias. For instance, Günther et al. (2020a), in their implementation, use a linear spline with break points every two weeks as a time-of-reporting effect. The time between break points is another model parameter to be considered, however, and unfortunately we did not have time to include an implementation modified for the Swedish fatality data. Particularly, break points every two weeks is likely to be too often, since the daily fatality counts are only assumed to be fully observed after a delay of five weeks. As such, a hastening in the reporting delay would not be "noticed", and instead would be attributed in the model to a higher number of daily fatalities. Another, simple,

option would be to add a categorical effect for the reporting period during the end of December and beginning of January.

A particularity of the Swedish data is that fatalities that had previously been classified as being COVID-19 related can be de-classified as such. We did not take this into account in our implementation of Günther et al. (2020a), nor do Altmejd et al. (2021b) in their method. A straight-forward way of taking this into account would be to add some negatively oriented binomial noise to the fatality counts reflecting that each fatality has a probability of being de-classified as COVID-19 related. That is, having predicted N_t fatalities in the original model, e.g. that of Günther et al. (2020a), one would quote $N_t' \sim \mathit{Binom}(N_t, 1 - p_{\mathrm{dC}})$ as the final predictive distribution, where p_{dC} is the probability of eventual de-classification. The rarity and long delay of de-classifications, however, poses a challenge with reliably estimating p_{dC}.

We mentioned in section 2 that the data set collected from FHM (2020-21) by Altmejd et al. (2021b) contains other indicators as to describe the dynamics of the COVID-19 outbreak and its impact in the population. These indicators include the number of daily newly confirmed cases of COVID-19 and new admissions to the intensive care unit (ICU) related to COVID-19. Since an increase, or decrease, in the number of daily COVID-19 related fatalities is necessarily preceded by a corresponding increase, or decrease, in the number of cases of COVID-19, and since a proportion of COVID-19 related fatalities occur in ICU admission, it might be possible to incorporate these other indicators in the now-casting model for fatalities, as to more accurately and quickly notice changes in the current trend.

6.2 Closing remarks

The Public Health Agency of Sweden publishes the time series for the daily number of COVID-19 related fatalities that have been reported to them at the time of publishing by the date of their occurrence. Bergholtz et al. (2020) argue that, because there is a delay in reporting, and since the most recent days are the most under-reported, this time series will always have it appear as though the number of daily fatalities is currently decreasing. As we saw in Figure 1, this is certainly the case, and it may give the wrong impression as to the current trend of the number of daily fatalities. Bergholtz et al. (2020) further argue that it would be better to report the time series for the number of daily COVID-19 related fatalities by the date of reporting. The number of fatalities reported per reporting occasion depends greatly on the weekday of the report, however, as we saw in Figure 4. Hence, some kind of smoothing would be necessary in this case, such as the 7-day average fatality count. This may be a good indicator for the current trend if the reporting delay is relatively short, e.g. with most fatalities being reported with a delay of a week or less, but we saw in section 2.2 that this is not generally the case. For example, during the period from September 14, 2020 until only 37.4% of fatalities were reported with a delay of one week or less. Also, a proper interpretation would require the reporting delay to be relatively stable over time, which we have seen is also not generally

the case.

Nowcasting offers an alternate way of presenting the current trend in the daily number of COVID-19 related fatalities, while still publishing the time series for the number of daily COVID-19 related fatalities by the date of their occurrence. Nowcasting does not rely on the reporting delay being short to provide timely estimates for the current number of daily fatalities, and the model of Günther et al. (2020a) provides the possibility to incorporate changes in the reporting delay distribution in the nowcasting model. As such, nowcasting may be more informative as to the current trend in the number of daily COVID-19 related fatalities. However, since the nowcast is a prediction, it includes an uncertainty, which may be a challenge to communicate to the public.

References

A. Altmejd, J. Rocklöv, and J. Wallin. Nowcasting Covid-19 statistics reported with delay: a case-study of Sweden. *Under review at Scientific Reports*, 2020. URL https://arxiv.org/abs/2006.06840.

A. Altmejd, J. Rocklöv, and J. Wallin. Adam Altmejd's COVID-19 statistics web page (retrieved 2021-05-07). 2021a. URL https://adamaltmejd.se/covid/.

A. Altmejd, J. Rocklöv, and J. Wallin. Adam Altmejd's GitHub covid-repository (retrieved 2021-05-07). 2021b. URL https://github.com/adamaltmejd/covid/.

E. J. Bergholtz, N. Brusselaers, L. Einhorn, A. Ewing, Å. Gustafsson, Å. Lundkvist, J. Stilhoff Sörensen, J. Svensson, and V. Anders. Opinion piece in Dagens Nyheter newspaper: Svensk coronastatistik ger en skev bild av smittspridningen. *Dagens Nyheter*, 2020. URL https://www.dn.se/debatt/svensk-coronastatistik-ger-en-skev-bild-av-smittspridningen/.

T. M. Cover and J. A. Thomas. Elements of Information Theory, Second Edition. *Wiley*, 2006.

C. Czado, T. Gneiting, and L. Held. Predictive Model Assessment for Count Data. *Biometrics*, 65(4):1254–1261, 2009. URL https://doi.org/10.1111/j.1541-0420.2009.01191.x.

FHM. The Public Health Authority of Sweden's COVID-19 data portal (latest retrieved 2021-05-07). 2020-21. URL https://www.folkhalsomyndigheten.se/smittskydd-beredskap/utbrott/aktuella-utbrott/covid-19/statistik-och-analyser/.

FHM. Jämförelse av olika mått på COVID-19-dödsfall (retrieved 2021-05-07). 2020a. URL https://www.folkhalsomyndigheten.se/contentassets/4b4dd8c7e15d48d2be744248794d1438/jamforelse-av-olika-matt-pa-covid-19-dodsfall.pdf.

FHM. Press release relating to changes in the Public Health Agency of Sweden's publication pattern. 2020b. URL https://www.folkhalsomyndigheten.se/nyheter-och-press/nyhetsarkiv/2020/september/forandrad-rapportering-for-statistiken-av-covid-19/.

F. Günther. Felix Günther's GitHub covid repository (retrieved 2021-02-19). 2020b. URL https://github.com/FelixGuenther/nc_covid19_bavaria.

F. Günther, A. Bender, K. Katz, H. Küchenhoff, and M. Höhle. Nowcasting the COVID-19 pandemic in Bavaria. *Biometric Journal*, 63:490–502, 2020a. URL https://doi.org/10.1002/bimj.202000112.

M. Höhle and an der Heiden. Bayesian Nowcasting during the STEC O104:H4 Outbreak in Germany, 2011. *Biometrics*, 70(4):993–1002, 2014. URL http://dx.doi.org/10.1111/biom.12194.

J. F. Lawless. Adjustments for Reporting Delays and the Prediction of Occurred but Not Reported Events. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, 22(1):15–31, 1994. URL https://www.jstor.org/stable/3315820.

T. Mack. Distribution-Free Calculation of the Standard Error of Chain Ladder Reserve Estimates. *ASTIN Bulletin*, 23(2):213–225, 1993. URL https://doi.org/10.2143/AST.23.2.2005092.

S. F. McGough, M. A. Johansson, M. Lipsitch, and N. A. Menzies. Nowcasting by Bayesian Smoothing: A flexible, generalizable model for real-time epidemic tracking. *PLoS Copmuational Biology*, 2020. URL https://doi.org/10.1371/journal.pcbi.1007735.

T. Mikosch. Non-Life Insurance Mathematics: An Introduction with the Poisson Process, Second Edition. *Springer*, 2009.

W. M. Morgan and J. W. Curran. Acquired Immunodeficiency Syndrome: Current and Future Trends. *Public Health Reports*, 101(5):459–465, 1986. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1477784.

K. H. Pollock. Modeling Capture, Recapture, and Removal Statistics for Estimation of Demographic Parameters for Fish and Wildlife Populations: Past, Present Future. *Journal of the American Statistical Association.*, 86(413):225–238, 1991. URL https://www.jstor.org/stable/2289733.

R Core Team. R: A Language and Environment for Statistical Computing. 2021. URL https://www.R-project.org/.

Soc. The National Board of Health and Welfare's COVID-19 fatality data portal (retrieved 2021-05-07). 2021a. URL https://www.socialstyrelsen.se/statistik-och-data/statistik/statistik-om-covid-19/statistik-over-antal-avlidna-i-covid-19/.

Soc. Datakällor för avlidna i COVID-19 (retrieved 2021-05-07). 2021b. URL https://www.socialstyrelsen.se/statistik-och-data/statistik/statistik-om-covid-19/statistik-over-antal-avlidna-i-covid-19/datakallor-for-avlidna-i-covid-19/.

Stan Development Team. Stan Modeling Language Users Guide and Reference Manual, Version 2.21.0. 2019a. URL https://mc-stan.org.

Stan Development Team. RStan: the R interface to Stan, R package version 2.21.0. 2019b. URL http://mc-stan.org/.

O. Stoner and T. Economou. Multivariate hierarchical frameworks for modeling delayed reporting in count data. *Biometrics*, 76:789–798, 2019. URL https://doi.org/10.1111/biom.13188.

Wikipedia. Matérn covariance function wiki (retreived 2021-05-07). 2021. URL https://en.wikipedia.org/wiki/Mat%C3%A9rn_covariance_function.

S. N. Wolf. Generalized Additive Models: An Introduction with R, Second Edition. *Taylor & Francis Group*, 2017.