Evaluating Semi-Parametric Nowcasts of COVID-19 Hospital Admissions in Germany

Authors

S. Abbott (1), S. Funk (1)

Correspondence to: sam.abbott@lshtm.ac.uk

Affiliations

1. Center for the Mathematical Modelling of Infectious Diseases, London School of Hygiene & Tropical Medicine, London WC1E 7HT, United Kingdom

Abstract

TODO: Summarise the project in a structured abstract (background, methods, results and conclusions)

Introduction

COVID-19 hospitalisations in Germany are released by date of positive test rather than by date of admission. This has some advantages when they are used as a tool for surveillance as these data are closer to the date of infection and so easier to link to underlying transmission dynamics and public health interventions. Unfortunately, however, when released in this way the latest data are right-censored meaning that final hospitalisations for a given day are initially underreported. This issue is often found in data sets used for the surveillance of infectious diseases and can lead to delayed or biased decision making. Fortunately, when data from a series of days is available we can estimate the level of censoring and provide estimates for the truncated hospitalisations adjusted for truncation with appropriate uncertainty. This is usually known as a nowcast.

TODO: What else has been done for nowcasting

In this work, we aim to evaluate a series of novel semi-parametric nowcasting model formulations in real-time and provide an example workflow to allow others to do similarly using German COVID-19 hospitalisations by date of positive test at the national level both overall and by age group, and at the state level. This project is part of a wider collaboration assessing a range of nowcasting methods whilst providing an ensemble nowcast of COVID-19 Hospital admissions in Germany by date of positive test. This ensemble should be used for any policy-related work rather than the nowcasts provided in this repository. See here for more on this nowcasting collaboration.

Methods

Data

TODO: Data source and processing details

Models

In the following sections we provide methodological and implementation details for the nowcasting framework implemented in `epinowcast` and applied here. Our approach is an extension of that proposed by Günther et al.[1] which was itself an extension of the model proposed by Höhle and Heiden[2] and implemented in the `surveillance` R package[3]. Compared to the model proposed in Günther et al.[1], `epinowcast` adds an optional parametric assumption for the underlying delay from occurrence to report, support for jointly nowcasting multiple related datasets, a flexible formula interface allowing for the specification of a large range of models, and an efficient implementation in `stan` which makes use of sparse design matrices and within chain parallelisation to reduce runtimes.[4,5]

Decomposition into expected final notifications and report delay components

We follow the approach of Höhle and Heiden[2] and consider the distribution of notifications (n_{gtd}) by time of occurrence (t) and reporting delay (d) conditional on the final observed count N_{gt} for each dataset (g) such that,

$$N_{gt} = \sum_{d=0}^{D} n_{gtd} \tag{1}$$

where D represents the maximum delay between time of occurrence and time of report which in theory could be infinite but in practice we set to a value in order to make the model identifiable and computationally feasible. This formulation means that $n_{gtd} \mid N_{gt}$ is multinomial with a probability vector (p_{gtd}) of length D for each t and g that needs to be estimated at the same time as estimating the expected number of final notifications $\mathbb{E}[N_{gt}] = \lambda_{gt}$.

An alternative approach, not explored here, is to consider each n_{gtd} independently at which point the model can be defined as a standard regression with the appropriate observation model and adjustment for reporting delay (i.e it becomes a Poisson or Negative Binomial regression)[6]. An implementation of this approach is available in Bastos et al.[6]. More work needs to be done to evaluate which of these approaches produces more accurate nowcasts for epidemiological count data.

Expected final notifications

Here we follow the approach of Günther et al.[1] and specify the model for expected final notifications as a first order random walk. This model can in principle be any model such as a more complex time-series approach, a gaussian process, or a mechanistic or semi-mechanistic compartmental model. Extending the flexibility of this model is an area of further work as is evaluating the benefits and tradeoffs of more complex approaches.

$$\log(\lambda_{gt}) \sim \text{Normal}\left(\log(\lambda_{gt-1}), \sigma_g^{\lambda}\right) \tag{2}$$

$$\log(\lambda_{g0}) \sim \text{Normal}\left(\log(N_{g0}), 1\right) \tag{3}$$

$$\sigma_g^{\lambda} \sim \text{Half-Normal}\left(0, 1\right) \tag{4}$$

Delay distribution

Again following the approach of Günther et al.[1] we define the delay distribution (p_{gtd}) as a discrete time hazard model ($h_{gtd} = \mathrm{P}\left(\text{delay} = d | \text{delay} \geq d, W_{gtd}\right)$) but we extend this model to decompose W_{gtd} into 3 components: hazard derived from a parametric delay distribution (γ_{gtd}) dependent on covariates at the date of occurrence, hazard not derived from a parametric distribution (δ_{gtd}) dependent on covariates at the date of occurrence, and hazard dependent on covariates referenced to the date of report (ϵ_{gtd}).

For first component (γ_{gtd}) we assume that the probability of reporting p'_{gtd} on a given date given follow a parametric distribution (in the baseline case a discretised log normal distribution) with the log mean and log standard deviation being defined using an intercept and arbitrary shared, reference date indexed, covariates with fixed (α_i) and random (β_i) coefficients,

$$p'_{gtd} \sim \text{LogNormal}\left(\mu_{gt}, \upsilon_{gt}\right) \tag{5}$$
$$\mu_{gt} = \mu_0 + \alpha_\mu X_\gamma + \beta_\mu Z_\gamma \tag{6}$$
$$\upsilon_{gt} = \exp\left(\upsilon_0 + \alpha_\upsilon X_\gamma + \beta_\upsilon Z_\gamma\right) \tag{7}$$

The parametric logit hazard (probability of report on a given date conditional on not already having reported) for this component of the model is then,

$$\gamma_{gtd} = \text{logit}\left(\frac{p'_{gtd}}{\left(1 - \sum_{d'=0}^{d-1} p'_{gtd'}\right)}\right) \tag{8}$$

The non-distributional logit hazarad components for the date of occurrence and report are then again defined using an intercept and arbitrary shared covariates with fixed (α_i) and random (β_i) coefficients.

$$\delta_{gtd} = \mu_0 + \alpha_\delta X_\delta + \beta_\delta Z_\delta \tag{9}$$
$$\epsilon_{gtd} = \epsilon_0 + \alpha_\epsilon X_\epsilon + \beta_\epsilon Z_\epsilon \tag{10}$$

The overall hazard for each group, occurrence time, and delay is then,

$$\text{logit}(h_{gtd}) = \gamma_{gtd} + \delta_{gtd} + \upsilon_{gtd},\ h_{gtD} = 1 \tag{11}$$

where the hazard on the final day has been assumed to be 1 in order to enforce the constraint that all reported observations are reported within the specified maximum delay. The probability of report for a given delay, occurrence date, and group is then as follows,

$$p_{gt0} = h_{gt0},\ p_{gtd} = \left(1 - \sum_{d'=0}^{d-1} p_{gtd'}\right) \times h_{gtd} \tag{12}$$

All (α_i) and random (β_i) coefficients have standard normal priors by default with standard half-normal priors for pooled standard deviations.

Observation model and nowcast

Expected notifications by time of occurrence (t) and reporting delay can now be found by multiplying expected final notifications for each t with the probability of reporting for each day of delay (p_{gtd}). We then assume a negative binomial observation model with a joint overdispersion parameter (with a 1 over square root standard half normal prior) and produce a nowcast of final observed notifications at each occurrence time by summing posterior estimates for each observed notification for that occurrence time.

$$n_{gtd} \mid \lambda_{gt}, p_{gtd} \sim \text{NB}(\lambda_{gt} \times p_{gtd}, \phi),\ t = 1, ..., T. \tag{13}$$

$$\frac{1}{\phi^2} \sim \text{Half-Normal}(0, 1) \tag{14}$$

$$N_{gt} = \sum_{d=0}^{D} n_{gtd} \tag{15}$$

Specific model formulations

TODO: Detail specific model formulations

We explore two primary models and submit nowcasts from these models to the nowcasting hub. The first of these is fit independently to each data set by age and location. Hospitalisations are modelled using a random walk on the log scale. Reporting delays are then modelled parametrically using a lognormal distribution with the log mean and log standard deviation each modelled using a weekly random walk with a pooled standard deviation and a random effect for the day of the week (introduced on the 6th of December 2021) with public holidays assumed to be reported like Sundays. Report date effects are again modelled using a random effect for day of the week with public holidays assumed to be reported like Sundays. The second model is fit jointly to age groups but is otherwise structured in the same way as the unpooled model except that report day of the week effects and the observation overdispersion are assumed to be joint across age groups, age groups are assumed to have a random intercept for both the log mean and the log standard deviation of the reporting delay distribution, and there is no random effect for reference day of the week. We also consider a series of pooled models which sequentially include the features of our most complex model. These are: age groups are fit jointly, day of the week reporting effects, a random intercept for each age group, and a random walk by positive test week shared across age groups.

Model | Formula Reference: Fixed, Report: Fixed | Reference: Fixed, Report: Day of week | Reference: Age, Report: Day of week | Reference: Age and week, Report: Day of week | Reference: Age and week by age, Report: Day of week | Independent by age, Reference: Week, Report: Day of week | Independent by age, Reference: Week and day of week, Report: Day of week |

Implementation

The model is implemented as part of the `epinowcast` R package[7] using `stan` and `cmdstanr`[4,5]. Sparse design matrices have been used for all covariates to limit the number of probability mass functions that need to be calculated. `epinowcast` incorporates additional functionality written in R[8] to enable plotting nowcasts and posterior predictions, summarising nowcasts, and scoring them using `scoringutils`[9]. All functionality is modular allowing users to extend and alter the underlying model whilst continuing to use the package framework.

Evaluation

TODO: Model evaluation. What targets. What scoring measures. What aggregation TODO: Evaluation of performance for submitted models vs others submitted to forecasting hub.

TODO: Add links to nowcasting hub and evaluation against those models.

We evaluate these models first visually across a range of nowcasting dates and then quantitatively using proper scoring rules[9] on both the natural and log scales (corresponding to absolute and relative performance) aggregating scores first across all targets and then stratifying in turn by age group, nowcast horizon, date of postive test, and date of report. We also explore other aspects of our models performance by highlighting models that have problematic fitting diagnostics and summarising the estimation time for each model. We provide a report of this evaluation that is updated in real-time as new data and nowcasts become available.

Reproducibility

All analysis was carried out using R version 4.1.2.[8] The analysis pipeline described here is available as a `targets` workflow[10] along with an archive of all interim results generated using `gittargets`[11] and stored using `piggyback`.[12] A Dockerfile, and prebuilt archived image, has been made available with the code to enhance reproducibility.[13]

Resullts

This project is still underway and so results are not finalised. Please see our real-time evaluation report for our initial findings.

Dicussion

TODO: Summarise project TODO: Discussion strengths and limitations of the evaluation TODO: Compare approach evaluated here to the literature TODO: Discuss further work TODO Summarise conclusions.

Software availability

Source code is available from: https://github.com/epiforecasts/eval-germany-sp-nowcasting/

License: MIT

Models are implemented in the `epinowcast` R package. This is available from: https://github.com/epiforecasts/epinowcast/

Data availability

Input data and summarised results are available from: https://github.com/epiforecasts/eval-germany-sp-nowcasting/tree/main/data

Data from all interim stages of the analysis is from: https://github.com/epiforecasts/eval-germany-sp-nowcasting/releases/tag/latest

License: MIT

Grant information

This work was supported by the Wellcome Trust through a Wellcome Senior Research Fellowship to SF [210758].

Competing interests

There are no competing interests.

References

1. Günther, F., Bender, A., Katz, K., Küchenhoff, H., & Höhle, M. (2021). Nowcasting the COVID-19 pandemic in Bavaria. *Biometrical Journal*, *63*(3), 490–502. https://doi.org/10.1002/bimj.202000112

2. Höhle, M., & Heiden, M. an der. (2014). Bayesian nowcasting during the STEC O104:H4 outbreak in Germany, 2011. *Biometrics*, *70*(4), 993–1002. https://doi.org/10.1111/biom.12194

3. Meyer, S., Held, L., & Höhle, M. (2017). Spatio-temporal analysis of epidemic phenomena using the R package surveillance. *Journal of Statistical Software*, *77*(11), 1–55. https://doi.org/10.18637/jss.v077.i11

4. Team, S. D. (2021). *Stan modeling language users guide and reference manual, 2.28.1.*

5. Gabry, J., & Češnovar, R. (2021). *Cmdstanr: R interface to 'cmdstan'.*

6. Bastos, L. S., Economou, T., Gomes, M. F. C., Villela, D. A. M., Coelho, F. C., Cruz, O. G., Stoner, O., Bailey, T., & Codeço, C. T. (2019). A modelling approach for correcting reporting delays in disease surveillance data. *Statistics in Medicine*, *38*(22), 4363–4377. https://doi.org/10.1002/sim.8303

7. Abbott, S. (2021). Epinowcast: Hierarchical nowcasting of right censored epidemological counts. *Zenodo*. https://doi.org/10.5281/zenodo.5637165

8. R Core Team. (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

9. Bosse, N. (2020). *Scoringutils: A collection of proper scoring rules and metrics to assess predictions*. https://github.com/epiforecasts/scoringutils

10. Landau, W. M. (2021). The targets r package: A dynamic make-like function-oriented pipeline toolkit for reproducibility and high-performance computing. *Journal of Open Source Software*, *6*(57), 2959. https://doi.org/10.21105/joss.02959

11. Landau, W. M. (2021). *Gittargets: Version control for the targets package.*

12. Boettiger, C. (2021). *Piggyback: Managing larger data on a github repository*. https://github.com/ropensci/piggyback

13. Boettiger, C. (2015). An introduction to Docker for reproducible research. *ACM SIGOPS Operating Systems Review*, *49*(1), 71–79.