Bayesian Hierarchical Modelling Frameworks for Correcting Reporting Delays in Disease Surveillance

Alba Halliday, MMath

SUBMITTED IN FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

SCHOOL OF MATHEMATICS AND STATISTICS

COLLEGE OF SCIENCE AND ENGINEERING

SEPTEMBER 2024

Dedicated to Granda'
& Granny Penny.

Abstract

Accurate and timely surveillance of infectious diseases is critical for effective public health responses. Up-to-date quantitative indicators for the prevalence of diseases in a population, e.g. case or death counts, can provide early warning of outbreaks, empowering public health bodies to develop targeted interventions, allocate limited resources, and communicate risks to influence public behaviour. However, data collection for such indicators often suffers from delays, for example due to administrative protocols, testing processes, or resource limitations. These delays mean that available information on outbreaks lags behind reality; delays also vary randomly and systematically in space and time, making it difficult to confidently detect disease outbreaks and provide timely, effective interventions.

From a statistical perspective, correcting delayed reporting is a compositional count data prediction problem. Compositional data, take the form of parts of some whole, in this case a set of non-negative counts reported after each delay that sum to a total count, such as the number of disease cases. In a nowcasting setting, the total count is not yet observed and we aim to predict it given the observed parts of the total for delays that have already elapsed. Applying appropriate statistical methodology for count data with this structure can yield models that learn about the properties of the delay distribution, to provide nowcasting predictions. At the same time, this means that methodological advancements in the field of correcting delayed reporting can potentially lead to innovation in the general field of modelling compositional counts, relevant to a wide range of research fields beyond disease epidemics.

Research carried out prior to this project developed a general multivariate Bayesian hierarchical framework, based on the Generalized-Dirichlet-Multinomial (GDM) family of distributions, that can flexibly account for the different sources of variability in count data suffering from delayed reporting. The framework was developed into a model for a time series of an individual disease in one geographic region. The model demonstrated theoretical and practical potential for the GDM method to provide more accurate and precise predictions, compared to alternative methods.

The work presented here is underpinned by two broad aims: to make the GDM approach more practical for real-time public health applications and to develop novel extensions to the methodology to account for more complex data challenges and features. For the first aim, we developed improvements in computational efficiency and in streamlining applications to real data. Then, we demonstrated the efficacy of the improved GDM model as a solution for nowcasting COVID-19 hospital deaths in different regions of England. Through an unprecedented rolling prediction experiment, we assessed the performance of the GDM against a cohort of competing methods representing the current state-of-the-art, finding that predictions from the GDM were the most accurate and most precise.

For the second aim, our work was informed by a collaboration with experts at Brazil's leading public health institute, the Oswaldo Cruz Foundation (Fiocruz). This offered unique insights into the specific data challenges affecting Brazil's current operational disease warning systems, while also supporting our understanding of more general issues in correcting delayed reporting. One component of this work was motivated by the challenge of nowcasting COVID-positive severe acute respiratory illness (SARI) cases, as an indicator of COVID outbreaks in Brazil. Here, we developed a joint modelling framework for nowcasting total SARI and COVID-positive SARI cases. The framework addressed the novel challenge of correcting delayed reporting of disease counts where information on the length of the reporting delay was not recorded. Applied to data spanning the whole of the Brazil, our approach allowed for predictions of COVID-positive cases, which

suffer from this data challenge, through leveraging the more timely and complete data for the total SARI cases. A rolling prediction experiment demonstrated improvements in predictive performance from incorporating links between overall SARI incidence and COVID-positive rates, as well as from accounting for patient age distributions.

The last major piece of work of the thesis explored potential effects of the level of a disease in the population on the severity of reporting delays. We investigated this issue in data for different diseases, offering new insights into potential capacity limitations or elasticity within the respective reporting processes. We propose a framework that flexibly models the effect of the prevalence of the disease on the delay distribution. Through a simulation study aiming to imitate real data, we demonstrated the framework's ability to disentangle the various sources of variability in the data, including the prevalence-delay interaction, and improve overall prediction accuracy. Since the existing statistical and biostatistical literature on correcting delayed reporting does not assume an explicit effect of disease prevalence on reporting delays, this work could represent the first step for a new paradigm of nowcasting frameworks.

Overall, the work in this thesis provides substantial methodological advancements in correcting reporting delays for disease surveillance, taking the initial proof-of-concept of the GDM framework and greatly enhancing its practicality and versatility. All aspects of the work were driven by and demonstrated using real-world data challenges, employing realistic prediction experiments to develop a robust evidence base for the potential of advanced methods based on the GDM framework to enhance public health responses and policy decisions.

Contents

List of Tables

List of Figures

Acknowledgements

First of all, thank you to my supervisors, Oliver Stoner, Theo Economou, and Duncan Lee; your guidance and insights throughout this journey have made it possible. Also, thank you to Leo Bastos for your vital collaboration.

On a more personal note: Thank you to Cassie for being there every step of the way. Thank you to Isabella for making Glasgow a home. Thank you to Iain for being perfect. Thank you to Daniella for being my rock. Thank you to Romy for being like a sister. Thank you to Iliria for being my best friend. Thank you to Mum for everything. Thank you to Merlin for leading the way. Finally, thank you to Dad for always being there, no matter what.

Declaration

I declare that, except where explicit reference is made to the contribution of others, that this thesis is the result of my own work.

Alba Halliday

Abbreviations

- `INLA` - Integrated Nested Laplace Approximation
- `MCMC` - Markov Chain Monte Carlo
- `DAG` - Directed Acyclic Graph
- `SARI` - Severe Acute Respiratory Illness
- `GDM` - Generalised-Dirichlet Multinomial
- `AR` - Auto-regressive
- `ess` - Elliptical Slice Sampling
- `AF` - Automated Factor
- `GLM` - Generalized Linear Model
- `GAM` - Generalized Additive Model
- `NHS` - National Health Service
- `NB` - Negative-Binomial
- `ESS` - Effective Sample Size
- `MLE` - Maximum Likelihood Estimate
- `EM` - Expectation–Maximisation
- `MAP` - Maximum A Posteriori
- `PSRF` - Potential Scale Reduction Factor

Chapter 1

Introduction to Delayed Reporting in Disease Surveillance

Disease surveillance is the continued monitoring of epidemic trends through the systematic collection and analysis of relevant data. Enabling public health action to then be taken to try and control or prevent the spread of the infectious disease. As well as aiding the implementation of new public health interventions, surveillance can also allow the evaluation of previous policies. Moreover, examining trends of infection within a population could help determine possible risk factors.

Many examples of disease surveillance have been carried out historically, especially in high-income countries. For example, diseases such as smallpox, influenza and cholera had surveillance systems in place in the 19^{th} century (Simonsen et al. (2016)). Monitoring of demographic and geographic variability is now common practice for a large number of infectious diseases. In the case of highly transmissible infections, timely outbreak detection can be critical in reducing the social and economic impact of the disease. A more recent example of this is the COVID-19 global pandemic. Not only was close surveillance vital for decision making, such as non-pharmaceutical interventions, it was also needed to communicate health risks to the public, as discussed in Kline et al. (2022). This included the identification and notification of high risk individuals and regions. Mathematical modelling was also a key tool for public health organisations. Panovska-Griffiths (2020) discusses how the mathematical models, that were used to predict COVID-19 epidemic curves in the UK, aided policy making in the early stages of the outbreak.

Surveillance that is representative of the current regional risk level relies on thorough data collection, which has improved significantly over the last century. The introduction of the International Classification of Disease (ICD) standardised the reporting of diseases globally to allow data to be comparable across hospitals, regions and countries, as well as over time. A version of the ICD was first introduced in 1893 and revisions to it were entrusted to The World Health Organisation (WHO) when it was founded in 1948 (World Health Organization (2020a)). The ICD/WHO also prioritises the easy storage and retrieval of information to allow for timely analysis and decision-making, this has been made more achievable with the advances in computer power and electronic reporting. However, there

is still room for improvements within current disease surveillance systems. One gap is the ability to monitor disease at local level to allow for a more detailed overview of disease patterns and more targeted public health policies. A second gap is the timeliness of disease surveillance which is often hindered by the presence of reporting delays (Simonsen et al. (2016)), which is the topic this thesis addresses.

The overall aim of this body of work is to provide statistical frameworks for carrying out effective and efficient predictions within an operational infectious disease surveillance context. The motivation behind the more specific directions of this work comes from two places. First, a desire to improve the general operational ability of disease surveillance systems. Second, through collaboration with the Oswaldo Cruz Foundation (Fiocruz) in Brazil, the need for novel approaches to real-world problems driven by data challenges that render existing approaches unsuitable for use. Addressing both motivations yields modelling frameworks that not only showcase theoretical improvement in correcting for delayed reporting but also lays a foundation for real-world impact of the work through applications.

In this chapter we first introduce the problem of reporting delays in disease counts and the benefits of correcting for them within an operational surveillance context (Section 1.1). This includes framing delayed reporting as two processes that combine to generate the available data; first the underlying processes that drives the observable disease outcomes (Section 1.1.1), and second the process that creates delays in the reporting of these outcomes (Section 1.1.2). Next, in Section 1.2 we discuss the main challenges of designing statistical frameworks for data subject to delayed reporting and explain why a modular Bayesian approach that models these processes as hierarchical layers is a compelling solution. Finally, in Section 1.3 we briefly outline the contents of each chapter in this thesis.

1.1 Delayed Reporting

Delayed reporting is where information about disease cases, hospitalisations or fatalities are not immediately available to public health decision makers due to lags in the reporting process. Kline et al. (2022) outline that the chains of reporting that can create these delays include: the time it takes for individuals to be subject to and notice symptoms, receiving confirmatory test results, administrative processing of the information, and finally the time it takes to evaluate data that has eventually been reported. Despite recent improvements in disease surveillance and data collection, for the COVID-19 pandemic which was first identified in December 2019 (World Health Organization (2020b)), significant delays were common in reporting health outcomes such as cases, hospitalisations and fatalities. A substantial body of work studies these delays in various countries, including Günther et al. (2020), Stoner, Halliday and Economou (2022) (Chapter 4), Kline et al. (2022), and Seaman et al. (2022).

Through conducting an online survey Gutierrez, Rubli and Tavares (2022) identifies that presenting individuals, daily COVID-19 deaths in Mexico by date reported, compared to date occurred, can reduce perception of the relative risk which is then reflected in public behaviours. Hence, if public health bodies, who often format and publicise disease data, are not accounting for reporting delays it could result in misinformation within a population. This could then lead to higher non-compliance with lockdown recommendations, put in place to help reduce transmission. Therefore, it is generally best practice within disease monitoring systems to present predictions of current disease levels alongside counts reported so far to communicate the uncertainty in our understanding of the current epidemic (Gutierrez, Rubli and Tavares (2022)).

As an illustrative example, Figure 1.1 shows the total weekly severe acute respiratory illness (SARI) hospitalisations reported across the whole of Brazil (points).

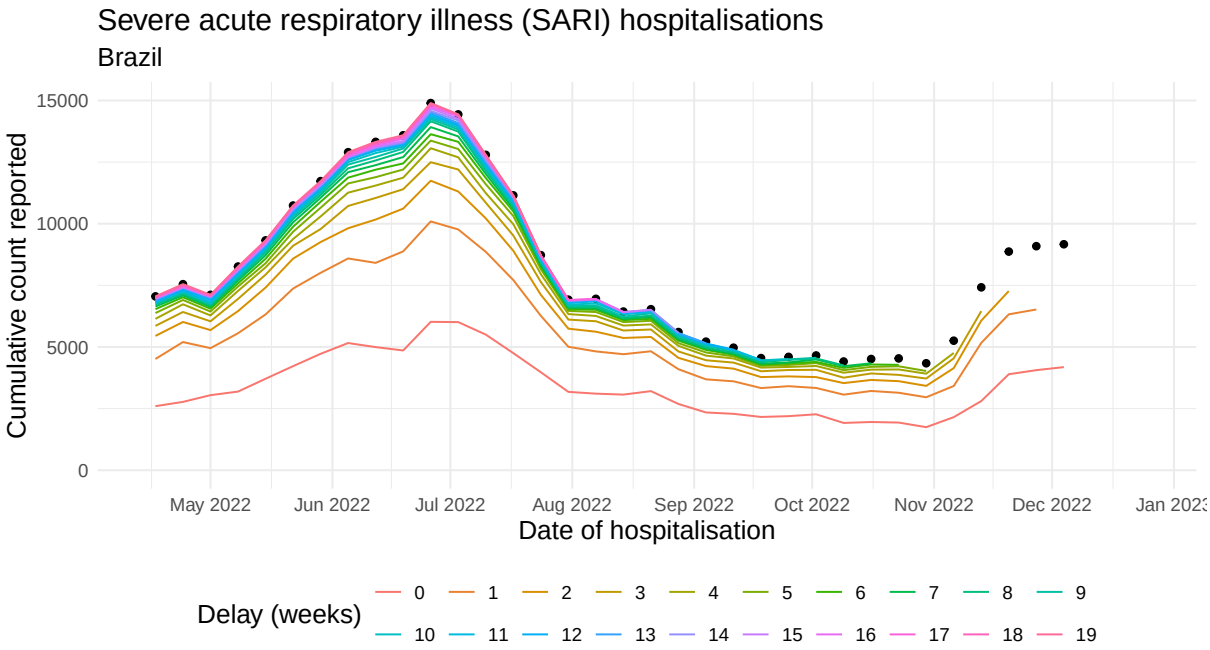

Figure 1.1: Total observed SARI hospitalisations for the whole of Brazil which will eventually be reported (points) against the date of hospitalisation. The lines depict the cumulative counts reported which occurred that date (x-axis) and were reported after the weeks of delay indicated by the line colour.

However, these totals suffer from reporting delays and will not be known in full until a later date. Instead the coloured lines indicate the cumulative counts that were reported at each delay after the date of occurrence (x-axis), where a delay of 0 indicates the counts that were reported the same week they occurred, a delay of 1 indicates the counts reported 1 week after they occurred and so on. For the most recent week, starting 4th of December 2022, we do not know the eventual total counts (represented by the points) that will be reported for the current week or previous weeks, the only information known is given by the coloured lines. Hence, for the most recent week we only have one set of partial counts that have been reported, for the previous week we have the partial counts that occurred last week and were reported last week plus the partial counts that were reported this week. Therefore, as we go back in time the number of delays that we know the number of cumulative counts reported for increases by one each week.

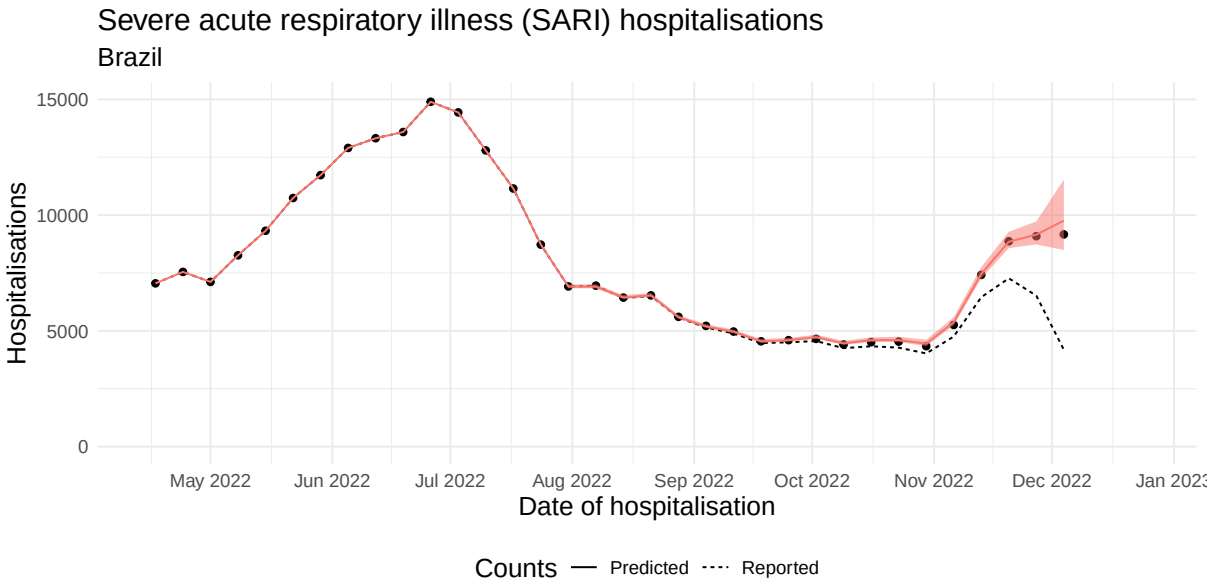

Figure 1.2: Total observed SARI hospitalisations for the whole of Brazil (points) which will eventually be reported. Dashed line gives the hospitalisations reported-so-far up to the week starting December 4th 2022. The solid line gives the posterior median predictions of the total counts from a Generalised-Dirichlet Multinomial model fitted to the available data. The shaded region gives the corresponding 95% prediction intervals.

Correcting for these delays in reporting up to the present time is known as "nowcasting", which allows decisions to be based on a more complete and less biased view of the current state of the outbreak. In the example of Figure 1.1, the task would be to generate accurate predictions of the total number of SARI hospitalisations that occurred each week (given by the points) using just the available data (coloured lines). Breaking down the number of counts reported so far by delay gives us more information than considering just the total counts reported-so-far over all delays, which is represented by the dashed line in Figure 1.2. This figure demonstrates the information nowcasting provides from capturing trends in the counts reported-so-far. The posterior predictive medians (solid line) and 95% prediction interval (shaded region) gives an accurate estimate of the unknown eventual total counts (points) and the uncertainty in the predictions. This allows public health organisations to appropriately prepare for the likely risks associated with different levels of disease outcomes within a population. Where possible, forecasting future counts can also give these decision makers a key outlook of upcoming circumstances allowing for preventative as well as reactive measures to be put in place.

Examples of currently operational nowcasting surveillance systems include infoDengue (https://info.dengue.mat.br) and infoGripe (http://info.gripe.fiocruz.br). These were developed through a collaboration between the Oswaldo Cruz Foundation (Fiocruz) and the University of Exeter (University of Exeter (2021)). In 2017, infoDengue was updated using methods developed by Bastos et al. (2019) to predict current levels of mosquito borne disease (arbovirus) cases in six Brazilian states, to facilitate early warnings and manage risks. Due to infoDengue reliably and accurately predicting the spread of arbovirus cases, allowing for more confident monitoring by public health bodies and the general public, the Brazilian Ministry of Health requested an equivalent surveillance system for severe acute respiratory illness (SARI). This system, infoGripe, was officially launched in 2018.

These systems have played a key role in mitigating the widespread public health risks and challenges associated with both SARI and arboviruses across Brazil, including improvements to population health and well being, e.g. as measured by potential reductions in loss of life, and more efficient use of public health resources. For example, state health authorities in Paraná changed their policy so that laboratory testing is restricted to severe dengue cases when the state is in a period of sustained transmission, as identified using corrected data from infoDengue. The Brazilian ministry of health reported that monitoring of the entire Brazilian population through infoGripe has enabled the introduction of preventative and control policies for influenza on a national level, the development of a targeted emergency SARI vaccine program, and the ability for local authorities to prepare/allocate resources based on potential demand. Meanwhile, it has been estimated that between 300K – 2million US dollars worth of resources were saved through the use of infoGripe by the public health system between 2019 – 2020 (University of Exeter (2021)).

However, the design of such systems is made complex by the need to understand the structure of available data as a convolution of two main processes: a first process that results in the overall level of disease in the population, that produces a potentially observable “total count” (e.g. cases, deaths) for a given time period and geographic area; and a second process that characterises flaws in the reporting of this total, including un-

avoidable delays. In other words, available data in the form of the partial counts reported at each delay, like those in Figure 1.1, are a product of the distribution of the total counts (e.g. number of hospitalisations per week) and the delay distribution (i.e. the breakdown of those counts over different delay intervals). In the following subsections we provide a closer insight into each of these data generating processes, using examples to illustrate the different sources of variability present in these two distributions.

1.1.1 Observed total counts

First, we introduce some notation that will be consistent throughout this thesis when referring to the observed total counts. These are the eventual counts which will be reported for a regular time period t and spatial region s, referred to as the total counts and denoted by $y_{t,s}$. Hence, we also define $\boldsymbol{y}_s$ as the vector of the total counts for region s for all time steps. For example, in Figure 1.1, t represents the week and y_t (given by the points) is the total hospitalisations that will eventually be observed that week, for the whole of Brazil. But, the time-steps could be any scale such as days, weeks or months. Additionally, the counts could represent any type of measurable disease outcome such as positive tests or fatalities. In this subsection we discuss the different potential drivers behind the distribution of the observed total counts.

Severe acute respiratory illness (SARI) is an umbrella term used to for an individual who has been hospitalised with a cough and fever that has onset within 10 days prior to hosipitalisation. The virus strain responsible for the respiratory disease can be determined by laboratory tests, potential viruses include COVID-19, Severe acute respiratory syndrome (SARS) and influenza, among others.

Figure 1.3 plots the COVID-19-positive SARI hospitalisations as well as the remaining SARI hospitalisations that may have positive tests for other viruses or no results and we refer to as the non-COVID-19 SARI hospitalisations.

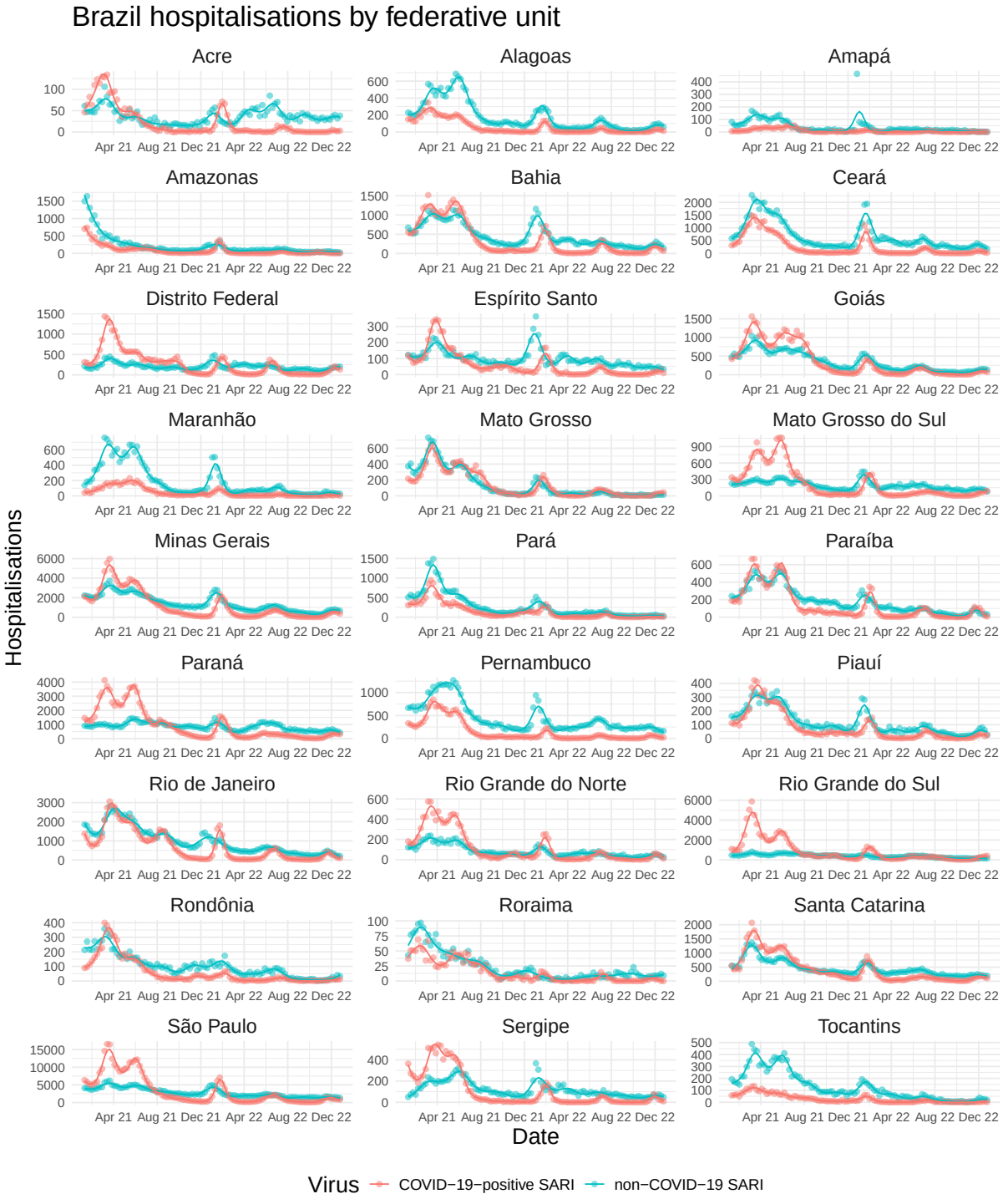

Figure 1.3: Data points represent the observed total counts for COVID-19-positive and non-COVID-19 SARI hospitalisations in each Brazilian federative units. To highlight the trend over time, smooth thin plate regression splines were generated, with a the maximum number of basis functions set to 20, for each count using a generalized additive model (GAM) with a Poisson distribution and log link function for each region. Note the y-axis scale is independent for each panel.

Figure 1.3 demonstrates a systematic trend in both types of hospitalisations over time. The most notable temporal trend is the peaks and troughs in hospitalisations that represent outbreaks of the different viruses. Difference in these two trends over region and time is a result of the systematic variability being driven by the exposure and transmission of the respective viruses over region and time, which will be influenced by the degree of contagiousness and natural immunity/vaccination coverage of each virus. Similarities between the two trends could be a result of seasonal effects such as weather and human behaviour or public health policies. However, in this example similarity in trends may also be influenced by the potential presence of undiagnosed COVID-19 cases in the non-COVID-19 SARI hospitalisations. It is worth noting the top right-hand panel shows that the region Amapa has a potential outlier in the number of non-COVID-19 SARI hospitalisations, where 400 cases were reported in a single week, whilst other weeks during that time period have tend to have a magnitude of under 200 for this federative unit. However, from looking across all the Brazilian federative units it is clear that there is a peak in SARI hospitalisations not associated with COVID-19 around this date, hence it is unclear whether this data point is an accurate representation of the increase in cases or whether hospitalisations may be over-reported due to administrative errors.

In Chapter 5 (Halliday, Stoner and Leonardo (2025)) we investigate modelling this data set to jointly model hospitalisations that can be classified as SARI and the nested proportion of these that can be classified as COVID-19 in order to gain insights and improve predictions of both. These temporal trends also show a clear systematic regional variability with some regions having consistently lower numbers of hospitalisations for outbreaks that appear across all regions. Spatial variability may occur due to difference in population size and structure, education availability and health care resources.

In general, as a disease spreads systematic variability occurs, interpreted as the differences in the prevalence of a disease that we can attribute to measurable effects, such as time and space in Figure 1.3. These are controlled by factors that drive infections. The random weekly deviations from the the overarching trends seen in Figure 1.3 represent the random

variability in the data. These unsystematic shifts are driven by the same processes that influence the systematic trend in the data but operating on a smaller scale, such as day-to-day and regional difference, which is difficult to measure and therefore perceived to be random. Examples of random events that can trigger transmission includes; large social gatherings, population movement, availability of local healthcare, spatial variations in the date of disease introduction, individual vaccination status and genetic mutations of the virus.

To summarise, the distribution of the total counts exhibit both systematic and random variability that needs to be accounted for in potential modelling frameworks. In a disease context the total counts monitor outcomes such as cases or fatalities, and are representative of the epidemic curve which is driven by the occurrence and transmission of a disease.

It is worth emphasizing, for disease surveillance we are actually considering the observed total counts, which are the counts that have not only occurred but have also been reported. Hence, Bastos et al. (2019) note that this observed count will also exhibit variability dependent on factors such as access to health care and reporting protocols and is potentially an under-representation on the true total counts. This may systematically vary over time and space due to government health expenditure and resources availability. Also, randomly on an individual level due to the presence of symptoms, willingness to seek medical care and adoption of reporting. Despite this, it is still worth considering observed total counts as indicator of the prevalence of a disease in a population to gain useful insights about outbreaks.

1.1.2 Delay distribution

For delayed reporting we will consistently use the following notation to describe the structure of the data. The total counts that will eventually be reported at time t and spatial regions s, $y_{t,s}$, can be split into delay specific partial counts $z_{t,d,s}$, which are the counts that occur at time t but aren't reported until delay d. We can then sum the partial counts over delay to obtain the total count $y_{t,s} = \sum_{d=0}^{D_{max}} z_{t,d,s}$. D_{max} is defined such that the total count $y_{t,s}$ is fully reported D_{max} time-steps after t. Since $y_{t,s}$ is then known, all $z_{t,d,s}$ will also be known by time $t + D_{max}$, where the possible delays are $d = 1, ..., D_{max}$. In this section we discuss the processes that determine the distribution of the reporting of these partial counts. We also define $\boldsymbol{z}_{d,s}$ as the vector of all time steps for the partial counts for delay d and region s.

Figure 1.4 shows, for severe acute respiratory illness (SARI) hospitalisations in Brazil, the proportion/percentage of cases reported at delay d (0 to 4 weeks as indicated by colour) against the date of hospitalisation t for each federative unit s. The "absolute proportions", denoted by $p_{t,d,s}$, are the counts reported at each delay d divided by the observed total counts $p_{t,d,s} = \frac{z_{t,d,s}}{y_{t,s}}$. They are called "absolute proportions" as the difference between any two values will give the absolute difference as the proportions are not relative to another value. Alongside the raw data (points), smooth splines have been fitted to the data (lines) and plotted to help identify trends. The reporting process is systematically different over time and delay for each federative unit. This is expected as the reporting procedure and associated resources are likely to vary by region, potentially with different systems and protocols in place. Similarly, the distribution of counts reported at each delay varies over time and space, but in general for Figure 1.4 higher absolute proportions are reported in earlier delays. This reflects a higher proportion of the eventual SARI hospitalisations often being reported in the first two weeks after occurring across all regional reporting systems. The overall trend in the absolute proportions reported over time for most regions is increasing for delay $d = 0$. Suggesting that, since a higher proportion of the total counts

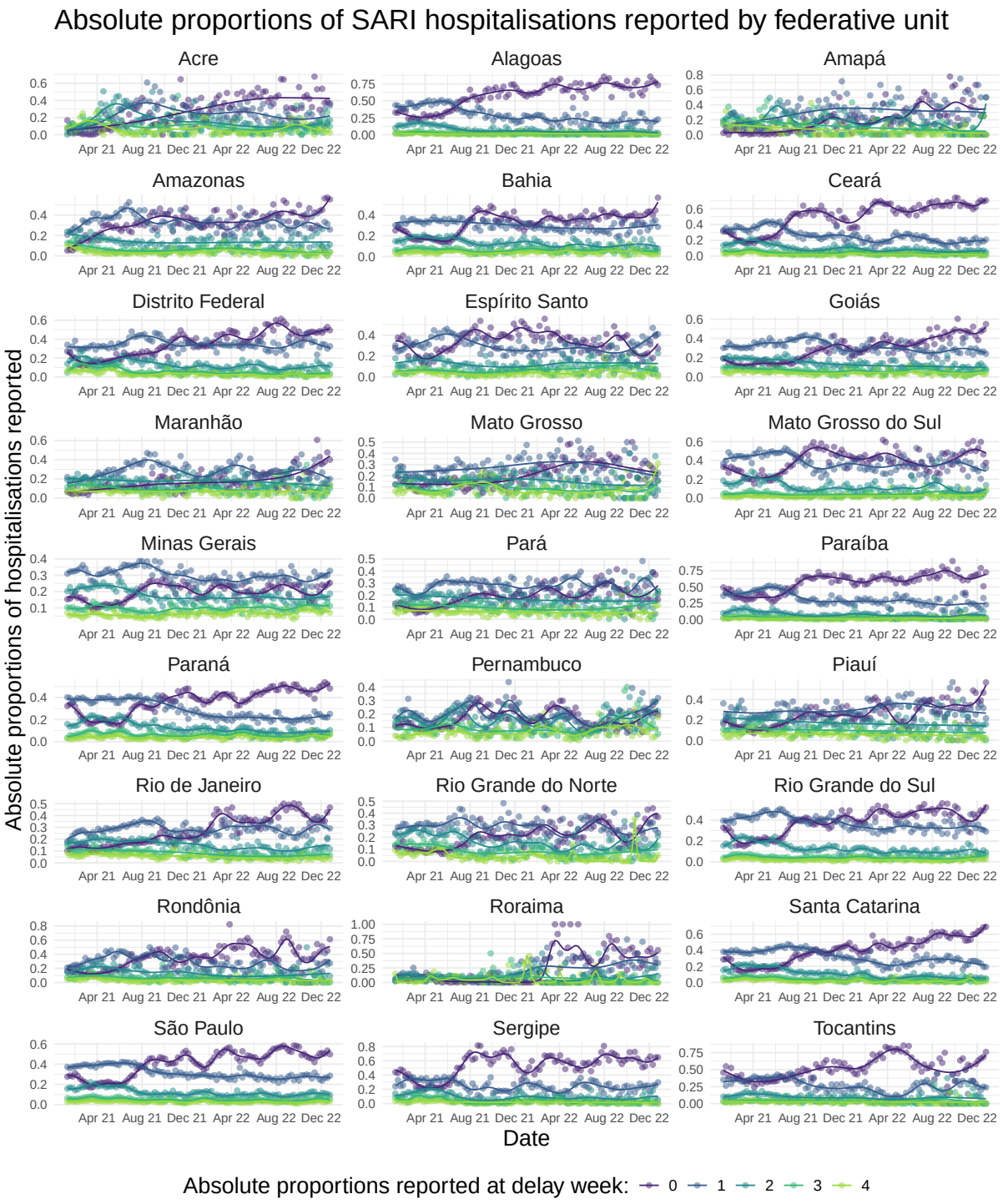

Figure 1.4: Data points represent the absolute proportions ($p_{t,d,s} = \frac{z_{t,d,s}}{y_{t,s}}$) of all SARI hospitalisations in each Brazilian federative units. To highlight the trend over time, smooth thin plate regression splines were generated using a generalized additive model (GAM) with a Gaussian distribution and log link function, specified as $y_t \sim s(t, k = 20)$ for each region and delay independently.

are reported within the first delay, delays are shorter on average over time. However, there is likely to be systematic trends in this phenomenon due to the regional differences over time in the potential causes of this improvement. This may include updating protocols, allocating resources, staff training and reductions in case load burdens.

The delay distribution explains the proportion of the total counts that will be reported at each delay. As we have seen in Figure 1.4, there is likely to be systematic variations due to changes in reporting process which will impact the proportion of counts reported over delay, time and space. Also, in Stoner, Halliday and Economou (2022) (Chapter 4) we observed the announced COVID-19 hospital deaths in England fluctuate in a weekly cycle due to a phenomena known as the "weekend effect", where announced deaths dip on Saturday and Sundays followed by a spike on Mondays. This occurs due to less reporting taking place over weekends as less administrative staff are working and this back-log then being dealt with on the subsequent Mondays. Hence, many factors will contribute to the resulting systematic trends in the delay distribution rendering them complex and hard to predict. Not only will reporting delays depend on the relative efficiency of the local reporting procedures in place, they may also depend on the incidence of the disease. Bastos et al. (2019) note that high incidence could lead to awareness and the prioritising of reducing delays in order to obtain more complete data. Alternatively, having to process a large number of cases could strain the health care system and lead to an increase in case backlogs. We explore and attempt to model this relationship between case load and reporting delay in Chapter 6.

Additionally, random variability induces fluctuations within this systematic variability of the delay distribution due to unknown processes. This can be visualised as the difference between the smooth splines (lines) and raw data (points) in Figure 1.4 driven by day-to-day difference in the local reporting processes. Any number of random events could trigger these variations at any point in the reporting process. For example, changes in staffing,

resource allocations and priorities within the public healthcare system, which will all vary frequently over time and region. In order for this random variability to be accounted for by the model the covariance structure of the partial counts, which explains their joint variability, will need to be well captured.

1.1.3 Summary of the sources of variability

Summarising the discussion of the total count process and delayed reporting process in Sections 1.1.1 and 1.1.2, we follow Stoner and Economou (2019) in characterising available data suffering from delayed reporting as having four main sources of variability:

(1) Systematic variability in the total count $\boldsymbol{y}$ (e.g. exponential growth/decay, seasonal patterns, regional variation).

(2) Random variability in $\boldsymbol{y}$ (e.g. day-to-day variation in the total count).

(3) Systematic variability in the reporting delay (e.g. weekly cycles, improvements in reporting efficiency over time, between-region differences).

(4) Random variability in the reporting delay (e.g. day-to-day variation in the reporting process).

In the next section, we will make theoretical arguments for why capturing these sources of variability in both processes appropriately is essential for reliable nowcasts and forecasts.

1.2 Modelling challenges

The combination of the heterogeneous sources of variability discussed in the previous section makes it difficult to draw empirical conclusions about the current level of disease and its trajectory from available data. This gives an idea of the data that decision makers have to work with, failing to correct the reporting delays to give the true epidemic curves can lead to biased understandings of disease prevalence. Therefore, the ultimate goal is to nowcast the unknown total counts using just the available data and, where possible, forecast into the future as well.

Given the history of disease counts broken up by reporting delay, statistical models can be developed to nowcast the as-of-yet unknown cases by capturing the complex trends in the available data. Predictions of infectious disease outcomes can then be used to help set public health policies. Moreover, insights into the reporting delay distribution gained through these models can help tackle the root issue of reporting delays by highlighting and targeting improvements within the reporting system.

First, we define $\boldsymbol{y}^{(obs)}$ as a vector of the totals counts for all time steps where these total counts have been observed. Similarly, $\boldsymbol{y}^{(unobs)}$ represents the vector of total counts for time steps where the eventually reported totals are not yet fully observed. If we consider the total counts to be made up of two vectors, $\boldsymbol{y} = (\boldsymbol{y}^{(obs)}, \boldsymbol{y}^{(unobs)})$, then we can summarise the goal of nowcasting as a prediction problem for the unobserved counts $\boldsymbol{y}^{(unobs)}$. In a Bayesian framework, we can conceptualise $\boldsymbol{y}^{(unobs)}$ as random quantities and quantify our uncertainty in them through posterior inference. If we consider $p(\boldsymbol{y})$ to be the probability distribution for the total counts and $p(\boldsymbol{z}|\boldsymbol{y})$ is the probability distribution for the delayed reporting process, then by Bayes theorem we obtain

$$p(\boldsymbol{y}|\boldsymbol{z}) \propto p(\boldsymbol{z}|\boldsymbol{y})p(\boldsymbol{y}). \tag{1.1}$$

This suggests that to achieve an accurate predictive distribution for $p(\boldsymbol{y}|\boldsymbol{z})$, we need an appropriate $p(\boldsymbol{z}_t|\boldsymbol{y})$ as well as explicitly capturing the systematic and structured variability of the total counts through $p(\boldsymbol{y})$, as informed by previous observed total counts. Alternative modelling approaches can predict the total counts without explicitly modelling the total counts in this way. But, if we wish to utilise all available data for predictions of the unknown total counts, comprising of the observed partial counts $\boldsymbol{z}^{(obs)}$ and observed total counts $\boldsymbol{y}^{(obs)}$, then a Bayesian modelling approach is a natural fit. Although, if $p(\boldsymbol{z}|\boldsymbol{y})$ and $p(\boldsymbol{y})$ aren't chosen appropriately to capture the observed data, then this could lead to predictions that are uncertain or inaccurate.

Constructing an appropriate model for these probabilities can be challenging due to the heterogeneous systematic and random variability in the data generating process of both the observed total counts and the distribution of the delayed reporting, that we have summarised in Section 1.1.3. Hence, models implemented for disease surveillance applications need to have the flexibility to be able to both capture and separate these four sources of variability in the data. This could include temporal trends, spatial trends, seasonality, interaction effects, auto-correlation as well as incorporating informative covariates to help aid model predictions. An additional consideration when endeavouring to jointly model the total and partial counts together, is to account for the compositional structure of the data. The partial counts are compositional elements that sum to a whole, given by the total counts. This sum constraint means that $z_{t,d}$ are all both bounded between 0 and y_t and not independent of each other across $\boldsymbol{d}$. Therefore, complex hierarchical frameworks and sophisticated techniques are required to model this data structure.

Furthermore, the nature of delayed reporting means dealing with missing data, which is an additional modelling challenge. Not only is there uncertainty in the reporting delays that hinders the understanding of when the eventual totals will fully be reported across regions, there is also incomplete information to predict these eventual totals. The combination of these various obstacles makes designing frameworks for nowcasting infectious disease a non-trivial task. On top of this, for real-time applications the practicality of models also has to be thoroughly considered including timeliness, robustness and usability.

Bayesian methods are well suited to nowcasting applications as they can enable a thoughtful and rigorous treatment of $p(y_t)$ and $p(\boldsymbol{z}_t \mid y_t)$, and they are naturally able to handle the aforementioned missing data structures. Stoner (2019) studied the challenge of nowcasting and correcting delayed reporting in the broader context of accounting for flawed observation mechanisms in modelling. Stoner (2019) framed such problems in terms of a "modular framework", beginning with a process represented by "Y" generating the quantity we are interested in, y. In the following equation, $\rightarrow$ denotes the generation process of "Y" given a set of model parameters and/or random effects, $\boldsymbol{\Theta}$:

$$Y(\boldsymbol{\Theta}) \rightarrow y. \tag{1.2}$$

Next, "Z" represents a flawed observation process for y, generating the quantity z. This process, depending on model parameters $\boldsymbol{\Pi}$, translates the original quantity y into z:

$$Y(\boldsymbol{\theta}) \rightarrow y \rightarrow Z(\boldsymbol{\Pi}) \rightarrow z. \tag{1.3}$$

Implementing these processes as modular layers in a Bayesian hierarchical framework, both model parameters (e.g. $\boldsymbol{\Theta}$, $\boldsymbol{\Pi}$) and data (y, z) are treated as random quantities we can learn about through posterior inference. This enables prediction of unobserved y given observed data for z (Stoner (2019)).

In the context of the challenge addressed in this thesis, "Y" can represent the process generating the total observable count time series of a disease, y_t, and "Z" can represent the delayed reporting mechanism(s) that generate partial counts $\boldsymbol{z}_t$. This framework provides a natural solution to the systemic missing data challenge arising from delayed reporting, where we can predict unobserved total counts y_t ($y_{(unobs)}$) based on observed partial counts $\boldsymbol{z}_t$. The flexibility of the framework derives from the way in which modules can be adapted, added or removed with ease, for instance to account for more than one flawed observation mechanism (e.g. under-reporting, misdiagnosis). Moreover, Stoner (2019) argues that the separation of parameters and models associated with each module/quantity simplifies their interpretation and design, since the role of each hierarchical layer in generating available data is conceptually clear. Notably, this can aid us in designing models for Y and Z that appropriately capture the different sources of variability in our data (Section 1.1.3). Information from experts or other sources pertaining to the original data generating process Y and/or any flawed observation mechanisms can also be included through informative prior distributions for related parts of $\boldsymbol{\Theta}$ or $\boldsymbol{\Pi}$.

In support of transparent disease surveillance systems that allow decision-makers to consider and prepare for all likely incidence levels of a disease, the modular Bayesian hierarchical approach allows for rigorous quantification of uncertainty in model parameters and predicted counts through their posterior (predictive) distributions (Gelman et al. (2013)). Meanwhile, we can test model assumptions thorough through posterior predictive checking (Gelman et al. (2013, Chapter 6)).

For these strengths – which we can summarise as flexibility, interpretability, and uncertainty quantification – we choose to follow this framework as a guide for the methods we develop and present in this thesis, and for critically evaluating work that came before. However, to achieve reliable predictions of y_t based on $p(y_t \mid \boldsymbol{z}_t)$, we must still ensure that $p(y_t)$ and $p(\boldsymbol{z}_t \mid y_t)$ are appropriate, which is equivalent to ensuring Y and Z can appropriately capture the main sources of variability associated with the disease counts and the delayed reporting, respectively.

1.3 Summary of thesis

This thesis aims to enhance the practicality and versatility of frameworks for correcting delayed disease reporting. It focuses on developing novel extensions to address complex data challenges and improving computational efficiency. Hence, providing substantial methodological advancements for aiding public health surveillance. The content of the thesis is arranged as follows:

Chapter 2: Reviews current literature on correcting delayed reporting; highlights the strengths and weaknesses of the Generalised-Dirichlet Multinomial (GDM) model, including its predictive power and computational demands; discusses extensions to nowcasting models found in the literature.

Chapter 3: Explores ways to improve the computational efficiency of the GDM model, such as using marginal models, efficient sampling methods, and direct optimization of the joint posterior; compares the impact of these improvements using case studies.

Chapter 4: Describes contributions to a publication, Stoner, Halliday and Economou (2022), on using the GDM model for correcting COVID-19 death reporting delays in England; examines the necessity of accounting for auto-correlation and the benefits of a moving window approach; presents a simulation experiment to evaluate the GDM's performance.

Chapter 5: Develops a framework for modeling nested disease data, using a novel link function with the GDM framework; applies this framework to SARI hospitalizations in Brazil, demonstrating improved prediction performance by linking overall SARI incidence with COVID-positive test rates and age distribution effects. The publication, Halliday, Stoner and Leonardo (2025), relating to the work covered in this chapter is, as of writing, under peer review.

Chapter 6: Investigates the potential of a case load effect on reporting delays and develops a general framework to model this effect; presents simulation and real-world data experiments that evaluate the suitability of the proposed framework.

Chapter 2

Review of Existing Bayesian Nowcasting Models for Delayed Reporting

In this chapter we provide background details on Bayesian modelling methods, inferential algorithms, and software: Section 2.1.1 gives details on Markov Chain Monte Carlo (MCMC) methods for general Bayesian inference; Section 2.1.2 describes use of the NIMBLE software for implementing flexible Bayesian hierarchical models; and Section 2.1.3 describes the Integrated Nested Laplace Approximation (INLA) approach for efficient implementation of latent Gaussian methods. These details provide a foundation for understanding the Bayesian approaches to nowcasting reviewed and developed in this thesis.

The main body of this chapter is then a critical review of existing approaches to nowcasting and correcting delayed reporting in the context of disease surveillance. In Section 1.2, we made arguments in favour of using the Bayesian approach for nowcasting, in terms of flexibility, interpretability, and quantification of uncertainty. As such, our review mainly focuses on Bayesian nowcasting methods. We illustrate how these can be broadly separated into two groups of approaches; joint models of the total counts and the reporting delay mechanism as two layers in a hierarchical framework (Section 2.2.1), and direct models for the partial delayed counts that rely on a conditional independence assumption (Section 2.2.2). For some approaches in the latter group, we give brief insights into potential frequentist equivalents – we also discuss the relatively new area of disease nowcasting using machine learning approaches in Section 2.4. We use a critical comparison of these groups to motivate the Generalized-Dirichlet Multinomial (GDM) method for correcting delayed reporting, which we describe in detail in Section 2.3.

Section 2.5 discusses extensions that have been to made to existing Bayesian frameworks. We showcase these to highlight the need to adapt methods for different applied data challenges arising in the real world. Finally, in Section 2.6 we summarise the criteria that will guide us in constructing nowcasting models that are computationally efficient and as precise as possible for prediction.

2.1 Bayesian methods

In the Bayesian approach, inference is based on the posterior distribution, which represents the probability distribution of the unknown quantities (θ) conditional on the data (x). The posterior is linked to prior beliefs and the data through Bayes' Theorem:

$$P(\theta|x) = \frac{P(x|\theta)P(\theta)}{\int P(x|\theta)P(\theta)d\theta}. \tag{2.1}$$

In Equation (2.1), the numerator is the product of the likelihood $P(x|\theta)P(\theta)$ and the prior distribution $P(\theta)$. The likelihood is the probability of generating data x given the parameters θ, and the prior represents the credibility of the parameter values without having yet seen the data. The denominator, known as the normalising constant or the "evidence" $P(x)$, derives from a multi-dimensional integral over all possible values of the parameters.

Gelman and Hill (2006, Chapter 2) outline that in classical modelling, data x is often treated as independent, with predictor X assumed to be the same for all observations. However, if the data x contains N groups or levels, a more suitable approach is to model group-specific predictors X_n for $n = 1,...,N$. This approach leads to a multilevel model where parameters can vary across groups, allowing the model to capture both within-group and between-group variation. Hierarchical models usually describe multilevel models where groups in x have a nested structure. These models involve layers of dependencies and multiple parameters that vary at different levels. By modeling the dependence structure within the data, hierarchical models achieve greater flexibility in capturing group-specific trends while avoiding over-fitting.

Prior distributions allow existing knowledge or assumptions to be incorporated into a Bayesian model, which can help inform the posterior distribution. This approach is often referred to as subjective Bayesian modelling. However, this can introduce subjective bias into the model, especially when strongly informative priors, that reflect strong beliefs about parameter values, dominate the posterior. On the other hand, weakly informative priors offer some structure (e.g. Normal distributions with large variances), with the expectation that a sufficiently large data size will reduce the influence of the prior and mitigate any bias it may introduce. Alternatively, non-informative priors aim to minimize the influence on the posterior, allowing the data to drive inference as much as possible. In Empirical Bayesian modelling, the priors are estimated directly from the data, adapting to the data-driven context. Meanwhile, in Objective Bayesian modelling, non-informative priors are constructed when prior information is unavailable, focusing on minimizing subjectivity and reflecting the lack of prior knowledge (Gelman et al. (2013)).

In a Bayesian hierarchical model, over-fitting can be controlled through prior distributions that pool information across groups. This partial pooling allows the model to estimate an overall mean while shrinking group-specific estimates towards it, this is particularly beneficial for modelling small or noisy groups. In contrast, non-hierarchical models that ignore group structure may require many parameters to fit existing data well, leading to over-fitting and rendering the model unsuitable for predicting new data (Gelman et al. (2013)). Hierarchical models, on the other hand, utilize parameters that reflect the data structure. Also, joint probability models are required, since one parameter value depends markedly on another, resulting in large, complex, multi-dimensional integration to calculate the posterior distributions. Hence, fitting hierarchical models typically requires a method that can approximate the posterior distribution.

The posterior distribution can be calculated using numerical methods if the range of possible parameter values is finite and can be divided into a sufficiently dense grid of values. Alternatively, if the distribution for the prior distribution is conjugate to the likelihood, meaning that their product has the same form as the prior distribution, then

the posterior can be solved analytically. However, it is often the case that the integral to calculate the evidence is analytically intractable. As a result, posterior inference is usually achieved through advanced computational methods, which either simulate samples from the posterior distribution or approximate the distribution (Kruschke (2015)). Specific methods for Bayesian inference and prediction are introduced in the next two subsections prior to reviewing literature about their implementation.

2.1.1 MCMC

Markov chain Monte Carlo (MCMC) is a general approach to fit Bayesian models no matter the form of the posterior. This is achieved by approximating the target posterior distribution using a two step process of drawing samples from an approximate distribution and then evaluating those samples to improve the approximation. The sampling method sets the posterior as the target distribution of a ergodic Markov chain $\boldsymbol{\theta}^0, \boldsymbol{\theta}^1, ..., \boldsymbol{\theta}^t$. New positions of the Markov chain are generated using a sampling algorithm that doesn't require the normalising constant, as defined in Section 2.1. Once these chains have converged, samples from their stationary distribution can be used to represent the posterior.

As discussed in Gelman and Rubin (1996), the curse of dimensionality for multivariate regression describes how when scaling up the number of predictors in a model, the relative number of basis functions required to model the data may simultaneously increase, which is computationally expensive to fit. Moreover, it encompasses the challenge of observations being more sparse across the domain of the predictors as the number of predictors increase. Hence, for multivariate modelling, larger amounts of data are needed to estimate the regression. Therefore, MCMC algorithms can be slow for complex problems, such as models with a large number of parameters or large data sets, as it can take a long time for the multiple chains to converge.

There are many different versions of the sampling algorithm, but most stem from the Metropolis-Hastings and Gibbs algorithm. Firstly, Metropolis-Hastings involves a proposal distribution $J_t(\boldsymbol{\theta}^*|\boldsymbol{\theta}^t)$ that at iteration $t+1$ suggests new positions $\boldsymbol{\theta}^*$ for the Markov chain, given the current position at iteration t is $\boldsymbol{\theta}^t$. The new position is then either accepted such that $\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^*$, or rejected so that $\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t$, with a probability α calculated by

$$\alpha = min\left(1, \frac{p(\boldsymbol{\theta}^*|x)}{J_t(\boldsymbol{\theta}^*|\boldsymbol{\theta}^t)} \frac{J_t(\boldsymbol{\theta}^t|\boldsymbol{\theta}^*)}{p(\boldsymbol{\theta}^t|x)}\right), \tag{2.2}$$

where $p(.|x)$ is the posterior distribution function. Since the posterior distribution is in both the numerator and the denominator the normalising constant will cancel out. Hence, we only need to be able to calculate the posterior up to a proportional constant, given by the product of the likelihood and the prior by Bayes rule (Equation 2.1). This process is then repeated for each iteration (Gelman et al. (2013)).

Gibbs sampling is an alternative sampling algorithm and is also beneficial for multidimensional sampling. It involves dividing the parameters $\boldsymbol{\theta}$ into d parts $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, ..., \boldsymbol{\theta}_d)$. These are sampled in the Gibbs algorithm such that each $\boldsymbol{\theta}_j$ is conditional on the values of all other sub-vectors, denoted $\boldsymbol{\theta}_{-j} = \boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_{j-1}, \boldsymbol{\theta}_{j+1}, ..., \boldsymbol{\theta}_d = (\boldsymbol{\theta}_{1:(j-1)}, \boldsymbol{\theta}_{(j+1):d})$, at their current value. At each iteration t, each sub-vector $\boldsymbol{\theta}_j^t$ is sampled from the conditional distribution $p\left(\boldsymbol{\theta}_j^t|\boldsymbol{\theta}_{1:(j-1)}^t, \boldsymbol{\theta}_{(j+1):(d)}^{t-1}, \boldsymbol{x}\right)$.

MCMC sampling being an iterative sampling method presents a few practical considerations. Firstly, the starting points of the MCMC chains have to be decided by setting initial values for each of the model parameters and any unobserved data. Random initial values, that still satisfy the priors and distributions of the respective parameters and data, are usually generated in order to help diagnose multi-modality (Brooks and Gelman (1998)), which we discuss in depth later in this subsection. Secondly, within sampling correlation is where one iteration of the MCMC chain is correlated to the previous iteration, which can reduce the efficiency of the simulation. This occurs due to MCMC chains using the

previous iteration as a starting point for the next. The resulting inefficiency can slow convergence and increase the number of iterations needed to ensure samples are representative of the target distribution. To quantify this, we can measure the "effective sample size" of MCMC chains.

The ESS of a given parameter in an MCMC model can be estimated using

$$\widehat{ESS} = \frac{n}{1+2\sum_{k=1}^{\infty}\hat{\rho}_k}, \tag{2.3}$$

where n is the number of total samples taken, and $\hat{\rho}_k$ is an estimate of autocorrelation at lag k. Therefore, n would be the ESS for uncorrelated samples. The estimated spectral density at frequency zero is used to calculate $\widehat{ESS}$ using the R function `effectiveSize`.

The effective sample size (ESS) is an indicator of how efficient MCMC sampling is by measuring the number of independent samples generated from the target distribution. MCMC chains that are well mixed, which means they don't have high auto-correlation, should have a higher ESS. The length of the Markov chain needed to approximate the posterior distribution to a reasonable degree of accuracy will therefore be shorter the higher the ESS (Harrington, Wishingrad and Thomson (2021)). Hence, improving the mixing of MCMC chains can reduce the time it takes to fit the model by reducing the number of iterations the chains need to be run for. Vats, Flegal and Jones (2019) advocate a lower bound for the estimated ESS to reach before terminating the MCMC sampling. For a large number of parameters, such as the GDM model, this lower bound converges to

$$\widehat{ESS} \geq \frac{2\pi e}{\varepsilon^2}. \tag{2.4}$$

The user is left to determine the relative precision ε, which is defined as the relative contribution of Monte Carlo error to the variability in the target distribution. For example, Vats, Flegal and Jones (2019) set $\varepsilon = 0.05$, meaning that the Monte Carlo error accounts for 5% of the variability in the target distribution.

Another important step of any MCMC implementation is setting the number of iterations and burn in. The initial iterations from a sampling algorithm will be dependent on the parameter starting values. To diminish the influence of both the chosen initial values and the stage of the MCMC chain prior to convergence, this initial period of iterations are discarded. The number of iterations which you discard is known as the "burn in" of the chain, it is best practice to set this conservatively large such that any influence from initial values and pre-convergence is unlikely. Two or more MCMC chains are often run in parallel with random initial values for each. Multiple chains allow for convergence to be assessed by ensuring that within chain variation is approximately equivalent to the between chain variation, which indicates all chains are approximating the target distribution. This can be measured using the the Potential Scale Reduction Factor (PSFR) equation from Gelman and Rubin (1992),

$$\hat{R} = \sqrt{\frac{\widehat{\text{Var}}(\theta)}{W}}. \tag{2.5}$$

For Equation 2.5, if we have m chains with n iterations and $\theta_{i,j}$ represents the j^{th} sample from the i^{th} chain then $W = \frac{1}{m}\sum_{i=1}^{m}\left(\frac{1}{n-1}\sum_{j=1}^{n}(\theta_{ij} - \bar{\theta}_i)^2\right)$ calculates the within-chain variance and $B = \frac{n}{m-1}\sum_{i=1}^{m}(\bar{\theta}_i - \bar{\theta})^2$ calculates the between-chain variance. We can then estimate the marginal posterior variance by $\widehat{\text{Var}}(\theta) = \frac{n-1}{n}W + \frac{1}{n}B$.

Moreover, starting multiple chains from multiple random points in the parameter space helps identify multi-modality in distributions, which would results in multiple chains converging to different maxima. Viewing trace plots of the multiple chains can help determine if each of the chains have converged to the same posterior mode. This is indicated by each chain exhibiting random samples with no trends across iterations, centred at the same posterior value. However, if multiple chains converge to the same maxima this is not sufficient evidence to discount multi-modality (Gelman et al. (2013, chapter 11)).

2.1.2 NIMBLE

There are various R packages (R Development Core Team (2011)) available to fit Bayesian models in R. Two popular choices are **rstan** by Carpenter et al. (2017) and **rjags** by Plummer (2003). Here, we introduce an R package NIMBLE (Numerical Inference for statistical Models using Bayesian and Likelihood Estimation) that we will utilise throughout this thesis, when developing any novel Bayesian models using MCMC. Developed by de Valpine et al. (2017), NIMBLE is a tool to facilitate the construction and implementation of complex hierarchical models. However, it is mainly suitable for models that can be defined as directed acyclic graphs (DAG) (some exceptions apply, i.e. ICAR models). Directed acyclic graphs are made up of nodes, representing the model variables (including data, parameters and latent variables), and directed edges that represent the marginal or conditional dependencies between nodes. Directed graphs are acyclic if following the directed edges between nodes never creates a closed loop (Thulasiraman and Swamy (1992, Chapter 1)). Fortunately, directed acyclic graphs encompasses a large variety of Bayesian models. The NIMBLE package can be summarised as comprising of three main components:

- An extension of the BUGS language for model definitions, which are then programmable objects in R.
- A catalogue of algorithms for models written in BUGS and a system for defining and executing customised functions.
- A compiler that generates more efficient C++ code for user-defined models and functions, which the user can easily interface with in R.

The aim of NIMBLE is to enable developments in MCMC algorithms and maximum likelihood methods to be accessible through the NIMBLE package, to close the gap between methodological advancements and available software that can be readily applied. Furthermore, it combines a thorough catalogue of sampling algorithms with flexibility in defining model structures and distributions. Hence, enabling a wider class of model specification and customisation of the methods used to fit them (de Valpine et al. (2021)).

The capacity to fit such a wide range models, and the ability to create custom functions, within NIMBLE makes it invaluable for our goals, no competing packages also allow for both modelling latent discrete parameters and creating custom distribution functions. In Section 2.1.2.1, we overview the available MCMC samplers in NIMBLE, including specialised samplers that we benefit from in later work. Section 2.1.2.2, then outlines how we design smooth spline modelling effects within the NIMBLE environment using the **mgcv** package.

2.1.2.1 MCMC samplers

Statistical frameworks, such as those we develop in later chapters, utilise NIMBLE to automatically build and run MCMC sampling algorithms for Bayesian models. A stochastic node is a component of a probabilistic graphical model that represents a random variable with an associated probability distribution. These nodes are central to the Bayesian framework, as they capture the uncertainty inherent in the model parameters or observed data. Samplers are automatically assigned using a list of nine criteria for the stochastic node; each criterion relates to one of the default MCMC samplers in an order of preference. This ensures that nodes with specific properties (e.g. continuous, discrete, conjugate) or distributions are assigned appropriate samplers. For example, nodes that have a conjugate relationship between their prior distribution and the distribution of their stochastic dependents are assigned a conjugate Gibbs sampler. Similarly, nodes with Multinomial or Dirichlet distributions would be given Multinomial random walk or Dirichlet random

walk samplers respectively (the full list of criteria is given in de Valpine et al. (2021, Chapter 7)). However, there are twenty samplers in the NIMBLE library that a user can choose from, in place of the default samplers, where desired. Most sampler options are versions of the random walk (RW) sampler which is an adaptive Metropolis-Hastings algorithm, discussed in Section 2.1.1, where the proposal distribution is Gaussian. For this thesis, the MCMC samplers used are the default NIMBLE samplers unless explicitly stated otherwise, see de Valpine et al. (2021, Chapter 7)) for full details of the default NIMBLE samplers.

Discrete valued scalar nodes are automatically assigned a slice sampler. Similarly, continuous valued scalar nodes can be manually assigned a slice sampler. The general concept of slice sampling can be outlined in the following steps which have been visualised in Figure 2.1:

1. Start with an initial x value x_0, and calculate $f(x_0)$ where the function $f(x)$ defines the target density. This step is usually achieved using Gibbs sampling as the density function may not be analytically attainable.
2. Take a "vertical" sample from the uniform distribution defined by the distance from zero to the curve of the density function $y = \text{Uniform}(0, f(x_0))$.
3. A horizontal window is then constructed such that it includes the slice $S = x' : f(x') > y$. There are different approaches to this step but the "stepping out" approach adopted by Neal (2003) is the algorithm available in NIMBLE. This approach involves widening the horizontal window until both ends are outside the horizontal area that satisfies that y is less than the density function before sampling a new horizontal x value.
4. A slice or "horizontal" sample is then taken conditional on this vertical y value. This is sampled by proposing points within the horizontal window.
5. If the proposed point x' does not satisfy $f(x') > y$ (lies outside the slice) then the proposed sample is discarded, and the window is updated to exclude this value.
6. If the proposed point x' satisfies $f(x') > y$ (lies within the slice) then it is accepted as x_1, and is the current position for the next iteration.

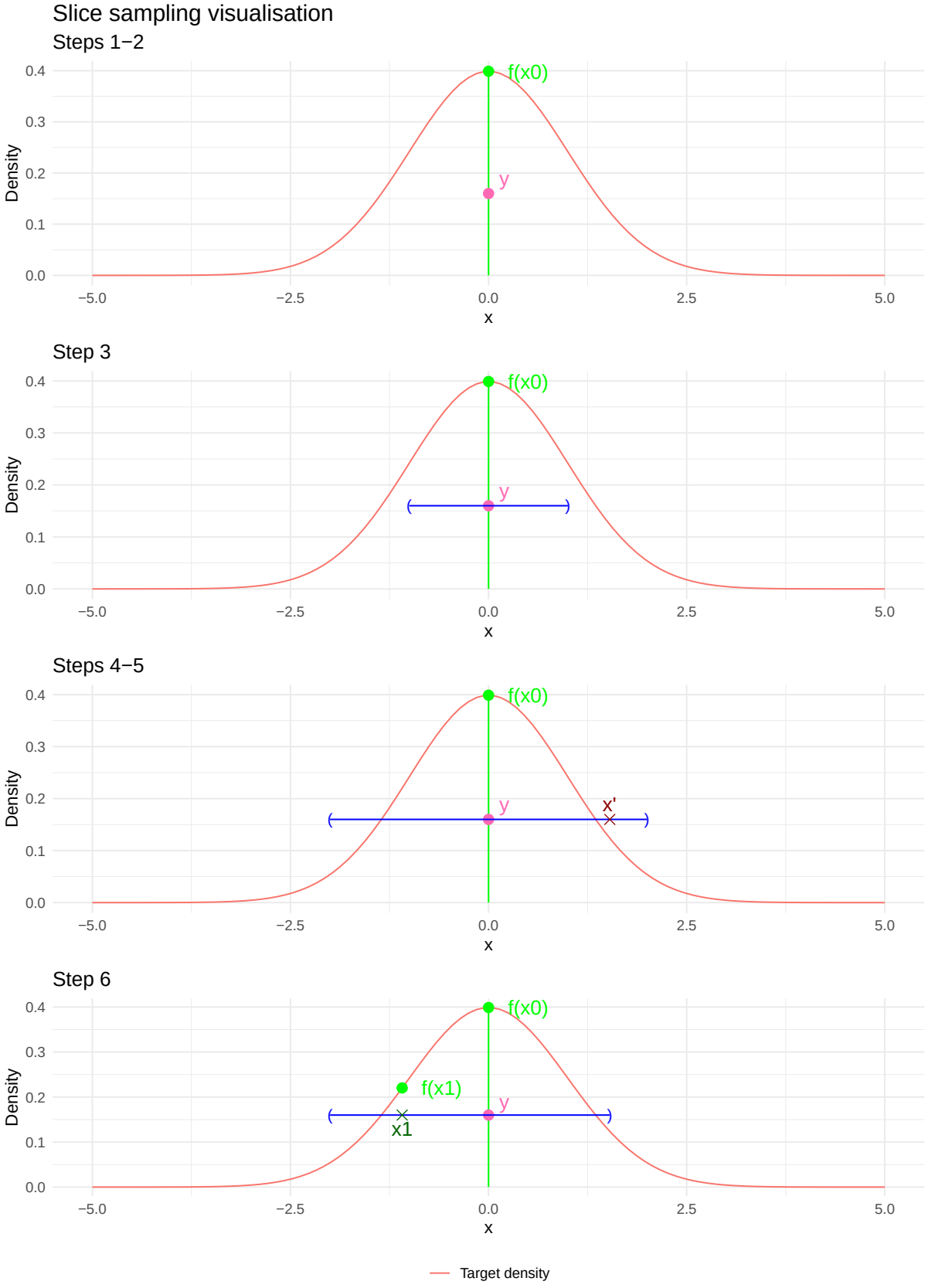

Figure 2.1: Visual representation of the first iteration of a slice sampling algorithm where the target distribution is assumed to be a Gaussian distribution with mean zero and standard deviation 1. The initial position of the algorithm is $x_0 = 0$. Steps correspond to the description of slice sampling given above.

The motivation for slice sampling methods is that they require less tuning than alternative MCMC methods. Tuning MCMC methods involves defining parameter values required by the algorithms. For example, the initial width of the horizontal slice window for slice sampling. Neal (2003) argues that that slice sampling is robust even when this parameter is set to a factor of 100 away from its optimal value.

Two additional specialised version of this slice sampling algorithm are also available in NIMBLE; elliptical slice sampling (ess) and automated factor slice sampling (afss). Elliptical slice sampling is an extension of slice sampling by Murray, Adams and MacKay (2010) that requires no tuning parameters and aims to improve sampling efficiency for Multivariate Gaussian distributions. On the other hand, automated factor (AF) slice sampling was developed by Tibbits et al. (2014) to cope with high dimensional and correlated target distributions by sampling from a transformed space. This addresses the issue of individually sampling multiple strongly related quantities, for example using Gibbs samplers, inducing high auto-correlation within those samples and thus inefficient sampling. The AF slice approach does not require prior knowledge of the correlation structure and sampling parameters are tuned automatically.

2.1.2.2 Smooth effects

When defining Bayesian models in NIMBLE or other software a wide range of modelling effects can be utilised. This is beneficial when constructing modelling frameworks as the most intuitive effects for the application can be selected. In particular, we utilise smooth model effects throughout this thesis to capture complex trends in data using NIMBLE. Here we outline the benefits of smooth modelling effects and how to implement them in NIMBLE.

Modelling effects are parameters that can describe and capture the trends present in the response variable of interest, and represent the underlying statistical assumptions. This includes, but is not limited to; linear effects, covariate effects, random effects, random walks, Gaussian processes, spatial effects and smooth effects. Smooth effects describe the statistical assumption that a trend is not fixed over a given explanatory variable. Hence, they allow for more flexible relationships to be captured and are particularly beneficial for modelling over time and space. Capturing trends with smooth modelling effects provides intuitive inference about potentially complex relationships and can allow for predictions of new data points.

In general smooth functions can be formatted in the following way:

$$f(x) = \sum_{j=1}^{k} \beta_j b_j(x). \tag{2.6}$$

Where $b_j(x)$ is a predetermined form of basis function (e.g. B-splines, thin plate splines) and β_j are the corresponding unknown model coefficients for k knots. For a frequentist approach, the smoothness of the function is penalised using a penalty parameter in the loss function. This smoothness parameter is then estimated by a method which minimises the penalised likelihood such as cross-validation or restricted maximum likelihood (REML).

On the other hand, the Bayesian equivalent of a penalised likelihood is to introduce smoothing constraints through improper Multivariate Gaussian priors for the unknown β_j coefficients:

$$\boldsymbol{\beta} \sim \text{Multivariate-Normal}(0, \boldsymbol{\Omega}^{-1}). \tag{2.7}$$

The penalty parameter τ is then introduced in the precision matrix, $\boldsymbol{\Omega}$, of these prior distributions:

$$\boldsymbol{\Omega} = \tau \boldsymbol{M}. \tag{2.8}$$

Where $\boldsymbol{M}$ is a known non-diagonal matrix and is scaled such that more smoothness in the coefficients is enforced for larger τ. The Multivariate Gaussian priors are improper as it does not integrate to a finite value over the entire parameter space of the coefficients β. This is because the smoothing penalty matrix M, which penalises certain properties of the smooth function e.g. a second-derivative penalty, includes constant or linear functions that relate to the un-penalised components of the model, making the precision matrix Ω singular. The smoothing parameter τ is then determined by the data when fitting the model using MCMC (or other Bayesian sampling method) to estimate the posterior of the parameter. In general, the spline, here denoted $\boldsymbol{\alpha}$, is again then formulated by multiplying the spline basis function matrix $\boldsymbol{X}$ by the model coefficients β:

$$\boldsymbol{\alpha} = \boldsymbol{X}\boldsymbol{\beta}. \tag{2.9}$$

Wood (2016) developed the function `jagam`, which was added to the **mgcv** R package, to conveniently define smooth effects for a number of different basis functions. This is useful when defining models in NIMBLE or other Bayesian modelling packages such as **rjags** by allowing users to set the desired basis functions and knots. The function `jagam` will then automatically generate the basis function matrix $\boldsymbol{X}$ as well as the non-diagonal matrix $\boldsymbol{M}$, depending on the specifications of the spline.

Two common types of splines used in this work are thin plate and cubic splines. Both types of splines are smooth, allowing for intuitive prediction of future data points using historical information. This is particularly useful for applications where underlying trends are complex and not easily captured by simple linear models. Thin plate splines provide smooth and flexible solutions, making them ideal for modelling spatial and temporal trends in data where smooth transitions are expected across multiple dimensions. On the other hand, cubic splines are often used for capturing one-dimensional trends (e.g. temporal effects) and offer a balance between flexibility and computational efficiency. The cubic spline basis functions are piecewise polynomials, ensuring smoothness across the knots with continuous first and second derivatives.

In a Bayesian framework, smoothness is controlled by introducing priors on the model coefficients, which implicitly enforce smoothness through the prior distribution, such as the Multivariate-Normal prior in Equation (2.7). Moreover, in the Bayesian context, all splines naturally exhibit some form of shrinkage because there are two penalty parameters: one for the linear trend and one for the non-linear trend. In models with shrinkage splines, a single penalty parameter can shrink both the linear and non-linear components simultaneously, which prevents overfitting, particularly in situations with limited data.

The number and placement of knots is an important aspect of spline construction. Knots define the upper limit of flexibility of the spline, which is then penalised by penalty parameters. The more knots that are placed in the model, the more flexible the spline can be. However, placing too many knots can lead to overfitting, as the spline will become overly sensitive to small variations in the data. Conversely, too few knots may result in an overly simplistic model that cannot capture the true underlying trend.

In a Bayesian setting, knots are penalised by the precision matrix $\mathbf{\Omega}$, which ensures that flexibility is controlled. This allows for a balance between flexibility and smoothness. However, a larger number of knots comes at a computational cost due to the Multivariate Normal formulation of the spline, so knots are often chosen to give a bare minimum of flexibility to avoid unnecessary computational costs. In practice, we choose the number of knots by fitting splines to historical data and assessing the ability of the model to capture both short-term and long-term trends.

2.1.3 INLA

One method for approximating the posterior, proposed by Rue, Martino and Chopin 2009 and discussed in more depth than we cover here, is Integrated Nested Laplace Approximations (INLA).

INLA is a fast and automatic method of model fitting which approximates the posterior by joint Gaussian approximations. It can be implemented using the R programming (R Development Core Team (2011)) package **r-inla** by Lindgren and Rue (2015). A compelling benefit of INLA is its reduction of computational cost; it allows many latent Gaussian models to be run within seconds or minutes that could otherwise take much longer using alternative sampling methods such as Markov chain Monte Carlo (MCMC). Additionally, it is relatively accessible and user friendly, owing to its use of a formula based syntax widely adopted in other R packages. This makes it an attractive choice for nowcasting frameworks.

One limitation of this method is that its only suitable for the subset of Bayesian additive models known as latent Gaussian models. This requires the response variable $\boldsymbol{y}$ to come from an exponential family and with Gaussian priors assigned to any unknown functions, linear effects, unstructured effects and intercepts. Moreover, software to implement the INLA method only supports certain families of probability distributions. There will also be some degree of approximation error when implementing INLA for any application due to the assumptions about the data it requires to yield results. However, Rue, Martino and Chopin (2009) show that this can be less than the equivalent MCMC sampling error for a number of examples.

The original INLA methodology, using nested laplace approximations, can be summarised as the following modelling approach. For a given model, let $\boldsymbol{y}$ be the response variable, $\boldsymbol{x}$ be a set of latent Gaussian variables and $\boldsymbol{\theta}$ be the hyper-parameters. Laplace approximation (first suggested by Tierney and Kadane (1986)) can approximate the conditional distribution of $\boldsymbol{\theta}|\boldsymbol{y}$ by dividing the joint posterior density of the Bayesian model by a Gaussian approximation of the marginals of the latent variables ($\tilde{\boldsymbol{\pi}}_G(\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{y})$), all evaluated at the mode of the conditional density function of $\boldsymbol{x}$ for a given $\boldsymbol{\theta}$. Using INLA to approximate the posterior marginals of the model of interest, can be roughly outlined in three steps:

1. Calculate a Laplace approximation to the posterior marginal of the model parameters, conditional on the model response variables $\boldsymbol{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$.
2. Calculate a Laplace approximation of $\boldsymbol{\pi}(x_i|\boldsymbol{y},\boldsymbol{\theta})$ for selected values of parameters $\boldsymbol{\theta}$ to improve the Gaussian approximation.
3. Use nested numerical integration (Equations (2.10)-(2.11)) to combine the approximations in the previous two steps to obtain an approximation of the posterior $\tilde{\boldsymbol{\pi}}(x_i|\boldsymbol{y})$.

$$\tilde{\boldsymbol{\pi}}(\boldsymbol{\theta}_j|\boldsymbol{y}) = \int \tilde{\boldsymbol{\pi}}(\boldsymbol{\theta}|\boldsymbol{y}) d\boldsymbol{\theta}_{-j} \tag{2.10}$$

$$\tilde{\boldsymbol{\pi}}(x_i|\boldsymbol{y}) = \int \tilde{\boldsymbol{\pi}}(x_i|\boldsymbol{\theta},\boldsymbol{y}) \tilde{\boldsymbol{\pi}}(\boldsymbol{\theta}|\boldsymbol{y}) d\boldsymbol{\theta} \tag{2.11}$$

As detailed in Van Niekerk et al. (2023), recent developments in the **r-inla** methodology have introduced improvements in computational efficiency, numerical stability, and scalability for large datasets. One key enhancement is the restructuring of the latent field to remove the inclusion of linear predictors. In the original INLA methodology, the linear predictors were part of the latent field, leading to increased computational costs. The updated formulation instead defines the linear predictors separately as deterministic functions of the latent field, which reduces the size of the augmented latent model and enhances efficiency, particularly for data-rich models.

Additionally, the updated INLA framework incorporates a Variational Bayes correction to the posterior means of the latent field. This correction refines the Gaussian approximations used in INLA, achieving accuracy comparable to MCMC methods while maintaining the computational speed benefits of INLA. By applying low-rank corrections to the mean estimates, this approach ensures that the removal of linear predictors is not detrimental to inference quality.

The updated INLA methodology now has the following steps:

1. Define the latent field without explicit inclusion of linear predictors, thereby reducing model complexity.
2. Approximate the posterior distribution of the hyperparameters $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ using a Laplace approximation, avoiding nested approximations where possible.
3. Use a Variational Bayes correction to refine the posterior means of the latent field, ensuring greater accuracy in inference.
4. Compute the posterior marginals $\pi(x_i|\boldsymbol{y})$ through improved numerical integration, leveraging parallelized computation where applicable.
5. If necessary, perform additional Monte Carlo sampling for predictions, especially in large-scale spatial and spatio-temporal models.

Since this has been implemented as the default methodology in for the `inla(.)` function in the **r-inla** package (Lindgren and Rue (2015)), this is the framework used for any application using INLA throughout this thesis.

2.2 Bayesian Nowcasting Models

The nowcasting frameworks in this review assume that, if we are at current time T_{now}, for all discrete time points $t \leq T_{now}$ in the past a total number of counts y_t occurred and will eventually be reported. Recall from Section 1.1 that the total counts y_t can then be split into partial counts $z_{t,d}$, which are the number of cases that occurred at t and were reported with delay d after t. Furthermore, a maximum delay D_{max} is usually chosen such that after D_{max} time-steps all counts are assumed known, therefore all y_t are fully reported for $t < T_{now} - D_{max}$. We define the cumulative counts, $c_{t,T_{now}} = \sum_{d=0}^{T_{now}-t} z_{t,d}$, as the counts that have been reported by the current time T_{now}, and occurred at time t, where $T_{now} - t < D_{max}$. If nowcasting applications include a spatial dimension (e.g. different geographic regions),

as opposed to focusing on a single time series, then we denote for spatial regions s the total as $y_{t,s}$ and the relative partial counts as $z_{t,d,s}$. In this section, readers are advised some notation for external methods have been rewritten in equivalent notation, to aid cohesion and comparison.

Modelling of data subject to delayed reporting was first carried out in the context of predicting insurance claims. For these early models, nowcasting involved modelling the delay distribution followed by modelling the estimated total counts as a second step (Brookmeyer and Gail (1994)). Methods that modelled total incidence and the delay distribution in one step were then developed for nowcasting AIDS cases, such as in Zeger, See and Diggle (1989), Lawless (1994) and Brookmeyer and Gail (1994). Also, Lawless (1994) noted that the delay distribution often changed un-systematically over time and modelled this random temporal variability by introducing random effects in the delay distribution. Subsequent existing approaches to Bayesian modelling of count data subject to reporting delays can be broadly summarised into these two main groups: 1) joint models and 2) conditionally independent models.

2.2.1 Jointly modelling the total and partial counts

Joint modelling approaches first assume that the total counts y_t come from some probabilistic model Y,

$$y_t \sim Y(\boldsymbol{\Theta}). \tag{2.12}$$

Then, they assume that the partial counts $\boldsymbol{z}_t$, given the total count y_t, come from a another probabilistic model Z:

$$\boldsymbol{z}_t \mid y_t \sim Z(\boldsymbol{\Pi}, y_t). \tag{2.13}$$

Joint models are called as such because we implement both of these layers together in a single hierarchical framework. In these general terms, we can draw a clear equivalence with the modular framework for flawed observation proposed by Stoner (2019) and discussed in Chapter 1. Here, Bayesian probability provides a direct approach to predicting unobserved y_t given any observed partial counts $z_{t,d}$, and the key challenge is in specifying appropriate models Y and Z to capture the main sources of variability associated with the disease counts and the delayed reporting appropriately (Section 1.2).

One example of a joint model is the proposed approach for "Joint Bayesian Modeling of Epidemic Curve and Delay Distribution" in Höhle and An Der Heiden (2014). This is a hierarchical Bayesian model that first assumes Y is a Poisson model for the total counts:

$$y_t \sim \text{Poisson}(\lambda_t), \tag{2.14}$$

$$\log(\lambda_t) = f(t). \tag{2.15}$$

The mean of the total counts λ_t is modelled by a selection of smooth spline effects and polynomial effects, which vary over time, in $f(t)$. The delay model Z is then given by the Multinomial model:

$$\boldsymbol{z_t}|y_t \sim \text{Multinomial}(\boldsymbol{p_t}, y_t), \tag{2.16}$$

$$p_{t,d} = (1 - \sum_{i=0}^{d-1} p_{t,i}) h_{t,d}, \tag{2.17}$$

$$\text{logit}(h_{t,d}) = \gamma_d + \boldsymbol{W}'_{t,d}\boldsymbol{\eta}. \tag{2.18}$$

The proportion reported at each delay, $p_{t,d}$, is modelled by the discrete time hazard $h_{t,d}$ and is able to vary across time and delay, owing to covariates that depend on time and delay, $\boldsymbol{W}_{t,d}$, and quadratic spline over delay γ_d. For, this quadratic spline knots are placed equidistant over time up to time $T_{now} - D/2$, to avoid over-extrapolation for the most recent dates where uncertainty is larger.

The main limitation with a model like this one is the assumption of simple and inflexible probability distributions for Y and Z, in this case Poisson and Multinomial distributions, which lack variance parameters to fit different real data problems well. Recall from Section 1.2 that if Y or Z are not appropriate, then predicted total counts based on $p(y_t \mid \boldsymbol{z}_t)$ will not be appropriate either.

Due to their hierarchical structure, a more general challenge associated with joint models is they usually involve hand-written code for MCMC software, rather than convenient "off the shelf" packages. In some cases we may be able to choose Y and Z such that we can obtain an exact marginal probability model for the $z_{t,d}$, $p(z_{t,d})$, through summation over y_t:

$$p(z_{t,d}) = \sum_{y_t} p(z_{t,d}|y_t)p(y_t). \tag{2.19}$$

In the case of the model from Höhle and An Der Heiden (2014) (Equations (2.14)–(2.18)), the exact marginal model for $z_{t,d}$ is the following Poisson model, as derived in Appendix B.2:

$$z_{t,d} \sim \text{Poisson}(\mu_{t,d}), \tag{2.20}$$

$$\log(\mu_{t,d}) = \log(\lambda_t) + \log(p_{t,d}). \tag{2.21}$$

Similarly, Salmon et al. (2015) replace the Poisson model for the total counts with a more flexible Negative-Binomial model, to allow for over-dispersion with dispersion parameter θ, however they assume the probability vector p is constant over time:

$$y_t|\lambda_t \sim \text{Negative-Binomial}(\lambda_t, \theta), \tag{2.22}$$

$$\boldsymbol{z}_t|y_t \sim \text{Multinomial}(\boldsymbol{p}, y_t). \tag{2.23}$$

The total counts can be summed out to give the following marginal model, as derived in Appendix B.3:

$$z_{t,d} \sim \text{Negative-Binomial}(\mu_{t,d}, \theta), \tag{2.24}$$

$$\log(\mu_{t,d}) = \log(\lambda_t p_d) = \beta_0 + \beta_1 t + \gamma(t) + \alpha_d, \tag{2.25}$$

$$\gamma(s(t)) \sim \text{Normal}(0,_s (t)^2), \tag{2.26}$$

$$\beta_i \sim \text{Normal}(0, \sigma^2_{\beta_i}) \quad i = \{0,1\}, \tag{2.27}$$

$$\alpha_d \sim \text{Normal}(0, \sigma^2_{\alpha_d}) \tag{2.28}$$

where $\gamma(t)$ is a seasonal factor effect, β_0 is an intercept term, β_1 is the linear coefficient of time t and α_d is factor variable to account for delay. Such models can be fit as non-hierarchical models, e.g. as a Generalized Linear Model (GLM) or Generalized Additive Model (GAM). However, obtaining and fitting an exact marginal model for the individual partial counts $z_{t,d}$ does not guarantee access to an appropriate predictive model for the totals given observed $z_{t,d}$, i.e. $p(y_t \mid \boldsymbol{z}_t)$. Typically, one relies on prediction of unobserved $z_{t,d}$ and summation of both observed and predicted $z_{t,d}$ to produce predictions of y_t, which relies on the assumption that $z_{t,d}$ are conditionally independent. This assumption is valid in the simple Poisson-Multinomial case but, generally, approaches motivated as a joint models but then implemented as marginal models fall within a broader family of approaches reliant on the suitability of this conditional independence assumption, which we discuss in the next subsection.

2.2.2 Conditionally independent models of the partial counts

The second group of approaches do not implement a model for the total counts y_t explicitly and instead only assume some model Z for the partial counts z, independent of the total counts:

$$z_{t,d} \sim Z(\boldsymbol{\Theta}, \boldsymbol{\Pi}), \tag{2.29}$$

where $\boldsymbol{\Theta}$ and $\boldsymbol{\Pi}$ represent parameters and/or random effects intended to capture disease incidence and the delay distribution, respectively.

These approaches make an assumption that the partial counts $z_{t,d}$ are independent of each other across delay d, conditional upon any covariates or temporal, spatial and delay structures, including for prediction of the total counts. As such, we call them conditional independence model. Such approaches can be motivated by summing the total counts out of an initial joint model, as discussed in the previous subsection, but many do not. Ideally, a model within this group should be designed so that the systematic variability of the total counts can still be inferred, since the predictant of interest is $y_{t,s}$.

The widely cited Bastos et al. (2019) assumes a Negative-Binomial model for spatio-temporally indexed partial counts $z_{t,d,s}$ and aims to capture systematic variability associated with the disease count generating process and the delay mechanism over time through the combination of various flexible functions of time, space, and delay in the expected mean $\mu_{t,d,s}$:

$$z_{t,d,s} \sim \text{Negative-Binomial}(\mu_{t,d,s}, \theta). \tag{2.30}$$

Functions within the mean $\mu_{t,d,s}$ include cyclical seasonal effects, to capture strong seasonal structures in the diseases the model is applied to. Predictions are then obtained for the unobserved partial counts which are summed to give estimates of the total counts by $y_{t,s} = \sum_{d=0}^{D} z_{t,d,s}$. In conjunction with the Negative-Binomial distribution given by Equa-

tion (2.30) the following equations fully specify the model:

$$\log(\mu_{t,d,s}) = \iota + \alpha_t + \beta_d + \gamma_{t,d} + \eta_{w(t)} + \psi_s + \beta_{d,s} + \boldsymbol{X}'_{t,d}\delta, \tag{2.31}$$

$$\beta_{d,s} \sim \text{Normal}(\beta_{d-1,s}, \omega^2_\beta), \tag{2.32}$$

$$\psi_s = \psi_s^{IAR} + \psi_s^{ind}, \tag{2.33}$$

$$\alpha_t \sim \text{Normal}(\alpha_{t-1}, \sigma^2_\alpha) \tag{2.34}$$

$$\beta_d \sim \text{Normal}(\beta_{d-1}, \sigma^2_{\beta 1}) \tag{2.35}$$

$$\beta_{d,s} \sim \text{Normal}(\beta_{d,s-1}, \sigma^2_{\beta 2}) \tag{2.36}$$

$$\gamma_{t,d} \sim \text{Normal}(\gamma_{t-1,d}, \sigma^2_\gamma) \tag{2.37}$$

$$\eta_{w(t)} \sim \text{Normal}(2\eta_{w-1} - \eta_{w-2}, \sigma^2_\eta). \tag{2.38}$$

In Equation (2.31), ι is the log of the mean overall total count, α_t and β_d are first order random walk effects that aim to capture the temporal trend in the total count and the mean proportion reported at each delay, respectively. An additional random walk term, $\beta_{d,s}$, is included to capture the way in which the delay structure varies across space. Then, $\gamma_{t,d}$ is a time-delay interaction which allows for temporal changes in the delay distribution with a first order random walk over time which is independent for each delay. This is key in capturing potential changes in the delay distribution that are dependent on time. The cyclical seasonal effect $\eta_{w(t)}$ is a second-order random effect that captures temporal variability in the total counts related to the week of the year $w(t)$. Finally, $\boldsymbol{X}'_{t,d}$ is a matrix of temporal and delay related covariates. We discuss features of the model that aim to capture spatial variation in both the disease count and the delay mechanism in detail, in Section 2.5.1. However, due to the nature of conditional independence approaches, the partial counts, $z_{t,d}$ are assumed to be independent given the model effects and any covariates. Therefore, there is no constraint that proportions reported at each delay sum to one. Furthermore, including the effects related to both the partial counts and the total counts in a single log link will limit the models ability to fully capture and identify the different sources of systematic and random variability. Hence, over-dispersion of the partial counts could be incorrectly absorbed by the total counts.

McGough et al. (2020) also assume a Negative-Binomial model for $z_{t,d}$ with dispersion parameter θ. But, unlike Bastos et al. (2019), they attempt to better separate the systematic variability of the epidemic curve and delay distribution by decomposing the mean of $z_{t,d}$, $\mu_{t,d}$, into a part relating to the incidence of the disease and a part capturing the delay distribution, i.e. the mean proportion p_d of y_t reported at delay index d:

$$z_{t,d} \sim \text{Negative-Binomial}(\mu_{t,d}, \theta), \tag{2.39}$$

$$\log(\mu_{t,d}) = \alpha_t + \log(p_d), \tag{2.40}$$

$$\alpha_t \sim \text{Normal}(\alpha_{t-1}, \sigma_\alpha^2), \tag{2.41}$$

$$p_d \sim \text{Dirichlet}(\boldsymbol{\beta}). \tag{2.42}$$

The mean of the total counts is captured by the first order random walk term α_t and a Dirichlet prior is assumed for $\boldsymbol{p}$.

Temporal variation in the delay mechanism is allowed outside of the model by only including data within a moving window: the size of this windows controls the amount of historic data that informs the parameters. The results in McGough et al. (2020) showed a smaller window resulted in faster computational speeds and improved delay distribution estimates but with a trade-off of "more volatile" estimates of weekly cases and decreasing nowcasting accuracy in certain periods. The R package **NobBS**, which supplements their paper, includes a function that allows users to fit the model using MCMC. However, it is worth noting that while this package allows for the model to be stratified by spatial regions or other variables (which is not discussed in McGough et al. (2020)) it does not allow the inclusion of possible informative covariates. For this approach, the partial counts are still assumed to be conditionally independent and there is no allowance for the delay distribution (as quantified by $\boldsymbol{p}$) to vary over time within the moving window. Also, this model does not account for spatial variation or interactions between time and delay.

Günther et al. (2020), also use a first order random walk prior for the expected mean of total counts at day t, λ_t, and applied the following model for COVID-19 cases in Bavaria. The expected mean of the totals, and the expected probability of reporting at delay d, are multiplied to give the expected mean ($\mu_{t,d}$) of the partial counts $z_{t,d}$, which are modelled by a Negative-Binomial distribution:

$$z_{t,d}|\lambda_t, p_{t,d} \sim \text{Negative-Binomial}(\lambda_{t,d}p_{t,d}, \theta), \tag{2.43}$$

$$\log(\lambda_0) \sim \text{N}(0,1), \tag{2.44}$$

$$\log(\lambda_t)|\lambda_{t-1} \sim \text{Normal}(\log(\lambda_{t-1}), \phi), \tag{2.45}$$

$$p_{t,d} = \left(1 - \sum_{i=0}^{d-1} p_{t,i}\right) h_{t,d}, \tag{2.46}$$

$$p_{t,0} = h_{t,0}. \tag{2.47}$$

Additionally, this model uses a discrete time hazard model for the probability vector $p_{t,d}$. Four different versions of the hazard function $h_{t,d}$ were considered for the Negative-Binomial model, Equation (2.43)–(2.47), for the partial counts of the COVID-19 cases in Bavaria. The first allowed for changes in the reporting delay with linear time effects with a two week change point $\text{logit}(h_{t,d}) = \gamma_d + \boldsymbol{W}'_{t,d}\boldsymbol{\eta}$, and the second used $\text{logit}(h_{t,d}) = \gamma_d + \alpha_t$, such that α_t is a first order random walk term, which will therefore be able to capture the daily changes in the reporting delay. Each of these models were then fitted with and without a reporting weekday effect to allow for the fact reporting of COVID-19 cases is reduced at the weekend. This approach attempts to separate the systematic variability in the totals and delay distribution, whilst allowing for more flexibility when modelling the expected proportion reported than McGough et al. (2020). However, it is similarly limited in capturing the covariance of the partial counts, since $p_{t,d}$ is fixed given the time-delay effects/covariates.

Recall from Section 1.2 that, to achieve an appropriate $p(y_t|\boldsymbol{z}_t)$ for prediction, we need appropriate $p(\boldsymbol{z}_t|y_t)$ and $p(y_t)$. As noted by Stoner and Economou (2019), an inappropriate conditional independence assumption means that $p(\boldsymbol{z}_t|y_t)$ is not appropriately specified, potentially leading to over-fitting and unreliable quantification of prediction uncertainty.

Notably, in their simulation experiment (Supporting information, Web Appendix A), Stoner and Economou (2019) found that predictions of y from a model designed to be conceptually similar to Salmon et al. (2015) had excessive variance compared to the total counts in the original data.

2.2.3 Comparison of joint and conditional independence models

The advantage of a "conditional independence" model is avoiding a hierarchical structure for $y_{t,s}$ and $z_{t,d,s}|y_{t,s}$. This allows for implementation using a wider variety of approaches, including frequentist approaches or other Bayesian methods e.g. Integrated Nested Laplace Approximations (INLA), detailed in Section 2.1.3. For example, the model from Salmon et al. (2015), Equation (2.28), is essentially a Generalized Linear Model, since the Negative-Binomial belongs to the exponential family given when the dispersion parameter θ is known (in practice, θ is estimated simultaneously in fitting algorithms). We can therefore conceive frequentist variations of this model based on e.g. Generalized Additive Models (GAMs) or Generalized Linear Models (GLMs). Such models may also be expressible as latent Guassian models, enabling inference using INLA (Section 2.1.3). Bastos et al. (2019) carried out a comparison where their proposed model was fitted using both INLA and MCMC methods: they found that MCMC was more accurate once fully converged, but in time-sensitive scenarios the speed and reduced user input of INLA made it more practical. Hence, the INLA models are currently operational in the warning systems infoDengue (https://info.dengue.mat.br) and infoGripe (http://info.gripe.fiocruz.br). These alert systems allow national and local authorities to make decisions based on current and future predictions of the number of dengue and SARI cases in Brazilian states. Section 1.1 outlines the impact of these surveillance systems in more detail.

Although they generally excel with respect to timeliness, the fact that marginal approaches don't explicitly model the total counts y_t means they are theoretically restricted in the capacity to capture the random variability in the total counts well since they are not explicitly included in the model. If, as in these approaches, y_t is defined as $y_t = \sum_{d=1}^{D_{max}} z_{t,d}$, for prediction, then we have that:

$$\text{Var}[y_t] = \text{Var}\left[\sum_{d=1}^{D_{max}} z_{t,d}\right] = \sum_{i=1}^{D_{max}} \sum_{j=1}^{D_{max}} \text{Cov}[z_{t,i}, z_{t,j}]. \tag{2.48}$$

This implies that, to capture $\text{Var}[y_t]$ well, we must capture the covariance of $z_{t,d}$ across delay well. However, models that assume that the partial counts are conditionally independent have much less flexibility/control to capturing their covariance structure, compared to e.g. a multivariate model for $\boldsymbol{z}_t$. They inherently rely on the belief that the conditional independence assumption holds or is at least a close approximation, which in turn means they are relying on covariate and/or random effects to explain all of the covariance.

Stoner and Economou (2019) argue that joint model designs generally allow for greater flexibility and control over appropriately capturing and separating the different sources of systematic and random variability in the data (Section 1.1). We can also remark that they yield posterior predictive inference for unseen total counts and that they can be conceptually explained based on the modular framework for flawed observation from Stoner and Economou (2019). As explained in Section 1.2, the main challenge is ensuring the models for the total counts and the delayed reporting, $\boldsymbol{Y}$ and $\boldsymbol{Z}$ respectively, are appropriate. The main limitation of existing works prior to Stoner and Economou (2019) was the reliance on Multinomial models for $\boldsymbol{Z}$, which have no flexibility to account for different levels of random variability in $z_{t,d}$, in absolute terms and relative to each other, since $\text{Cov}[z_{t,i}, z_{t,j}]$ is fixed given $p_{t,d}$, and y_t, such that we are relying on model effects/covariates to fully explain the covariance.

To summarise, all modelling frameworks discussed so far all have at least one theoretical limitation in appropriately capturing and separating the four sources of variability in data suffering from delayed reporting (Section 1.1):

- Conditional independence approaches may generally struggle to capture the random variability of the total counts y_t – the model never sees them – and may not produce reliable predictive uncertainty, when the conditional independence assumption is not a close approximation.
- Joint modelling approaches offer greater potential to appropriately separate and capture the different sources of variability, but existing methods are held back by assuming a Multinomial model for the partial counts.

2.3 Generalized-Dirichlet Multinomial Model

To address the lack of a general framework that is able to appropriately separate and capture the four main sources of variability in data suffering from delayed reporting, as discussed in Chapter 1.2, Stoner and Economou (2019) proposed the Generalized-Dirichlet Multinomial (GDM) method. As a joint modelling framework (Section 2.3), the GDM (Equations (2.49)–(2.53)) separates the processes generating total counts and capturing the delay mechanism into hierarchical layers, hence capturing all sources of variability in the data.

Like other approaches that came before, e.g. Salmon et al. (2015), the GDM assumes a flexible Negative-Binomial (acting as Y in the modular framework from Section 1.2):

$$y_t|\lambda_t, \theta \sim \text{Negative-Binomial}(\lambda_t, \theta), \tag{2.49}$$

$$\log(\lambda_t) = f(t). \tag{2.50}$$

Here, $f(t)$ could be any combination of effects or covariates to capture the mean temporal trend in the total counts. We then assume a Multinomial distribution for $p(\boldsymbol{z_t} \mid \boldsymbol{p}_t, y_t)$:

$$\boldsymbol{z}_t \mid \boldsymbol{p}_t, y_t \sim \text{Multinomial}(\boldsymbol{p}_t, y_t). \tag{2.51}$$

However, where this framework deviates from what came before (e.g. Höhle and An Der Heiden (2014)) is by then assuming that the Multinomial follow an i.i.d. Generalized-Dirichlet (GD) distribution:

$$\boldsymbol{p}_t \sim \text{Generalized-Dirichlet}(\boldsymbol{\alpha}_t, \boldsymbol{\beta}_t). \tag{2.52}$$

Here, the GD acts as additional source of variability, introducing more degrees of freedom to better capture the random variability and covariance in the delayed reporting mechanism. Finally, assuming a Generalized-Dirichlet (GD) prior for $\boldsymbol{p}_t$ yields a Generalized-Dirichlet Multinomial (GDM) model for $\boldsymbol{z_t} \mid y_t$ (acting as $\boldsymbol{Z}$ in the modular framework from Chapter 1.2), as shown in Appendix B.4:

$$\boldsymbol{z}_t | \boldsymbol{p}_t, y_t \sim \text{GDM}(\boldsymbol{p}_t, y_t). \tag{2.53}$$

This introduces greater flexibility for the partial counts than a Multinomial model alone, recalling that the variance of which is fixed given $\boldsymbol{p}_t$ and y_t.

As an aside, the GD has as many degrees of freedom for the variance of $\boldsymbol{p}_t$ as the length of $\boldsymbol{p}_t$ minus 1. Thus, the GD has much greater flexibility than a Dirichlet model for $\boldsymbol{p}_t$, which only has one degree of freedom for the variance.

Overall, the GDM offers considerably more flexibility in capturing the covariance structure of the partial counts, compared to alternative choices discussed in the previous subsections. This is achieved by not assuming conditional independence between delays and enabling the variance to not be fixed given the total counts and the expected probabilities. Hence, reducing the risk of the model confounding variability between the totals and the delay distribution, where over-dispersion the covariance structure of $p(\boldsymbol{z_t}|y_t)$ (e.g. with respect to a Multinomial) could be absorbed by $p(y_t)$.

2.3.1 Conditional series

The GD distribution used in the context of delayed reporting is constructed as a series of independent Beta distributions in a "stick-breaking" fashion:

$$p_{t,1} \sim \text{Beta}(\nu_{t,1}, \phi_{t,1}); \tag{2.54}$$

$$\frac{p_{t,2}}{1-p_{t,1}} \sim \text{Beta}(\nu_{t,2}, \phi_{t,2}); \tag{2.55}$$

$$\dots \tag{2.56}$$

$$\frac{p_{t,d}}{1-\sum_{i=1}^{d-1} p_{t,i}} \sim \text{Beta}(\nu_{t,d}, \phi_{t,d}); \tag{2.57}$$

$$\dots \tag{2.58}$$

$$p_{t,D_{max}} = 1 - \sum_{i=1}^{D_{max}-1} p_{t,i}, \tag{2.59}$$

where each Beta is parameterised in terms of its mean $\nu_{t,d}$ and dispersion parameter $\phi_{t,d}$. Hence, for this parametrisation the mean of the Beta distribution is $\mathbb{E}[X] = \nu$ and the variance is $\text{Var}[X] = \frac{\nu(1-\nu)}{1+\phi}$, see Appendix A.2 for a more formal definition of the Beta distribution.

Meanwhile, the Multinomial can be expressed as a series conditional Binomial distributions:

$$z_{t,1} \mid p_{t,1}, y_t \sim \text{Binomial}(p_{t,1}, y_t); \tag{2.60}$$

$$z_{t,2} \mid p_{t,2}, z_{t,1}, y_t \sim \text{Binomial}(p_{t,2}, y_t - z_{t,1}); \tag{2.61}$$

$$\dots \tag{2.62}$$

$$z_{t,d} \mid p_{t,d}, z_{t,<d}, y_t \sim \text{Binomial}(p_{t,d}, N_{t,d}); \tag{2.63}$$

$$\dots \tag{2.64}$$

$$z_{t,D_{max}} = N_{t,D_{max}}, \tag{2.65}$$

where $N_{t,d} = y_t - \sum_{j=1}^{d-1} z_{t,j}$. Note that the partial counts for the final delay is equal to the total counts minus the partial counts up to $D_{max}-1$, $z_{t,D_{max}} = y_t - \sum_{j=1}^{D_{max}-1} z_{t,j}$. Hence, for the GDM model, if only D delays are explicitly modelled, where $D < D_{max}$, the final partial count $z_{t,D} = y_t - \sum_{j=1}^{D-1} z_{t,j}$ will capture all remaining partial counts reported between delay $D+1$ and D_{max}. As such, we define $D < D_{max}$ as the maximum number of delays explicitly modelled that are less than the maximum number of possible delays.

Therefore, it can be shown that the GDM can be expressed as a series of Beta-Binomial distributions, as shown in Appendix B.5:

$$z_{t,d} \mid z_{t,<d}, y_t \sim \text{Beta-Binomial}(\nu_{t,d}, \phi_{t,d}, N_{t,d}). \tag{2.66}$$

The interpretation of $\nu_{t,d}$, which we call "expected proportions" here is the proportion of $N_{t,d}$ we expect to be reported at delay d, i.e. are given by $\nu_{t,d} = E\left[\frac{z_{t,d}}{N_{t,d}}\right]$. The dispersion parameters $\phi_{t,d}$ controls the variability around the expected relative proportions. As $\phi_{t,d} \to \inf$ the Beta-Binomial reduces to a Binomial and, as such, if $\phi_{t,d} \to \inf$ the GDM reduces to a Multinomial.

The Beta-Binomial series representation of the GDM is very convenient for implementation in MCMC software. For t where only some $z_{t,d}$ are observed, we need only include the Beta-Binomials corresponding to observed $z_{t,d}$. For such t, y_t is unobserved and need only be sampled from an appropriate algorithm, generating predictions from the posterior distribution given any observed $z_{t,d}$. If desired, posterior predictions of unobserved $z_{t,d}$ can be produced using Monte Carlo simulation.

In summary, the GDM framework is a compelling option for separating and capturing the systematic and random variability in both the total counts and the delay distribution. In the next subsection, we explain how the parameters of the GDM can be characterised in terms of structures explaining variation in time and across delays. Further details on including spatial effects within the GDM are given in Section 2.5.1. Later in the thesis (Chapter 4), we extend the GDM to allow for systematic variability over space and time, as well as space-time interactions, in the delay distribution and in the total counts.

However, the GDM framework still has both practical and theoretical limitations. First, as we will discuss in more detail in Section 2.5.4, a barrier to the implementing the GDM in an operational setting is its relatively slow computational speed. This is a result of its complex structure and the requirement of MCMC sampling methods. We attempt to address this issue in Chapter 3. Furthermore, neither the GDM nor any other approach currently in the literature accounts for the potential relationship between the reporting process and the incidence of the disease. As we discussed in Section 1.1.2 this could induce a systematic trend in the delay distribution. Thus, not accounting for this could limit the information models have to improve predictive precision. We investigate the potential relationship between incidence and delay more thoroughly in Chapter 6.

2.3.2 Existing GDM link functions

When defining a GDM modelling framework Stoner and Economou (2019) provide two options for linking covariate/random effects to the Beta-Binomial mean parameters, $\nu_{t,d}$ (the expected relative proportions), to capture systematic variability over time and delay in the delayed reporting model.

First, they suggest a "hazard" model:

$$\log\left(\frac{\nu_{t,d}}{1-\nu_{t,d}}\right) = g(t,d), \tag{2.67}$$

which directly models the relative proportions at the logit scale. The function $g(t,d)$ is given as a general combination of covariates and random effects. This option allows for relatively flexible choices for introducing time and delay effects or covariates in $g(t,d)$, since the logit function will translate any value of $g(t,d)$ on the unbounded real line to the appropriate parameter space $\nu_{t,d} \in (0,1)$.

However, intuitively thinking about structured variability within the relative proportions can be difficult. This is particularly true when considering $\nu_{t,d}$ for later delays indices. Recall that $\nu_{t,d} = E\left[\frac{z_{t,d}}{N_{t,d}}\right]$, is the expected part of y_t that is yet to be reported ($N_{t,d}$) that we expect to be reported at delay d. For later delays, both $z_{t,d}$ (the numerator) and $N_{t,d}$ (the denominator) are likely to be small, such that small changes in either could change the value of the fraction considerably, thus muddying our intuition for designing structured variability in $\nu_{t,d}$.

Second, Stoner and Economou (2019) suggested a "survivor" model:

$$\text{probit}(S_{t,d}) = g(t,d), \tag{2.68}$$

where $g(t,d)$ is again a general term. Here, $S_{t,d} = E\left[\frac{\sum_{i=0}^{d} z_{t,i}}{y_t}\right]$ is the cumulative proportion of y_t expected to be reported up to and including delay d. The relative proportions are derived as $\nu_{t,d} = \frac{S_{t,d} - S_{t,d-1}}{1 - S_{t,d-1}}$.

Since the cumulative proportions are being modelled the choices for $g(t,d)$ are restricted to be monotonically increasing over delay, to ensure the condition that the cumulative counts similarly increase over delay. However, considering structures for cumulative proportions is potentially more intuitive than for the relative proportions. Temporal trends are likely to be similar across delays as they get carried from one delay to the next in the cumulative counts reported (the numerator of $S_{t,d}$).

For their application to dengue cases in Rio de Janeiro, Stoner and Economou (2019) made arguments in favour of the survivor version based on intuition, simplicity, and computational expedience. However, this is not a trivial decision as this choice potentially determines how well the model is able to capture the delay distribution. A closer fit will depend on whether the relative or cumulative proportions can be appropriately characterised by the combination of the link function and the structural effects inside $g(t,d)$.

2.3.3 Forecasting with the GDM model

The GDM model can also be used to generate forecasts as well as nowcasts. This is beneficial in disease surveillance contexts as it enables public health bodies to prepare tailored strategies for potential upcoming scenarios. Hence, allowing interventions to be put in place to avoid worse case scenarios rather than just implementing reactive interventions as the scenarios unfold. However, certain alterations are required for the GDM framework to be suitable for forecasting. These are relatively straightforward to implement in part due to the flexibility of the R package NIMBLE (de Valpine et al. (2017)) which can be used to fit the GDM since it is a directed acyclic graph (See Appendix C.1 for

the GDM graph). Stoner and Economou (2019) introduce the GDM alongside its forecasting capabilities. Here we present a survivor version of the forecasting framework with no spatial element, but forecasting can be similarly carried out with the hazard version (both versions are defined in Section 2.3.2) or any other link function. As discussed in Section 2.1.2.2, we often fit the GDM with smooth splines to capture variability over time in the data. Here we will consider a model with cubic splines, but the same considerations are necessary when using any smooth spline within the framework. The GDM model can be defined with the following equations:

$$y_t|\lambda_t,\theta \sim \text{Negative-Binomial}(\lambda_t,\theta), \tag{2.69}$$

$$\log(\lambda_t) = \iota + \alpha_t, \tag{2.70}$$

$$z_{t,d}|\nu_{t,d},\phi_{t,d},N_{t,d} \sim \text{Beta-Binomial}(\nu_{t,d},\phi_{t,d},N_{t,d}), \tag{2.71}$$

$$\text{probit}(S_{t,d}) = \psi_d + \beta_t, \tag{2.72}$$

$$\boldsymbol{\kappa}^{(\alpha)} \sim \text{Multivariate-Normal}(0,\boldsymbol{\Omega}^{(\alpha)}), \tag{2.73}$$

$$\alpha_t = \boldsymbol{X}_t\boldsymbol{\kappa}^{(\alpha)} \tag{2.74}$$

$$\boldsymbol{\kappa}^{(\beta)} \sim \text{Multivariate-Normal}(0,\boldsymbol{\Omega}^{(\beta)}), \tag{2.75}$$

$$\beta_t = \boldsymbol{X}_t\boldsymbol{\kappa}^{(\beta)} \tag{2.76}$$

For the total counts y_t, ι is a intercept and α_t is a cubic spline. Similarly, for the partial counts $z_{t,d}$, ψ_d is a delay specific intercept which is defined by a first order random walk that is constrained to increase over delay and β_t is a cubic spline of time. Additionally, to make the intercept terms more interpretable, the cubic splines are centred to have zero mean. The cubic splines are constructed using the `jagam(.)` function from Wood (2016), which produces JAGS objects for model specification including the spline basis matrix X_t. The spline basis function is constructed such that it is linear during the forecasting period by placing the last knot of the spline at the end of the nowcasting period (often interpreted as the current time we are nowcasting up to). This knot placement is vital to allow for forecasting as knots can't be placed where there is no data as there is no way for

the model to compute or constrain the splines behaviour at that point. Hence, this knot placements is needed for any type of spline when forecasting. For cubic splines, the spline is constrained to be linear in the forecast period beyond the last knot but this will differ with different spline choices.

2.4 Machine learning nowcasting approaches

A recent development in the methods for nowcasing delayed reporting is the adoption of machine learning techniques to gain improvements in computational speed over existing Bayesian and frequentist methods. For example, Sahai et al. (2022) utilise an ensemble machine learning approach that uses a random forest regression model to nowcast incidence of COVID-19 infection in Ohio. A random forest is a collection of trees, each tree is fitted using a process called binary recursive partitioning to create a base learner function. The average of these functions is then taken to obtain the predictor function for the regression (Cutler, Cutler and Stevens (2012)). They report that the machine learning approach had a run time of seconds compared to the 20 hour run time outlined in Kline et al. (2022). Hence, machine learning techniques are a possible avenue for improving nowcasting efficiency.

However, their real-world desirability is conditional on their predictive performance compared to the current state-of-the-art models, which has not yet been thoroughly investigated. For instance, it is unclear whether they are able to offer reliable predictive uncertainty. For this approach, the actual reported counts seem to lie outside the predicted uncertainty for prediction time difference close to zero, as shown in the Sahai et al. (2022) figures comparing the model predictions to the true data for four nowcast dates in November. This approach was compared to the Bayesian Modelling approach presented by Kline et al. (2022) for a single nowcasting date in September, both models over-predict and do not capture the true data in their 95% prediction intervals for the most recent nowcasting

dates. Whilst they found that the medians of the random forest predictions were closer to the eventually observed number of cases than the Bayesian model medians, the lower 95% prediction interval for the Bayesian approach often appear closer to the data than that of the random forest approach. Meanwhile, the choice to base their comparison with existing work on only one historic nowcast date means that the presented results may not be representative of all potential nowcast dates for daily COVID-19 infections in Ohio, in the time period considered. Hence, a more thorough comparison is required to determine the differences in the Sahai et al. (2022) and Kline et al. (2022) approaches for a range of trends in the total counts and delay distribution of the data being considered.

It is worth noting these findings are not representative of all machine learning applications. However, in general a possible disadvantage to these approaches is that inference about the reporting delay, covariate effects and possible temporal, spatial, and seasonal trends in the data are difficult to obtain.

2.5 Model Extensions

The nature of designing operational surveillance systems means that modelling frameworks need to be flexible and adaptable to be suitable for a range of real-world applications. It is common for variations in the data or in requirements for being operational to create scenarios where standard versions of frameworks are not fit for purpose. Where necessary, extension of published methods can help to overcome these obstacles.

In this section we discuss ways that existing nowcasting frameworks have been adapted, and highlight common situations where model flexibility is an advantageous quality. In particular, we will argue that the GDM model is especially well equipped for extension to novel data challenges. This is in part due to its structure clearly separating different sources of variability in the data, as discussed in Section 2.3, but also due to the versatility of the modelling software NIMBLE used to fit it which was discussed in Section 2.1.2.

2.5.1 Spatial variation

In disease surveillance, it is very common for data to be grouped into spatial regions of various scales. Variability across regions is often present in both the epidemic curve and the reporting delay process, however this is not always considered in nowcasting models. In some applications it may be crucial to capture spatio-temporal variation in the total counts, to understand how the disease may be spreading, as well as the spatial variation in the delay distribution, to possibly identify regions with longer reporting delays. As we discuss in Stoner, Halliday and Economou (2022) (Chapter 4), while most nowcasting frameworks for time series can be applied to each spatial data group independently, extending the models to account for spatial variation may aid estimation, especially in spatial regions with low counts, due to the pooling of information across locations (Gelman and Hill (2006)). Moreover, jointly modelling the regions also may allow for the covariance of the total counts across regions to be captured more accurately.

In Equations (2.30)–(2.38), Bastos et al. (2019) consider the spatial variation in the reporting delays and the total counts through the inclusion of a first order random walk (over delay) space-delay term $\beta_{s,d}$, which is independent for each spatial region, as well as the inclusion of structured ψ_s^{IAR} and unstructured ψ_s^{ind} spatial effects.

The structured spatial effect ψ_s^{IAR} is given by an intrinsic conditional autoregressive (ICAR) process,

$$\psi_s^{IAR}|\psi_{-s}^{IAR} \sim \text{Normal}\left(\frac{\sum_{j\neq s}\omega_{s,j}\psi_j^{IAR}}{\sum_{j\neq s}\omega_{s,j}}, \frac{\sigma_{IAR}^2}{\sum_{j\neq s}\omega_{s,j}}\right), \tag{2.77}$$

where $\omega_{s,j}$ indicates whether regions s and j are neighbouring, and σ_{IAR}^2 is the variance which controls the extent to which neighbouring spatial regions have similar temporal trends. The spatially unstructured effect is a random effect $\psi_s^{ind} \sim \text{Normal}(0, \sigma_{ind}^2)$ with variance σ_{ind}^2.

Autoregressive models determine spatial dependencies through the network of neighbouring spatial regions. This is more intuitive for modelling disease data divided by regions, compared to geostatistical models that require a point to represent the entire region to be specified. Spatial autoregressive models that could potentially be included in this framework include simultaneous autoregressive (SAR) or various conditional autoregressive (CAR) models, including intrinsic conditional autoregressive (ICAR) models. These spatial models differ in the way the covariance matrix is calculated (Ver Hoef et al. (2018)).

On the other hand, Stoner, Halliday and Economou (2022) (Chapter 4) propose a version of the GDM model that pools information across regions using nested spline structures:

$$y_{t,s}|\lambda_{t,s},\theta_s \sim \text{Negative-Binomial}(\lambda_{t,s},\theta_s), \tag{2.78}$$

$$\log(\lambda_{t,s}) = f(t,s), \tag{2.79}$$

$$z_{t,d,s}|\nu_{t,d,s},\phi_{t,d,s},N_{t,d,s} \sim \text{Beta-Binomial},(\nu_{t,d,s},\phi_{t,d,s},N_{t,d,s}), \tag{2.80}$$

$$\text{probit}(S_{t,d,s}) = g(t,d,s). \tag{2.81}$$

The splines are created by multiplying $\boldsymbol{X}_t$, which is a model matrix of the basis function at each time point, and $\boldsymbol{\kappa}$, a vector of coefficients. These model matrices and coefficients are defined using the `jagam` function by Wood (2016) (more details about how we define smooth effects in NIMBLE is given in Section 2.1.2.2). The resulting model for the expected mean of the total counts can be defined by,

$$\log(\lambda_{t,s}) = f(t,s) = \iota_s + \delta_{t,s}, \tag{2.82}$$

$$\delta_{t,s} = \boldsymbol{X}_t \boldsymbol{\kappa}_s^{(\delta)}, \tag{2.83}$$

$$\boldsymbol{\kappa}^{(\alpha)} \sim \text{Multivariate-Normal}(0, \boldsymbol{\Omega}^{(\alpha)}), \tag{2.84}$$

$$\boldsymbol{\kappa}_s^{(\delta)} \sim \text{Multivariate-Normal}(\boldsymbol{\kappa}_s^{(\alpha)}, \boldsymbol{\Omega}^{(\delta)}). \tag{2.85}$$

A common zero-mean temporal trend across all regions is captured by α_t, with intercept ι_s. This is not explicitly included in the model but could be calculated by $\alpha_t = \boldsymbol{X}_t \boldsymbol{\kappa}^{(\alpha)}$. To centre $\delta_{t,s}$ on α_t the basis function coefficients of α_t ($\boldsymbol{\kappa}^{(\alpha)}$) are used as the mean of the multivariate Gaussian prior for the basis function coefficients of $\delta_{t,s}$ ($\boldsymbol{\kappa}^{(\delta)}$) for each region. Hence, $\delta_{t,s}$ encompasses both the common trend α_t and the regional differences from this common trend for each region. The spline coefficient precision matrices, such as $\boldsymbol{\Omega}_s^{(\delta)} = \tau_s^{(\delta)} \boldsymbol{M}^{(\delta)}$ and $\boldsymbol{\Omega}^{(\alpha)} = \tau^{(\alpha)} \boldsymbol{M}^{(\alpha)}$, are calculated by scaling known non-diagonal matrices $\boldsymbol{M}$ by the corresponding smoothing parameters τ. Hence, the splines are penalised to prevent over-fitting. The same structure is used for modelling the expected cumulative proportion reported at each delay $S_{t,d,s}$,

$$\text{probit}(S_{t,d,s}) = g(t,d,s) = \beta_{s,d} + \gamma_{t,s} + \xi_{t,s} \tag{2.86}$$

$$\gamma_{t,s} = \boldsymbol{X}_t \boldsymbol{\kappa}_s^{(\gamma)}, \tag{2.87}$$

$$\boldsymbol{\kappa}^{(\psi)} \sim \text{Multivariate-Normal}(0, \boldsymbol{\Omega}^{(\psi)}), \tag{2.88}$$

$$\boldsymbol{\kappa}_s^{(\gamma)} \sim \text{Multivariate-Normal}(\boldsymbol{\kappa}_s^{(\psi)}, \boldsymbol{\Omega}^{(\gamma)}), \tag{2.89}$$

$$\xi_{t,s} = \boldsymbol{X}_t \boldsymbol{\kappa}_s^{(\xi)}, \tag{2.90}$$

$$\boldsymbol{\kappa}^{(\eta)} \sim \text{Multivariate-Normal}(0, \boldsymbol{\Omega}^{(\eta)}), \tag{2.91}$$

$$\boldsymbol{\kappa}_s^{(\xi)} \sim \text{Multivariate-Normal}(\boldsymbol{\kappa}_s^{(\eta)}, \boldsymbol{\Omega}^{(\xi)}). \tag{2.92}$$

The parameters $\gamma_{t,s}$ and $\xi_{t,s}$, capture the regional temporal trend and regional week day effect of the expected cumulative proportion reported respectively. These are centred on ψ_t and η_t (in the same way that δ_t is centred on α_t in Equations (2.82)–(2.85)), and can be calculated by $\psi_t = \boldsymbol{X}_t\boldsymbol{\kappa}^{(\psi)}$ and $\eta_t = \boldsymbol{X}_t\boldsymbol{\kappa}^{(\eta)}$ respectively. Where ψ_t capture the overall temporal trend across all regions and η_t captures the overall weekly effect across all regions. The delay effect, $\beta_{s,d}$, is independent for each region and is a first order random walk conditioned such that $\beta_{s,d} > \beta_{s,d-1}$. This ensures the expected cumulative proportion reported for each time step will increase as delay d increases.

A GDM and spline approach was also used to nowcast deaths in England by region and age group in Seaman et al. (2022). However, they apply their model to each spatial region independently. Also, a few other adjustments were made to the framework, including a different modelling of the weekday and temporal effects. On the other hand, Kline et al. (2022) predicted COVID-19 cases in Ohio using a GDM model as well, but instead of the spline approach they used an autoregressive spatial structural time series as defined by Equation (2.77). Either splines or an IAR process can reasonably capture spatial trends in most situations, especially if the spatial scale is relatively large. One limitation of the nested spline approach is that it doesn't assume any kind of neighbouring structure. Hence, it assumes trends will be similarly related across all regions regardless of distance from each other. For applications where cases between neighbouring regions are strongly correlated this assumption fails to incorporate this information into the model. For such cases, potentially including problems at finer spatial resolutions, more complex spatial modelling structures may be needed to capture the more complex and intricate relationships between regions. Other approaches reviewed in Section 2.2 don't explicitly consider the spatial structure of the data and instead are implemented by fitting the respective models to each spatial region independently.

2.5.2 Incorporating covariates

The majority of modelling techniques discussed so far are able to incorporate covariates in various parts of their frameworks. Covariates can be a powerful tool to help models pick up trends and capture otherwise unknown data generating processes. We can also gain key insights into these processes through inference on the covariate effects. Potentially informative covariates for disease surveillance could include social media, search engine data, or weather forecasts. For example, Bastos et al. (2019) suggest weather patterns, which are informative for diseases such a dengue which are mosquito borne, could be included as covariates in their model given by Equation (2.30). However, currently weather forecasts and twitter feeds are used separately by operational warning systems alongside model predictions.

One attempt at utilising real time covariates in the model is Aiken et al. (2020), they combine an autoregressive (AR) model of the total counts with a linear regression of Google query volumes for a series of search terms chosen using a machine learning framework. The selection of google search queries is carried out using a LASSO (least absolute shrinkage and selection operator) regression technique. This supervised machine learning model was trained on a subset of the data (the training data) and the LASSO cost function then shrinks data values towards a central point to prevent the model from over-fitting. Over-fitting occurs when the model fits the noise in the training data too closely, such that prediction accuracy is worse for unseen data. However, whilst these automatic variable selection tools are efficient they may leave out queries for which there is strong physical intuition for using to inform the model. Incorporating these possibly informative covariates aimed to improve nowcasting and forecasting in cases where there is limited historic data of a disease to train models. Although this model was able to outperform the simple AR model for 3 out of 5 of the diseases investigated, both these models lack the flexibility and robustness of the nowcasting models previously discussed in this chapter.

The presence of delayed reporting in the data is not considered and the assumption that the total counts come from an first order AR process is potentially restrictive. Hence, this approach has limited ability to capture the random and systematic variability in the total counts.

The Google trends data used is also biased to only take into account those who have internet access, happen to use Google as their search engine, and use the selected search terms. Aiken et al. (2020) found that Google trends were most useful as predictors during early stages of an outbreak, but raised possible issues of media coverage reducing the interpretability of search volumes being related to those suffering from infection. Two previous Google trends disease tracking projects have both been discontinued due to inaccurate predictions (Aiken et al. (2020)). Therefore, this type of internet search data shows some potential in aiding disease surveillance models, especially when there is a lack of other data available, but they are not yet operationally viable.

In general, if the covariates are readily available it is fairly trivial to add them into many of the statistical frameworks discussed in this review. Although, this relies on covariates data being available at the same spatial and temporal resolutions as the disease data of interest. Furthermore, where informative covariates are time-varying, practical use for operational nowcasting will depend on how immediately up-to-date data can be obtained/downloaded. Meanwhile, accurately predicting future covariate values may be required to perform disease forecasts.

2.5.3 Under-reporting

This thesis largely focuses on the problem of predicting disease counts (e.g. cases) $y_{t,s}$ that will eventually be reported after any delays have passed. However, in many situations, some cases in a population are never reported, e.g. due to asymptomatic individuals or a lack of resources such as testing materials or staff. Hence, the observed total counts of eventually reported cases, and therefore the predicted nowcasts of these counts, are under-representations of the actual disease counts occurring. We can refer to this problem as "under-reporting".

A general framework for correcting under-reporting is introduced in Stoner, Economou and Drummond Marques da Silva (2019) and Stoner and Economou (2019) outlines how this can be integrated into the GDM framework. Recall that $y_{t,s}$ are the total observed counts and $z_{t,s,d}$ are the delayed counts. Now, let $x_{t,s}$ be the true counts that actually occurred (and are usually unobserved), such that $y_{t,s} \leq x_{t,s}$. Further recall the modular framework for flawed observation discussed in Chapter 1.2, which suggests that we can describe the generation of $y_{t,s}$ and $\boldsymbol{z}_{t,s}$ as a sequence of two processed Y and Z, i.e.

$$Y(\boldsymbol{\Theta}) \rightarrow y \rightarrow Z(\boldsymbol{\Pi}) \rightarrow z. \tag{2.93}$$

We can modify this such that we have a new model/module $X(\boldsymbol{\Phi})$ that describes the true count generating process, depending on some parameters $\boldsymbol{\Phi}$, and where Y now describes the under-reporting mechanism translating $x_{t,s}$ into $y_{t,s}$:

$$X(\boldsymbol{\Phi}) \rightarrow x \rightarrow Y(\boldsymbol{\Theta}) \rightarrow y \rightarrow Z(\boldsymbol{\Pi}) \rightarrow z. \tag{2.94}$$

Stoner, Economou and Drummond Marques da Silva (2019) assume a Binomial model for Y, conditional on $x_{t,s}$ the true counts (that actually occurred and are unobserved),

$$y_{t,s}|x_{t,s} \sim \text{Binomial}(\kappa_{t,s}, x_{t,s}). \tag{2.95}$$

The mean of the Binomial distribution is then $x_{t,s}\kappa_{t,s}$, where $\kappa_{t,s}$ is the proportion expected to be reported. Then $x_{t,s}$ are modelled with a Poisson distribution,

$$x_{t,s} \sim \text{Poisson}(\lambda_{t,s}). \tag{2.96}$$

None of the other papers included in this review of approaches explicitly discuss under-reporting in conjunction with delayed reporting. However, the **NobBS** package does allow users to set a "proportion reported" when implementing the model. The resulting nowcasts are then multiplied by a constant which corrects for proportions that have been set to less than one. This approach requires users to be certain about reporting rates which are constant over time and spatial regions. On the other hand, Stoner, Economou and Drummond Marques da Silva (2019) allows $\kappa_{t,s}$ to be modelled by covariates and random effects that depend on time or spatial regions. If all $y_{t,s}$ are assumed to be potentially under-reported, there is non-identifiability between the reporting rate $\kappa_{t,s}$ and the incidence rate of $x_{t,s}$. The model is identifiable if some of the $y_{t,s}$ are assumed to be fully reported or using informative prior information (Stoner, Economou and Drummond Marques da Silva (2019)).

2.5.4 Operational ability

As discussed in Section 1, the overarching aim of this body of work is to develop efficient and effective operational warning systems with a focus on tailoring these to address real-world data challenges put forward by the Oswaldo Cruz Foundation in Brazil. In general, surveillance models must be judged by their ability to inform and aid decision makers who are using them, particularly in the context of disease data. More accurate nowcasts and forecasts would allow decision makers to make better informed decisions, which could be crucial in an outbreak by allowing for preventative measures to be implemented.

However, another important characteristic of models used in operational contexts is timeliness. Predictions need to be produced and analysed in time for public health policies to be assessed and executed. For example, in Stoner, Halliday and Economou (2022) (Chapter 4) we compare the GDM to competing models and found it to be relatively slow. For predicting COVID-19 hospital deaths in England the GDM takes just under 20 minutes when fitted using MCMC methods in nimble, whereas the Bastos et al. (2019) INLA model takes under 7 minutes. But for larger data sets, like predicting SARI cases in Brazil, the GDM takes approximately 12 hours to run. This makes it less viable for operational application, especially since such a long run time would possibly make it unsuitable for daily predictions.

In relation to the models discussed so far, an "ideal modelling framework" would have the flexibility and performance of the GDM model in Stoner and Economou (2019) but with an operational speed more comparable to methods such Bastos et al. (2019), which are aided by using the INLA method for fitting Bayesian models. Furthermore, there is motivation to develop a generalised framework that could be applied to any disease surveillance setting experiencing reporting delays, and in an easy-to-use format that allows those with more field specific expertise to implement it, would enable these models to be utilised in a more efficient and effective way. For example, the R packages that supplement Höhle and

An Der Heiden (2014) and McGough et al. (2020) created more user-friendly functions which run the respective models. As discussed by Rivers et al. (2019), there is a need to bridge the gap between model designers and those who will eventually be using the models. Model improvements and outputs should be tailored to aid decision makers in minimising the impact of infectious disease. For example, Höhle and An Der Heiden (2014) recognise that the trend state of the outbreak ("up", "stable", "down") can be inferred from the spline for the total counts and communicated to users of their package, this type of qualitative communication could be highly useful in aiding quick decisions during time-sensitive situations.

2.6 Discussion

In this chapter we reviewed the current literature on correcting reporting delays for disease data. We presented this literature as arranged into two broad groups of nowcasting methodology; jointly modelling the total and partial counts together (Section 2.2.1) and direct models of the partial counts that rely on a conditional independence assumption (Section 2.2.2). However, whilst the joint models are potentially able to better separate and capture all sources of variability in the data, as summarised in Section 1.1.3, due to their more complex hierarchical structure, implementation is generally more complex than for the conditional independence models.

Due to the relative simplicity of modelling just the partial counts to capture all sources of variability in the data, the conditional models can designed within more efficient model fitting frameworks. This is because their lack of a multivariate hierarchical structure means they can be expressed as a frequentist model, or a latent Gaussian model that can be fit using INLA. The main downside of the conditional independence group is that they can't separate random variability in the total counts from the random variability in the reporting delay. As such, achieving appropriate model prediction uncertainty (as

measured by the prediction interval widths) relies on capturing the covariance structure of the partial counts which, due to the conditional independence assumption, in turn relies on random and covariate effects fully explaining the covariance. In some situations, excessive uncertainty results from summing over the predictions of the partial count to generate predictions of the total counts.

Hence, a more robust framework is required such as the GDM method, as discussed in Section 2.3. This model is able to separate the four sources of variability in the data, summarised in Section 1.1.3, allowing for intuitive additions of spatial modelling effects (Section 2.5.1) and informative covariates (Section 2.5.2). Moreover, the choice of both the Negative-Binomial distribution for the total counts and the GDM for the partial counts offers the flexibility to capture the variance and covariance structures present in the data. This hierarchical framework is also straightforward to extend to consider additional data challenged such as under-reporting (Section 2.5.3). However, the GDM could be made more suitable for operational purpose with improvement in computational speed and user accessibility. As discussed in Section 2.5.4, the GDM is relatively slow compared to competing approaches due to the need to fit models using MCMC methods. Also, it requires fundamental understanding of Bayesian modelling procedures, and both R and NIMBLE software, in order to be implemented effectively.

In this thesis we endeavour to improve the computational efficiency of the GDM in Chapter 3 and then apply this to nowcasting COVID-19 deaths in England in Chapter 4 (Stoner, Halliday and Economou 2022). This includes a thorough comparison between the GDM model and both the Bastos et al. (2019) and McGough et al. (2020) methods that we have discussed in this literature review. Later, we develop novel versions of the GDM for applications where none of the existing methods and extensions that we have covered here are suitable.

Chapter 3

Computational Efficiency of the GDM Framework

In this chapter, we explore and assess potential avenues for improving the computational efficiency of the Generalised-Dirichlet Multinomial (GDM) method introduced by Stoner and Economou (2019). In Section 2.3 the Generalised-Dirichlet Multinomial (GDM) model was presented as a framework that is theoretically better able to capture all types of variability in delayed reporting data compared to competing approaches. We described the sources of this variability in Section 1.1 and outlined why capturing it results in improvements in predictive precision in Section 1.2. However, in Section 2.5.4 it was noted that in terms of operational ability the GDM model may not be suitable for more time sensitive disease surveillance applications.

In Section 2.1, we introduced both INLA (Lindgren and Rue (2015)) and NIMBLE (de Valpine et al. (2017)) as two options for fitting Bayesian models and discuss the relative drawbacks and benefits of each. Initially, in Section 3.1, to try and achieve improvements in the timeliness of the GDM model we attempt to approximate it with a marginal model. Hence, enabling it to be fit using the more computationally efficient method and software, INLA. However, due to the approximation not exhibiting as consistently precise predictions as the GDM, we instead shift focus to improving the run time of the original GDM model.

All subsequent versions of the GDM are instead fitted using MCMC techniques in NIMBLE, the flexibility of this software helps facilitate updating the computational efficiency of the GDM. Avenues to increase computational speed that we explore include: more efficient MCMC sampling, through parallel processing and model formatting choices (Section 3.2); and directly optimising the joint posterior of the GDM model, in order to better initialise MCMC chains (Section 3.3). We then compare the relative impact on run time of all improvements for two case studies in Section 3.4.

3.1 Approximating the GDM

At the beginning of the PhD project, we conceived and investigated one possible approach to improving the computational efficiency of the GDM, centering around developing an alternative latent Gaussian modelling framework that approximates the characteristics of the GDM method and can be fitted using Integrated Nested Laplace approximations (INLA). As detailed in Section 2.1.3, INLA is an approach to fitting Bayesian latent Gaussian models that can be considerably faster than MCMC. However, the GDM method combines a Negative-Binomial model with a series of Beta-Binomial models with unknown total. In MCMC, the unknown total counts y_t are sampled as unknown quantities. This situation, involving complex hierarchical and missing data has not yet been implemented in INLA, and it is not clear if it is even possible.

As explained in Section 2.2.2, one group of approaches to nowcasting, which we call "conditional independence" models, assumes a direct model for the partial counts $z_{t,d}$ (with no model for the total counts y_t), assuming that $z_{t,d}$ are independent of each other, conditional on all covariate and random effects. We further explained in Section 2.2.3 that capturing the random variability in the total counts well relies on capturing the covariance of the partial counts across the delays well, and that the conditional independence assumption means we must rely on the covariate and random effects to fully describe the covariance. Thus, we might suppose that thoughtful design of the covariate and random effects, to mimic the flexibility of the GDM, could prove fruitful.

First, consider the hierarchical structure of the GDM, where we express the Negative-Binomial model for y_t as a Poisson-Gamma mixture, as derived in Appendix B.1:

$$y_t \mid \kappa_t \sim \text{Poisson}(\kappa_t); \tag{3.1}$$

$$\kappa_t \mid \theta, \lambda_t \sim \text{Gamma}(\theta, \lambda_t,); \tag{3.2}$$

$$\boldsymbol{z}_t \mid \boldsymbol{p}_t, y_t \sim \text{Multinomial}(\boldsymbol{p}_t, y_t); \tag{3.3}$$

$$\boldsymbol{p}_t \sim \text{Generalized-Dirichlet}(\boldsymbol{\alpha}, \boldsymbol{\beta}). \tag{3.4}$$

Here, λ_t is the mean parameter of the Gamma distribution and θ is a shape parameter equivalent to the Negative-Binomial dispersion parameter. Hence, the mean is given by $\mathbb{E}[X] = \lambda$ and variance by $\text{Var}[X] = \frac{\lambda^2}{\theta}$, for more details on the Gamma distribution see Appendix A.3. We can further sum out y_t to obtain the marginal distribution of $z_{t,d}$:

$$z_{t,d} \mid \kappa_t, p_{t,d} \sim \text{Poisson}(p_{t,d}\kappa_t). \tag{3.5}$$

Thus, given $p_{t,d}$ and κ_t, the marginal models for the $z_{t,d}$ are Poisson. As shown in Appendix B.2, these $z_{t,d}$ (conditional on $p_{t,d}$ and κ_t) are independent of one another, and summing together observed and independently predicted $z_{t,d}$ would yield the appropriate predictive distribution for unobserved y_t.

3.1.1 Approximation framework

Now, suppose we could design an approximate version such that $\mu_{t,d} \approx p_{t,d}\kappa_t$ is the expected mean of a Poisson distribution of the partial counts. This $\mu_{t,d}$ would have to account for the same systematic structures as in $p_{t,d}$ and κ_t. It would also have to account for the positive covariance induced by the Gamma distributed κ_t appearing in the model for all delayed partial counts, and the multivariate random variability induced by the GD distributed $p_{t,d}$. This motivates a Poisson model for $z_{t,d}$ combining multiple random effect structures in $\mu_{t,d}$ that play different roles in capturing the different sources of random and

systematic variability, as inspired by these insights into the GDM. Here, our proposed approximation to the GDM, referred to in later sections as the "GDM approximation INLA", model is given by

$$z_{t,d} \sim \text{Poisson}(\mu_{t,d}); \tag{3.6}$$

$$\log(\mu_{t,d}) = \iota + \delta_t + \tilde{\theta}_t + \zeta_d + \gamma_{t,d} + \pi_d t, \tag{3.7}$$

$$\delta_t \sim \text{Normal}(2\delta_{t-1} - \delta_{t-2}, \sigma_\delta^2), \tag{3.8}$$

$$\delta_1, \delta_2 \sim \text{Normal}(0, \sigma_\delta^2), \tag{3.9}$$

$$\tilde{\theta} \sim \text{Normal}(0, \sigma_{\tilde{\theta}}^2), \tag{3.10}$$

$$\zeta_d \sim \text{Normal}(\zeta_{d-1}, \sigma_\psi^2) \tag{3.11}$$

$$\pi_d \sim \text{Normal}(0, \sigma_\pi^2), \tag{3.12}$$

$$\zeta_1 \sim \text{Normal}(0, \sigma_\zeta^2), \tag{3.13}$$

$$\gamma_t \sim \text{MVN}(\mathbf{0}, \boldsymbol{W}_\gamma^{-1}). \tag{3.14}$$

In the log link of the expected mean partial counts reported $\mu_{t,d}$, we first include an intercept term ι. Alongside this, we include a second order random walk δ_t, a independent identically distributed (i.i.d.) random effect $\tilde{\theta}_t$ for each time step and a first order random walk ζ_d over delay. Finally, we include a delay specific random slope over time π_d and a three dimensional i.i.d. correlated random effect γ_t. This correlated random effect has a Wishart prior with 3 degrees of freedom and scale matrix equal to the identity matrix $\boldsymbol{I}$.

We endeavor to capture the different sources of variability in the data using these random effects in the above Poisson model. Firstly, if we consider the systematic and random variability in the total counts, within the GDM framework this is partly captured by the expected mean λ_t and dispersion parameter θ of the Gamma distribution (Equation (3.2)) respectively. In our framework to approximate the GDM (Equations (3.6)–(3.14)), we approximate λ_t with the random walk term δ_t and we approximate θ with the random effect $\tilde{\theta}_t$ to capture the random variability at each time step.

However, the total counts are not explicitly included in our approximation of the GDM model. But, for the GDM framework the expected partial counts reported at each delay are conditional on the Generalised-Dirichlet for expected proportions reported, $p_{t,d}$, as well as the expected total counts y_t with expected mean $\kappa_t \mid \theta, \lambda_t \sim \text{Gamma}(\theta, \lambda_t)$. This dependence on the Gamma distribution for κ_t allows the model for the partial counts to capture the positive covariance present in $z_{t,1:D}$, over delay for a given time. This covariance between the partial counts at time step t is due to $z_{t,1:D}$ being compositional components of the total counts y_t. Hence, in our approximation model we attempt to capture the positive covariance in the partial counts with $\tilde{\theta}$.

For the original GDM framework, the parameters α and β of the Generalised-Dirichlet distribution (Equation (3.4)) are then able to capture the mean and covariance over delay of $z_{t,1:D}$. Some degree of negative covariance is induced by the sum to one constraint of the proportions reported, $\sum_{d=0}^{D} p_{t,d} = 1$, thus if one proportion increases in one delay the sum of proportions in the remaining delays will in turn decrease and vice versa. Our proposed INLA model attempts to capture the extra covariance structure of the partial counts, including this source of negative covariance, with a three dimensional i.i.d. correlated random effect, γ_t. Here, it is six dimensional as this is the number of delays in our simulated data, but in general this could be a n^{th} dimensional correlated random effect.

Finally, to capture the mean systematic variability in the delay distribution we include a first order random walk ζ_d, as well as delay specific random slopes π_d. This choice was motivated by the fact the data (described in the next subsection) is simulated to have a linear trend over time in the delay distribution, but in practice could be any combination of model effects to capture the systematic trend in the expected proportions reported.

Due to the latent Gaussian structure of this GDM approximation model, it can be fit using the Bayesian modelling approach INLA, which is a relatively computationally efficient method. We set the following priors for each of the model effects to fully specify the model:

$$\iota \sim \text{Normal}\left(0, \sqrt{\frac{1}{10^{-6}}}^2\right), \tag{3.15}$$

$$\tilde{\theta}_t \sim \text{Normal}(0, \sigma^2_{\tilde{\theta}}), \tag{3.16}$$

$$\sigma^2_{\delta} \sim \text{Half-Normal}(0, 0.1^2), \tag{3.17}$$

$$\sigma^2_{\zeta} \sim \text{Half-Normal}(0, 1^2), \tag{3.18}$$

$$\sigma^2_{\pi} \sim \text{Half-Normal}(0, 1^2), \tag{3.19}$$

$$\boldsymbol{W}_{\gamma}^{-1} \sim \text{Wishart}(3, \boldsymbol{I}). \tag{3.20}$$

The choice of priors and parameters for our approximation model were chosen to reflect those used in Bastos et al. (2019), which we define following subsection. This was done to ensure the INLA approaches are fitted effectively in a similar way to how they are currently operating, and to ensure the comparison between the three models we introduce in this section is as fair as possible. Furthermore, this is the model our approximation model attempts to improve upon, as the goal is to achieve a computationally efficient model using INLA, as in Bastos et al. (2019), which can better capture the different sources of variability in the data than this current model. We additionally compare the Stoner and Economou (2019) GDM model, which we define below, as a model that is able to separate the different sources of variability in the data.

3.1.2 Existing frameworks

We compare our GDM approximation to an equivalent GDM model, fit using MCMC in NIMBLE, and referred to in later sections as the "GDM NIMBLE" model, from Stoner and Economou (2019):

$$y_t \sim \text{Negative-Binomial}(\lambda_t, \theta); \tag{3.21}$$

$$\log(\lambda_t) = \tau + \delta_t \tag{3.22}$$

$$\delta_t \sim \text{Normal}(2\delta_{t-1} - \delta_{t-2}, \sigma_\delta^2), \tag{3.23}$$

$$\delta_1, \delta_2 \sim \text{Normal}(0, \sigma_\delta^2), \tag{3.24}$$

$$z_{t,d}|y_t, z_{t,1:(d-1)} \sim \text{Beta-Binomial}(\nu_{t,d}, \phi_d, y_t - \sum_{j=1}^{d-1} z_{t,j}); \tag{3.25}$$

$$\text{logit}(\nu_{t,d}) = \psi_d + \beta_{t,d} + \pi_d t, \tag{3.26}$$

$$\beta_{t,d} \sim \text{Normal}(\beta_{t-1,d}, \sigma_\beta^2), \tag{3.27}$$

$$\beta_{1,d} \sim \text{Normal}(0, \sigma_\beta^2), \tag{3.28}$$

$$\pi_d \sim \text{Normal}(0, \sigma_\pi^2). \tag{3.29}$$

$$\tag{3.30}$$

Here, we opt for the hazard version of the GDM model as defined in Section 2.3.2. In order to make fair comparisons with the GDM approximation model, given by Equations (3.6)–(3.14), we attempt to keep as many of the modelling effects as similar as possible and set the priors to be equivalent. Since the model for the total counts and the delay distribution is now separate, we have effectively split up the intercept ι, hence we have ψ_d as independent intercept for each delay d for the expected relative proportions reported, and τ as an intercept for the expected mean of the total counts. The other addition is a Beta-Binomial dispersion parameter, ϕ_d, and the Negative-Binomial dispersion parameter θ_t. Similarly to Equations (3.6)–(3.14), we include a second order random walk δ_t to capture the trend in the expected mean λ_t. Also, we include a delay specific random slope over time π_d and a first order random walk ζ_d to capture the mean systematic variability in the expected relative proportions reported. Finally, we have removed the three dimensional i.i.d cor-

related random effect γ_t, since both the positive and negative covariance structure of the expected partial counts can be captured by the Beta-Binomial distribution for $z_{t,d}|y_t$. To fully specify the GDM model, the following priors were used:

$$\tau \sim \text{Normal}(0, 10^2), \tag{3.31}$$

$$\sigma_\delta^2 \sim \frac{1}{\text{half-Normal}(0, 0.1^2)}, \tag{3.32}$$

$$\theta \sim \text{Gamma}(2, 0.02), \tag{3.33}$$

$$\sigma_\beta^2 \sim \frac{1}{\text{half-Normal}(0, 1^2)}, \tag{3.34}$$

$$\sigma_\pi^2 \sim \frac{1}{\text{half-Normal}(0, 1^2)}, \tag{3.35}$$

$$\psi_d \sim \text{Normal}(0, 10^2), \tag{3.36}$$

$$\phi_d \sim \text{Gamma}(2, 0.02). \tag{3.37}$$

These were chosen to be comparable to those selected for the GDM approximation model priors (Equations (3.15)–(3.20)) where parameters are identical. Since both θ and ϕ do not have equivalent parameters in the approximation model, these priors were instead chosen as $Gamma(r, \theta)$ distributions, with rate parameter r and shape parameter θ, to ensure they are strictly positive and reflecting the prior choices for the equivalent GDM dispersion parameters in Stoner and Economou (2019).

Finally, we also compare our GDM approximation in INLA approach to the existing and highly cited INLA approach by Bastos et al. (2019), which we reviewed in detail in Section 2.2.2. This model does not attempt to capture the random correlated variability between the partial counts or explicitly model the total counts. Instead, the partial counts are modelled by a Negative-Binomial (NB) distribution with dispersion parameter $\tilde{\phi}$, and is implemented by fitting a Negative-Binomial model in INLA. The following equations

specify this model, which we later refer to as the "NB INLA" model,

$$z_{t,d} \sim \text{Negative-Binomial}(\mu_{t,d}, \tilde{\phi}); \tag{3.38}$$

$$\log(\mu_{t,d}) = \iota + \delta_t + \zeta_d + \pi_d t; \tag{3.39}$$

$$\delta_t \sim \text{Normal}(2 * \delta_{t-1} - \delta_{t-2}, \sigma_\delta^2), \tag{3.40}$$

$$\delta_1, \delta_2 \sim \text{Normal}(0, \sigma_\delta^2), \tag{3.41}$$

$$\zeta_d \sim \text{Normal}(\zeta_{d-1}, \sigma_\zeta^2), \tag{3.42}$$

$$\zeta_1 \sim \text{Normal}(0, \sigma_\zeta^2), \tag{3.43}$$

$$\pi_d \sim \text{Normal}(0, \sigma_\pi^2). \tag{3.44}$$

where $\tilde{\phi}$ is the dispersion parameter of the Negative-Binomial distribution of the partial counts. Similarly to GDM approximation model (Equations (3.6)–(3.14)), for the log of the expected mean number of partial counts reported at each delay $\mu_{t,d}$ we define an intercept term ι. Also, a second order random walk over time δ_t to capture the systematic variability in the total counts. To capture the systematic variability in the delay distribution, ζ_d is a first order random walk over delay and π_d is an independent random slope of time for each delay. Unlike the GDM approximation model, this model does not attempt to capture the positive covariance of the partial counts with $\tilde{\theta}_t$ or the negative covariance structure of the partial counts using $\gamma_{t,d}$. The priors for the "NB INLA" model are given by,

$$log(\tilde{\phi}) \sim \text{Gamma}(1, 0.1), \tag{3.45}$$

$$\iota \sim \text{Normal}\left(0, \sqrt{\frac{1}{10^{-6}}}^2\right), \tag{3.46}$$

$$\sigma_\delta^2 \sim \text{Half-Normal}(0, 0.1^2), \tag{3.47}$$

$$\sigma_\zeta^2 \sim \text{Half-Normal}(0, 1^2), \tag{3.48}$$

$$\sigma_\pi^2 \sim \text{Half-Normal}(0, 1^2). \tag{3.49}$$

The choice of priors for this model is comparable to those fitted to the previous two models and was informed by the choices in Bastos et al. (2019). This was done to ensure the INLA approaches are fitted effectively and to reflect how the Negative-Binomial INLA model, Equations (3.38)–(3.44), is currently being used operationally.

3.1.3 Simulation experiment

For our simulation experiment, we fit all three models to simulated data sets, here we detail how we generate each data set. First, we set the length of the simulated data to be $T = 100$, and the number of delays to be $D_{max} = 6$. The maximum number of delays for this simulation study reflects those sometimes observed in real data and is consistent with the number of delays often modelled throughout this thesis. Then, to specify a smooth temporal trend in the mean of the total cases (λ_t), which we believe reflects potential real world scenarios, we first generate a cubic spline ρ_t,

$$\rho_t = \boldsymbol{X}_t \boldsymbol{\upsilon}^{(\rho)}, \tag{3.50}$$

$$\boldsymbol{\Omega}^{(\rho)} = \boldsymbol{M}_1 / \left(\sigma_\rho^{(1)}\right)^2 + \boldsymbol{M}_2 / \left(\sigma_\rho^{(2)}\right)^2, \tag{3.51}$$

$$\boldsymbol{\upsilon}^{(\rho)} \sim \text{Multivariate-Normal}(0, \boldsymbol{\Omega}^{(\rho)}), \tag{3.52}$$

where $\sigma_\rho^{(1)} = 5$ and $\sigma_\rho^{(2)} = 5$. We use the **mgcv** package to generate the spline basis matrix $\boldsymbol{X}_t$, and the penalty matrices $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$. The cubic plate spline ρ_t has a zero mean and we subtract the linear trend (as estimated using linear regression). We then take the exponential of the de-linearised trend, ρ_t, plus an intercept term to generate the data we then use a Negative-Binomial distribution for the total counts y_t,

$$y_t \mid \lambda_t, \theta \sim \text{Negative-Binomial}(\lambda_t, \theta); \tag{3.53}$$

$$\log(\lambda_t) = \tau + \rho_t. \tag{3.54}$$

The intercept is randomly simulated from $\tau \sim \text{Normal}(5, 0.25^2)$, to reflect that the average number of simulated cases is $\exp(5)$ across all simulations since the cubic spline ρ_t has a zero mean. A relatively small variance was chosen because we are not primarily interested in recapturing the intercept in this simulation experiment, so we do not require it to vary much between simulations. On the other hand, a randomly generated dispersion parameter with a large variance, $\theta \sim \text{Gamma}(2, 0.02)$, was used for each simulation to

induce more variability in the dispersion of the total counts between simulations. Here, the Gamma distribution is parameterised in terms of shape and rate, see Appendix A.3 for more details. The dispersion parameter of the Negative-Binomial model of the total counts has a mean $\frac{2}{0.02} = 100$, and variance $\frac{2}{0.02^2} = 5000$ across simulations.

Next, for the partial counts we simulate using a GDM model with a maximum possible delay of $D_{max} = 6$,

$$z_{t,d}|\nu_{t,d},\phi,N_{t,d} \sim \text{Beta-Binomial}(\nu_{t,d},\phi_d,N_{t,d} = y_t - \sum_{j=1}^{d-1} z_{t,j}); \tag{3.55}$$

$$CLR(p_{t,d}) = \psi_d + \eta t^*, \tag{3.56}$$

$$p_{t,D_{max}} = -\sum_{d=1}^{D_{max}-1} p_{t,d}, \tag{3.57}$$

$$\nu_{t,d} = \frac{exp(p_{t,d})}{\sum_{d=1}^{D_{max}} exp(p_{t,d})}, \tag{3.58}$$

with intercept ψ_d and linear temporal trend ηt^*, where t^* is a scaled variable of time to improve the interpretability of η and aid the efficiency of the MCMC algorithm. The absolute proportions reported ($p_{t,d}$) have been generated using a centralised log ratio (CLR) transform:

$$\text{CLR}(\boldsymbol{p}_{t,1:(D_{max})}) = \left[\log\left(\frac{p_{t,1}}{g(\boldsymbol{p}_t)}\right), ..., \log\left(\frac{p_{t,D_{max}}}{g(\boldsymbol{p}_t)}\right)\right], \tag{3.59}$$

where $g(\boldsymbol{p}_t)$ is the geometric mean of $\boldsymbol{p}_t$. Simulating the trends in the partial counts in the absolute proportions instead of the relative proportions, which are modelled in "GDM NIMBLE" (Equations (3.6)–(3.14)), is to attempt not to give the GDM model an unfair advantage when compared to "NB INLA" and "GDM approximation INLA" which do not model the relative proportions.

For each simulation, the delay specific intercept is generated by $\psi_d \sim \text{Normal}(0, 0.25^2)$ to reflect no overall trend across simulations in the absolute proportions over delay, and the average absolute proportion being similar between simulations as this is not a parameter of interest. The linear slope parameter for the scaled time variable is generated by $\eta \sim \text{Normal}(0.2, 0.1^2)$, hence the simulations reflect a potential real-world scenario where reporting in the delay distribution improves over time at different rates for each simulation. The relative proportions reported are then calculated by Equation (3.58), which we derive in Section 5.3.3. For the Beta-Binomial dispersion parameter, for each simulation we generate the partial counts for three separate values, $\phi \in \{2, 6, 10\}$, resulting in three separate data sets, all of length $T = 100$. These values of ϕ were chosen to show how the three models behave when different amounts of the total random variability for the partial counts comes from the Generalised-Dirichlet distribution. We randomly generate these data sets as outlined above for 100 simulations, giving a total of 300 individual data sets.

The simulated total counts are split over six delays to give the partial counts; partial counts reported in the first delay are observed for the whole time series up to $t = 100$, and all partial counts are reported by $t = 95$. This reflects a scenario where the minimum delay is $d = 1$ (0 days) and the maximum delay length is $D_{max} = 6$ (5 days) which is consistent for how we model real world data in Section 3.4.

The results in Figure 3.1 show the model's ability to capture the trend in the total counts in three (of the 100) randomly simulated data sets (rows) for each of the models of interest, for the three chosen values of the Beta-Binomial dispersion parameter ϕ (columns).

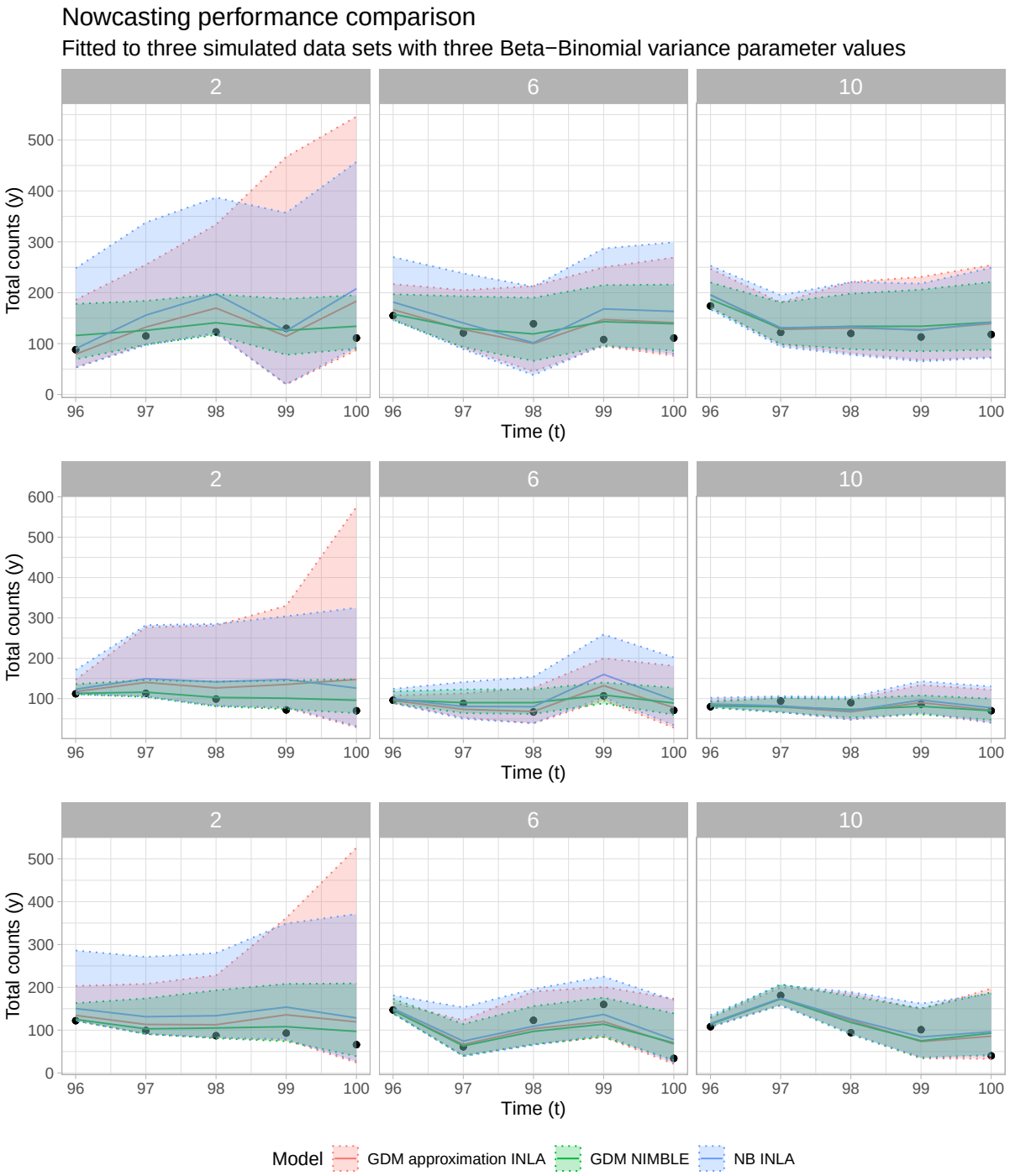

Figure 3.1: Median posterior predictions (solid lines) and 95% predictions intervals (dotted shaded regions) for the GDM approximation model fitted in INLA, the GDM model fitted in NIMBLE, and the Bastos et al. (2019) Negative-Binomial (NB) model fitted in INLA. The nine panels show for the three randomly selected simulated data sets (rows) the affect of three chosen values of the Beta-Binomial dispersion parameter (columns), $\phi \in \{2, 6, 10\}$, where the points give the simulated counts.

In Figure 3.1, both the INLA models exhibit more uncertainty in predictions for smaller values of ϕ, the Beta-Binomial dispersion parameter in the simulation model. In practice, the positive covariance in the partial counts, induced by the dependence on the total counts, and the negative covariance, induced by the sum constraint, may approximately cancel out. Hence, considering smaller values of ϕ may be forcing the negative covariance in the partial counts to be greater, thus unbalancing this cancellation. For example, when $\phi = 2$ (the smallest value of ϕ considered) this induces more variability in the probabilities of the delay distribution than larger values of ϕ, and hence more competition in the composition of the total counts y_t which the partial counts $\boldsymbol{z}_{t,1:D}$ are constrained to sum to. Thus, smaller ϕ values result in greater negative covariance between the partial counts. Therefore, when $\phi = 2$ the "GDM NIMBLE" model (Equations (3.21)–(3.30)), which includes a Generalised-Dirichlet prior for the probabilities of the delay distribution, is able to capture this negative covariance. This is reflected by the "GDM NIMBLE" in Figure 3.1 displaying more certain prediction intervals for $\phi = 2$ compared to the other two models. We would also expect the "GDM approximation INLA" model (Equations (3.15)–(3.20)) to be able to capture this negative covariance due to the inclusion of the random effect $\boldsymbol{\gamma}_t \sim \text{MVN}(\mathbf{0}, \boldsymbol{W}_{\gamma}^{-1})$. However, it is clear it is not as effective at separating the positive and negative covariance of the data as the GDM, since its prediction intervals are wider across all values of ϕ, and more uncertain for smaller values of ϕ. Similarly, the "NB INLA" model (Equations (3.38)–(3.44)) is restricted in its ability to capture unusual covariance structures (as discussed in Section 2.2) due to the assumption the partial counts are conditionally independent Negative-Binomial distributions.

Occasionally in Figure 3.1 the "GDM approximation INLA" model appears to have slightly narrower prediction intervals compared to the "NB INLA" model, which may identify scenarios where the introduction of the correlated random effect is allowing the GDM approximation model to capture some of the negative covariance in the partial counts. However, both the INLA models exhibit relatively wide and non-smooth 95% prediction intervals compared to the GDM model. Both the INLA models do not explicitly model the total counts (where as the GDM does), which makes it difficult for the models to separate

the random variability of the total counts and the partial counts. The variability in the partial counts over delay also comprises of a negative and positive covariance structure which is induced by summing to the total counts and being compositional dependent on the total counts. Therefore, the Multivariate-Normal random effect $\boldsymbol{\gamma}_t$ in the "GDM approximation INLA" model appears to be unable to separate and capture the negative covariance from the positive covariance.

To quantify the difference in prediction performance across all simulations, we compare three model metrics in Figure 3.2. First, we calculate the mean absolute error of the median posterior predictions compared to the true simulated data, to quantify the accuracy of the point predictions. Next we calculate the mean difference between the upper 97.5% and lower 2.5% quantiles of the posterior predictions, giving the mean 95% prediction interval width, to measure the precision of predictions. Finally, the coverage measures the proportion of the simulated totals that fall within the respective 95% prediction interval width, to quantify the ability of the model uncertainty to capture the true data. We calculate each metric (columns) by model (colour) and the chosen values of $\boldsymbol{\phi}$, the Beta-Binomial dispersion parameter, used to generate the data (rows). The x-axis gives the prediction time difference (PTD) that the metric is calculated for, which is the difference between when the counts we are predicting occurred (t) and the current time we are nowcasting for ($T_{now} = 100$).

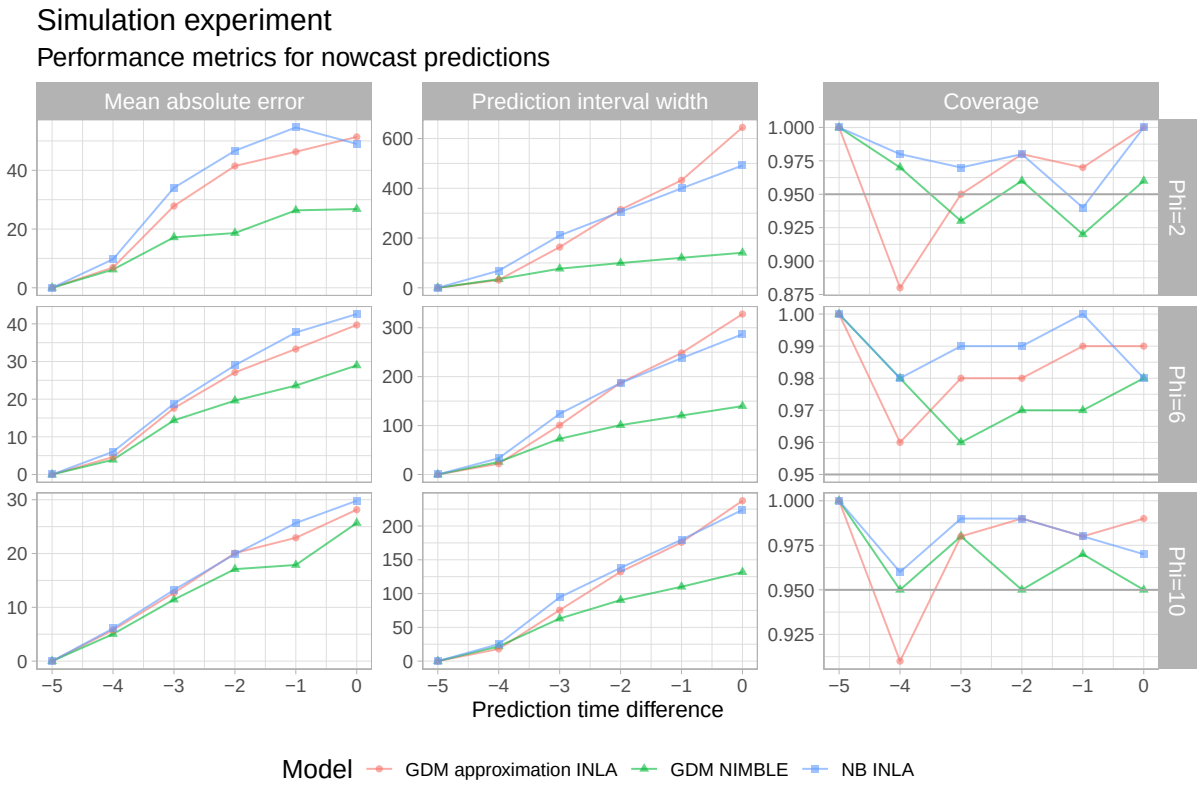

Figure 3.2: From left to right, the calculated mean absolute error, mean 95% prediction interval width and mean 95% prediction interval coverage for the posterior predictions of the 100 simulated data sets. For each chosen value of $\phi \in \{2, 6, 10\}$ (rows), the metrics are calculated across the prediction time difference (x-axis) of the nowcasts. The models, indicated by colour and shape, are; the GDM approximation model fitted in INLA, the GDM model fitted in NIMBLE, and the Negative-Binomial (NB) model fitted in INLA.

In Figure 3.2, on average the mean absolute error appears to be smallest for the "GDM NIMBLE" model for all three chosen values of $\phi \in \{2, 6, 10\}$. But, as the chosen value of ϕ increases the mean absolute error between models becomes more comparable. This may suggest that the positive random variability induced into the partial counts by being conditional on the total counts, through the Negative-Binomial dispersion parameter θ, is similar in magnitude to the negative covariance, due to the sum to a total constraint of the partial counts, when ϕ (the Generalised-Dirichlet dispersion parameter) used for simulation is smaller. This could mean for smaller values of ϕ, the GDM is able to predict the total counts more accurately as it has more flexibility to be able to capture and separate both types of random variability, owing to explicitly modelling the total counts and capturing the compositional structure of the partial counts. However, the "GDM

approximation INLA" model does have a smaller mean absolute error than the "NB INLA" model across all PTD's for $\phi = 6$ and $\phi = 10$. Hence, the correlated multivariate normal effect may be able to capture some of the negative random variability in the total counts. But, it appears to not do so when $\phi = 2$, again due to the random variability being particularly large or similar in magnitude to the positive variability in the partial counts.

For prediction interval width, the "NB INLA" model has wider intervals than the "GDM NIMBLE" model and often the "GDM approximation INLA" model, this may reflect excessive model uncertainty compared to the alternative models, and results in the higher coverage observed for "NB INLA". Meanwhile, during the prediction time difference period of $-20 < PTD \leq -10$, the "GDM approximation INLA" models has similar or lower prediction interval widths compared to the "GDM NIMBLE" for $\phi = 6$ and $\phi = 10$ respectively, but this is in conjunction with lower coverage in that same period. Hence, the model uncertainty of "GDM approximation INLA" is less likely to be capturing the true data for $-20 < PTD \leq -10$, and has larger prediction interval width for $-10 < PTD \leq 0$, reflecting more uncertainty when less of the partial counts have been observed. Therefore, the "GDM NIMBLE" model has the most consistently narrow prediction interval widths across the prediction time differences for all chosen values of ϕ.

This simulation experiment shows that explicitly modelling the total counts within the modelling framework does appear to be crucial in obtaining optimal precision in model nowcast predictions. But, a significant drawback of this approach is its run time compared to the two INLA models. The "GDM NIMBLE" models takes approximately 18 minutes to run per (simulated) data set where as the "GDM approximation INLA" takes approximately 43 seconds and the "NB INLA" takes approximately 31 seconds. The rest of this chapter investigates avenues for improving the computational speed of the GDM model.

3.2 Improving MCMC Sampling

To the best of our knowledge, in all published literature so far the GDM method has only been fitted using Markov Chain Monte Carlo (MCMC) techniques, specifically with the R package NIMBLE. This software package can be used to fit general statistical models that can be represented as a direct acyclic graph, as we discussed in Section 2.1.2. The GDM is not a member of the standard exponential family of distribution, and the non-linearity of modelling the relative proportions reported means that a full predictive MCMC method for inference is required (Stoner and Economou (2019)). However, the more extensive model run times of MCMC methods potentially render the GDM less operationally viable, especially within disease nowcasting contexts. As a consequence, possible alterations could be considered to try to increase the computational speed of the MCMC model fitting.

One approach is to increase the speed of which samples are taken for a given number of iterations. In fact, Seaman et al. (2022) showed that by formatting the GDM model such that there are no missing partial counts reduced the run time of the GDM model. In Stoner and Economou (2019), the original GDM fitted in NIMBLE included data with missing values for the unobserved partial counts ($z_{t,d}$) as well as the missing total counts (y_t) in the MCMC algorithm. Hence, NIMBLE assigned a Beta-Binomial distribution for each of the unobserved partial counts, and the missing values of $z_{t,d}$ were sampled by the GDM model. Since the sampling of other parameters, such as the Beta-Binomial mean and dispersion parameters, depend of the $z_{t,d}$ being sampled, sampling the missing partial counts increases the auto-correlation in the model (as explained below). However, the missing partial counts are not needed to predict the total counts and therefore are unnecessarily sampled when included in the model as missing data values. Intuitively, the unobserved partial counts can't provide additional information in the model as they are unobserved.

By including Beta-Binomial models for unseen parts of $\boldsymbol{z}_t$, $\boldsymbol{z}_t^{(unobs)}$, and sampling them as part of the MCMC algorithm, we are introducing needless dependency structures in the sampling of y_t. In this situation, the sampling of y_t will involve calculations for $p(\boldsymbol{z}_t^{(obs)}, \boldsymbol{z}_t^{(unobs)} \mid y_t)$ at each iteration, which will depend on the sampled values of $\boldsymbol{z}^{(unobs)}$. This dependence on the sampling of $\boldsymbol{z}^{(unobs)}$ induces auto-correlation in the MCMC sampling, which compounds the computational expense of sampling $\boldsymbol{z}^{(unobs)}$.

By including all of these in the MCMC, we are generating samples from $p(y \mid \boldsymbol{z}_t^{(obs)})$, where the MCMC is marginalising over $z_t^{(unobs)}$. I.e., we are generating samples from: $p(y_t \mid \boldsymbol{z}_t^{(obs)}) = \sum_{\boldsymbol{z}_t^{(unobs)}} p(y_t \mid \boldsymbol{z}_t)$. Thus, we would obtain the same predictions by only including probability models for $\boldsymbol{z}_t^{(obs)}$.

We know that the GDM can be expressed as a series of conditional Beta-Binomials, i.e.

$$p(\boldsymbol{z}_t \mid y_t) = p(z_{t,1} \mid y_t) p(z_{t,2} \mid y_t, z_{t,1}) \dots p(z_{t,D} \mid y_t, z_{t,1} \dots, z_{t,D-1}). \tag{3.60}$$

As long as $\boldsymbol{z}_t^{(unobs)}$ come strictly after $\boldsymbol{z}_t^{(obs)}$ in the series $z_{t,1}, \dots, z_{t,D}$, we can drop the models for $\boldsymbol{z}_t^{(unobs)}$, since all of the terms on the right hand side of the above equation depend only on y_t and $z_{t,d}$ that come before in the series. This situation holds in the situation where unobserved $z_{t,d}$ only arise due to reporting delays and the delayed counts are observed sequentially, by definition.

In summary, we only need to include Beta-Binomial models for $\boldsymbol{z}_t^{(obs)}$, which avoids a source of posterior autocorrelation and computational expense associated with sampling $\boldsymbol{z}_t^{(unobs)}$. Thus, modifying the GDM implementation from Stoner and Economou (2019) to achieve this, as done by Seaman et al. (2022), will reduce the computation time per iteration and, likely, the number of iterations needed to obtain a converged set of posterior samples of sufficient quality (e.g. as quantified by effective sample sizes, as detailed in Section 2.1.1).

3.2.1 Parallel processing

Further reductions in computation time may be achievable by making better use of parallel computing. This means, where possible, carrying out multiple computation tasks simultaneously on separate processing cores, rather than sequentially.

In practical use of MCMC methods, multiple Markov chains are usually run when obtaining samples from the posterior. This is to determine convergence and to help diagnose multi-modality by assessing whether chains starting from different points across the parameter space all converge to the same maxima, as discussed in Section 2.1.1. Before running the MCMC chains, the model must be specified, including the likelihood and the priors for all parameters. This step ensures the sampler has all the necessary information to explore the posterior distribution. In NIMBLE, the model and the MCMC algorithm are compiled separately into C++ code to enhance performance, especially for complex models. The compilation step is done for each MCMC chain separately to improve the efficiency of running the MCMC. Once compiled, the MCMC algorithm is run for each chain.

Since the chains are independent of each other, we can greatly reduce computation times by running them in parallel simultaneously, as opposed to running them sequentially. This functionality is not built into the NIMBLE software package we rely on (Section 2.1.2), and the choice of parallel processing method is limited because each chain requires its own instance of the compiled model and MCMC objects, which cannot be shared across parallel processes due to memory isolation and the interaction with compiled C++ code.

There are two main approaches of parallel processing for running the MCMC chains; copying the current version of R to each core, or launching a new version of R on each core. Both of these methods are possible using the function included in the R package **parallel** (R Core Team (2023)). The first approach, known as the forking approach, is where

processes share a parent memory space and the child processes are then implemented separately. The forking approach is only used for running parallel process across cores on a single machine, and is more straightforward to implement than the second approach (Peng (2020)), using the `mclapply` function.

The second approach, known as the socket approach, can be used on any computing system. Each process is run on a separate network with its own memory space, so this is ideal for large scale parallel computations and can even be set up to run over different machines. This makes it more convenient for the applications covered in this thesis as it will be easier to transfer to novel and potentially operational contexts that could utilise multiple machines. As shown in de Valpine et al. (2021), the socket approach can be implemented by creating a socket cluster with the `makeCluster` function in R, which assigns the number of cores the task will be run across. The NIMBLE model code is then built into a function that is executed, in parallel using the cluster, with the `parLapply` function.

An alternative implementation of the socket approach, using the R package **doParallel** (Corporation and Weston (2022)), is opted for in Stoner and Economou (2019) using the `dopar(.)` function. However, this approach does not use clusters to run the entirety of the NIMBLE code in parallel (including compilation and MCMC chains) and instead only runs the MCMC chain sampling in parallel. As previously mentioned, the NIMBLE compilation step is done for each MCMC chain separately, so this may not be the most efficient parallelisation technique to adopt. In Section 3.3, we compare the original socket approach used in Stoner and Economou (2019) (with R package **doParallel**) to the cluster based approach (R package **parallel**) to formally determine the potential gain in computational speed of updating the parallelisation method used. A further potential benefit of the cluster based approach is the ability to parallelise models over quantities other than the number of MCMC chains. For example, if the same model is being fit independently to

multiple regions or age groups then this could be done in parallel instead. This could be beneficial for operational surveillance systems by further reducing computational costs, especially if the number of regions or other quantity being parallelised over is greater than the number of MCMC chains used.

Both versions of the socket approach were tested when implementing the GDM framework in NIMBLE for the case studies in Section 3.4, where the original approach is used for all initial implementations and the cluster approach is used when explicitly stated in the version name. As discussed further in Section 3.4.1, we found that the **parallel** socket approach was more efficient than the **doParallel** for both case studies covered in this chapter. Hence, this is the method we adopted in the final iterations of this model, and for the remainder of this thesis for models that require an MCMC approach, to reduce the computational cost.

3.3 Direct Optimisation of the Joint Posterior

Recall from Section 2.1.1 that we must select initial values for all parameters and unobserved data for MCMC algorithms.

Gelman and Rubin (1996) advise to have some idea of the location of the posterior within the parameter space before fitting a Bayesian model, to help check that MCMC output is sensible and help set the starting points of MCMC algorithms. Point estimates for starting points can be calculated using maximum likelihood estimates (MLE) or posterior modes. MLEs are obtained by maximising the likelihood function $p(x \mid \boldsymbol{\theta})$, and thus finds values

of $\boldsymbol{\theta}$ that maximises the probability of the observed data x. On the other hand, Maximum A Posteriori (MAP) estimates, estimate $\boldsymbol{\theta}$ by maximising the posterior $p(\boldsymbol{\theta} \mid x)$, and hence obtain estimations using both the likelihood and the prior, since by Bayes theorem we have $p(\boldsymbol{\theta} \mid x) = \frac{p(x|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(x)}$.

However, analytically calculating the MLE or MAP estimates can be challenging for multivariate problems due to complex likelihood functions, and iterative methods may be needed to find the maximum of the log-likelihood or the posterior modes, such as expectation–maximisation (EM) algorithms. Gelman and Rubin (1996) argue that using approximation methods as a starting point for iterative methods gives users an initial idea of the location and scale of the posterior distributions and facilitates obtaining samples for the initial values of the MCMC algorithms. These initial values are otherwise time consuming to set for complex models due to hierarchical structures with potentially missing, longitudinal and/or multivariate outcomes that have to be considered by programmers.

In a example application given by Harms and Roebroeck (2018), they set out to make MCMC sampling in diffusion MRI analysis faster and more robust. Since the MLEs were obtainable in this case, the point estimates were used to initialise the Markov chains. However, since the GDM framework is fully Bayesian, we believe that MAP estimates are more appropriate. Hence, we focus on improving MCMC computational efficiency by setting the initial values of the model parameter for MCMC chains closer to their eventual, converged distribution using the posterior modes. Since the chains converge to the posterior we can attempt to estimate the parameter values of model effects and unobserved data in the model by directly optimising the joint posterior to find its maximum. Moreover, these estimates of the parameters are usually obtained in a much quicker time frame than the equivalent MCMC posterior samples for a given model.

Due to the nature of operational disease surveillance systems, it is likely that models will be run periodically to monitor the progression of an outbreak. For example, they may be re-run every time new counts are reported to gain the most up-to-date and informed predictions. In this scenario, it would be possible to inform parameter starting values based on the posterior distribution of those parameters from previous model runs to potentially improve model speed. Hence, the methodology we suggest here has been envisioned for use in scenarios where no current operational surveillance system has been set up.

In addition to providing sensible initial values for MCMC chains, MAP estimates could also prove to be a useful output in their own right. In Section 1.2, we argued that measures of uncertainty are important in the context of disease surveillance, so that public health officials can appropriately prepare when utilising the nowcasting methods. However, point estimates, including MAP estimates, that are more readily available than the measures of uncertainty that will eventually be available, as an output of the MCMC procedure, could provide a timely "early look" at eventual nowcasts and forecasts of the total counts. This could prove vital in a fast-moving situation, giving decision makers initial insights into e.g. a rapid outbreak while awaiting the full picture from a potentially time consuming MCMC run. The MAP estimates could also flag any potential issues or strange results from the model, giving the technical practitioner responsible for the modelling time to potentially address problems (e.g. incorrectly recorded data) before committing to an MCMC run.

3.3.1 Optimisation algorithm

There are many different R functions that carry out different optimisation algorithms. Trialling different methods included the `optim` function available in the base R distribution, as well as the `tnewton` and `varmetric` functions from **nloptr** (Ypma (2014)) package. But, I found that the **trustOptim** (Braun (2014a)) package, using the function of the same name (`trust.optim`), gave estimates of the nowcasted total counts closest to the true totals reported.

Braun (2014b) explains the difference between the trust region algorithm that is used by `trust.optim` and the line search methods used in alternative non-linear optimisation. Line search algorithms choose the distance along a given line, usually determined by the current gradient, that results in the greatest improvement in the objective function being optimised. However, if the objective function is non-concave or ill-conditioned the algorithm may be inefficient or fail. On the other hand, the trust region approach sets a maximum radius around the current point, which creates the trust region. The next point is the minimum of a quadratic approximation of the objective function that remains in this trust region, the approximation involves the gradient and Hessian of the objective function. If the next point is not finite, or leads to an inefficient step (with a worse or insufficiently better objective function), then this new point is rejected, the trust region is shrunk, and the algorithm step is repeated. Hence, trust region approaches are more robust and capable of finding posterior modes (Braun (2014b)). However, this method will converge to a local optimum. To determine potential multi-modality in the objective function this method would need to be employed with multiple starting points.

The three different available methods in the function `trustOptim(.)` determine how the Hessian is computed and stored. One requires the function to have a sparse Hessian structure, which also needs to be calculated. Since this cannot always be assumed, such as for the GDM model, two alternative quasi-Newton methods that estimate the inverse

of the Hessian are given. The Broyden–Fletcher–Goldfarb–Shanno ("BFGS") method was found to perform faster and maximise the objective function more effectively than the Symmetric Rank 1 ("SR1") method when applied to the GDM. "BFGS" uses the repeated estimates of the gradient to trace the curvature of the objective function and approximate the Hessian. However, due to the storing of the Hessians this method can be slow for problems with a large number of parameters. This issue was reduced by optimising each spatial region in a data set independently, and so this independent optimisation by spatial region was adopted for our approach.

In this context, the objective function we are maximising is the summed log probability density of the parameters we are estimating, and the gradient of this function is given by a numerical approximation of the first derivative using a simple epsilon difference calculation. This is done by the `grad` function from the **numDeriv** (Gilbert and Varadhan (2019)) package. Even though more complex gradient approximation calculations can be done with this function, it was found that these increased the run time of the optimisation without resulting in noticeably more accurate estimates.

3.3.2 Surrogate model

The aim of the optimisation is to obtain approximations of the Maximum A Posteriori (MAP) estimates by maximising the log joint posterior, given by $\log(p(\boldsymbol{\theta} \mid x)$. Therefore, obtaining the values for all unknown random quantities in the GDM model that have the highest posterior probability, i.e. the highest probability given the observed data and the priors. This includes the unobserved y values, dispersion parameters, model intercepts, and the multiple parameters required to construct each spline. Since the total counts in the GDM model are discrete this creates difficulties when trying to optimise them as the optimisation algorithm will have to be altered to identify the discrete parameters as well as assign them discrete values. To sidestep this issue, we define a surrogate model that has most of the same parameters as the GDM, while not including the total counts.

Hence the surrogate model is a marginal Negative-Binomial model of the partial counts,

$$\log(\lambda_{t,s}) = f(t,s), \tag{3.61}$$

$$z_{t,d,s}|\tilde{\phi}_{d,s}, p_{t,d,s}, \lambda_{t,s} \sim \text{Negative-Binomial}(p_{t,d,s}\lambda_{t,s}, \tilde{\phi}_{t,d,s}), \tag{3.62}$$

$$\text{probit}(S_{t,d,s}) = g(t,d,s), \tag{3.63}$$

$$S_{t,d,s} = \sum_{i=0}^{d} p_{t,i,s}. \tag{3.64}$$

The mean number of counts reported at delay d is given by $p_{t,d,s}\lambda_{t,s}$, where $\lambda_{t,s}$ is the expected mean of the total counts and $p_{t,d,s}$ are the expected proportions reported. The dispersion of the Negative-Binomial distribution is captured by $\tilde{\phi}_{t,d,s}$. Crucially, this model has the same systematic model effects for the total count and the reporting delay as the GDM but the total counts are not included in the model. The GDM model is given by,

$$y_{t,s} \sim \text{Negative-Binomial}(\lambda_{t,s}, \theta_s) \tag{3.65}$$

$$\log(\lambda_{t,s}) = f(t,s) \tag{3.66}$$

$$z_{t,d,s} \mid N_{t,d,s} \sim \text{Beta-Binomial}(\nu_{t,d,s}, \phi_{d,s}, N_{t,d,s}) \tag{3.67}$$

$$\nu_{t,d,s} = \frac{S_{t,d,s} - S_{t,d-1,s}}{1 - S_{t,d-1,s}} \tag{3.68}$$

$$\text{probit}(S_{t,d,s}) = g(t,d,s), \tag{3.69}$$

which does include the total counts as $N_{t,d,s} = y_{t,s} - \sum_{j=1}^{d-1} z_{t,j,s}$. Therefore, there are no dispersion parameters for the totals to be optimised in the surrogate model (Equations (3.61)–(3.64)), and the Negative-Binomial dispersion parameter $\tilde{\phi}_{t,d,s}$ for the delay distribution is not directly comparable to that of the Beta-Binomial dispersion parameter $\phi_{d,s}$ of the GDM model. Therefore, the MAP estimates for the GDM will not be exact since this model is only a close approximation, and the original GDM dispersion parameters will require separate optimisation.

For the nowcast period, the approximate MAP estimates of the total counts (y_t), are obtained by summing together the MAP estimates of the missing partial counts and any observed partial counts over delay. For the forecast period, estimates of y_t are set to be the approximate MAP estimates of the expected mean of the total counts ($\lambda_{t,s}$) rounded up to the nearest integer. Hence, optimising the surrogate model given by Equations (3.61)–(3.64), compiled with random initial values, results in approximate MAP estimates of all parameters except the dispersion parameters.

The GDM survivor model (Equations (3.65)–(3.69)), can then be compiled using these estimate values of the parameters, from the optimisation of the surrogate model, as the initial values of the MCMC chains. But, before the MCMC is run, the compiled GDM model is optimised for just estimates of the dispersion parameters of the partial counts (ϕ) and total counts (θ). This can be done as all other parameters are now fixed at their initial values (the approximate MAP estimates). As a consequence, when the GDM model is then run with the initial parameters now all set to their respective approximate MAP estimates, the chains should require less burn-in. This is because the chains should converge quicker as they start closer to their eventual converged distribution, so we would expect the number of iterations discarded before convergence (the burn in) to be less.

3.4 Case Studies

To determine the effectiveness of optimising the joint posterior of the GDM model in reducing the total run time of the framework, we applied this strategy when modelling two real-world case studies. The respective data sets are; the number of COVID-19 hospital deaths in England and the number of SARI hospitalisations in Paraná, Brazil. The COVID-19 data spans from April 2020 to July 2020, and the reporting delay of interest is the difference in days between the date of death and the date the deaths are published by NHS England (n.d.). The deaths are divided into 7 broad regions of England. On the

other hand, the SARI data is weekly spans from January 2013 to May 2017 and hospitalisations are given by the 22 health regions in Paraná. The delay we are modelling is the weeks between the onset of symptoms of the hospitalisations reported and when the case files for these hospitalisations are digitalised.

To assess the convergence and the quality of samples obtained from our optimisation strategy, compared to randomly generating initial values for each chain that satisfy any parameter restrictions (Gelman et al. (2013)), we computed a number of convergence diagnostic metrics. First, we calculated the potential scale reduction factor (PSRF), given by Equation (2.5) in Section 2.1.1, which measures the ratio of between sample variance to within sample variance. If this is close to 1 then there is evidence the MCMC chains have all found the same maximum, although this may be a local not global maximum. Brooks and Gelman (1998) note that it is beneficial to set random initial values to ensure samples between chains are not related, and hence multi-modality is more likely to be diagnosed. Additionally, Gelman et al. (2013, Chapter 11) recommend setting starting points with a wide range across the multiple MCMC chains so that the chains explore the posterior distribution more effectively. However, due to the choice of initial parameter values being determined by optimisation this is not possible for this application. Also, the effective sample size (ESS), defined in Section 2.1.1, was compared to ensure that the MCMC chains had been run for a sufficient number of iterations and burn-in. The PSRF and ESS for the unobserved total counts (y), the expected mean of the total counts ($\boldsymbol{\lambda}$) and the dispersion parameter of the total counts ($\boldsymbol{\theta}$) are of particular interest as the nowcasts of the total counts are the prediction of interest.

Five versions of the GDM model were fitted to each data set. Each version adds an incremental change to the last, to assess the individual and combined influence of the various changes:

1. Firstly, we fit the *Original* GDM survivor framework as outlined in Equations (3.65)–(3.69), which uses the **doParallel** package (Corporation and Weston (2022)) socket technique for parallel processing.
2. Second, we modified the implementation of the original GDM so that no missing partial counts are sampled during MCMC, as suggested by Seaman et al. (2022) and detailed in Section 3.2, which gives the *No missing z* version.
3. Third, we altered the parameterization of several quantities in the *No missing z* version so that no hyper-parameters are strictly positive, which gives the *No missing z + reparameterized* version. For example, we can assume that the log of any parameter that is required to be positive, such as dispersion parameters, is a random quantity that we assign a prior and sampler. The sampling will then take place on the unbounded real line, but when the exponential of the random quantity is taken it will still be strictly positive. This reparameterisation, so that all random quantities in the MCMC algorithm exist in the unbounded real space, ensures that the log joint posterior is well conditioned for compatibility with the optimisation algorithm described in Section 3.3.
4. Next, to determine the impact of running the MCMC chains in parallel with the **parallel** package (R Core Team (2023)) socket technique using clusters, compared to the alternative socket approach used in Stoner and Economou (2019), we ran the *No missing z + reparameterized* version using clusters in the *No missing z + reparameterized + clusters* version.
5. Finally, the *No missing z + reparameterized + clusters + optimisation* version includes the compilation and optimisation of the model given by Equations (3.61)–(3.64) to get the approximate MAP parameter estimates, which is done in parallel with one cluster. Once this is completed, in a second subsequent cluster, the estimates of the parameters are used as initial values for the same GDM as in *No missing z + reparameterized + clusters*. The full GDM is optimised for the relevant dispersion parameters in this second cluster before the MCMC is run. Due to the optimisation step the model has to be fit to each spation region independently, hence there is no nested hierarchical structure between regions as there is in all previous implementations.

3.4.1 Results

To compare the MCMC outputs, the ESS and PSRF were calculated for all parameters. However, we focus here on the parameters for the expected mean total counts, λ, and the predicted unobserved total counts themselves, since these are the predictions of interest. The *Original* version acts as a baseline for these metrics. Although, using Equation (2.4) with precision $\varepsilon = 0.1$ gives a lower bound of $\widehat{ESS} \geq 1708$, which we instead use as a rough guide more robust baseline for effective sampling if the ESS of the *Original* model is particularly less than this lower bound. All versions of the GDM are detailed in Section 3.4, and are fitted using the R package NIMBLE (de Valpine et al. (2017)) with the default MCMC samplers.

It is important to note that for all the GDM models nested cubic splines are used, as suggested in Stoner and Economou (2019) (Section 5.3.1). Hence, we assume that counts are similar between regions with a mean trend for the total spatial area considered that exhibits regional differences. However, for the final *No missing z + reparameterized + clusters + optimisation* model, we fit a surrogate GDM model that assumes each regions are independent and there is no spatial dependencies between them. As discussed in Section 3.3.1, this reduced the run time of the optimisation step by reducing the number of parameters in the model, and hence reducing the size of the hessian matrix that must be stored.

3.4.1.1 COVID-19 case study

First, we fit the 5 versions of the GDM to the COVID-19 deaths in England data set. The data was censored such that only the deaths reported up to 33rd day were seen by the model, where the data has been assumed to be fully reported after the cut off $D_{max} = 14$ days. The model GDM models then nowcast up to the 33rd day, and the number of delays modelled by the GDM was $D = 6$ days (where the remaining counts for $6 < d \leq 14$ are implicitly modelled). We also include a forecasting period of 7 days, which is achievable given the model considerations discussed in Section 2.3.3.

For our models of the daily COVID-19 hospital deaths in England we use the following effects for the GDM versions that do not include an optimisation step. First, for the expected mean of the total counts, we have

$$\log(\lambda_{t,s}) = f(t,s) = \iota_s + \delta_{t_s}, \tag{3.70}$$

where ι_s is a independent intercept for each region in England s. To capture the trend in the total hospital deaths over time we have an nested cubic spline $\delta_{t,s}$ for each region, which is centred on the cubic spline α_t. Next, for the probit transform of the expected cumulative proportions with

$$\text{probit}(S_{t,d,s}) = g(t,d,s) = \psi_{d,s} + \gamma_{t,s} + \zeta_{day[t],s}. \tag{3.71}$$

Here, $\gamma_{t,s}$ is also a nested cubic spline over time for each region, centred on the cubic spline η_t, to capture the change in the delay distribution over time. To capture the weekend effect in reporting of COVID-19 deaths in England, $\zeta_{day[t],s}$ is a cyclic cubic regression spline for the week day (Monday to Sunday) of day t ($day[t]$), which is centred on $\beta_{day[t]}$. Finally, $\psi_{d,s}$ is an independent delay effect for each region, which captures the intercept of the expected cumulative proportions reported, and increases over delay to satisfy the required monotonicity of the expected cumulative proportions, $S_{t,d,s}$, also increase over delay. Each

spline is fitted using the **mgcv** package, as outlined in Section 2.1.2.2, in the nested formulation defined in Section 5.3.1. To fully specify the model, the following priors were used for the *Original* and *No missing z model*, for Equation (3.70):

$$\iota_s \sim \text{Normal}(0, 10^2), \tag{3.72}$$

$$\boldsymbol{\alpha} = \mathbf{X}_t \boldsymbol{\kappa}_\alpha, \tag{3.73}$$

$$\boldsymbol{\Omega}_\alpha = \mathbf{S}_t \tau_\alpha, \tag{3.74}$$

$$\boldsymbol{\kappa}_\alpha \sim \text{Multivariate-Normal}(\mathbf{0}, \boldsymbol{\Omega}_\alpha), \tag{3.75}$$

$$\tau_\alpha \sim \text{Inv-Gamma}(0.5, 0.5), \tag{3.76}$$

$$\boldsymbol{\delta}_{t,s} = \mathbf{X}_t \boldsymbol{\kappa}_{\delta_s}, \tag{3.77}$$

$$\boldsymbol{\Omega}_{\delta_s} = \mathbf{S}_t \tau_{\delta_s}, \tag{3.78}$$

$$\boldsymbol{\kappa}_{\delta_s} \sim \text{Multivariate-Normal}(\boldsymbol{\kappa}_\alpha, \boldsymbol{\Omega}_{\delta_s}), \tag{3.79}$$

$$\tau_{\delta_s} \sim \text{Inv-Gamma}(0.5, 0.5), \tag{3.80}$$

$$\theta_s \sim \text{Gamma}(2, 0.02). \tag{3.81}$$

Similarly, the following priors were used for Equation (3.71):

$$\psi_{s,1} \sim \text{Normal}(0, 10^2), \tag{3.82}$$

$$\psi_{s,d} \sim \text{Normal}(0, \psi_{s,d-1}, 10^2), \tag{3.83}$$

$$\boldsymbol{\eta} = \mathbf{X}_t \boldsymbol{\kappa}_\eta, \tag{3.84}$$

$$\boldsymbol{\Omega}_\eta = \mathbf{S}_t \tau_\eta, \tag{3.85}$$

$$\boldsymbol{\kappa}_\eta \sim \text{Multivariate-Normal}(\mathbf{0}, \boldsymbol{\Omega}_\eta), \tag{3.86}$$

$$\boldsymbol{\gamma}^{(s)} = \mathbf{X}_t \boldsymbol{\kappa}_\gamma^{(s)}, \tag{3.87}$$

$$\boldsymbol{\Omega}_\gamma^{(s)} = \mathbf{S}_t \tau_\gamma^{(s)}, \tag{3.88}$$

$$\boldsymbol{\kappa}_\gamma^{(s)} \sim \text{Multivariate-Normal}(\boldsymbol{\kappa}_\eta, \boldsymbol{\Omega}_\gamma^{(s)}), \tag{3.89}$$

$$\boldsymbol{\beta} = \mathbf{X}_w \boldsymbol{\kappa}_\beta, \tag{3.90}$$

$$\boldsymbol{\Omega}_\beta = \mathbf{S}_w \tau_\beta, \tag{3.91}$$

$$\boldsymbol{\kappa}_\beta \sim \text{Multivariate-Normal}(0, \boldsymbol{\Omega}_\beta), \tag{3.92}$$

$$\boldsymbol{\zeta}_{\boldsymbol{s}} = \mathbf{X}_w \boldsymbol{\kappa}_{\zeta_s}, \tag{3.93}$$

$$\boldsymbol{\Omega}_\zeta^{(s)} = \mathbf{S}_w \tau_{\zeta_s}, \tag{3.94}$$

$$\boldsymbol{\kappa}_{\zeta_s} \sim \text{Multivariate-Normal}(\boldsymbol{\kappa}_\beta, \boldsymbol{\Omega}_{\zeta_s}), \tag{3.95}$$

$$\tau_{\gamma_s} \sim \text{Inv-Gamma}(0.5, 0.5), \tag{3.96}$$

$$\tau_{\zeta_s} \sim \text{Inv-Gamma}(0.5, 0.5), \tag{3.97}$$

$$\tau_\eta \sim \text{Inv-Gamma}(0.5, 0.5), \tag{3.98}$$

$$\tau_\beta \sim \text{Inv-Gamma}(0.5, 0.5), \tag{3.99}$$

$$\phi_{s,d} \sim \text{Gamma}(2, 0.02). \tag{3.100}$$

The splines were derived using the `jagam(.)` function from Wood (2016). In the above equations, for the temporal cubic spline; X_t is the design matrix of the splines, S_t is the penalty matrix, where $\tau_\delta^{(s)}$ and $\tau_\gamma^{(s)}$ determines the smoothness of the splines, and 8 knots were evenly placed from day 1 up to day 33 (where day 33 is the end of the nowcasting period). For the day-of-week cyclic cubic spline; X_w is the design matrix of the splines, S_w is the penalty matrix where $\tau_\zeta^{(s)}$ determines the smoothness of the splines and 8 knots were evenly placed to capture the changes over the 7 days. The Gamma and inverse-

Gamma priors above are used for strictly positive parameters. The Gamma distribution is parameterised by shape and rate parameters, and the inverse-Gamma is parameterised in terms of shape and scale parameters (See Appendix A.3 & A.4 respectively for more details). In the subsequent GDM versions which have been *reparameterised* to remove any strictly positive model parameters, the following changes are made:

$$\log(\tau_\alpha) \sim \text{Normal}(1.26, 2.23^2), \tag{3.101}$$

$$\log(\tau_{\delta_s}) \sim \text{Normal}(1.26, 2.23^2), \tag{3.102}$$

$$\log(\theta_s) \sim \text{Normal}(4.3, 0.8^2), \tag{3.103}$$

$$\log(\tau_\eta) \sim \text{Normal}(1.26, 2.23^2), \tag{3.104}$$

$$\log(\tau_{\gamma_s}) \sim \text{Normal}(1.26, 2.23^2), \tag{3.105}$$

$$\log(\tau_\beta) \sim \text{Normal}(1.26, 2.23^2), \tag{3.106}$$

$$\log(\tau_{\zeta_s}) \sim \text{Normal}(1.26, 2.23^2), \tag{3.107}$$

$$\log(\phi_{d,s}) \sim \text{Normal}(4.3, 0.8^2). \tag{3.108}$$

For the GDM model (Equations (3.65)–(3.69)) in the optimisation method we use the modelling effects given in Equations (3.70)–(3.71) reparameterised priors that are strictly positive as given above. However, we following effects in the surrogate model (Equations (3.61)–(3.64)) to reduce the number of parameters in the model. First, for the expected mean number of COVID-19 deaths:

$$\log(\lambda_{t,s}) = f(t,s) = \iota_s + \delta_t^{(s)}, \tag{3.109}$$

where ι_s is a independent intercept for each region in England s. To capture the trend in the total hospital deaths over time we have an independent cubic spline $\delta_t^{(s)}$ for each region. Next, for the probit transform of the expected cumulative proportions with

$$\text{probit}(S_{t,d,s}) = g(t,d,s) = \psi_{d,s} + \gamma_t^{(s)} + \zeta_{day[t]}^{(s)}. \tag{3.110}$$

Here, $\gamma_t^{(s)}$ is also a independent cubic spline over time for each region and $\zeta_{day[t]}^{(s)}$ is a cyclic cubic regression spline for the week-day effect. Finally, $\psi_{d,s}$ is still an independent delay effect for each region.

Table 3.1: Table of the total run time and diagnostic measures for different versions of the GDM model fitted to COVID-19 hospital deaths in England. The columns from left to right denote; the model version, the total run time in minutes, the number of MCMC iterations, the number of MCMC burn in, the estimated mean ESS for all parameters, the estimated mean ESS for just the λ parameters and the estimated mean ESS for all unobserved y parameters. All GDM models were fitted using MCMC with 4 chains and a thinning of 10. The top half of the table (rows 1-5) show model diagnostics for all models where the MCMC iterations have been set to 200 000 and the burn-in has been set to 100 000, as used in Stoner and Economou (2019). The lower half of the table (rows 6-10) differ in iteration and burn-in length for each model version as they have been set to try and achieve comparable mean ESS and PSRF diagnostics. This was determined by changing reducing the burn-in in increments and ensuring convergence has still occurred through visually inspecting trace plots and ensuring the PSRF values didn't worsen. Then, iterations were reduced in increments until the ESS columns in the lower half were approximately the same magnitude as the upper half.

Model Version	Time	Iterations	Burn In	ESS	ESS_λ	ESS_y
Original	19.45	200k	100k	3061	1714	7429
No missing z	19.00	200k	100k	3145	1712	7665
No missing z + reparameterized	25.58	200k	100k	7606	4900	11042
No missing z + reparameterized + clusters	16.87	200k	100k	7578	4946	10983
No missing z + reparameterized + clusters + optimisation	18.70	200k	100k	7577	4926	10983
Original	17.97	180k	80k	3133	1857	7337
No missing z	17.78	180k	80k	3193	1834	7626
No missing z + reparameterized	17.85	100k	30k	5284	3478	7706
No missing z + reparameterized + clusters	9.07	100k	30k	5311	3527	7688
No missing z + reparameterized + clusters + optimisation	9.85	90k	20k	5205	3397	7586

Initially, in the top half of Table 3.1, each model version was run for the same number of iterations (200,000), burn in (100,000), and thinning (10). Thinning is used to reduce the auto-correlation within chains and preserve computer memory: a thinning of k means only every k^{th} iteration is saved. The three ESS columns are the computed average ESS (after computing the ESS for each parameter separately): for all parameters; just the expected mean of the total counts (λ); and just the unobserved total counts ($y^{(unobs)}$), respectively.

Looking at the top half of Table 3.1, for 200 000 iterations with 100 000 burn in, comparing *No missing z* to the *Original* version shows that removing the missing partial counts reduced the run time. However, the *No missing z + reparameterized* compared to the *Original* version had much higher ESS values, but at the cost of a longer run time. This tells us that the mixing of the MCMC chains was improved by modifying the model such that there were no strictly positive parameters. This is likely as the MCMC algorithm no longer had to reject negative samples that were proposed. Therefore, the MCMC chains for these versions with this alteration don't need the iteration length (iterations − burn in) of 100 000 that the *Original* does in order to obtain the same ESS values.

To compare the time needed to achieve the same quality outputs, in the bottom half of Table 3.1, the number of burn in and iterations were each altered in increments until the PSRF and ESS metrics of the modified implementations approximately matched the *Original* benchmark values. Convergence was assessed by PRSF values, examining trace plots for a sample of each of the parameters, and comparing the predictions and uncertainty of the total counts. The proportion of the PSRF less than 1.05 was 1 for all model versions for $\boldsymbol{\lambda}$, for $\boldsymbol{\theta}$ and for the unobserved total counts, so these are not included in Table 3.1.

From comparing rows two and seven of Table 3.1, an iteration length (calculated by $\textit{iterations} - \textit{burnin}$) of 70,000, instead of 100,000, was sufficient to obtain equivalent MCMC mixing, as reflected in the similar ESS metrics, for the model that had been reparameterized to not include strictly positive parameters. Furthermore, the *No missing z + reparameterized + clusters + optimisation* version converged with less burn in, as determined from both trace plots and PSRF calculations. Therefore, less iterations were required to gain an iteration length of 70,000. A modest amount of burn in (20,000) was still needed, despite initialising chains at the (approximate) MAP estimates, to allow the chains to travel the (hopefully small) distance to the actual posterior of the GDM.

With respect to run times, the results from the lower half of Table 3.1 show that the *No missing z + reparameterized + clusters* version, which has a run time of 9.07 minutes, is actually just under a minute quicker than the *No missing z + reparameterized + clusters + optimisation* version, which took 9.85 minutes to run, despite the number of iterations being higher for *No missing z + reparameterized + clusters*. This is due to the time it takes to compile the surrogate Negative-Binomial model which is optimised in *No missing z + reparameterized + clusters + optimisation*.

However, a potential benefit to the *No missing z + reparameterized + clusters + optimisation* version is the optimised parameter estimates are available in under two minutes, before the MCMC is then run. For the nowcast period, the approximate MAP estimates of the total counts (y_t), are obtained by summing together the MAP estimates of the missing partial counts and any observed partial counts over delay. For the forecast period, estimates of y_t are set to be the approximate MAP estimates of the expected mean of the total counts ($\lambda_{t,s}$) rounded up to the nearest integer. These estimates from the optimisation are shown in Figure 3.3 alongside the eventual MCMC estimates from the GDM model.

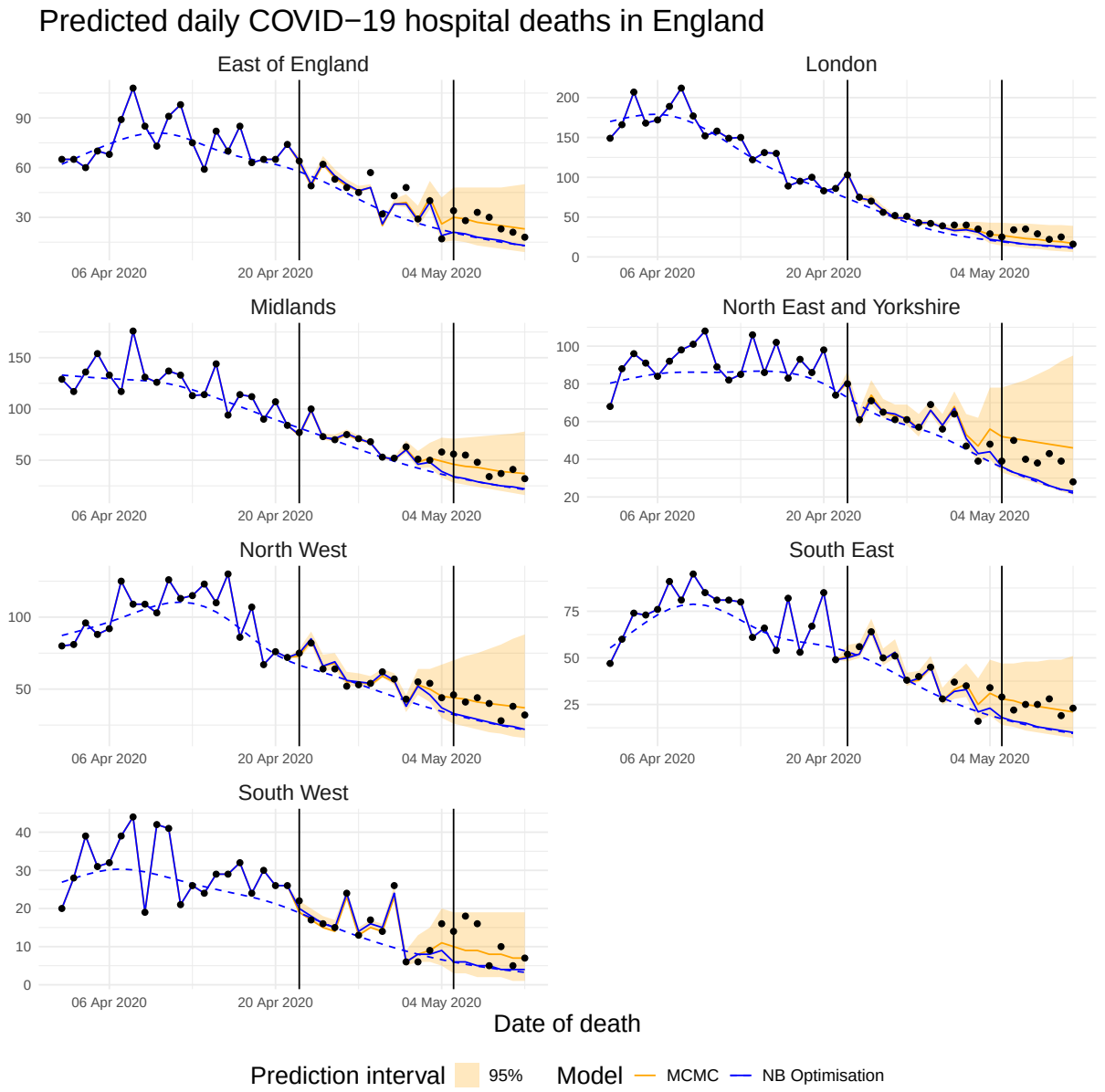

Figure 3.3: The daily COVID-19 hospital deaths in seven regions of England (points). The first vertical black line denotes the start of the nowcasting period, prior to which all the total hospital deaths are known. The second line on May 5^{th} indicates the start of the forecasting period, after which no counts of hospital deaths are known for the models to use. The orange line shows the GDM model median posterior predictions obtained by MCMC with the 95% prediction intervals given by the shaded region. The estimates of the total counts obtained by optimising the surrogate Negative-Binomial model are given by solid blue lines. Similarly, the dashed blue lines give the estimates of the Negative-Binomial model expected mean.

The optimisation estimates are close to the MCMC medians, and could therefore provide decision makers with a crucial and relatively quick first look at the possible levels of the disease while the MCMC is running. The full GDM model will then provide complete predictive distributions and potentially an increase in accuracy through not using the surrogate model.

3.4.1.2 SARI case study

For our second case study, we fit the 5 versions of the GDM (detailed in Section 3.4) to the SARI hospitalisations in Paraná data set. The data has been assumed to be fully reported after $D_{max} = 3$ weeks (in which approximately 85% of the total counts have been reported historically), and GDM models $D = 2$ weeks. For this data we nowcast up to 224 weeks and forecasts up to 230 weeks (using the forecasting framework outlined in Section 2.3.3).

For the weekly SARI cases the state of Paraná in Brazil, we use the following effects. For the log link of the expected mean of the total counts we have

$$\log(\lambda_{t,s}) = f(t,s) = \log(population_s) + \iota_s + \delta_{t,s} + \xi_{week[t],s}, \tag{3.111}$$

where a population offset $\log(population_s)$ is included for each of the health districts s in Paraná. The mean SARI rate can vary considerably from region to region due to differences in population sizes. Including $\log(population_s)$ as an offset accounts for this variability so that ι_s only needs to capture differences in the per capita rate. This gives us the option to use more constrained/informative priors but we don't exploit that in this example as the prior is given by $\iota_s \sim \text{Normal}(0, 10^2)$. Additionally, including $\log(population_s)$ also would give the option to use a hierarchical prior or a spatially-structured prior for ι.

To capture the temporal trend in the total weekly cases, we have a nested cubic spline $\delta_{t,s}$ for each region. This is centred on α_t, which is a cubic spline that captures the mean trend across all regions. Similarly, $\xi_{week[t],s}$ is a nested cyclic cubic regression spline of the seasonal trend of the SARI cases. The seasonal cycle of this spline is over the week of the year (1 to 52) given by $(week[t] \in 1,...,52)$ for time t, and is therefore assumed to be the same for each year such that 1 and 52 join. This nested spline $\xi_{week[t],s}$, is centred on a cyclic cubic spline for all regions $\beta_{week[t]}$. Hence, $\xi_{week[t],s}$ captures the overall trend and the regional difference. For the expected cumulative proportions reported, we then model

$$\text{probit}(S_{t,d,s}) = g(t,d,s) = \psi_{d,s} + \gamma_{t,s}. \tag{3.112}$$

Here, $\gamma_{t,s}$ is also a nested cubic spline over time for each region to capture the change in the delay distribution compared to the overall mean, which is captured by a cubic spline η_t. Finally, $\psi_{d,s}$ is an independent delay effect for each region. As in Section 3.4.1.1, when defining the surrogate model (Equations (3.61)–(3.64)) we use the following model effects where all temporal and seasonal splines are independent for each region and not nested:

$$\log(\lambda_{t,s}) = \log(population_s) + \iota_s + \delta_t^{(s)} + \xi_{week[t]}^{(s)}, \tag{3.113}$$

$$\text{probit}(S_{t,d,s}) = \psi_{d,s} + \gamma_t^{(s)}. \tag{3.114}$$

Hence, for the surrogate model it is assumed that counts are independent for each region. The priors as the same as for the COVID-19 model, apart from the seasonal spline which is not included in the COVID-19 hospital deaths model. For all GDM model the priors for the seasonal spline is given by:

$$\boldsymbol{\beta} = \mathbf{X}_w \boldsymbol{\kappa}_\beta, \tag{3.115}$$

$$\boldsymbol{\Omega}_\beta = \mathbf{S}_w \tau_\beta, \tag{3.116}$$

$$\boldsymbol{\kappa}_\beta \sim \text{Multivariate-Normal}(0, \boldsymbol{\Omega}_\beta), \tag{3.117}$$

$$\boldsymbol{\xi_s} = \mathbf{X}_w \boldsymbol{\kappa}_{\xi_s}, \tag{3.118}$$

$$\boldsymbol{\Omega}_{\xi_s} = \mathbf{S}_w \tau_{\xi_s}, \tag{3.119}$$

$$\boldsymbol{\kappa}_{\xi_s} \sim \text{Multivariate-Normal}(\boldsymbol{\kappa}_\beta, \boldsymbol{\Omega}_{\xi_s}), \tag{3.120}$$

$$\log(\tau_{\beta_s}) \sim \text{Normal}(1.26, 2.23^2), \tag{3.121}$$

$$\log(\tau_{\xi_s}) \sim \text{Normal}(1.26, 2.23^2). \tag{3.122}$$

For the surrogate model in the optimisation step the priors for the regionally independent splines $\xi^{(s)}_{week[t]}$ are:

$$\xi^{(s)}_{week[t]} = X_w \boldsymbol{\kappa}^{(s)}_\xi, \tag{3.123}$$

$$\boldsymbol{\Omega}^{(s)}_\xi = \mathbf{S}_w \boldsymbol{\tau}^{(s)}_\xi, \tag{3.124}$$

$$\boldsymbol{\kappa}^{(s)}_\xi \sim \text{Multivariate-Normal}(\mathbf{0}, \boldsymbol{\Omega}^{(s)}_\xi), \tag{3.125}$$

$$\log(\tau^{(s)}_\xi) \sim \text{Normal}(1.26, 2.23^2). \tag{3.126}$$

For the temporal cubic splines we evenly place 18 knots up to week 224, and for the seasonal cyclic cubic spline we evenly place 9 knots to capture the 52 week yearly cycle.

The weekly SARI cases in the 22 regional health districts in Paraná, Brazil, is a much larger data set compared to the COVID-19 hospital deaths in the previous section. Hence, the models have a much longer run time and are measured in hours instead of minutes. The top half of Table 3.2 consists of each model version being run for 2,000,000 iterations with 1,000,000 burn in and a thinning of 1,000. Note that the proportion of the PSRF less than 1.2 for $\boldsymbol{\lambda}$, and less than 1.05 for $\boldsymbol{\theta}$, were both 1 for all models considered so are not included in the table.

Table 3.2: Table of total run time and diagnostic measures for different versions of the GDM model fitted to SARI cases in Paraná, Brazil. The columns from left to right denote; the model version, the total run time in hours, the number of MCMC iterations, the number of MCMC burn in, the estimated mean ESS for all parameters, the estimated mean ESS for just the λ parameters and the estimated mean ESS for all unobserved y parameters, the proportion of the PSRF less than 1.05 for λ and less than 1.2 for the unobserved y parameters. All GDM models were fitted using MCMC with 4 chains and a thinning of 1000. The top half of the table (rows 1-5) show model diagnostics for all models where the MCMC iterations have been set to 2 000 000 and the burn-in has been set to 1 000 000. The lower half of the table (rows 6-10) differ in iteration and burn-in length for each model version as they have been set to try and achieve comparable diagnostics. This was determined by changing reducing the burn-in in increments and ensuring convergence has still occurred through visually inspecting trace plots and ensuring the PSRF values didn't worsen. Then, iterations were reduced in increments until the ESS columns in the lower half were approximately the same magnitude as the upper half or close to a rough lower bound of $\widehat{ESS} \geq 1708$. This lower bound was calculated with Equation (2.4) where $\varepsilon = 0.1$ is the user specified precision.

Version	Time	Iterations	Burn in	ESS	ESS_{λ}	ESS_{y}	$\text{PRSF}_{\lambda<1.05}$	$\text{PRSF}_{y<1.2}$
Original	13.26	2000k	1000k	1270	2518	3510	0.986	0.990
No missing z	13.14	2000k	1000k	1288	2500	3496	0.991	0.977
No missing z + reparameterized	14.46	2000k	1000k	1704	3511	3627	1	0.982
No missing z + reparameterized + clusters	13.72	2000k	1000k	1703	3496	3647	1	0.981
No missing z + reparameterized + clusters + optimisation	13.83	2000k	1000k	1701	3470	3656	1	0.991
Original	18.84	2600k	1000k	1692	3585	5695	0.999	0.982
No missing z	18.29	2600k	1000k	1663	3443	3327	0.993	0.988
No missing z + reparameterized	8.28	1200k	200k	1718	3545	3677	1	0.984
No missing z + reparameterized + clusters	8.30	1200k	200k	1715	3578	3662	1	0.963
No missing z + reparameterized + clusters + *optimisation*	7.28	1050k	50k	1751	3571	3662	1	0.982

The results in Table 3.2 convey a similar story to the COVID-19 case study, it is clear that the removal of any strictly positive parameters has improved the mixing of the MCMC chains due to the increase in ESS. Furthermore, removing missing partial counts and running the chains in parallel using clusters improves the speed of the model. The bottom half of the table, the burn in and then iterations of the different model versions were altered until the ESS and PSRF values reflect similar values to those achieved by the *Original* version with the iterations of 2,000,000 and burn in of 1,000,000. However, for this SARI case study, since the ESS averaged over all parameters is only 1270, the iterations and burn in were chosen to such that this value was closer to 1700 instead. The lower half of Table 3.2, shows that only 50,000 burn-in is required for the *No missing z + reparameterized + clusters + optimisation* version to achieve convergence, compared to *No missing z + reparameterized + clusters* requiring 200,000 burn in. Also, the optimisation model is the faster despite the additional time it takes to compile the marginal Negative-Binomial model and carry out the optimisation. This shows that when the compilation time takes up a smaller proportion of the overall run time due to a longer time series over more spatial regions, the impact of this extra compilation time is negligible, unlike in Section 3.4.1.1.

Once again an initial outline of the nowcasts and forecasts are given after completing the optimisation before the full GDM model is run. The estimates in Figure 3.4 took approximately 6 minutes to obtain and are relatively close to the final MCMC predictions.

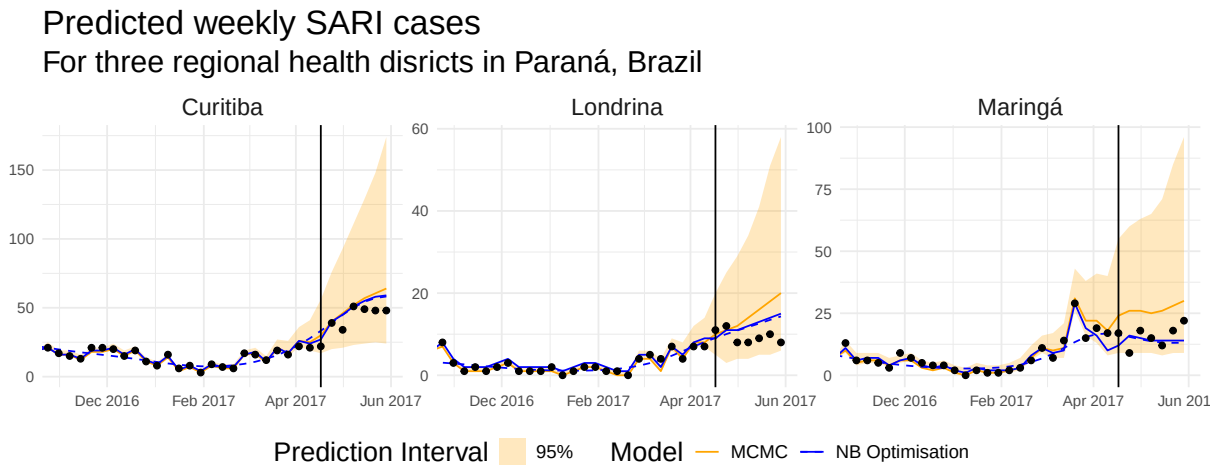

Figure 3.4: The weekly number of SARI cases in the three most populated health regions of Paraná, Brazil (determined by the 2010 census) are given by the points on each plot. The orange line shows the GDM model median posterior predictions obtained by MCMC with the 95% prediction intervals given by the shaded region. The estimates of the total counts obtained by optimising the surrogate Negative-Binomial model are given by the solid blue lines, where dashed blue lines give the estimates of the Negative-Binomial model expected mean.

As a result of the chains being run in parallel, increasing the number of MCMC chains can increase the Effective Sample Size (ESS) of a model, with comparatively smaller increases in run times compared to increasing the length of the chains. Table 3.3 compares the *Original* model to the *No missing z + reparameterized + clusters + optimisation* version, where both are using 8 chains instead of 4, for the same iterations and burn in (top two rows). Then for row three of Table 3.2, the burn in and iterations of the *No missing z + reparameterized + clusters + optimisation* version were reduced simultaneously, to maintain an iteration length of 1 000 000, such that the PRSF values were still comparable to the same *No missing z + reparameterized + clusters + optimisation* model version run with only 4 chains (Table 3.2 row five). Just the total number of iterations were then reduced until the ESS values were comparable to those calculated for the *Original* version with 8 chains (Table 3.3 row one).

Table 3.3: Table of total run time and diagnostic measures for different versions of the GDM model fitted to SARI cases in Paraná, Brazil. The columns from left to right denote; the model version, the total run time in hours, the number of MCMC iterations, the number of MCMC burn in, the estimated ESS mean for all parameters, the estimated ESS mean for $\boldsymbol{\lambda}$, the estimated ESS mean for unobserved y, the proportion of the PSRF less than 1.05 for $\boldsymbol{\lambda}$ and less than 1.2 for the unobserved y parameters. All GDM models were fitted using MCMC with 8 chains and a thinning of 1000. The proportion of the PSRF less than 1.2 for $\boldsymbol{\lambda}$ and less than 1.05 for $\boldsymbol{\theta}$ were both 1 for all models in this table.

Version	Time	Iterations	Burn in	ESS	$\text{ESS}_{\boldsymbol{\lambda}}$	ESS_{y}	$\text{PRSF}_{\boldsymbol{\lambda}<1.05}$	$\text{PRSF}_{y<1.2}$
Original	18.62	2000k	1000k	2555	4824	7078	0.997	0.988
No missing z + reparameterized + clusters + optimisation	17.85	2000k	1000k	3454	7030	7356	1	0.995
No missing z + reparameterized + clusters + optimisation	9.42	1050k	50k	3514	7107	7318	1	0.992
No missing z + reparameterized + clusters + optimisation	6.70	750k	50k	2525	5036	5068	1	0.993

As expected, Table 3.3 shows that with more MCMC chains less iterations are needed before suitable ESS are obtained and therefore the overall run time is approximately half an hour less than for 4 chains. The improvement in run time has also been increased as the optimisation is quicker when carried out over 8 computer cores instead of 4. Hence, the time that the initial optimised parameter estimates are available was reduced from 6 to 4 minutes.

3.5 Discussion

In this chapter, we have investigated potential avenues for improving the computational efficiency, and therefore the operational practicality, of the GDM method. The GDM is introduced in Section 2.3 as a competitive alternative to current state-of-the-art models in terms of predictive precision. However, due to the current need to fit the GDM model using MCMC techniques this makes it significantly computationally slower compared to methods that utilise an INLA approach for fitting Bayesian models.

In Section 3.1 we presented one possible approach for approximating the GDM with a latent Gaussian model that we can fit using the INLA approach. A strong motivation for this was due to the fact that the Negative-Binomial INLA model proposed by Bastos et al. (2019) does not fully capture the different sources of correlation within the partial counts and is therefore unable to properly account for the variability in the data. However, whilst our proposed GDM approximation in INLA model seemed to have less uncertainty compared the Negative-Binomial model, its uncertainty was still excessive compared to the GDM model, as shown by Figure 3.1. Hence, in applications where timeliness is a key priority, the GDM approximation INLA model could give more reliable nowcasts than the existing Bastos et al. (2019) Negative-Binomial INLA model. However, the GDM model is still optimal for applications where decision makers require a precise picture of the potential risk levels in a population.

Consequently, we then focused on improving the computational speed of the full GDM, which can currently only be fitted using MCMC. First of all, we managed to improve MCMC sampling efficiency by implementing some relatively straightforward changes. This included; not modelling the unknown partial counts as advised by Seaman et al. (2022), reparameterizing parameters such that none where strictly positive and using a cluster

parallel processing technique to simultaneously carry out MCMC sampling. Results in Section 3.4, Tables 3.1 & 3.2, show that all three of these changes resulted in improvements in model run times. Due to the ease of implementation, whenever employing the GDM model in subsequent chapters all three of these adjustments are adopted.

Next, we attempted to improve MCMC sampling efficiency by reducing the number of iterations required as burn in for the chains. This was achieved by starting the parameters in each chain at initial values that were likely to be close to the relative parameter values once it eventually converged to the posterior. We determined these initial values by directly optimising the joint posterior of the model. The results were promising as the run time of the optimisation model was more than halved compared to the original model, as shown for the case study of COVID-19 deaths in England by Table 3.1. Also, both the initial optimised parameter values and the eventual MCMC posterior medians were close to the true total counts they were nowcasting, as seen in Figures 3.3 & 3.4. Additionally, the two step process of this approach could allow for decision makers to gain a quick preliminary insight to the possible predictions from the optimised values and then examine the uncertainty and get more precise predictions once the full GDM is fitted with MCMC.

Before this approach can be confidently adopted, a more thorough investigation into its robustness, with an extensive simulation study, would need to be carried out. Since the initial values of MCMC chains are not started from random values across the parameter space there is a potential that it will be harder to diagnose multi-modality and convergence in models (Gelman and Rubin (1996)) and could have a non trivial impact on final model estimates. For this reasons, we do not utilise this technique in subsequent chapters.

However, an additional strategy for reducing computation time that we have not covered in this chapter is a moving data window. This has previously been adopted by nowcasting methods such as Kline et al. (2022) and McGough et al. (2020), and we fully explore the benefits of applying moving windows to nowcasting COVID-19 hospital deaths in England for Stoner, Halliday and Economou (2022) in Section 4.2.4.

Chapter 4

Nowcasting COVID-19 Fatalities in England

In this chapter, we outline work that contributed to the Stoner, Halliday and Economou (2022) paper, where we implement the GDM framework for delayed COVID-19 deaths in England. This follows on from our investigation into the efficiency of the GDM model in Chapter 3 where we refined the GDM framework. Key updates included omitting the modeling of unknown partial counts (as noted by Seaman et al. (2022)), reformatting model parameters to avoid strictly positive sampling, and implementing socket parallelization using clusters. These improvements were applied in Stoner, Halliday and Economou (2022), enhancing the GDM framework's competitiveness against existing approaches. The motivation for Stoner, Halliday and Economou (2022) was to determine the GDM's ability to capture all sources of variability in COVID-19 deaths in England suffering from delayed reporting, and showcase it as a viable option for general operational nowcasting. We summarise the four sources of variability that is present in delayed reporting data in Section 1.1.3 and discuss existing nowcasting approaches in Section 2.2.

For Stoner, Halliday and Economou (2022) we applied the GDM framework to the English fatalities data set and compared it to the most cited alternative approaches for nowcasting using a 15-month rolling prediction experiment, detailed in Section 4.1. The data is provided by the National Health Service for England (NHS England) and published as daily counts of deaths occurring in hospitals within England by UK Government (2023). This included patients that had either tested positive for COVID-19 or had COVID-19 mentioned on their death certificate. Daily counts were grouped in time by date of death and in space by seven regions in England (e.g. London, South East, Midlands) and captured deaths reported from 4pm two days prior to 4pm one day prior to publication. The data spans from April 2020 to November 2021.

The remainder of the work covered in this chapter outlines my specific contributions to Stoner, Halliday and Economou (2022) as well as related work not published. Firstly, Section 4.2 discusses the impact of auto-regressive structures, tensor product time-week day interaction effects and moving data windows on the performance of the GDM model. Whilst the investigation into auto-regressive effects and tensor product interactions were

not included in the final paper manuscript it impacted final model decisions for the rolling prediction experiment and was used to address reviewer comments regarding the necessity of auto-regressive structures. On the other hand, implementing moving data windows improved the computational speed of the GDM model and was applied to all model versions in the paper for a fair comparison. Section 4.3 covers the simulation study that I designed and carried out for the supporting information (Appendix C) of Stoner, Halliday and Economou (2022) to determine the framework's performance as a public health tool in terms of parameter inference.

4.1 Correcting Delayed Reporting of COVID-19 Using the GDM Method

The ultimate goal of the statistical models used to aid disease surveillance is to make accurate predictions of the total counts, i.e. cases or deaths, with reliable levels of certainty which can be quantified. In Stoner, Halliday and Economou (2022), we investigate the predictive performance of GDM for both the survivor ("GDM_Survivor") and hazard ("GDM_Hazard") versions, as introduced in Section 2.3 by comparing to two approaches which we reviewed in Section 2.2.2. First, we compare to the Negative-Binomial model fitted in INLA from Bastos et al. (2019), and refer to here as the "INLA" model. Finally, we compare to two version of the McGough et al. (2020) model, "NobBS" & "NobBS-14", where the latter is specified to have a 14 day moving window. Moreover, we also fit a marginal Negative-Binomial version of the Survivor GDM, which would come under the 'conditional independence' group of model approaches that we introduce and review in Section 2.2.2. This is to determine whether improvements over the other models are a consequence of the conditional full GDM framework being able to capture all random variability in the delay distribution, and not because of differences in the spatio-temporal

structures used. To further ensure a fair comparison between models, all models were fitted to each spatial region separately. We used COVID-19 hospital deaths data from UK Government (2023) to carry out a rolling predictions experiment where the model was fitted 20 times over 15-months.

This rolling prediction experiment includes 7 days of forecasting after each nowcast date, where the forecasting GDM framework discussed in Section 2.3.3 is used. We calculate the totals for each day by using a cut-off for delay of $D_{max} = 14$ days. Also, we calculate the prediction time difference (PTD), where a PTD greater than 0 correspond to dates after the current date (T_{now}) we are considering (i.e. forecasts). Meanwhile, a PTD of 0 days or less corresponds to predictions made leading up to the current date T_{now}, when at least one part of the total deaths has been observed (i.e. nowcasts).

As with previous implementations, the GDM was fitted in NIMBLE (de Valpine et al. (2017)) where the default MCMC samplers were used. For each nowcast date a MCMC chain was run with 20,000 iterations and 15,000 burn-in. The GDM was fitted to model $D = 6$ days of delay to data in a 70 day moving window. To be comparable to competing methods, the GDM assumed an indepedent model for each region. Hence, temporal and week-day effect cubic splines were defined indepdently for each of the seven regions of England, with no hierarchical nested structure. Within each 70 day moving window 10 knots where evenly placed for each of the temporal splines (where day 71 to 77 is the forecasting period where no knots were placed). For the week-day cyclical cubic spline 8 knots where evenly placed to capture the closed 7 day cycle.

We summarise prediction performance by calculating three metrics given in Figure 4.1. First, the mean absolute error (MAE), over region, of the posterior median predicted number of deaths occurring on each day, to quantify how accurate point estimates are. Second, is the mean 95% prediction interval width for the total number of deaths on each day, which quantifies how precise/uncertain predictions are. Third, the 95% PI coverage,

which checks whether uncertainty is adequately quantified by the model. Here, we use the word "coverage" to describe the proportion of data points contained within their corresponding 95% prediction intervals. Four days of nowcasting and forecasting were chosen as a reasonable period that may be of interest to public health researchers when monitoring the COVID-19 hospital deaths in England. In this period the nowcasts are informative for correcting the reported counts and the forecasts prediction intervals are still meaningful, where as they may become too uncertain to interpret if predicting too far into the future.

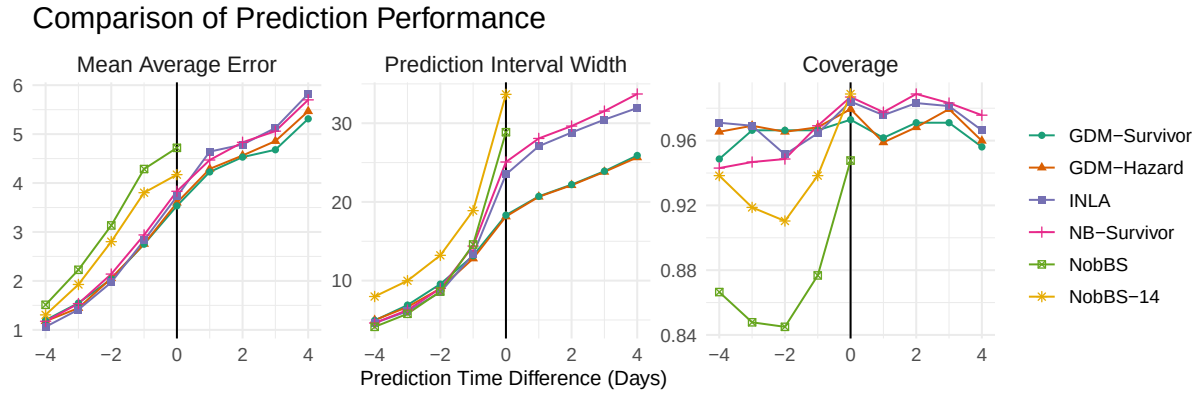

Figure 4.1: Mean absolute errors (left), mean 95% prediction interval widths (centre), and 95% prediction i9nterval coverage values (right) for daily COVID-19 deaths in the rolling prediction experiment. Performance metrics are arranged on the x-axis by prediction time difference (PTD), from -4 days up to +4 days, chosen as a period that may be informative to public health practitioners when monitoring the COVID-19 hospital deaths. The different models used to generate predictions are represented by different colours and shapes. ***Source***: Stoner, Halliday and Economou (2022).

The results, Figure 4.1, show that the both versions of the GDM out-performed the competing models in all metrics measured. Specifically, more accurate point estimation (smallest mean average error), higher precision (narrowest 95% prediction interval mean widths) and non-excessive uncertainty (95% prediction interval coverage closest to 0.95). Reasoning for this improvement in predictive power is likely to be due to the following factors:

- Appropriate separation of systematic variability in the total count and in the delay.

- Use of splines to capture all systematic temporal variation – this led to less over-prediction in the total counts than the alternative approach of first order random walk terms.
- Appropriate handling of the variability in the delay distribution with the flexibility of the GDM model.

Moreover, McGough et al. (2020) exhibited increased uncertainty in predictions during periods of particularly high cases. This kind of inconsistency can cause problems for decision makers, especially since predictions are more vital during epidemic outbreaks when case numbers are likely to be high.

4.2 Exploring Model Choices

4.2.1 Auto-regressive effect

Kline et al. (2022) adopt the modelling framework from Stoner and Economou (2019), when modelling COVID-19 counts in Ohio. They suggest that not only does this improve forecasting capabilities of the model, it also better imitates infectious disease dynamics. Furthermore, they argue the AR terms can be more easily included in the hierarchical spatio-temporal structures and avoid the need to specify knots, unlike the spline approach.

Hence, I investigated whether the GDM would perform better with a first order auto-regressive (AR) term when modelling English COVID-19 deaths. However, through inspecting auto-correlation function (ACF) and partial auto-correlation function (PACF) plots, it was determined that there was no signs of auto correlation in the current model residuals that needed to be captured. As you can see in Figure 4.2, the addition of an AR term as well as the spline has little impact on the model scaled residuals. Similarly,

just an AR term with no spline also has little effect on model residuals for nowcasting (to the left of the vertical line) but leads to larger residuals for the period of forecasting (to the right of the vertical line). Furthermore, the AR effect could theoretically absorb too much of the variance, making the nested splines effectively drop out of the model when both effects were included.

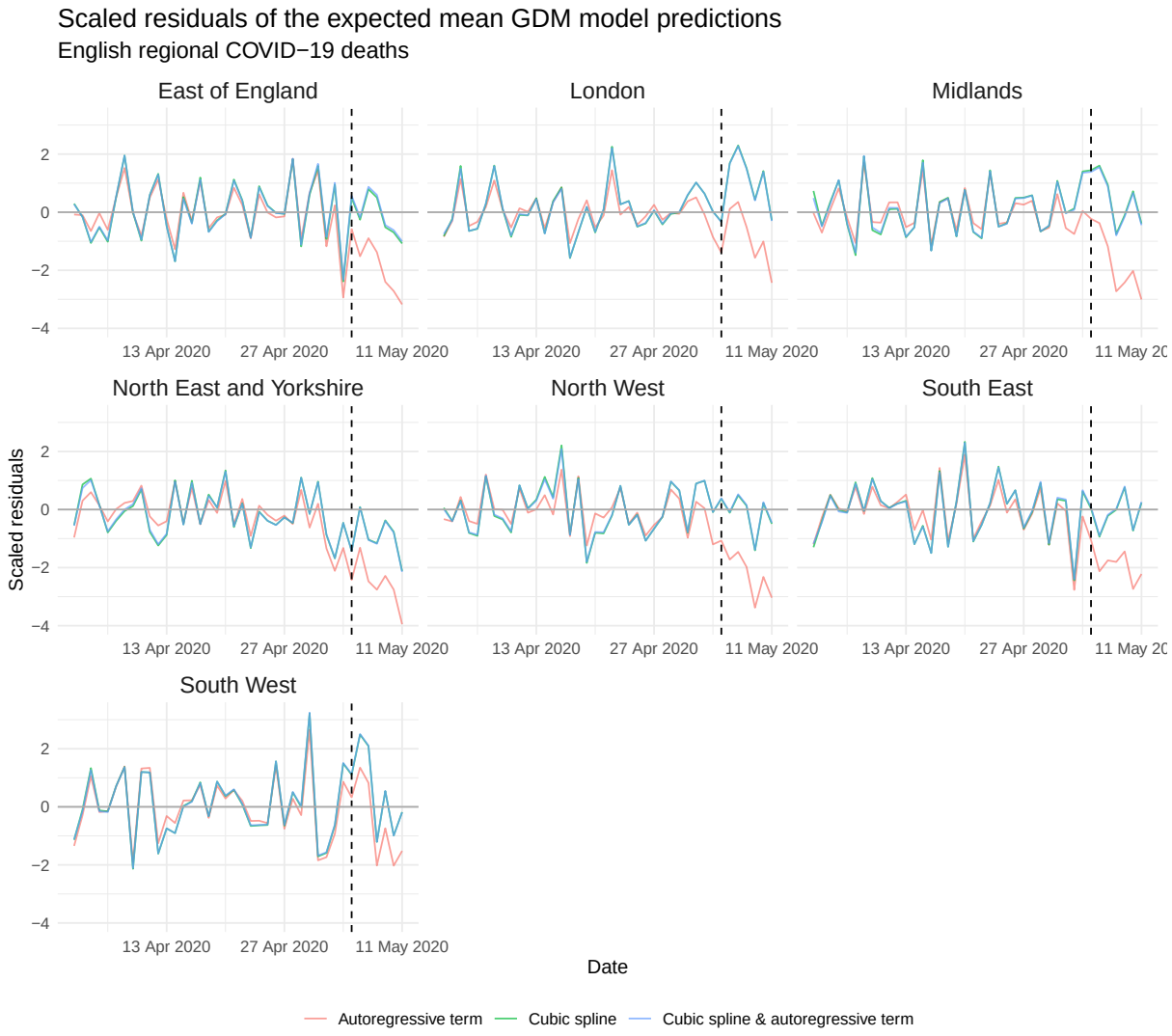

Figure 4.2: Plot of the expected mean COVID-19 deaths scaled residuals for three different specifications of the survivor GDM model effects. The vertical line indicates the date the theoretical "current day" (May 5^{th} 2020) the nowcasting is performed up to and the forecasts are predicted after.

4.2.2 Tensor product smooth interactions

The choice of link function for the delay distribution is not trivial and can have a large impact on prediction performance. The two options given in Stoner, Halliday and Economou (2022), and outlined in Section 2.3, are the 'Hazard' and 'Survivor' versions of the model. Ultimately, we advocated for the survivor version for this application of COVID-19 deaths in England as it involves modelling the more intuitive cumulative proportions. Hence, it is easier to interpret and fit effect structures than when considering relative proportions reported, especially for later delays. However, in order to settle on this choice it was important to investigate whether utilising the hazard version to its full flexibility was superior to the survivor version. Since effect structures for the hazard version are not constrained to be monotonically increasing over delay it is more straight forward to fit models that could capture more complex structures in the delay distribution of the data.

Hence, we investigated whether these complexities were present in the COVID-19 mortality data by fitting a flexible delay structure using the hazard version of the GDM. We model the logit-transformed expected relative proportion reported,

$$\text{logit}(\nu_{t,d,s}) = \beta_s^{(t,d)} + \gamma_{t,s}, \tag{4.1}$$

where $\beta_s^{(t,d)}$ represents an independent cubic spline with shrinkage applied separately for each delay (d) and region (s). To capture both the independent temporal and weekday trends, as well as their interactions, $\gamma_{t,s}$ is modeled using a tensor product spline. This tensor product interaction of time and weekday allows for flexible modeling of any complex relationships between these variables. All splines were fitted using the `jagam` function from the **mgcv** package (Wood (2016)).

Despite this extra flexibility, the model results were directly comparable to the hazard version with independent penalised splines for time and day-of-the-week and no interaction term, suggesting this interaction effect was not utilised and may not describe the data well. Although, the time-week day tensor product interaction hazard model was not included in the final version of the paper, developing the GDM to include tensor product interactions further highlights its flexibility.

Hence, the simpler hazard version, with no interaction, was chosen for the final comparison in Stoner, Halliday and Economou (2022) due to its less complex structure requiring fewer parameters and having a more practical run time. Furthermore, the hazard version run times benefited from carrying out MCMC using elliptical slice sampling (ess) for the delay distribution parameters. The MCMC ess samplers available in NIMBLE are designed to be able to efficiently sample latent variables with a multivariate Gaussian prior where there are strong correlations between these variables, as discussed in Section 2.1.2.2. Traditional MCMC samplers such as the Gibbs or Metropolis-Hastings samplers are not able to perform well under these conditions (Murray, Adams and MacKay (2010)). However, the survivor version was still more viable as it took just over 1 hour to run compared to the simpler hazard version taking approximately 3 hours. On the other hand, in this chapter and in Stoner, Halliday and Economou 2022, we used the default NIMBLE MCMC samplers for all other implementations of the GDM.

4.2.3 Improvements to application and implementation

Here we detail considerations that were undertaken to effectively implement our modelling framework for COVID-19 hospital deaths in England, which was mainly inspired by the work carried out in Chapter 3. Additionally, there was a need to ascertain the robustness of the GDM for more varied data trends extending beyond the relatively short time period of COVID-19 hospital deaths data our model was fitted to in earlier versions of our paper. As a consequence, I developed an R script which could automatically download the latest

daily COVID-19 hospital deaths data from the NHS England website. This allowed us to fit the GDM model to a much longer time series of data, which therefore exhibited more variation, which we then modelled using a rolling prediction experiment. I also carried out work so that the modified INLA model based on Bastos et al. (2019) and the NobBS model from McGough et al. (2020) were both fitted to this extended time series, to enable the thorough comparison presented in Stoner, Halliday and Economou (2022).

As investigated in Section 3.4.1, to improve the computational speed of the all GDM model versions, I implemented the recommendation in Seaman et al. (2022) of formatting the model such that no missing partial counts are included to improve computational speed. The missing partial counts are not needed to predict the total counts and therefore are unnecessarily sampled when inputted in the model as missing data values, as explained in Section 3.2. Furthermore, I ensured no model parameters being sampled were strictly positive and used a cluster parallelisation approach, summarised in Section 3.2.1, which reduced run times even further. The effectiveness of these alterations are fully investigated in Section 3.4.1.

However, despite these improvements to the GDM framework, the computational run times for the updated longer time series of hospital deaths in England weren't desirable. Following on from Kline et al. (2022) and McGough et al. (2020) who both use moving windows when formulating their nowcasting frameworks, to improve computationally efficiency, I applied similar moving windows to the GDM framework in Stoner, Halliday and Economou (2022). Not only does this improve the speed of our approach but it also makes it more generalised. Longer time series would require more knots for splines of time (which would increase run times further by increasing the number of parameters in the model and thus its overall complexity), needing to be set by end users prior to model running.

What follows in the remainder of this section is text from Web Appendix E of Stoner, Halliday and Economou (2022), which I contributed to jointly together with the other authors.

4.2.4 Moving windows

As surveillance of a disease persists, available data may eventually span months or years. Fitting complex models like those described in this article to long data time series can prove burdensome. For example, when models involve splines of time, as time series grow it may become necessary to increase the number of knots and therefore the number of coefficients to be estimated. It is therefore increasingly common to see models for correcting delayed reporting trained against only the most recent data, falling within a specified moving data "window". For example, Kline et al. (2022) used a moving data window of 90 days to reduce computation time when applying the GDM approach to COVID-19 cases in Ohio. Moving windows also have the potential to improve prediction performance in methods where the delay distribution is assumed constant in time, e.g. McGough et al. (2020), as estimation is only informed by more recent data, which may be more representative of the prediction period than older data.

Here we define a moving window in terms of the date of death, t, and the most recent date of death for which data is available T_{now} (called the data cutoff, C, in Stoner, Halliday and Economou (2022)). Given a moving window size W (days), we only fit models to partial count data $z_{t,s,d}$ and total count data $y_{t,s}$ falling within the range $T_{now} - W + 1 \leq t \leq T_{now}$.

We repeated the rolling prediction experiment for 3 different moving window lengths (W) to help determine general patterns in prediction performance and computation time. For the GDM Survivor model, we compared a moving windows of 5 weeks with 5 knots for temporal splines ($W = 35$ days), 10 week with 10 knots for temporal splines ($W = 70$ days) and 15 weeks with 15 knots for temporal splines ($W = 105$ days).

To investigate differences in accuracy, precision, and reliability, we computed mean average errors of predictions, mean 95% prediction interval widths, and 95% prediction interval coverage values. Figure 4.3, from Appendix E of Stoner, Halliday and Economou (2022), presents these metrics for different prediction time differences. Recall positive time differences indicate forecasting and negative time differences indicate prediction for dates of death in the past, relative to the model fit. The mean average errors are virtually identical across all 3 moving window sizes, meaning a moving window size of 5 weeks ($W = 35$) would be sufficient for accurate point estimates. On the other hand, while window sizes of 10 and 15 weeks had identical mean prediction interval widths, predictions with a 5 week window had substantially greater widths. This means predictions from the model with a 5 week window were more uncertain, especially when forecasting. For all window sizes, 95% prediction interval coverages were consistently above 0.95. Therefore, for this application we would recommend a 10 week moving window out of the three sizes we have investigated here. However, further window length could be tested to find an optimal length that reduces computational cost without increased uncertainty.

The average computation times were 30 minutes per day with a 5 week window ($W = 35$), 75 minutes with a 10 week window ($W = 70$), and 155 minutes with a 15 week window ($W = 105$). The computation time therefore more than doubled in both increments of 5 weeks to the window size. This makes sense, since both the number of data points in the model and the number of spline coefficients to estimate increase as W increases.

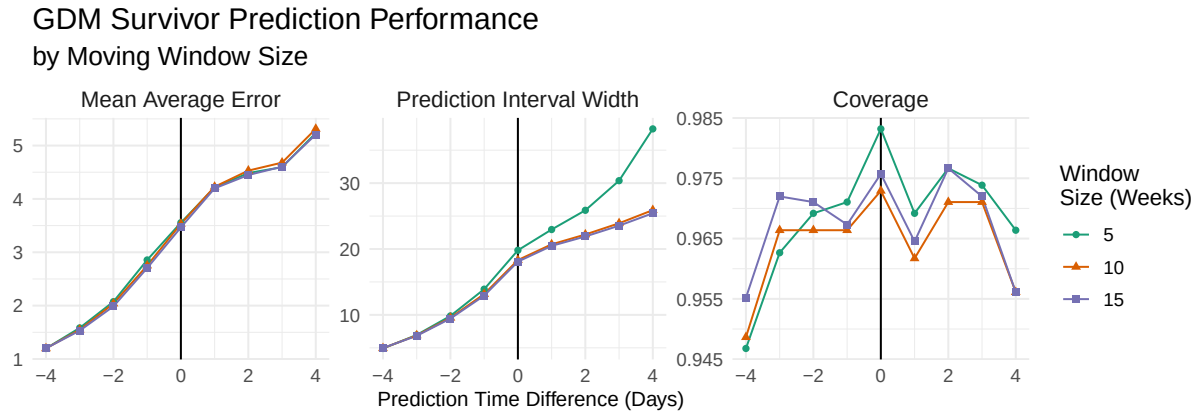

Figure 4.3: Mean average errors (left), mean 95% prediction interval widths (centre), and 95% prediction interval coverage values (right) for predicted daily COVID-19 deaths from GDM Survivor models with different moving window sizes (weeks). Performance metrics are arranged on the x-axis by prediction time difference, from -4 days up to +4 days, and different moving window sizes are represented by different colours and shapes. ***Source***: Appendix E Stoner, Halliday and Economou (2022).

In conclusion, we believe a moving window size of $W = 70$ is a reasonable choice for this data, as increasing the window size substantially increased daily computation time with no clear gains in prediction accuracy or precision. Smaller values of W could be tested to identify a window size with similar precision and even faster computation times.

4.3 Simulation Experiment

Alongside Stoner, Halliday and Economou (2022) we included a simulation experiment in the supplementary materials to investigate the ability of the GDM to provide insights into factors determining the structure of the reporting delay and variability in the total count. We tested this by simulating data such that inference can be compared against the known effect of covariates. What follows in the remainder of this section is text from Web Appendix C of Stoner, Halliday and Economou (2022), which I contributed to jointly together with other authors.

4.3.1 Simulated data

The data simulated are completely synthetic and are intended to represent a hypothetical disease epidemic, with covariates imitating real-world drivers of change in the disease progression and reporting performance. Note that repeated simulation would be necessary to assess bias, variance, prediction interval coverage etc. Here we simulate one set of data as an example and we use the phrase "true value(s)" to mean the coefficient values chosen in this one example.

Let $y_{t,s}$ be the number of disease cases occurring on day $t \in 1,\ldots,100$ in region $s \in 1,2,3$, and let $z_{t,d,s}$ be the parts of $y_{t,s}$ observed at each delay $d \in 1,\ldots,D_{max}$. The experiment focuses on the quality of inference rather than prediction performance, so we assume all $y_{t,s}$ and $z_{t,d,s}$ are known. First, we simulate $y_{t,s}$ from a Negative-Binomial model with mean $\lambda_{t,s}$ and scale parameter θ_s:

$$y_{t,s} \mid \lambda_{t,s}, \theta_s \sim \text{Negative-Binomial}(\lambda_{t,s}, \theta_s); \tag{4.2}$$

$$\log(\lambda_{t,s}) = \iota_s + \delta_{t,s} + \alpha_1 V_{t,s}^{(*)} + \alpha_2 W_{t,s}^{(*)} \tag{4.3}$$

The mean number of cases is comprised first of a fourth-order orthogonal polynomials in time $\delta_{t,s}$. The polynomials, shown in the left panel of Figure 4.4, are distinct for each region to reflect differences in disease progression and non-pharmaceutical interventions. In all three regions, the simulated polynomials have an "M" shape, representing two waves of the disease. The mean is also affected by the proportion of the population successfully administered a vaccine in each region, $V_{t,s}$, shown in the right panel of Figure 4.4. The vaccine becomes available after 40 days, and is then administered at a different rate in each region. After scaling to have 0 mean and standard deviation 1 (let $V_{t,s}^{(*)}$ be the scaled version), the vaccination covariate has coefficient $\alpha_1 = -1.75$, meaning the case rate reduces substantially as more people are vaccinated.

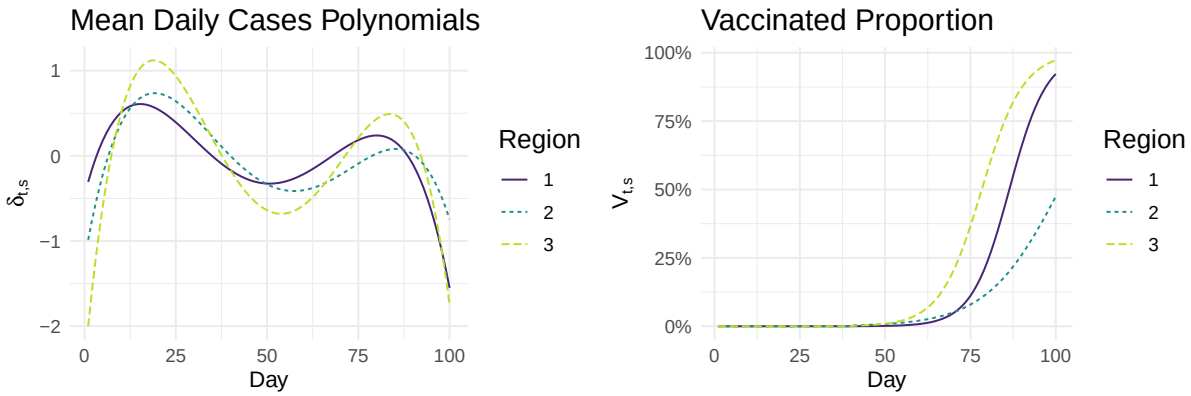

Figure 4.4: Left: simulated polynomial trends $\delta_{t,s}$ in the mean daily cases. Right: time series of the simulated percentages of the population administered with a vaccine $V_{t,s}$. ***Source***: Appendix C Stoner, Halliday and Economou (2022).

Another effect on the mean number of cases of our simulated disease is the relative prevalance of two different variants of the disease in question, how these variants progress over time is illustrated in Figure 4.5. Initially, only Variant 1 is present in the population but then Variant 2 begins to spread at a later time which is dependent on the region. The rate of growth of each variant is also region dependent. Let $W_{t,s}$ be the percentage prevalence of Variant 2, taking values between 0 and 1, and let $W_{t,s}^{(*)}$ be the scaled version. Its associated coefficient is $\alpha_2 = 0.5$, meaning Variant 2 causes more cases than Variant 1 as in a real world context would be considered more infectious. The coefficients of $V_{t,s}^{(*)}$ and $W_{t,s}^{(*)}$ are both assumed constant across regions, meaning we assume the protection of the vaccine and infectious properties of the new variant to be the same across regions.

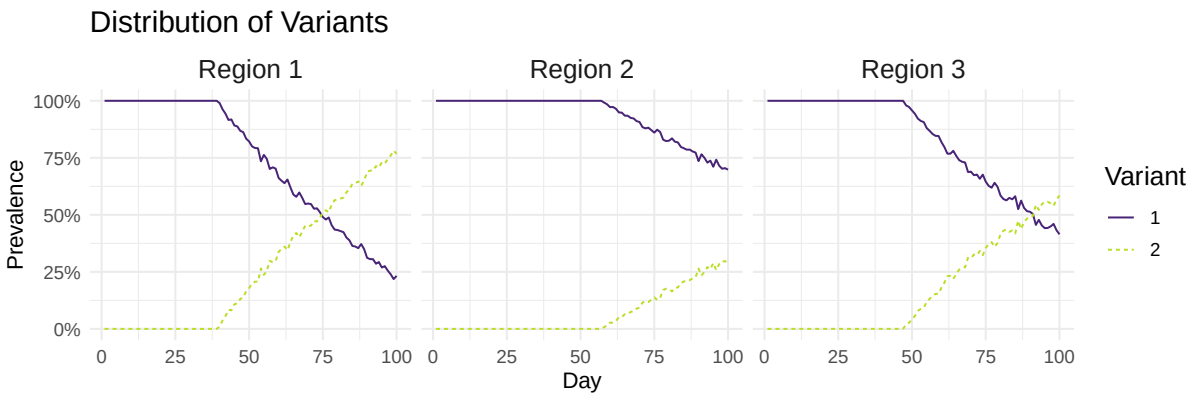

Figure 4.5: Time series of the simulated relative prevalence of two disease variants in the three regions. ***Source***: Appendix C Stoner, Halliday and Economou (2022).

Combining regional intercept terms ι_s, polynomials $\delta_{t,s}$, the effect of vaccination, and the effect prevalence of variant 2, Figure 4.6 shows the overall mean daily cases for each region (lines) and then the simulated daily cases $y_{t,s}$ (shapes). The M shape from the polynomials is still visible, but dampened by the effect of the vaccination covariate.

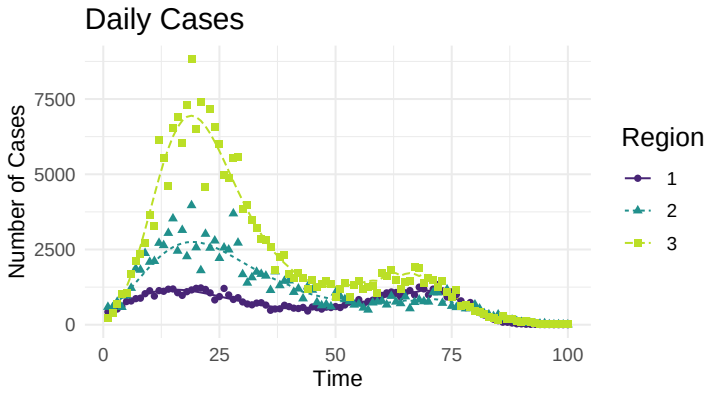

Figure 4.6: Times series of the mean daily cases $\lambda_{t,s}$ (lines) and daily cases $y_{t,s}$ simulated from the Negative-Binomial (shapes). ***Source***: Appendix C Stoner, Halliday and Economou (2022).

Now, it remains to simulate $z_{t,d,s}$: the parts of $y_{t,s}$ reported at each delay. Here, we simulate the first $D = 7 < D_{max}$ delays for modelling. Since we assume all $z_{t,d,s}$ are observed, the model will be able to determine the remainder term for any $D < d \leq D_{max}$ without explicitly modelling it's delay structure. To define the mean reporting distribution, we start with a probit model for the expected cumulative proportion reported $S_{t,d,s}$, as in the Survivor variant of the GDM (described in Section 3 of the main article):

$$\text{probit}(S_{t,d,s}) = \psi_{s,d} + \beta_{1,s} X_t^{(*)} + \beta_{2,s} A_{t,s}^{(*)} + \beta_{3,s} \lambda_{t,s}^{(*)}. \tag{4.4}$$

Temporal variation in $S_{t,d,s}$ is characterised first by monotonically increasing "delay curve" effects $\psi_{s,d}$ and upwards linear trends in (scaled) time $X_t^{(*)}$. We also included staff absence percentage covariates $A_{t,s}$ (Figure 4.7), simulated as first order auto-regressive variables. The coefficient for scaled time is given a positive value $\beta_{1,s} = [0.2, 0.15, 0.1]$ for the three simulated regions respectively, reflecting a hypothetical scenario where reporting efficiency

improved over time at different rates for each region. The coefficients of the scaled absence percentage $A^{(*)}_{t,s}$, are randomly generated for each region by $\beta_{2,s} \sim \text{Normal}(-0.2, 0.025^2)$, where they are intuitively negative so that a higher percentage of absences slows down reporting.

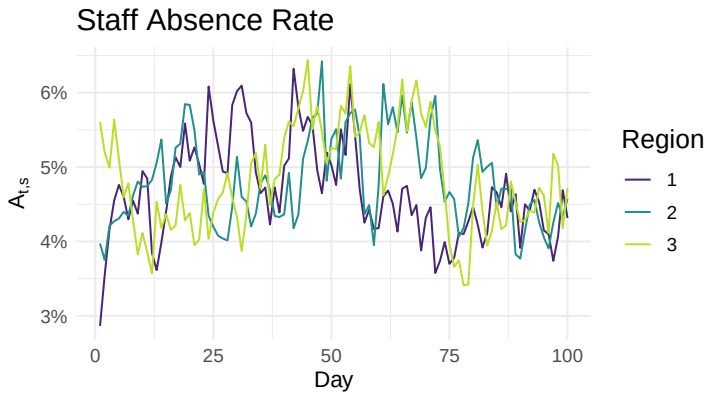

Figure 4.7: Time series of simulated staff absence percentage covariate $A_{t,s}$ for each of the three regions. ***Source***: Appendix C Stoner, Halliday and Economou (2022).

Finally, we have also included the scaled mean daily number of cases $\lambda^{(*)}_{t,s}$ as a covariate in the cumulative proportion reported. The coefficients of this effect, $\beta_{3,s} \sim \text{Normal}(-0.35, 0.05^2)$, are also randomly generated for each region to be negative such that, when cases are high, reporting slows down. Inclusion of this covariate is notable because: a) to the best of our knowledge, this is the first attempt (albeit for simulated data) to fit a model for delayed reporting which considers the effect of the same covariates on both the mean total cases/deaths and the reporting delay; and b) effectively including the same covariates in both parts of the GDM hierarchy is more likely to cause inferential problems such as non-identifiability of the covariate coefficients. Combining all of the effects in Equation (4.4) yields the mean cumulative proportion reported at each delay $S_{t,d,s}$, as shown in Figure 4.8. Notably, reporting performance visibly slows down in the first 30 days or so as cases surge (as seen in Figure 4.6). One can imagine this reflecting health systems prioritising treatment over administrative tasks when cases are high.

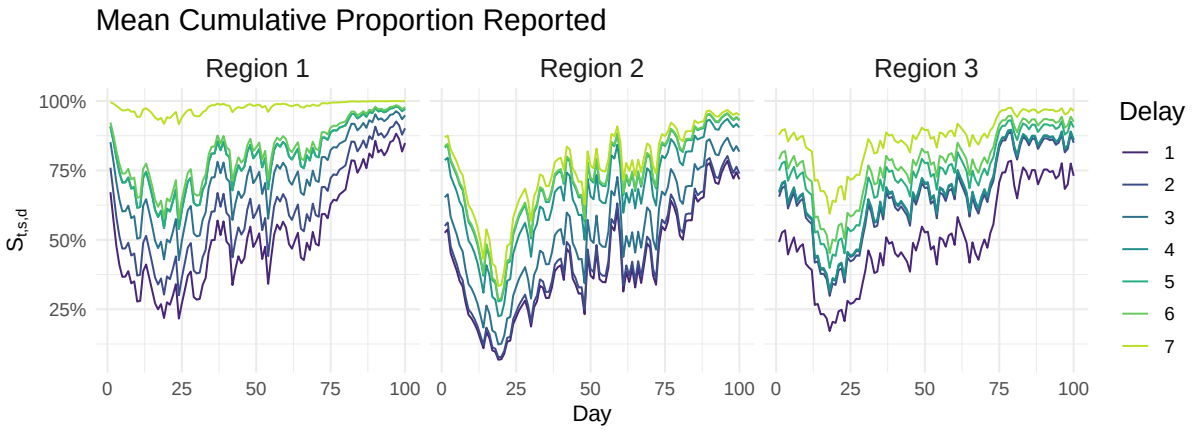

Figure 4.8: Time series of the mean cumulative proportion of cases reported, for each of the three regions, by delay. ***Source***: Appendix C Stoner, Halliday and Economou (2022).

Using these simulated $S_{t,d,s}$, we then simulated two different sets of delayed counts $z_{t,d,s}$, to be fitted separately. We simulated the first set, $z^{(1)}_{t,d,s}$, from the Generalised-Dirichlet Multinomial (GDM) itself:

$$\boldsymbol{z}^{(1)}_{t,1:D,s} \mid y_{t,s}, \boldsymbol{\nu}_{t,s}, \boldsymbol{\phi}_s \sim \text{GDM}(\boldsymbol{\nu}_{t,s}, \boldsymbol{\phi}_s, y_{t,s}) \tag{4.5}$$

$$\boldsymbol{\nu}_{t,d,s} = (S_{t,d,s} - S_{t,d-1,s})/(1 - S_{t,d-1,s}). \tag{4.6}$$

This allows us to assess inference when fitting the model to data generated from it. Then, we also simulated a second set, $z^{(2)}_{t,d,s}$, from a series of Binomial-Gaussian mixtures. First, we transform the mean proportion reported at each delay $p_{t,d,s} = S_{t,d,s} - S_{t,d-1,s}$ into a new set of real number values $u_{t,d,s} \in (-\infty, \infty)$, using a Center Log-ratio (CLR) transformation:

$$u_{t,d,s} = \log\left(\frac{p_{t,d,s}}{\prod_{j=1}^{D_{max}} (p_{t,j,s})^{\frac{1}{D_{max}}}}\right). \tag{4.7}$$

We then add an i.i.d. Gaussian noise term $\varepsilon_{t,d,s} \sim \text{Normal}(0, 0.25^2)$ to each $u_{t,d,s}$ ($\tilde{u}_{t,d,s} = u_{t,d,s} + \varepsilon_{t,d,s}$) and input this into an inverse CLR transformation to produce a new noisier set of proportions $p^{(2)}_{t,d,s}$,

$$p^{(2)}_{t,d,s} = clr^{-1}(\tilde{u}_{t,d,s}) = \frac{\exp \tilde{u}_{t,d,s}}{\sum_{j=1}^{D} \exp \tilde{u}_{t,j,s}} \tag{4.8}$$

A standard deviation of 0.25 was chosen for $\varepsilon_{t,d,s}$ so that the variance of $z^{(2)}$ is fairly close to the variance of $z^{(1)}$. From these new (noisier) proportions reported at each delay, we can compute relative proportions $\nu^{(2)}_{t,d,s}$:

$$\nu^{(2)}_{t,d,s} = \frac{p^{(2)}_{t,d,s}}{1 - \sum_{j=1}^{d-1} p^{(2)}_{t,j,s}}, \tag{4.9}$$

which can then be used as the mean of Binomial conditional distributions to simulate $z^{(2)}_{t,d,s}$:

$$z^{(2)}_{t,d,s} \sim \text{Binomial}\left(\nu^{(2)}_{t,d,s}, y_{t,s} - \sum_{j=1}^{d-1} z^{(2)}_{t,j,s}\right). \tag{4.10}$$

Fitting a GDM model to $z^{(2)}_{t,d,s}$ then allows us to test the flexibility of the GDM to capture non-GDM variance structures, through posterior predictive checking Gelman et al. (2013).

4.3.2 Results

We fit the GDM model outlined in Equations (4.2)–(4.6) to the simulated two data sets. We use the NIMBLE package de Valpine et al. (2017) for MCMC and weakly-informative prior distributions (e.g. Normal$(0, 10^2)$ for coefficients), as in the main article.

A key question of interest is whether the known trends and covariate effects in the mean daily cases and in the reporting delay are appropriately captured by the model. To assess this, we should examine outputs from the model fit to $z^{(1)}_{t,d,s}$. We can first look at the polynomial trends in the mean daily cases. Figure 4.9 shows the posterior median estimates of $\delta_{t,s}$ with 95% credible intervals (CIs). The polynomials are reproduced very closely, with the true values (dashed lines) captured completely by the 95% CIs.

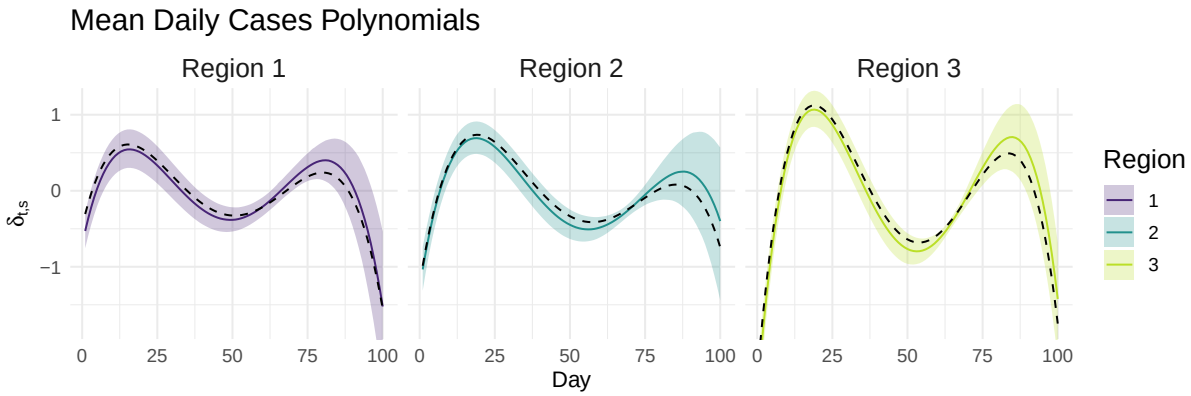

Figure 4.9: Posterior medians (solid lines) and 95% credible intervals (shaded areas) of the polynomial trends $\delta_{t,s}$ in the mean daily cases. True values are shown as dashed lines. ***Source***: Appendix C Stoner, Halliday and Economou (2022).

Next, we can look at the coefficients in the mean daily cases for the vaccine (α_1) and for the prevalence of Variant 2 (α_2). Figure 4.10 shows density estimates of the posterior samples of these two coefficients, with vertical lines representing the true values. In these plots, we are looking to see whether the true values are extreme with respect to their corresponding posterior distributions. Here, the true values for both coefficients are towards the centre of the distributions, indicating the models captures the effect of these covariates well. The model over predicts both the vaccine and variant effect on the mean daily cases as these cancel out; over predicting the vaccine effect results in lower mean daily cases, and over predicting the variant effect result in higher mean daily cases being predicted. Hence, in this simulation study there is not enough data for the model to accurately separate these two confounding effects.

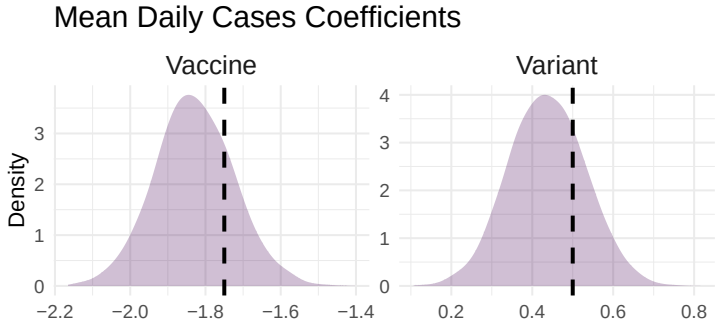

Figure 4.10: Posterior densities for the vaccine (left) and variant (right) coefficients (α_1 and α_2) in the mean daily cases, with dashed lines showing the true values. ***Source***: Appendix C Stoner, Halliday and Economou (2022).

Then, Figure 4.11 shows the posterior densities for the cumulative proportion reported coefficients of time ($\beta_{1,s}$), staff absence ($\beta_{2,s}$), and the daily case rate ($\beta_{3,s}$), respectively. In most cases, the true values are well within the bulk of the distributions, and after computing 95% CIs we determined that they all contain their corresponding true values. The model was fitted to each simulated "region" independently. For region 2 the coefficients for the scaled time variable and for staff absence are both under-predicted, where as for region 3 they are both over predicted. There may be some cancellation between these two coefficients as staff absences are generated using a first-order auto-regressive variable over time. However, for both, the model does accurately capture that the coefficient for the linear temporal trend is positive and for the staff absences is negative.

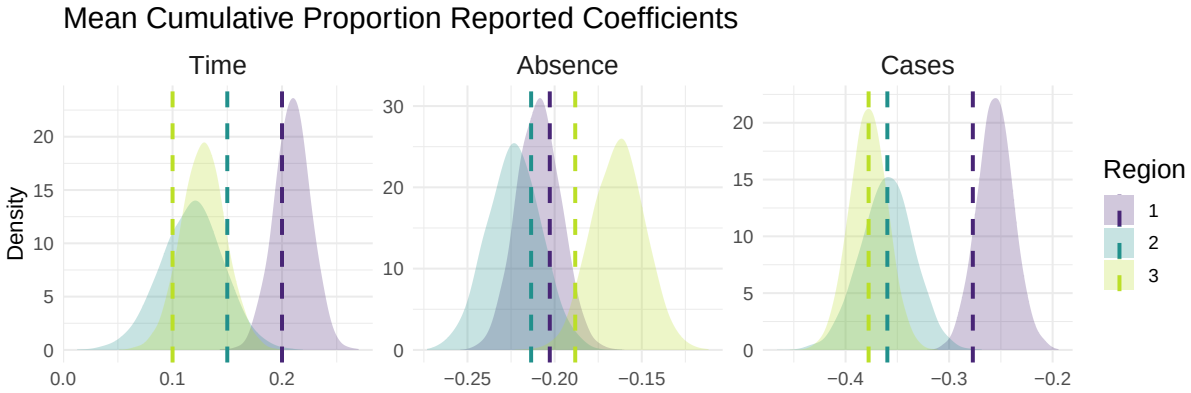

Figure 4.11: Posterior densities for the time (left), staff absence (centre) and variant (right) coefficients in the mean cumulative proportion reported, with dashed lines showing the true values and colours representing each of the three regions. ***Source***: Appendix C Stoner, Halliday and Economou (2022).

Finally, we can assess whether the GDM can appropriately capture the variance of the second set of delay counts $z^{(2)}_{t,d,s}$. Recall that these were simulated from Binomial-Gaussian mixture distributions, rather than a GDM. To achieve this, we use the MCMC samples to simulate posterior predictive replicates of the original set of $z^{(2)}_{t,d,s}$, let's call these replicates $\tilde{z}^{(2)}_{t,d,s}$. This results in one set of $\tilde{z}^{(2)}_{t,d,s}$ per saved MCMC iteration. We then compute the sample standard deviation of each set of $\tilde{z}^{(2)}_{t,d,s}$ for each region s and delay d. This results in a distribution of sample standard deviations for each s and d, which we can compare against the corresponding true values from the original $z^{(2)}_{t,d,s}$. Figure 4.12 shows the posterior predictive sample standard deviations for the first six delays. The true values are all within the bulk of the replicate distributions, indicating the model has captured the variances of the delayed counts well despite being generated from a different model.

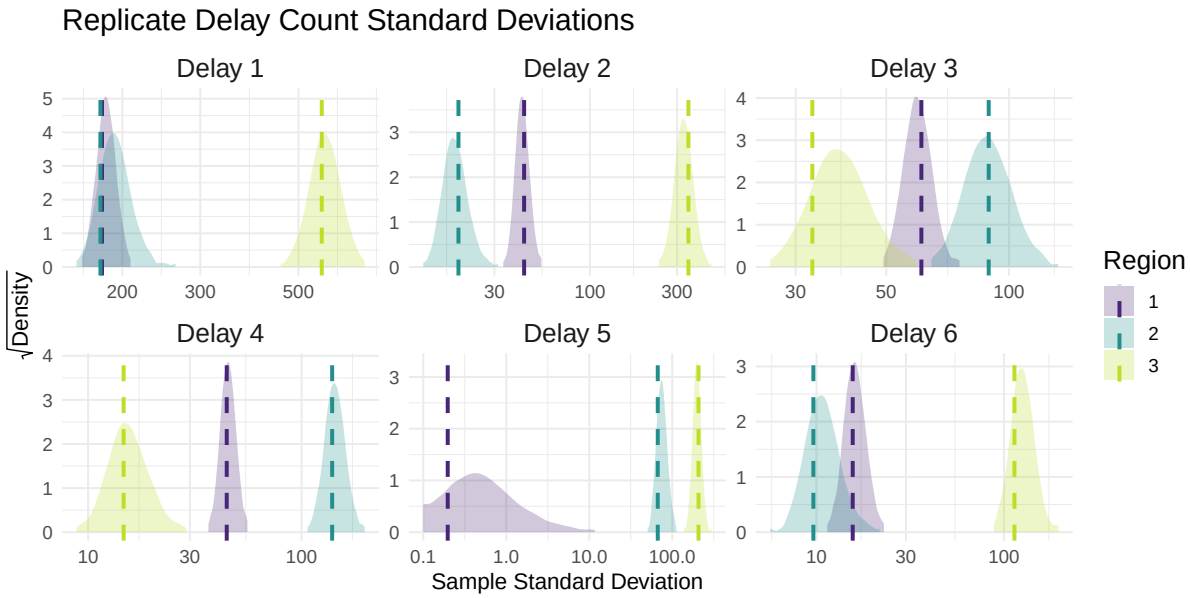

Figure 4.12: Density estimates of the posterior predictive sample standard deviations of the non-GDM delay counts ($z^{(2)}_{t,d,s}$), for the first six delays and each of the three regions. Dashed lines show the true values. A log-10 transformation was used for the x-axis, and a square root transformation was used for the y-axis. ***Source***: Appendix C Stoner, Halliday and Economou (2022).

4.3.3 Conclusions

Here we simulated a data set of daily disease case counts in three regions, with covariates imitating real-world drivers of disease (vaccination, proliferation of different variants) and reporting delays (staff absences, pressure from high case rates). Investigating the latter effect is particularly unusual, because it means we tested a model design that effectively included the same covariates in both the model for the total counts and the model for the reporting delay. Despite this, the GDM was able to reproduce all of the known covariate effects well.

We also tested the fit of the GDM to delay counts $z^{(2)}_{t,d,s}$ simulated from a different model, in this case a Binomial-Gaussian mixture. Compellingly, the flexibility of the GDM meant that it was able to capture the variance of these alternative counts very closely.

Chapter 5

Framework for Nested Disease Structures with Delayed Reporting

5.1 Introduction

Delayed reporting is treated as a specific modelling challenge in the existing literature. However, from a general statistical perspective, correcting reporting delays can be framed as a compositional time series modelling (of count data) challenge. As outlined in Section 1.1, data are structured as a total count (per time step) which is conceptually a sum of partial counts, each reported some time after they occurred. Besides the disease data applications we focus on for this thesis, delayed reporting is a common data problem affecting several branches of society, including insurance claims (Lawless (1994)), the reporting of crimes (Wesselbaum (2023)) and the monitoring of migration and asylum data (Singleton (2016)). In general, compositional time series count data is challenging to model due to a) the partial counts being constrained to sum to the total, b) complex correlation structures between the partial counts, and c) having to account for temporal trends and variability in the data appropriately (as we discuss in Chapter 1, and recall in Section 5.2). A further complication in the case of delayed reporting is that recent total counts have not yet been observed, prohibiting the use of "off-the-shelf" statistical methods for compositional data.

For this chapter, our motivation is still the issue of disease surveillance, where the total counts refer to the total number of observable occurrences of a health outcome (infections, fatalities, etc.) within a particular time period (e.g. one week). The partial count is then the number of these total cases that were reported with a certain temporal delay. For instance, the total count might be 50, of which 20 were reported during the week they occurred, 10 of which were reported a week later and so on until all 50 are reported. Disease surveillance involves mitigating the uncertainty introduced by the nature of delayed reporting, as we discussed in Section 1.1.

The aim here is develop a novel approach for the application of jointly monitoring the severe acute respiratory illness (SARI) outbreaks in Brazil alongside the severe COVID-19 positive patients within this group. SARI is an example of a disease requiring ongoing surveillance in Brazil. A SARI case is defined as a hospitalised individual with both a fever and cough, where the symptoms have onset within the last 10 days. SARI can be caused by several respiratory viruses, e.g. severe acute respiratory syndrome (SARS) caused by the corona-virus SARS-CoV, or COVID-19 caused by the corona-virus SARS-CoV-2. In Brazil, reporting of SARI can be delayed, often by several weeks, due to local municipality health authorities awaiting reports of individual SARI cases from hospitals.

Previous work aiming to nowcast SARI cases involve flexible statistical models that account for reporting delays explicitly. Notably, the framework in Bastos et al. (2019) is being used operationally by the Fiocruz institute in Brazil as a surveillance system for SARI (InfoGripe). In addition, a group of independent Brazilian scientists are performing ongoing surveillance for severe COVID-19 and SARI separately (Observatório COVID-19 BR (2024)), using the modelling framework by McGough et al. (2020). The underlying statistical frameworks used for these surveillance systems are reviewed alongside further existing approaches in Section 2.2.

SARI hospitalisations in Brazil are published as open data by Ministry of Health Brazil (2022). The data involves individual SARI cases, which are typically reported after some delay. Where patients are tested for COVID-19, the test result(s) appear later on, usually after a further delay. Here we aim to jointly model SARI hospitalisations as well as the nested portion of these hospitalisations that are COVID-19 positive. In relation to the current operational system, our goal is to improve the nowcasting of SARI using more flexible statistical methods, and to simultaneously produce more accurate predictions of severe COVID-19 through pooling of information with the SARI case counts.

This work aims to address the novel challenge of the length of delays associated with COVID-19 lab results being unknown. Patients records are updated manually with no time-stamp for when the confirmed lab result was recorded. Hence, there is no historical information in the data to inform modelling of trends in the delay distribution. However, we can deduce that the delays in COVID-19 reporting will be greater than the delays for SARI reporting – as they are subject to the same reporting processes but with the additional delay of waiting for the lab test results. We take this attribute into account in the design of our joint model for both types of hospitalisation.

By framing this application of Brazilian SARI hospitalisations as a compositional count data problem in Section 5.2.1, we aim to highlight our contribution of new methodological insights and extensions for general applications with similar data structures. This includes modelling nested structures with potentially unknown reporting delay lengths.

In this chapter, we extend the Generalised-Dirichlet Multinomial (GDM) framework to address a real-world disease surveillance problem faced by the Oswaldo Cruz Foundation (Fiocruz) in Brazil, which we introduce in Section 5.1. The modelling framework presented in this chapter represents a substantial extension of the GDM statistical framework. Here, its focal point is a framework for correcting reporting delays, but in Section 5.2.1 we present this as a single application in the context of general compositional data problems. In Section 5.2.2, we overview the GDM framework before extending it to our proposed modelling framework in Section 5.3. Our proposed framework aims to allow for multiple and nested disease occurrences (Section 5.3.1) that are subject to unknown reporting delays (Section 5.3.2). Furthermore, we expand the choice of model formulation for the delay structure to allow a more intuitive representation over time and delay (Section 5.3.3). In Section 5.3.4 we include age demographic covariate effects to improve nowcasting accuracy for severe COVID-19. In Section 5.4, we evaluate model performance through a series of rolling out-of-sample prediction experiments for federative-unit level data in Brazil. Finally, we summarise our findings and discuss possible future work in Section 5.5.

5.2 Background

We begin by recalling, from Section 1.1, some notation to describe the general structure of count data subject to delayed reporting so we can then extend this to include a nested count of the observed totals. First, we call the true number of counts (e.g. disease cases) occurring within time period t and region s the "total counts" and denote these as $y_{t,s}$. These total counts can be broken down into partial counts, $z_{t,d,s}$, reported at different delay intervals d – e.g. d could be the number of weeks since occurrence, with $d = 0$ denoting "no delay". Therefore, we can sum the partial counts over all possible delays to obtain the total count $y_{t,s} = \sum_{d=0}^{D_{max}} z_{t,d,s}$, where D_{max} is assumed to be the maximum delay. Together, the mean, variance, and covariance of the partial counts $z_{t,d,s}$ can be thought of as properties of the "delay distribution". In our data from Brazil, $y_{t,s}$ are the SARI hospitalisations occurring in week t and federative unit s, while $x_{t,s}$ are the nested portion of $y_{t,s}$ that test positive for COVID-19.

In Section 1.1, we highlight that the delay distribution may be subject to systematic variation over time and space due to changes in reporting efficiency and resources. For example, Bastos et al. (2019) note that for Brazil, reporting delays can be greatly impacted by fluctuations in hospital staff absence and workload over the course of an outbreak. Conversely, interventions and awareness associated with the progression of an infectious disease could support local reporting procedures. Similarly, the temporal evolution of the total counts $y_{t,s}$ will consist of both systematic and random variability, both of which need to be accounted for. Systematic variability in the prevalence of the disease could occur over spatial regions due to differing population structures, and movement between regions by infected individuals. Additionally, systematic variability will occur over time due to the natural spread or mutation of the disease, and possible seasonal effects.

In terms of viewing this problem as a modular framework for delayed reporting, introduced in Section 1.2, we first wish to model the epidemic process Y which generates a quantity y, which in this case is the total number of SARI hospitalisations:

$$Y(\Theta) \to y. \tag{5.1}$$

Where Θ is a set of random effects and/or model parameters. The counts subject to delayed reporting are then represented by quantity z, which is generated by the flawed reporting process Z with model parameters Π:

$$Y(\Theta) \to y \to Z(\Pi) \to z. \tag{5.2}$$

Parallel to this, we also generate a quantity x for the nested number of SARI hospitalisations that are COVID-19 positive, with X representing the process of testing SARI cases to confirm the COVID-19 virus and Λ representing the model parameters:

$$Y(\Theta) \to y \to X(\Lambda) \to x. \tag{5.3}$$

A secondary process then gives us $\widetilde{x}$, the COVID-positive SARI cases that are observed due to flawed process of reporting the confirmed COVID-19 tests W, which we capture with model parameters Ω:

$$Y(\Theta) \to y \to X(\Lambda) \to x \to W(\Omega) \to \widetilde{x}. \tag{5.4}$$

Hence, there is a need for a flexible modelling framework that can be designed to replicate these hierarchical, nested and parallel processes.

Capturing all these sources of variability in the data is vital for optimal nowcasting performance but requires models with appropriate flexibility, as discussed in Section 1.2. In Section 2.2, we reviewed existing methods for correcting delayed reporting that aim to achieve this. We focus on approaches within the Bayesian framework to allow for full

predictive distributions for $y_{t,s}$ and/or $x_{t,s}$ given all the available data. This quantification of predictive uncertainty is important in the context of infectious disease surveillance, so that the risks associated with different incidence levels of the disease can be considered and prepared for.

5.2.1 Delayed reporting as a compositional time series problem

Compositional data is defined as a set of non-negative parts that sum to a total. Examples of such data include: voting intention polls that measure the percentage of people intending to vote for different political parties (Vos (1998)); household surveys that capture the percentage of people using different fuels as their main energy source for cooking (Stoner et al. (2020)); or forensic data capturing the elemental composition of fragments at crime scenes (Campbell et al. (2009)). Models for such data must account for both the sum-to-a-total and the non-negativity constraint. Log-ratio transformations are commonly used to map continuous compositional data onto the real space, allowing the use of standard statistical techniques. An added challenge arises when either there are counts equal to 0 such that the log ratios are undefined, or when counts are small (e.g. rare diseases). In the latter case, transformed data will be strongly bounded and non-smooth so assuming a continuous probability model (e.g. Normal) would be inappropriate. Here we instead focus on probability models for the original count structure.

Another research area that involves modelling compositional counts is the analysis of the make-up of microbiomes. Here, one common choice is the Dirichlet multinomial (DM) model (Chen and Li (2013), Koslovsky (2023)), which describes the counts directly while also allowing extra variance relative to the multinomial. However, the DM restricts all cross-correlations to be negative (Stoner and Economou (2019)). A more flexible option is the generalised Dirichlet multinomial (GDM), proposed for microbiome modelling in Tang and Chen (2018), as it has a more general correlation structure.

A compelling alternative is the logistic normal multinomial (LNM) proposed in Xia et al. (2013). This assumes a Multinomial($\boldsymbol{p}_{t,s}, y_{t,s}$) model for $\boldsymbol{z}_{t,s} \mid \boldsymbol{p}_{t,s}, y_{t,s}$, where the multinomial probabilities $\boldsymbol{p}_{t,s}$ are linked to a vector $\boldsymbol{q}_{t,s} \sim$ Multivariate-Normal($\boldsymbol{\mu}_{t,s}, \boldsymbol{\Sigma}$) through an additive log-ratio (ALR) transform link function,

$$\boldsymbol{q}_{t,s} = \text{ALR}(\boldsymbol{p}_{t,s}) = \ln\left(\frac{p_{t,1,s}}{p_{t,D_{max},s}}\right), ..., \ln\left(\frac{p_{t,D_{max}-1,s}}{p_{t,D_{max},s}}\right). \tag{5.5}$$

This model is arguably more flexible than the GDM since the covariance matrix $\boldsymbol{\Sigma}$ can allow for general covariance structures in $\boldsymbol{z}_{t,s}$. However, computing the likelihood of $\boldsymbol{z}_{t,s}$ and inferring the whole covariance matrix $\boldsymbol{\Sigma}$ in the LNM is challenging (the number of elements in $\boldsymbol{\Sigma}$ grows with the square of the number of compositions (Zhang and Lin (2019))). In the GDM, the number of parameters scales linearly with the number of compositions and the likelihood is tractable both in its full joint form or as a series of Beta-Binomial conditional likelihoods, the latter of which is convenient when there are missing values in $\boldsymbol{z}_{t,s}$. To our knowledge, the LNM has only been explored in the context of known total counts $y_{t,s}$, and not within a joint hierarchical model for both $y_{t,s}$ and $\boldsymbol{z}_{t,s}$.

5.2.2 The Generalized-Dirichlet Multinomial method

Viewing delayed reporting as a challenge of modelling compositional time series of counts motivates an approach using the Generalized-Dirichlet Multinomial (GDM) family of distributions, which we review in Chapter 2. Stoner and Economou (2019) introduce the GDM as a way to model the partial counts $z_{t,d,s}$ conditional on a Negative-Binomial model for total counts $y_{t,s}$, to flexibly capture the different sources of random and systematic variability in the data:

$$y_{t,s}|\boldsymbol{\lambda}_{t,s}, \boldsymbol{\theta}_s \sim \text{Negative-Binomial}(\boldsymbol{\lambda}_{t,s}, \boldsymbol{\theta}_s); \qquad \log(\boldsymbol{\lambda}_{t,s}) = f(t,s), \tag{5.6}$$

$$\boldsymbol{z}_{t,s}|\boldsymbol{\nu}_{t,s}, y_{t,s}, \boldsymbol{\phi}_s \sim \text{GDM}(\boldsymbol{\nu}_{t,s}, y_{t,s}, \boldsymbol{\phi}_s). \tag{5.7}$$

The GDM is a conditional Multinomial($\boldsymbol{p}_{t,s}, y_{t,s}$) distribution, where the mean proportion of $y_{t,s}$ reported at each delay is modeled by $\boldsymbol{p}_{t,s} \sim$ Generalized-Dirichlet($\boldsymbol{\nu}_{t,s}, \boldsymbol{\phi}_s$). As discussed in Section 2.3, this increases the number of parameters of the model (compared to a Multinomial model) such that the mean, variance and covariance of the partial counts are no longer all determined by just $\boldsymbol{p}_{t,s}$. Stoner and Economou (2019) argue that this additional flexibility improves predictive performance both theoretically and in practice.

An alternative representation of the GDM is as a series of Beta-Binomial distributions for the partial counts $z_{t,d,s}$, conditional on the already observed partial counts $z_{t,1,s}, .., z_{t,d-1,s}$ and the total counts $y_{t,s}$ given by $z_{t,d,s}|\nu_{t,d,s}, \phi_{s,d}, N_{t,d,s} \sim$ Beta-Binomial($\nu_{t,d,s}, \phi_{s,d}, N_{t,d,s}$). Here, $N_{t,d,s} = y_{t,s} - \sum_{j=1}^{d-1} z_{t,s,j}$ are the remaining counts yet to be reported at delay d. This formulation is used in practice for an intuitive way of introducing covariates and temporal/spatial structures in the delay, and as a more straightforward implementation using MCMC. Here, variance parameters $\phi_{s,d} > 0$ inflate the variance of each Beta-Binomial component in turn (compared to Binomial models, which the Beta-Binomial reduces to as $\phi_{s,d} \to \infty$), allowing for a flexible covariance structure in $\boldsymbol{z}_{t,s}$.

The resulting expected relative proportions are then defined as $\nu_{t,d,s} = E\left[\frac{z_{t,d,s}}{N_{t,d,s}}\right]$. These are the number of counts expected to be reported at delay d divided by the as of yet unreported counts. The variance parameter can in principle also be modeled as a general function of time, space, and delay $\log(\phi_{t,d,s}) = h(t,d,s)$.

Stoner, Halliday and Economou (2022) (Chapter 4) presents an extensive comparison of the predictive performance of the GDM and the models presented in Bastos et al. (2019) and McGough et al. (2020) through a 15-month rolling prediction experiment on COVID-19 deaths in England, which we summarised in Section 4.1. The results demonstrated the separation and modelling of all sources of variability in the GDM model made it capable of accurately capturing all variability in the data and therefore allowing for better predictions with reduced uncertainty.

On this basis, and the challenges discussed in Section 5.2, the GDM framework is the necessary choice of foundation to build on for our joint model for related diseases (in this case SARI and COVID-19). Additionally, it has been proposed as an operationally feasible framework in Chapter 4, following improvement to its computational efficiency in Chapter 3. Also, we develop an alternative variation of the GDM which aims to bring together strengths and mitigate the weaknesses of the two previously studied variants of the GDM, which differ in the choice of link function for the relative means $\nu_{t,d,s}$ of the Beta-Binomial model, a choice that can have non-trivial impacts on predictive performance and implementation.

5.2.2.1 Link functions

For the GDM model, Stoner and Economou (2019) (Chapter 4) offer two alternative link functions. The "hazard" version:

$$\log\left(\frac{\nu_{t,d,s}}{1-\nu_{t,d,s}}\right) = i(t,d,s), \tag{5.8}$$

and the "survivor" version:

$$\nu_{t,d,s} = \frac{S_{t,d,s} - S_{t,d-1,s}}{1 - S_{t,d-1,s}}; \qquad \text{probit}(S_{t,d,s}) = i(t,d,s). \tag{5.9}$$

Here $S_{t,d,s} = E\left[\frac{\Sigma_{j=1}^{d} z_{t,j,s}}{y_{t,s}}\right]$ are the expected cumulative proportions defined by the cumulative number of counts reported up to delay d divided by the total number of counts for time t. In Equations (5.8) - (5.9), the general function $i(\cdot)$ is open to user choice and represents the systematic time-delay effects on the relative proportions. Possible options include random effects, covariates, Gaussian processes and auto-regressive terms. Stoner, Halliday and Economou (2022) (Chapter 4) opted for hierarchically-structured penalised cubic splines of time with shrinkage.

The hazard version (Equation (5.8)) offers a relatively straightforward way to fit flexible models, as it uses a simple logistic transformation to directly model the relative proportions. However, conceptualising intuitive models for the relative proportions is challenging, especially for later delays when the small number of remaining cases to be reported mean the expected relative proportion could plausibly be very small or very large. At the same time, modelling the relative proportions directly limits options for structures that pool information across delays (e.g. penalised smooth time-delay interactions), to potentially result in a simpler model that is better at predicting out-of-sample. This is because we might reasonably expect the relative proportions to vary in a less smooth or otherwise less predictable way across d, compared to alternative representations of the mean proportions, e.g. the expected proportion for each d, $p_{t,d,s} = E\left[\frac{z_{t,d,s}}{y_{t,s}}\right]$.

Meanwhile, the survivor version (Equation (5.9)) is more intuitive with respect to conceptualising models for the cumulative proportions (e.g. by exploring the data). However, any terms in $i(\cdot)$ must be constrained to monotonically increase as d increases. Aside from potentially prohibiting some options for such effects (e.g. smooth time-delay interactions), the monotonic constraint may cause inefficiencies for MCMC sampling. The best choice for the link function in the GDM may be application dependent and therefore one should consider the alternative link functions carefully (Stoner, Halliday and Economou (2022) (Chapter 4)).

In the next Section, we present our modelling framework for correcting delayed reporting of two related diseases, as well as a novel choice of link function for the GDM (Section 5.3.3) – to compliment the existing "hazard" and "survivor" versions – relevant to both disease surveillance and hierarchical models for general compositional count time series.

5.3 General Framework

The SARI and COVID-positive SARI data in Brazil is collected in the following way, if a patient is hospitalised with SARI symptoms the medical team will fill out a form where the onset of symptoms is recorded, and a sample will be taken to test for COVID-19 or other viruses. The delay for SARI hospitalisations is measured as the time between the onset of symptoms and when these forms are digitalised and hence made available to the national notification system, which we refer to as the reporting date of SARI. One entry of the form is the test result of COVID-19 which could be positive, negative, inconclusive or not filled in. If the COVID-19 test result has been recorded in the form before it is digitalised then the COVID-19 reporting delay is technically the same length as the SARI reporting delay. Alternatively, the COVID-19 reporting delay will be greater than the SARI reporting delay if the digitalised form is manually updated at a later date. Hence, we know that the delay for COVID-19 test results is greater than or equal to the reporting delay of SARI hospitalisations. However, since there is no time stamp as to when the COVID-19 test result field is filled in, the delay length of the severe COVID-19 hospitalisations (which are a subset of the SARI hospitalisations) is unknown. Additionally, for the most recent dates, the number of positive COVID-19 tests are right censored as the true number of positive COVID-19 tests will be greater or equal to the amount that have been reported so far once further test results become available (for both SARI patients that have and haven't yet been reported).

We begin by assuming the same Negative-Binomial and GDM conditional model presented in Section 5.2.2. As explained in Section 5.1, our motivation for this framework is to improve nowcasting predictions for SARI hospitalisations (which have a known reporting delay), and the number of SARI hospitalisations that tested positive for COVID-19 (which has an unknown reporting delay), in Brazil. To the best of our knowledge, methods for correcting such "unmarked" reporting delays have not been previously studied, therefore we believe it is worthwhile addressing this challenge directly.

5.3.1 Model for nested structures

First, we model the mean SARI hospitalisations ($\lambda_{t,s}$) in each region s independently, using a temporal cubic regression spline with shrinkage $\zeta_t^{(s)}$,

$$\log(\lambda_{t,s}) = f(t,s) = \zeta_{0,s} + \zeta_t^{(s)}, \tag{5.10}$$

where $\zeta_{0,s}$ are regional intercepts. Here, shrinkage means that $\zeta_t^{(s)}$ is penalised by a single penalty parameter for both smoothness and overall magnitude.

To model the nested disease, we can introduce a new Beta-Binomial layer to our model hierarchy for $x_{t,s} \leq y_{t,s}$, which in this case is the number of SARI cases that tested positive for COVID-19. In general, $x_{t,s}$ could represent any systematic subset of a disease or family of diseases $y_{t,s}$, e.g. one strain/variant of a disease. This is given by:

$$x_{t,s}|y_{t,s} \sim \text{Beta-Binomial}(\mu_{t,s}, \chi_s, y_{t,s}); \qquad \text{logit}(\mu_{t,s}) = \beta_{0,s} + \beta_t^{(s)} + \delta_s \zeta_t^{(s)}. \tag{5.11}$$

In Equations (5.11), we capture systematic variability in $\mu_{t,s}$ through the addition of the spline for the mean SARI cases $\zeta_t^{(s)}$ (as included in Equation (5.10) for the linear predictor of $\log(\lambda_{t,s})$), scaled by a coefficient δ_s. An additional penalised cubic spline of time $\beta_t^{(s)}$ is also included, as well as an intercept terms $\beta_{0,s}$. Modelling $\mu_{t,s}$ in this way allows us to capture any link between the total number of SARI cases and the proportion testing positive for COVID-19, e.g. in the case where recent peaks in SARI cases are predominantly driven by COVID-19 "waves". We discuss the effect of including this link on prediction performance for unknown $x_{t,s}$ based on results from our rolling prediction experiment in Section 5.4.3. The Beta-Binomial dispersion parameter is given by χ_s, and the expected mean of $x_{t,s}$ is $\mu_{t,s} y_{t,s}$, where $\mu_{t,s}$ is the expected proportion of SARI counts $y_{t,s}$ that test positive for COVID-19. We choose the Beta-Binomial because it offers more flexibility relative to the Binomial, e.g. to account for unmeasured covariate effects.

This new layer allows us to potentially include covariate effects or trends specific to COVID-19, e.g. vaccination rates or the emergence of new COVID-19 variants. We exploit this in Section 5.3.4 by including covariates that indicate the age demographic of reported SARI cases. Moreover, modelling the relationship between $x_{t,s}$ and $y_{t,s}$ hierarchically, pools information across the two sets of case counts. This is particularly advantageous in this application as information on the SARI cases is substantially more timely than the COVID-19 test results, as discussed in Section 5.3.2.

5.3.2 Delayed reporting of available COVID-19 case counts

The COVID-19 positive counts $x_{t,s}$ are subject to the same reporting delay mechanism as the total SARI counts $y_{t,s}$ (since they are both derived from data for the same individuals), in addition to a secondary delay in obtaining COVID-19 test results. For the SARI cases, a reporting date is recorded in the data. However, there is no time-stamp in the data indicating when the COVID-19 test results for those cases were received. There is only information on the total reported so far, $\widetilde{x}_{t,s} \leq x_{t,s}$, which we can interpret as right-censored.

The problem can also be thought of as an under-reporting one, which can be addressed either by explicitly modelling the under-reporting mechanism (Stoner, Economou and Drummond Marques da Silva (2019) & Arima et al. (2023)) or by using a censored likelihood approach (Bailey et al. (2005)). The latter does not allow us to model the under-reporting in any structured way, or take into account realistic constraints in the design of this model (Stoner, Economou and Drummond Marques da Silva (2019)). E.g. for our COVID-positive SARI cases we might plausibly assume that the degree of under-reporting is worse on average for more recent weeks. Here, we opt to model the under-reporting through a further Beta-Binomial model with variance parameter υ_s and a reporting rate $0 < \pi_{t,s} < 1$:

$$\widetilde{x}_{t,s} | x_{t,s}, \pi_{t,s}, \upsilon_s \sim \text{Beta-Binomial}(\pi_{t,s}, \upsilon_s, x_{t,s}). \tag{5.12}$$

We assume a-priori, that $\pi_{t,s}$ decreases linearly at the log scale, at a rate parameterised via slope parameter $\omega_s < 0$. Specifically, we assume that the reporting rate decreases from 100%, from the historic time-point we believe under-reporting starts, M, to the present time point, T_{now}. So, $\log(\pi_{t,s}) = \omega_s(t-M)$ for $t > M$ and $\pi_{t,s} = 1$ otherwise. We obtain predictions of COVID-positive cases by treating not-yet-observed $x_{t,s}$ as unknown quantities in the model, yielding posterior predictive samples given all available data.

5.3.3 Extending the GDM framework

In Section 5.2.2, we discussed two link function options for including systematic variability in the relative proportions $\nu_{t,d,s}$, and their respective drawbacks. In this section we introduce an alternative link function for the absolute mean proportion $p_{t,d,s}$ of the total counts reported at each delay – arguably a more intuitive quantity compared to relative proportions. The purpose of this link function in the GDM model is to map $p_{t,d,s}$ from a vector of proportions summing to 1 onto $\mathbb{R}$, enabling the use of general functions to capture systematic trends.

For both the existing "hazard" and "survivor" link function options (Section 5.2.2), it can be challenging to specify intuitive and general time-delay interactions (or other interactions with delay e.g. space-delay) without running into constraints, including assuming independence across delay (hazard version) or monotonicity (survivor version). Here we propose the centre log ratio (CLR) transformation as a novel link function for the GDM, to directly model the mean proportions $p_{t,d,s}$:

$$\text{CLR}(\boldsymbol{p}_{t,1:(D+1),s}) = \left[\log\left(\frac{p_{t,1,s}}{g(\boldsymbol{p}_{t,s})}\right), ..., \log\left(\frac{p_{t,D+1,s}}{g(\boldsymbol{p}_{t,s})}\right)\right], \tag{5.13}$$

where $g(\boldsymbol{p}_{t,s})$ is the geometric mean of $\boldsymbol{p}_{t,s}$, D is the number of delays we will model explicitly, and $p(t, D+1, s)$ – which we call the "remainder" term – is the mean proportion of $y_{t,d,s}$ reported in the remaining delays $D < d \leq D_{max}$.

As with other parts of the framework, we first suggest that $CLR(p_{t,d,s})$ can be modeled by some general function of t, s, and $1 \leq d \leq D$, in this case $i(\cdot)$:

$$CLR(p_{t,d,s}) = i(t,d,s); \qquad p_{t,D+1,s} = -\sum_{j=1}^{D} p_{t,j,s}. \tag{5.14}$$

The second part of Equation (5.14) introduces a sum-to-zero constraint across $d,\ldots,D+1$, which is needed for $i(t,d,s)$, since the inverse CLR function is invariant to an additive constant (this comes from the original sum-to-one constraint for $p_{t,d,s}$).

We can then derive, by substitution, the relative proportions $\nu_{t,d,s}$ – needed for the Beta-Binomials – as $\nu_{t,d,s} = \exp(i(t,d,s))/\sum_{j=d}^{D+1}\exp(i(t,j,s))$. Unlike the relative proportions $\nu_{t,d,s}$, we might assume that $p_{t,d,s}$ is similar across adjacent delays, motivating structures such as 2D tensor product smooth functions across time and delay, penalised in both dimensions against over-fitting, potentially resulting in better predictive performance. The CLR could therefore be a compelling new option for the GDM method of correcting delayed-reporting, while also providing greater flexibility in model design for general compositional count data with arbitrary dimensions (or compositional proportions by simply using the Generalized-Dirichlet in place of the GDM).

In the following Sections, we will investigate potential trade-offs of the GDM-CLR in practical use for correcting delayed reporting, in comparison to established GDM versions.

5.3.4 Informative population demographics

The flexible hierarchical structure of our model can be adapted to include covariate effects, as we note in Section 2.5.2. A notable available covariate for COVID-19 cases is the age distribution of the SARI cases reported so far: 1) $\text{Children}_{t,s}$, the proportion of the as-yet reported SARI patients aged between 0 and 18; 2) $\text{Adults}_{t,s}$, the proportion aged between

19 and 60 and 3) Elderly$_{t,s}$, the proportion aged over 60. These three variables sum to one for any given t, s (before standardisation to have mean 0 and variance 1), so we exclude the latter. Where SARI cases ($y_{t,s}$) are not yet fully reported, we assume these are indicative of the age distribution of all the SARI cases. This assumption, i.e. that all age groups will be reported at the same rate, may not be sensible, especially if reporting does differ widely across age, which may introduce bias in model predictions.

As noted in a report by the World Health Organization (2020b), individuals aged 18 and under have lower risk of contracting and developing severe COVID-19 symptoms; they are therefore less likely to be hospitalised. This, in conjunction with COVID-19 vaccination strategies often targeting the more vulnerable, including the elderly, means differences in COVID-19 hospitalisation between age groups are likely to be notable. We aim to capture potential effects of the age distribution on the proportion of SARI cases that test positive for COVID-19 in our model "Survivor Age", using cubic polynomials:

$$\text{logit}(\mu_{t,s}) = \beta_{0,s} + \beta_t^{(s)} + \delta_s \zeta_t^{(s)} + \sum_{j=1}^{3} \left[\beta_{1,j}^{poly} \text{Children}_{t,s}^{j} + \beta_{2,j}^{poly} \text{Adults}_{t,s}^{j} \right]. \tag{5.15}$$

5.4 Severe Acute Respiratory Illness in Brazil

In this section we compare the predictive performance of a cohort of models in their application to the data from Brazil. First, we aim to compare GDM models based on the CLR link function to models based on the survivor and hazard versions of the GDM method (Section 5.2.2). In the CLR models, we capture variability in the mean proportions reported at each delay using a 2D tensor product $\gamma_{t,d}^{(s)}$, with thin-plate spline marginal bases,

$$\text{CLR}(p_{t,d,s}) = \gamma_{0,s} + \gamma_{t,d}^{(s)}. \tag{5.16}$$

We include a 2D tensor product in the CLR model to investigate the flexibility of the CLR transform as a link function to capture possible interaction effects between time and delay, which is more intuitive to model when dealing with absolute proportions. This CLR model include an explicit link between the expected mean SARI cases and the proportion of SARI cases that are COVID-19 positive (Equation (5.11)), but does not include any age covariates.

In addition to the "CLR" models, we tested a "Hazard" and "Survivor" model, based on the respective version of the GDM introduced by Stoner and Economou (2019) and discussed in Section 5.2.2. Specifically, these model is comparable to the models that include no age distribution effects. For the "Hazard" version, we model the Beta-Binomial means $\nu_{t,d,s}$ as follows:

$$\log\left(\frac{\nu_{t,d,s}}{1-\nu_{t,d,s}}\right) = i(t,d,s), \tag{5.17}$$

$$i(t,d,s) = \psi_{d,s} + \eta_t^{(d,s)}, \tag{5.18}$$

where $\psi_{d,s}$ is an independent intercept for each delay, and $\eta_t^{(d,s)}$ are independent penalised cubic splines of time (with shrinkage) for each delay d and region s. Then, for the "Survivor" version we model the cumulative proportions reported $S_{t,d,s}$:

$$\text{probit}(S_{t,d,s}) = i(t,d,s) \tag{5.19}$$

$$i(t,d,s) = \psi_{d,s} + \eta_t^{(s)}. \tag{5.20}$$

Here, $\psi_{d,s}$ is a monotonically increasing random walk over delay and $\eta_t^{(s)}$ is a cubic shrinkage spline of time, both terms are independent for each region s.

To investigate whether the difference between the CLR and the established Stoner and Economou 2019 model versions is more attributable to the use of the CLR link function over the alternatives, or instead more attributable to the 2D tensor product, we developed an alternative "CLR no tensor" model:

$$clr(p_{t,d,s}) = \psi_{d,s} + \eta_t^{(d,s)}. \tag{5.21}$$

Much like the "Hazard" model, here $\psi_{d,s}$ is an intercept for each delay and $\eta_t^{(d,s)}$ is an independent penalised cubic spline of time (with shrinkage) for each delay and region. This CLR version also includes the shared parameter term ($\delta_s \zeta_t^{(s)}$) when modelling the proportion of SARI cases that are COVID-positive.

To assess the impact of the shared parameter on predictive performance, we fit another model "Survivor no shared parameter" that does not include scaled shared parameter $\delta_s \zeta_t^{(s)}$ in linear predictor for the COVID-positive proportion. Then, we fit a final model, "Survivor Age", that includes both the shared parameter and the age covariates. As defined in Section 5.3.4, we assume the cubic polynomials for age are the same across all federative units,

$$\text{logit}(\mu_{t,s}) = \beta_{0,s} + \beta_t^{(s)} + \delta_s \zeta_t^{(s)} + \sum_{j=1}^{3} \left[\beta_{1,j}^{poly} \text{Children}_{t,s}^{j} + \beta_{2,j}^{poly} \text{Adults}_{t,s}^{j} \right]. \tag{5.22}$$

All models considered in our study are outlined in Table 5.1, including the model structure, the model run time and the mean absolute errors (MAE).

Table 5.1: Outline of all GDM model versions compared in our rolling prediction experiment. The run time in hours of each model version is given in brackets below the name. Nowcasting mean absolute error (MAE) is for the most recent nowcast date (T_{now}) such that the prediction time difference (PTD) equals zero.

Model (hrs to run)	**Linear predictor for COVID-positive proportions**	**Linear predictor for SARI delay distribution**	**Nowcasting MAE (SARI, COVID-19)**	
Survivor no shared parameter (4.07)	$\text{logit}(\mu_{t,s}) = \beta_{0,s} + \beta_t^{(s)}$	$i(t,d,s) = \psi_{d,s} + \eta_t^{(s)}$	99	182
Survivor (4.82)	$\text{logit}(\mu_{t,s}) = \beta_{0,s} + \beta_t^{(s)} + \delta_s \zeta_t^{(s)}$	$i(t,d,s) = \psi_{d,s} + \eta_t^{(s)}$	90	51
Survivor age (9.78)	$\text{logit}(\mu_{t,s}) = \beta_{0,s} + \beta_t^{(s)} + \delta_s \zeta_t^{(s)} + \sum_{j=1}^{3} \left[\beta_{1,j}^{poly} \text{Children}_{t,s}^{j} + \beta_{2,j}^{poly} \text{Adults}_{t,s}^{j}\right]$	$i(t,d,s) = \psi_{d,s} + \eta_t^{(s)}$	84	43
CLR no tensor (25.37)	$\text{logit}(\mu_{t,s}) = \beta_{0,s} + \beta_t^{(s)} + \delta_s \zeta_t^{(s)}$	$i(t,d,s) = \psi_{d,s} + \eta_t^{(s)(d)}$	136	51
CLR (33.96)	$\text{logit}(\mu_{t,s}) = \beta_{0,s} + \beta_t^{(s)} + \delta_s \zeta_t^{(s)}$	$i(t,d,s) = \gamma_{0,s} + \gamma_{t,d}^{(s)}$	168	47
Hazard (21.06)	$\text{logit}(\mu_{t,s}) = \beta_{0,s} + \beta_t^{(s)} + \delta_s \zeta_t^{(s)}$	$i(t,d,s) = \psi_{d,s} + \eta_t^{(d,s)}$	85	56

To compare different models, we carried out a rolling prediction experiment using weekly data starting from July 2021. We selected 16 dates between February and July 2022. Our goal is to fit each of the models at each of these dates, only using data that would have been available for modelling at those time points. Specifically, the available data at each nowcasting date, T_{now}, includes the number of SARI hospitalisations reported up to that date (through observed $z_{t,d,s}$) and simulated under-reported COVID-19 counts $\tilde{x}_{t,s}$. The "true" under-reported COVID-19 counts that would have been available at each date T_{now} is unknown since the historic data lacks time-stamps for reporting delays. To simulate under-reported counts, we multiply the fully reported $x_{t,s}$ by a value that starts at 0.95 and decreases linearly to 0.25 over the 30 weeks leading up to the nowcasting date T_{now}. We then round the result down to the nearest integer. The study of simulated counts limits the realism of our experiment, but on the other hand it allows checks that the simulated under-reporting mechanism is captured by the models adequately.

We fit all the models to data from each of the 27 federative units in Brazil,such that we have $16 \times 27 = 432$ sets of predictions for both SARI and COVID-positive SARI. This allows us to assess their performance when applied to a wide variety of time series, each with different shapes and scales in the levels of SARI and COVID-positive SARI cases,

and each with differing systematic delay mechanisms. Apart from the "Survivor age" model, which assumes a common polynomial for the age covariates across all federative units, there is no hierarchical or spatial effect between the 27 regions in this application of the GDM. We fit models in 60-week moving windows, i.e. rather than fitting models to the whole time series of historic data, we discarded historic data before the 60-week window, up to and including T_{now}. Moving windows are common in recent literature on disease nowcasting, e.g. to improve computational feasibility in complex methods (Stoner, Halliday and Economou (2022) (Chapter 4)), or to keep estimation of time-invariant model parameters relevant to more recent data (McGough et al. (2020)). For each model fit, we generated predictions of recent $y_{t,s}$ and $x_{t,s}$ from MCMC sampling.

5.4.1 Implementation and Prior Distributions

Here we detail the model choices for the rolling prediction experiment. Choices include priors for model parameters as well as constants that represent certain qualities in the data, such as maximum delays. Although these decisions will effect the results of our experiment to some degree we believe that our choices reflect suitable assumptions and would be chosen similarly in an operational context and therefore allow us to judge if our method is fit for surveillance purposes.

For the SARI hospitalisations we have set a maximum reporting delay of $D_{max} = 20$ weeks, this was chosen as at 20 weeks over 97% of the SARI cases are usually reported. Within the GDM models we explicitly model $D = 8$ weeks delay. Alternatively, for the unknown COVID-19 reporting delay we know it is likely to be longer than the SARI delay as once a SARI hospitalisation is reported there is a further length of time to receive test confirmation of whether the patient has COVID-19. Therefore, we have chosen to set the maximum COVID-19 reporting delay as 30 weeks which we believe to be a conservative selection. This allows us to simulate possible COVID-19 cases that would have been reported $\widetilde{x}_{t,s}$ to retrospectively test our model. Specifically, the simulated reported

proportions $\pi_{t,s}$ of the eventual total COVID-19 hospitalisations $x_{t,s}$ are constructed such that they monotonically decrease in the 30 weeks leading up to the nowcasting date T_{now} and are 1 (fully reported) for all weeks prior to $M = T_{now} - 30$. We explore the GDM's ability to recapture this reporting proportion in Section 5.4.4. We model the reported proportions by $\log(\pi_{t,s}) = -\omega_s(t - M)$ for $t > M$. We choose the prior,

$$\omega_s \sim \text{Gamma}(10, 200), \tag{5.23}$$

which is parameterised in terms of shape and rate with mean 0.05 and variance 0.00025, for the degree of censoring of the COVID-positive counts. In particular, it results in the degree of under-reporting of COVID-19 (π) to be between 0 and 100% with the prior distribution density peaking at around 25% for the current nowcast date (T_{now}). Whilst we know this to be sensible from simulating this censoring, the prior could be adjusted to reflect alternative beliefs about the real-world censoring or left as is without drastically impacting predictions as it would still be sensible for any under-reporting rate.

For the coefficient of the mean SARI spline effect on the positive COVID-19 proportions we assumed a prior,

$$\delta_s \sim \text{Normal}(0, 5^2), \tag{5.24}$$

$$\tag{5.25}$$

as a non-informative prior of the effect of the mean number of SARI cases on μ. When there is an influx in COVID-19 cases this will cause both the mean SARI cases, $\log(\lambda_{t,s}) = \zeta_{0,s} + \zeta_t^{(s)}$, and the proportion of COVID cases, μ, to increase. This is the relationship we hope to capture with $\delta_s \zeta_t^{(s)}$ within $\text{logit}(\mu_{t,s}) = \beta_{0,s} + \beta_t^{(s)} + \delta_s \zeta_t^{(s)}$.

The choice of prior for the SARI intercept term is $\zeta_0 \sim \text{Normal}(\log(\overline{SARI}), 10^2)$ where $\overline{SARI}$ is the mean of the SARI cases that have been reported up to the week $T_{now} - 20$, where T_{now} is the nowcast date we are predicting up to and 20 weeks is the maximum delay we expect for SARI. On the other hand, the prior for the intercept of the proportions is $\beta_0 \sim \text{Normal}(0, 5^2)$ since it is the intercept for a log-odds link function. Finally, for our time-delay tensor product interaction spline for the absolute proportions of the partial counts in our 'CLR' models we have a prior $\gamma_0 \sim \text{Normal}(0, 10^2)$ for the intercept.

All Beta-Binomial and the Negative-Binomial (θ) dispersion parameters were given the prior distribution Gamma$(2, 0.02)$,which is also parameterised in terms of shape and rate with mean 100 and variance 5000,

$$\theta_s \sim \text{Gamma}(2, 0.02), \tag{5.26}$$

$$\chi_s \sim \text{Gamma}(2, 0.02), \tag{5.27}$$

$$\upsilon_s \sim \text{Gamma}(2, 0.02), \tag{5.28}$$

$$\phi_{d,s} \sim \text{Gamma}(2, 0.02). \tag{5.29}$$

When including age polynomial coefficients in the model, as discussed in Section 5.3.4, we use the following priors:

$$\beta^{poly}_{1,1:3} \sim \text{Normal}(0, 2^2), \tag{5.30}$$

$$\beta^{poly}_{2,1:3} \sim \text{Normal}(0, 2^2). \tag{5.31}$$

Here we use common polynomial coefficients across all regions. However, it is worth noting that all the splines used when investigating the nested GDM framework introduced in this chapter are independent for each region (with no nested hierarchical structure). We specify and define the splines used throughout our framework in Section 5.4.1.1.

5.4.1.1 Spline and tensor product terms

Within the CLR model for our rolling prediction experiment we utilise a 2D tensor product spine for time and delay:

$$\gamma_{t,d}^{(s)} = X_{td}\boldsymbol{\kappa}_{\gamma}^{(s)}, \tag{5.32}$$

$$\boldsymbol{\kappa}_{\gamma}^{(s)} \sim \text{Multivariate-Normal}(\mathbf{0}, \Omega_{\gamma,s}), \tag{5.33}$$

$$\Omega_{\gamma,s} = \frac{S_{td}^{(1)}}{\sigma_{\gamma,1,s}^2} + \frac{S_{td}^{(2)}}{\sigma_{\gamma,2,s}^2} + \frac{S_{td}^{(3)}}{\sigma_{\gamma,3,s}^2}, \tag{5.34}$$

$$\sigma_{\gamma,1,s}, \sigma_{\gamma,2,s}, \sigma_{\gamma,3,s} \sim \text{Half-Normal}(0, 1^2). \tag{5.35}$$

The basis function for the 2D tensor product X_{td} was generated using the R package **mgcv**, with the `jagam(.)` function. This was also used to obtain the penalty matrices $S_{td}^{(1)}$, $S_{td}^{(2)}$ and $S_{td}^{(3)}$. Here we apply the tensor product to each spatial region independently.

For both the Hazard and Survivor models we instead opt for a temporal cubic spline with shrinkage for a given region in any version of the model is defined by:

$$\eta_t^{(s)} = \boldsymbol{X}_t\boldsymbol{\kappa}_{\eta}^{(s)} \tag{5.36}$$

$$\boldsymbol{\Omega}_{\eta} = \boldsymbol{S}/\sigma_{\eta,s}^2 \tag{5.37}$$

$$\boldsymbol{\kappa}_{\eta}^{(s)} \sim \text{Multivariate-Normal}(0, \boldsymbol{\Omega}_{\eta}) \tag{5.38}$$

$$\sigma_{\eta,s} \sim \text{Half-Normal}(0, 1^2). \tag{5.39}$$

The model matrix of the basis function $\boldsymbol{X}_t$ at each time point and a vector of coefficients $\boldsymbol{\kappa}$ are used to define the temporal splines. To avoid over-fitting, the precision matrix $\boldsymbol{\Omega}$ is equal to a cubic spline penalty matrix $\boldsymbol{S}$ and scaled by the penalty parameter σ_{η}. Small σ_{η} values correspond to harsher penalties (more smoothing). Implementing smooth splines in NIMBLE is introduced in more detail in Section 2.1.2.2. This η_t spline can

then be independently applied for each region $\eta_t^{(s)}$ (Survivor version) or independently for each region and delay $\eta_t^{(s,d)}$ (Hazard and CLR no tensor versions). The same priors can be formulated for the temporal splines β_t and ζ_t, which can similarly be applied independently for each region (all model versions).

For the rolling prediction experiment, we fit a moving window of length 60 weeks. Within each 60 week window the the cubic splines with shrinkage $\eta_t^{(s)}$ and $\zeta_t^{(s)}$ are assigned 15 knots which are evenly placed up to the nowcast date. For the cubic spline with shrinkage $\beta_t^{(s)}$, which captures the proportion of SARI cases that are COVID-positive, 8 knots were evenly placed up to the nowcast date. For the tensor product spline in the CLR model 15 knots are similarly evenly placed over the 60 week temporal trend and 6 knots are evenly placed to capture the delay over the $d = 1, ..., 8$ delays

5.4.1.2 Software and convergence

All models were written in the statistical programming language R (R Development Core Team (2011)) and were implemented using the NIMBLE software package (de Valpine et al. (2017)), which provides flexible implementations of Markov Chain Monte Carlo (MCMC) algorithms for Bayesian inference. This software is suitable for this application as we can represent our model as directed acyclic graph (DAG), as shown in Appendix C.2 for the survivor version of our nested framework.

In general we use the default MCMC samplers given by NIMBLE as discussed in Section 2.1.2.1, here we implemented automated factor (AF) slice sampling to improve sampling efficiency, where model effects are grouped into a single AF slice where there is intuition for them having a strong correlation in the posterior distribution. Hence, we used the default nimble MCMC samplers for all parameters, except for $\kappa_\beta^{(s)}$, $\kappa_\zeta^{(s)}$, $\sigma_{\beta,s}$, $\sigma_{\zeta,s}$, $\beta_{0,s}$ and $\zeta_{0,s}$ which are all assigned a single AF slice sampler for each region s. Then,

for the CLR model $\kappa_\gamma^{(s)}$, $\sigma_{\gamma,1:3,s}$ and $\gamma_{0,s}$ are all assigned a separate single AF slice for each region. Similarly, for the hazard and survivor versions $\kappa_\eta^{(s)}$ and $\sigma_{\eta,s}$ are assigned a AF slice sampler for each region. For the nested models with age $\beta_{1,1:3}^{poly}$ and $\beta_{2,1:3}^{poly}$ are all assigned an additional single AF slice parameter. Each model implementation is run for 2 MCMC chains with 15,000 iterations and 10,000 burn-in, with a thinning of 5.

In order to gauge whether the models converged we calculated the potential scale reduction factor (PSRF) for all model parameters to confirm it to be less than 1.10, which suggests a lack of evidence that the MCMC chains haven't converged. Although there is no guarantee that this convergence is to a unique posterior distribution we endeavoured to reduce the chances of multi-modality by running multiple chains from random initial values as advised in Brooks and Gelman (1998).

5.4.2 Age effects and the impact of COVID-19

To uncover insights into the effect of patient age distributions on the COVID-positive proportion in SARI cases, and the role of COVID in driving recent SARI outbreaks, we examine posterior inference for relevant parameters.

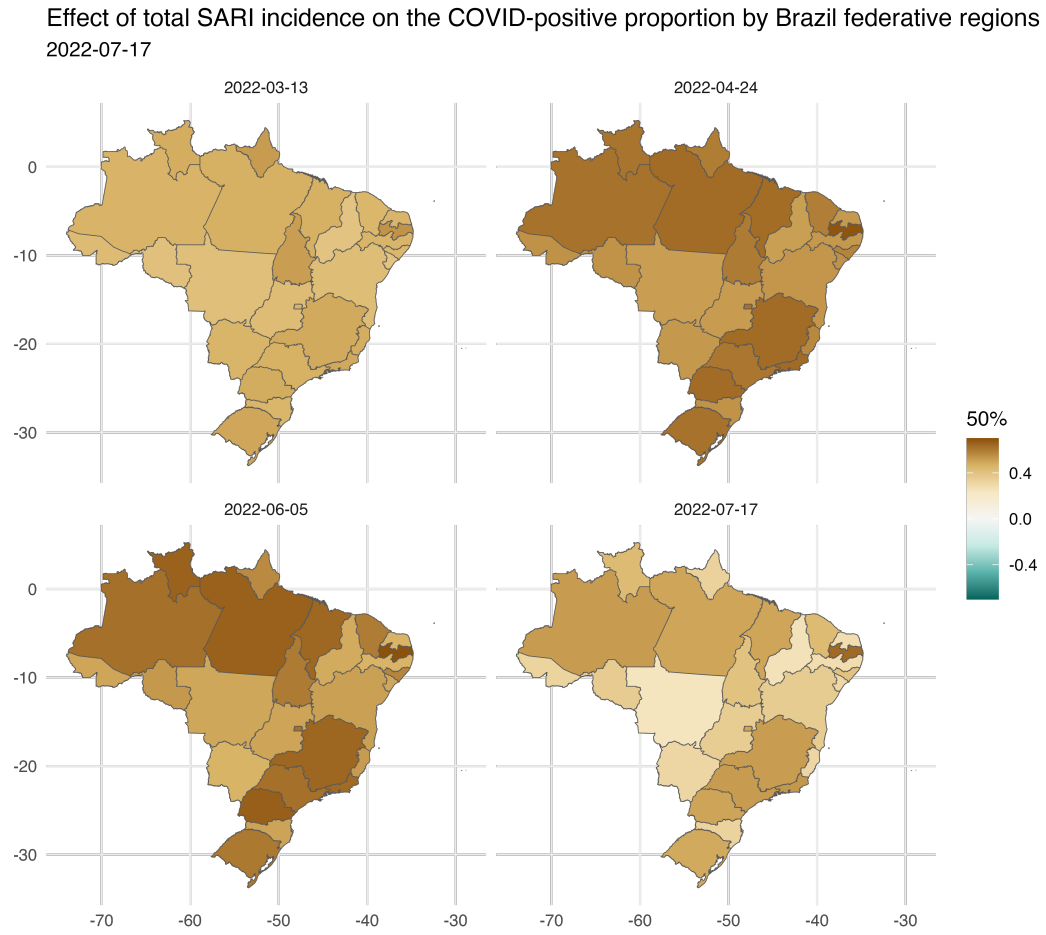

Figure 5.1: Posterior medians for the SARI-COVID link coefficient δ_s from the "Survivor age" model.

Figure 5.1 shows the posterior distributions of the coefficients δ_s, which links variation over time in the total SARI cases to the COVID-positive proportion (Equation (5.11)) for four nowcast dates. The posterior medians for δ_s are independently positive for the vast majority of federative units – generally in the region of about 0 to 2. These positive values imply that the higher the total SARI cases, the more of those cases will test positive for COVID-19 on average. However, mainly in the eastern part of the country, there are some negative values – these suggest that viruses other than SARS-CoV-2 are playing a proportionally larger role when overall SARI case levels are higher in these regions. A similar pattern is observed across all nowcast dates.

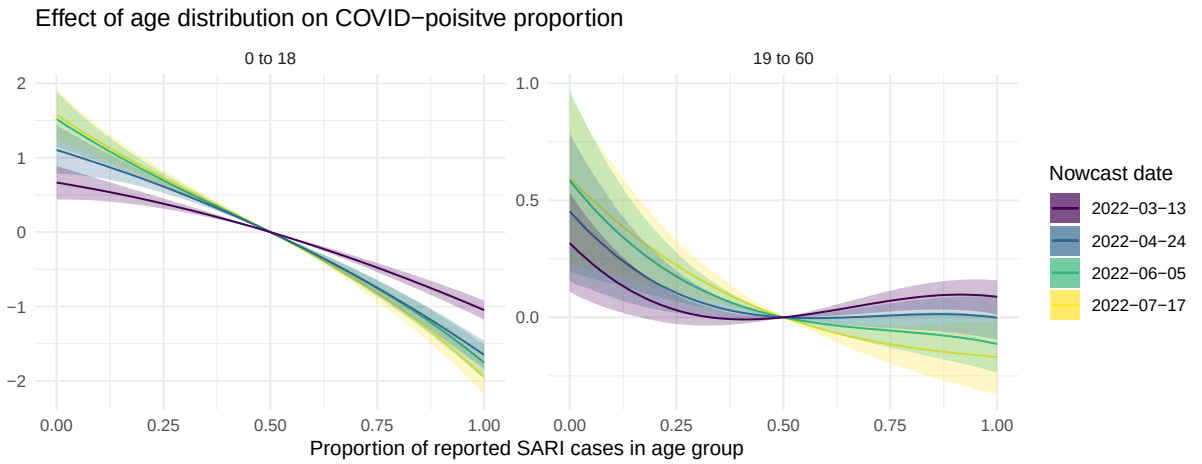

Figure 5.2: Posterior medians (lines) and 95% credible intervals (shaded areas) for the cubic polynomial effects of age distribution on the COVID-positive proportion of SARI patients from the "Survivor age" model.

Meanwhile, Figure 5.2 shows the predicted effect of the percentage of SARI patients aged 0-18 (left), or 19-60 (right), on the percentage of SARI patients that test positive. Here, we can see that the COVID-positive proportion decreases considerably where more of the SARI patients are in the child age group. We can also see decreasing effects for the adult age group. Implicitly, then, the COVID-positive proportion increases as the percentage of SARI patients in the "elderly" age group (60+ years old) increases. In all cases, the 95% posterior credible intervals are very narrow, reflecting the weight of evidence behind these associations.

5.4.3 Results from the rolling prediction experiment

To evaluate predictive performance, we compared posterior predictive samples for SARI and COVID-positive SARI to the counts that were eventually reported. We base conclusions on three prediction performance metrics: the mean absolute error (MAE), to capture accuracy of point predictions; mean 95% prediction interval width (PIW), to capture the prediction precision; and 95% prediction interval coverage, to asses whether uncertainty is quantified appropriately. Here, by coverage, we mean the proportion of observed values

that fall within the 95% prediction interval. We compute these metrics for each model, combining predictions made across the 27 federative units but separating predictions by the "prediction time difference" (PTD) – this is the difference in time between the now-casting date T_{now} and the recorded date of occurrence of cases. I.e. a PTD of 0 means predicting for the same week that data is available up to ("contemporaneous nowcasts"), and PTD<0 means predicting for past weeks, for which the counts have not yet been fully reported.

First, we aim to compare GDM models based on the CLR, survivor and hazard link function of the GDM method to determine which is best suited for this application. Additionally we compare the "CLR no tensor" model to determine the merits of the "CLR" model choice of a 2D tensor producrt to the "CLR no tensor" model that instead has an independent temporal spline for each delay $\mathrm{CLR}(p_{t,d,s}) = \psi_{d,s} + \eta_t^{(s,d)}$.

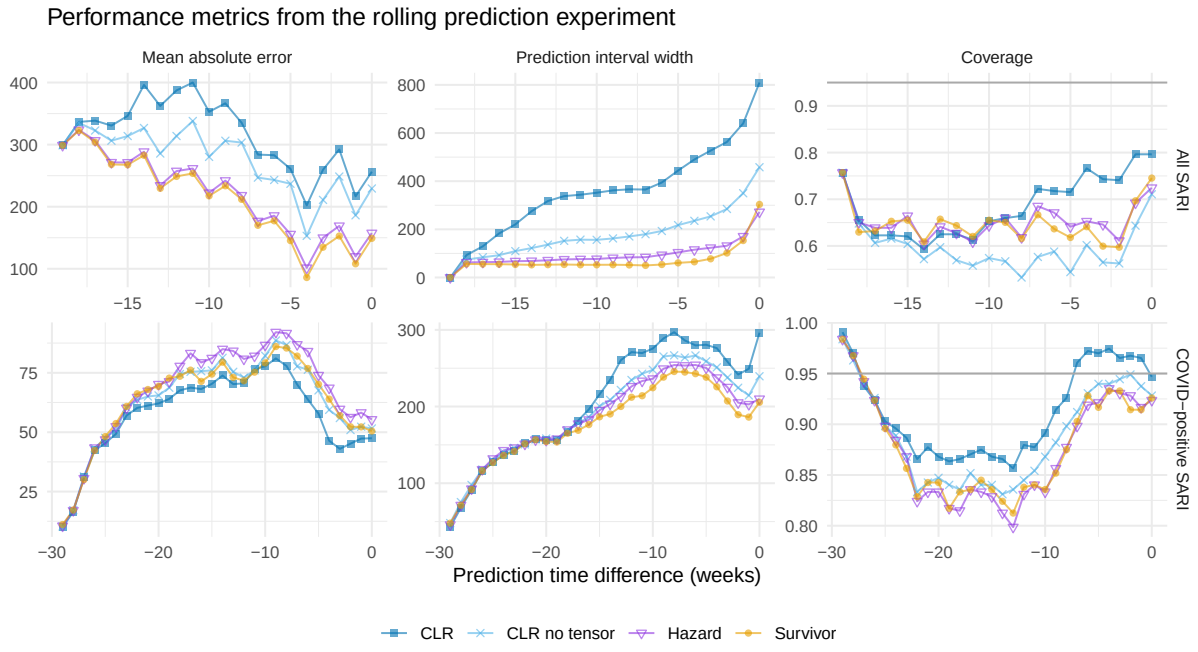

Figure 5.3: Prediction performance metrics from the rolling experiment for all SARI (top) and COVID-positive SARI (bottom): mean absolute error (left), prediction interval width (center), and 95% prediction interval coverage (right).

The upper row of Figure 5.3 shows the three performance metrics for predicting all SARI. Here, we can see that the "Survivor" model greatly outperform both the "CLR" models. Notably, the mean absolute error of point predictions for contemporaneous nowcasts (PTD$= 0$) is about 50% lower for the Survivor models. It is clear that the "Hazard" and "Survivor" variants of the GDM appear to be better suited for this application with the "Survivor" outperforming the "Hazard" slightly in terms of mean absolute error (MAE) and prediction interval width (PIW). Thus, in our main atricle we focus on different survivor versions to asses our proposed framework.

The hazard version may outperform the CLR versions, as with the survivor, as the "Hazard" has an independent temporal spline for each delay. In order to achieve a similar level of flexibility for the "CLR" model the number of knots could be increased but, this could cause over-fitting and therefore issues with uncertainty and accuracy when forecasting future counts. As we can see, the "CLR no tensor" model which has a more flexible delay structure outperforms the other CLR version for SARI predictions in terms of MAE and PIW, but has an even worse coverage. However, the fact that the other link functions still out-preform the more flexible "'CLR no tensor" version suggests that differences in performance may be due to modelling the relative or cumulative proportions being more appropriate in this particular data set. Possibly indicating that the GDM-CLR may still be beneficial for alternative applications.

A plausible reason for the poorer performance in the CLR models – compared to the survivor and hazard versions – is that here we assumed systematic variation in the reporting delay could be captured by a smooth 2D tensor product over time and, notably, delay. Assuming smoothness over the delay dimension may not be realistic (e.g., in Stoner, Halliday and Economou 2022 (Chapter 4) there was a spike in reporting of COVID-19 deaths in the second delay), which may be why the survivor models, which assume a non-smooth random effect for each delay, perform better here. Hence, we may have seen better performance had we constructed our tensor product using a less smooth marginal basis for

delay (e.g., spherical-covariance Gaussian-process basis functions, available in the `mgcv` package from Wood 2017). Meanwhile, coverage for the 95% prediction intervals was low across all models, for both SARI and COVID-positive SARI. We discuss the practical implications of this in Section 5 of the main article.

The rolling prediction experiment for sixteen nowcast dates took 5 hours to run for the "Survivor", 21 hours to run for the "Hazard" and 25 hours to run for "CLR no tensor". In a operational setting for a single nowcast date they would all take less time to run. But, run time can be crucial in applications where timeliness of disease surveillance is a priority. Therefore, the survivor link function has the most practical benefits in this application due to both prediction precision and computational performance.

Next, due to the superior performance of the survivor model in Figure 5.3, we compare the predictive performance of different "Survivor" versions of the GDM model. This included a "Survivor no shared parameter" model without the shared parameter for the mean number of SARI cases and the expected relative proportions reported, and a "Survivor no shared parameter" model with age covariate effects for the proportion of SARI cases that are COVID-19 positive.

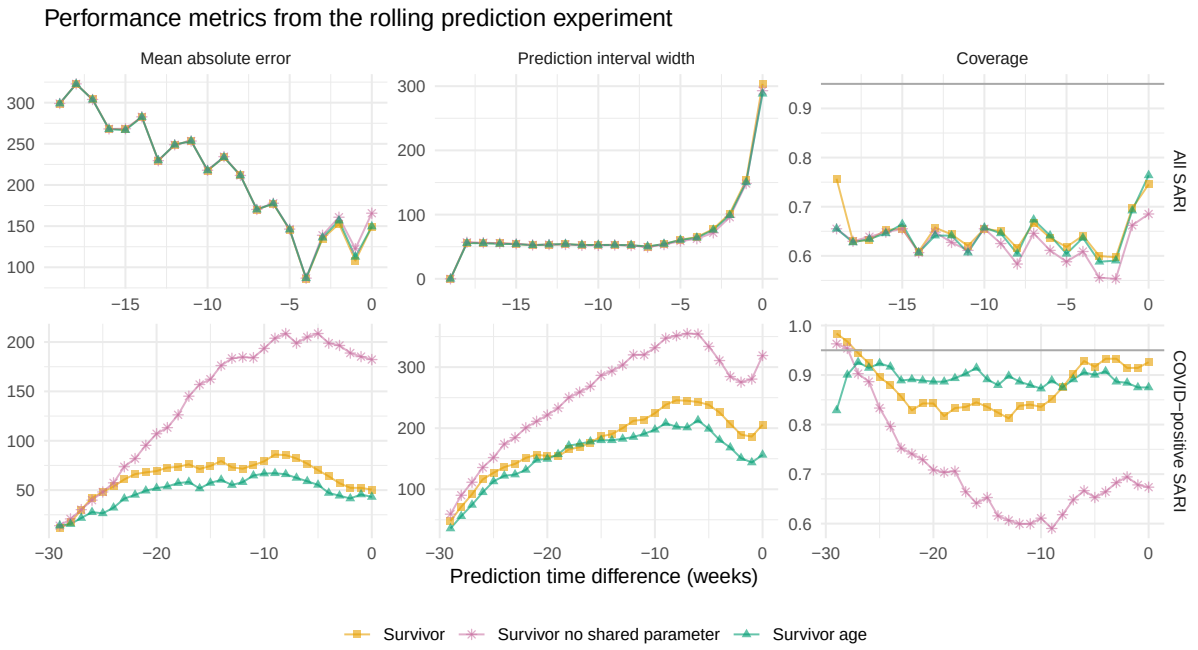

Figure 5.4: Prediction performance metrics from the rolling experiment for all SARI (top) and COVID-positive SARI (bottom) hospitalisation: mean absolute error (left), prediction interval width (centre), and 95% prediction interval coverage (right).

The lower panels in Figure 5.4 shows the performance metrics for predicting COVID-positive SARI cases, for the three survivor models, chosen to highlight the comparative performance when including age covariates and a shared parameter for the survivor versions of the model. The addition of age distribution polynomial effects greatly improves detection of trends in severe COVID-19 hospitalisation, which is reflected in improvements across all three performance metrics.

However, it is important to highlight that the trends in the metrics in both Figure 5.3 and Figure 5.4 are not as we would expect from a GDM model. For example, if we consider Figure 4.1 for comparison, we would expect to see the mean absolute error increase for more recent predictions (larger prediction time difference) as there will be less observed data at these times. This is not present in Figures 5.3 & 5.4, where we instead see the mean absolute error falling over prediction time difference for SARI and falling in the most recent 10 days of nowcasting. Also, while coverage may fluctuate we would expect

values to be much closer to the desired 95% level, where as here we consistently have coverages below the 70% level. There also seems to be an inverse relationship between mean absolute error and coverage for both the upper and lower panels, this suggest that the model is systematically not capturing the trend in the data.

Figure 5.4 also demonstrates that not including a shared parameter in the survivor model which between the mean expected SARI cases and the proportion on those SARI cases that are COVID-positive results in poorer performance across all three metrics (columns). To investigate this further, we zoom-in on one region for a closer inspection of how including a shared parameter between SARI and COVID-positive cases affects prediction performance. Figure 5.5 shows predictions of the proportion of SARI cases that test COVID-positive ($x_{t,s}/y_{t,s}$), for the São Paulo federative unit, from the "Survivor no shared parameter" and "Survivor" models fitted to 4 of the 16 nowcasting dates. It should be noted that, in our framework, accurately predicting the COVID-positive counts $x_{t,s}$ is dependent on capturing both $y_{t,s}$ and $x_{t,s}/y_{t,s}$ appropriately. These two models have an identical GDM structure for the total SARI counts $y_{t,s}$ and their respective partial counts $z_{t,d,s}$, meaning any differences in their performance predicting COVID-positive SARI should only arise from their respective ability to capture $x_{t,s}/y_{t,s}$ well (through $\mu_{t,s}$). The left

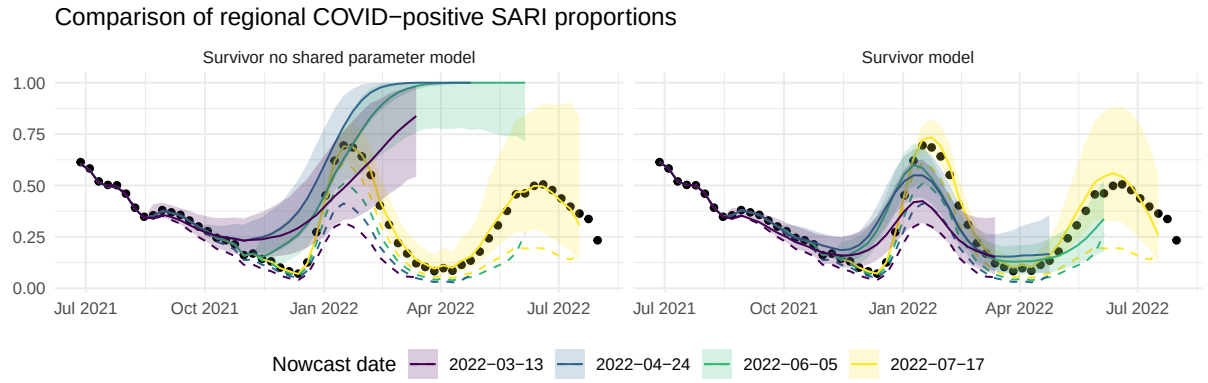

Figure 5.5: Posterior median predicted (lines) and 95% prediction intervals (shaded areas) for the proportion of SARI cases that test positive for COVID-19, from the "Survivor no shared parameter" and "Survivor" models. Dashed lines show the simulated under-reported COVID-positive proportions of SARI.

panel of Figure 5.5 shows predictions from "Survivor no shared parameter", we can see that the predictions are basically linear where the $x_{t,s}$ are not yet reported with wider

prediction intervals that fail to capture the true data (points). This is likely due to both the COVID-positive cases $x_{t,s}$ and the SARI hospitalisations $y_{t,s}$ being unknown for the most recent time steps creating identifiability issues for the proportion $x_{t,s}/y_{t,s}$. Hence, including the scaled shared parameter in the proportion of COVID-positive cases, as in the "Survivor" model (right panel), allows the general shape $x_{t,s}/y_{t,s}$ to be captured a lot better by pooling information from the model of the expected mean SARI cases.

Using results from the "Survivor Age" model, Figure 5.6 shows how the SARI and COVID-positive hospitalisation could be nowcasted for São Paulo. In a regular operational setting, we would see the model predictions (solid lines and shaded intervals) and the total reported so far (dashed lines), but not the total eventually reported (points).

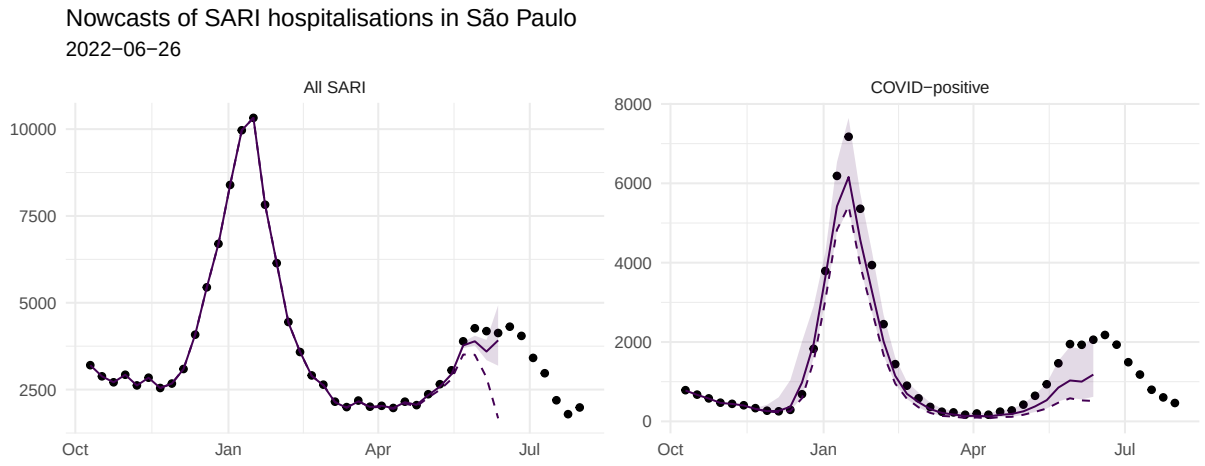

Figure 5.6: Posterior median predicted (lines) and 95% prediction intervals (shaded areas) for the nowcasted SARI and COVID-positive hospitalisations. Dashed coloured lines show available hospitalisations reported at time of the nowcast date.

In this example, the model predictions are very close for SARI but the point estimates under-predict the most recent COVID-positive counts. Furthermore, the prediction intervals fail to capture the majority of the true data (points) for the severe COVID-19 hospitalisations. This poor coverage is a result of model not capturing the true proportions of the COVID-positive proportions propagating into the COVID-19 predictions, again this is likely due to the identifiability issues of incomplete data for the total SARI cases and censored COVID-positive counts.

5.4.4 Capturing COVID-19 censoring

Figure 5.7 demonstrates the GDMs ability to capture the reporting rates of COVID-positive SARI which we constructed (grey dashed lines) prior to modelling for simulation purposes in our retrospective experiment.

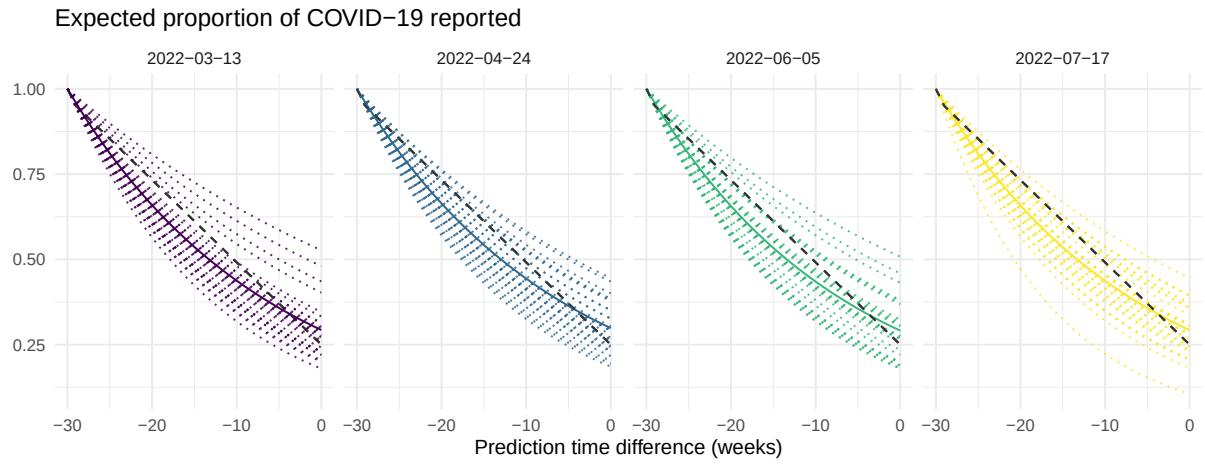

Figure 5.7: Posterior medians from the "Survivor Age" model of expected proportion of COVID-positive SARI cases reported (dotted lines) for 27 federative regions, the solid line shows the Brazil level mean. Dashed lines indicate the artificial reported proportions, used to simulate the COVID-19 censoring, that we hope to capture.

The across region means (solid lines) shows that on average these reporting proportions are fairly well captured by the GDM. However, some regions appear to over or under predict the reporting rate more than others (dotted lines). Like existing nowcasting frameworks we assume the delay and number of cases are independent. Figure 5.7 also shows the discrepancy between how the censoring of COVID-19 hospitalisations have been simulated and modelled. We simulate the censoring linearly with $\widetilde{x}_{t,s} = \pi_t^{sim} x_{t,s}$ where the simulated reporting rates π_t^{sim} decrease linearly from 1 to 0.25 from the time we simulate the COVID-19 hospitalisations to be censored. On the other hand, we modelled the censoring with a Beta-Binomial distribution with expected proportion $\log(\pi_{t,s}) = \omega_s(t - M)$, as defined in

Section 5.3.2. This was done as in practice the censoring will not have to be simulated and we have no prior information about how this censoring may be structured. Hence, we hope to demonstrate the abilities of the model despite potentially misspecifying the censoring assumption.

5.5 Discussion

The emergence of the SARS-CoV-2 virus has introduced a complexity in the incidence and management of severe acute respiratory illness (SARI) cases in Brazil. Supplementing existing surveillance of SARI with nowcasting of the number of cases caused by SARS-CoV-2 – as indicated by positive lab test results – is useful as an indicator of COVID-19 prevalence in the wider population, since more severe cases are more likely to receive a test. However, existing operational early warning systems in Brazil do not model all and COVID-positive SARI together. We argue that modelling them jointly can lead to more accurate predictions of COVID-positive SARI, drawing on the more timely and complete information available for SARI cases before COVID-19 test results are recorded.

In this chapter we aimed to develop a general framework for correcting delayed reporting of diseases with a nested structure (Section 5.3). To achieve this, in Section 5.3.1, we added an extra Beta-Binomial layer in the hierarchy for the nested disease to an established method based on a hierarchy of Negative-Binomial and Generalized-Dirichlet Multinomial models (Stoner and Economou (2019)). We further modified this new Beta-Binomial layer to account for under-reporting of recent confirmed COVID-positive SARI cases (Section 5.3.2). In doing so, we have proposed the first modelling approach (to the best of our knowledge) that addresses the issue of "unmarked" reporting delays. Such scenarios, where the length of the reporting delay is unknown, constitute a common challenge in

real-world data collection. For instance, in Höhle and An Der Heiden (2014) only 79.6% of hemolytic uremic syndrome (HUS) cases had a known date of hospitalisation, since recording it was not mandatory. Previously, such unmarked data have been removed prior to analysis where possible (Höhle and An Der Heiden (2014)).

The flexibility/generality of the framework allowed us to estimate the effects of overall SARI incidence and polynomial effects of the age distribution of SARI cases on the proportion of SARI cases testing positive for COVID-19, in Section 5.3.4. Here, we found that the COVID-positive proportion increases when the overall incidence of SARI increases *and* when the SARI patients tend to include more elderly people (Section 5.4.2). Through a comprehensive rolling prediction experiment in Section 5.4.3, we quantitatively assessed and compared the potential performance of joint predictive models for SARI cases and COVID-positive SARI cases. Here, we demonstrated that including age distribution polynomials and the explicit link between overall SARI incidence and the COVID-positive proportion led to the most convincing performance, e.g. in terms of point estimate accuracy (mean absolute error).

However, we also found that 95% prediction interval coverage was relatively low with an average of 76% for all SARI and 87% for COVID-positive SARI across all models tested. Prediction interval coverage has generally been appropriate in previous applications of the GDM method to correcting delayed reporting, e.g. Stoner, Halliday and Economou (2022) (Chapter 4), so the lower coverage for all SARI cases is especially curious. On reflection, we believe that this issue needs more thorough investigation that is outside the scope for this thesis. So far, we have considered both the joint modelling of SARI and COVID-19 as well as the specification of the age effect as potential sources of concern within the model, but both have been ruled at as the source of the problem. In practical applications, where it would be beneficial to present uncertainty intervals to public health policy makers with a desired coverage (for example of 95%) of the eventually reported data, more extreme quantiles for the prediction interval could be considered. However, this would require generating the posterior predictive samples for SARI and COVID-positive

SARI for times the true data is now available, to calibrate the choice of quantiles that are likely to encompass the desired proportion of the eventual counts. In the context of our work this would be straightforward to carry out as the rolling prediction experiment allows us to determine the quantiles needed to capture 95% of the true data for multiple nowcasts dates.

Although our motivation for developing this framework was capturing COVID-positive SARI cases nested within the broader cohort of all SARI cases, we believe it has strong prospects for application to related problems. In the Brazilian context, predictive performance of nowcasts for SARI hospitalisations caused by Respiratory syncytial virus (RSV) could also be improved through joint modelling with the total SARI cases. The distribution of ages within these total SARI counts are also likely to be informative here due to RSV occurring more frequently in small children (0-2 year-old's) Freitas and Donalisio (2016). Alternatively, our application could be extended to incorporate modelling multiple viruses that cause SARI (e.g. adding influenza and/or RSV-positive SARI) simultaneously. Alternatively, our framework could be applied to considering nested structures arising from different viruses, nested structures arising from different virus strains, or nesting from the severity of a disease (e.g. mild cases, severe cases, hospitalisations, fatalities).

Our assumptions for both modelling and simulating the censoring of the severe COVID-19 hospitalisations could benefit from a more in-depth simulation study in future work to systematically identify the impact of different assumptions in both. For simplicity, here we have simulated censoring linearly since this is not the focus of our investigation as there is no historic data to inform the censoring structure or to check our assumptions against. Currently, the only way to check this model assumption is to monitor the reported data over a period of time until enough historic data is collected such that the delay structure can be inferred. However, with the addition of more expert knowledge, or in a scenario where modelers have a greater intuition about the reporting process, the modelling of the unknown delay structure will hopefully be more aligned with the data (and hence to any simulations carried out to test the model). Thanks to the flexibility of the Beta-Binomial

model used to capture the censoring process in our framework, a wide range of censoring assumptions could be implemented in this model. For example, if the delay structure of the COVID-positive cases were in fact known we could add a delay process to the censoring model, Equation (5.12) which would then be equivalent to the GDM model,

$$\widetilde{x}_{t,d,s}|x_{t,s},\pi_{t,d,s},\upsilon_{d,s} \sim \text{Beta-Binomial}(\pi_{t,d,s},\upsilon_{d,s},x_{t,s}); \tag{5.40}$$

$$\pi_{t,d,s} = j(t,d,s), \tag{5.41}$$

where $j(t,d,s)$ captures the COVID-19 delay distribution in the same way $i(t,d,s)$ captures the SARI delay distribution. Any of the three link functions would be appropriate for this additional GDM model in the framework. The final framework would then continue to pool information between the SARI and COVID-19 hospitalisations under the hierarchical structure, but there would be less identifiability issues in capturing the proportion of SARI cases that are COVID-positive, as more data for the COVID-19 hospitalisations will have been observed. In general, this framework could be applied to different nested data sets with or without an unknown reporting delay structure in one of the counts.

We also aimed to extend available methodology for flexible modelling of general compositional (count) time series data, based on hierarchical Generalised-Dirichlet Multinomial methods. Here, we proposed an alternative approach to modelling the mean proportion belonging to each composition, using the centre log ratio as a link function for the GDM. We argued that this may avoid the major drawbacks of the two established "survivor" and "hazard" approaches, i.e. avoiding monotonicity constraints when instead modelling the cumulative proportions over the compositions, and avoiding a lack of intuition when modelling the relative proportions.

We demonstrated this extended framework in the context of a complex hierarchical data problem, with both a stochastic total for the compositions ($y_{t,s}$) and another stochastic variable ($x_{t,s}$) sharing the same parent as the count compositions $z_{t,d,s}$. Furthermore, we investigated the use of a 2D tensor product smooth term – as one possible option within

this general approach – to capture systematic variation over the time (t) and delay (d) dimensions relevant to our Brazilian SARI application. Our rationale for using a 2D tensor product was the potential to be flexible enough to capture non-linear changes in the mean delay distribution over time, with the smoothness penalty terms preventing over-fitting.

Ultimately, our models based on the CLR did not outperform established alternatives from Stoner and Economou (2019) in the rolling prediction experiment. Notably, we tested both a "CLR" model with a 2D tensor product of time and delay and an alternative "CLR No Tensor" based on independent splines. For the former, we constructed the tensor product using thin plate spline marginal bases, meaning it is constrained to be smooth in both dimensions. Meanwhile, the latter does not assume any smoothness in systematic time-delay variability across the delay dimension. The "CLR" model with the tensor product performed worse, such that we tentatively conclude that it may be inappropriate to assume smoothness across the delay dimension in this application. We may have seen better performance had we constructed our tensor product using a less smooth marginal basis for delay (e.g. spherical-covariance Gaussian-process basis functions, available in the `mgcv` package from Wood (2017)).

Chapter 6

Investigating the effect of case load on the delay in reporting infectious diseases

In Chapter 1, we discussed what we consider to be the four main types of variability in data suffering from delayed reporting and the need to appropriately capture these to develop reliable nowcasting models. One of these is systematic variability in the delayed reporting process. Often in analysis of real-world data, we see reporting rates (e.g. as measured by the percentage of cases reported within a delay period) trending over time, due to changes in the reporting process. In Stoner, Halliday and Economou (2022) (Chapter 4), we also identified a specific mechanism known as the "weekly cycle" in COVID-19 hospital deaths data for England, where reporting delays were systematically longer at weekends due to lower staffing levels reducing reporting capacity. Hence, this is an example of the limit of a reporting system capacity being met, and having detrimental effect on reporting efficiency. Related to this is the theoretical possibility that capacity might be increased in response to a greater anticipated or observed stress on reporting systems. Alternatively, a sudden surge in cases might overwhelm a reporting system and the fixed capacity may induce inefficiencies in reporting.

Furthermore, Gutierrez, Rubli and Tavares (2022) note that correcting these delays through statistical methods is more vital in low- and middle-income countries, due to limited capacity for improving the reporting process to reduce delays, often as a result of finite resources and funds being targeted towards patient care instead. Recent contributions to the literature, such as Bastos et al. (2019), Gutierrez, Rubli and Tavares (2020) and Harris (2022), have also highlighted that these resource limitations, and therefore reporting capacity, may be highly influential on the length of reporting delays when systems are overwhelmed by an uptick of incoming reports. This is somewhat intuitive as within any type of processing system there will be a capacity limit that has the potential to be overwhelmed by unexpected surges.

Broadly, the nowcasting frameworks reviewed in Chapter 2, and those developed in previous chapters, allow for changes over time in reporting performance through flexible temporal structures (e.g. random walks, splines) and/or through fitting the model to moving windows. These models can thus potentially "react" to sudden changes in reporting per-

formance relating to the level of incidence of the disease, but lack any structure that could inform a likely change in the delay distribution and thus provide more accurate correction of these delays for nowcasting. Indeed, to the best of our knowledge, no existing nowcasting methods explicitly account for any relationship between incidence and delay, which could be a source of bias and/or uncertainty in predictive performance. For example, in a situation where reported counts are lower than expected due to an overwhelmed reporting system, such models may interpret this information as a decrease in the level of the disease, resulting in under-prediction and potentially inadequate measures.

In this chapter, we aim to further investigate "case load" effects in real data from Brazil, and propose the first nowcasting framework (we are aware of) that can directly account for them, as a foundation for the development of less biased and more accurate disease surveillance systems in the future. In Section 6.1 we introduce the current literature on the potential effect of the level of the disease on delayed reporting – which we will refer to as the "case load" effect or "incidence-delay" effect – and investigate whether it is detectable in data for severe acute respiratory illness (SARI) hospitalisations and for arbovirus cases (including dengue and chikungunya) from Brazil (Section 6.1.1). In Section 6.2 we introduce and extend the GDM framework to quantify the effect of the total counts on the delay distribution, hence capturing the relationship between the total cases at a given time and the efficiency of reporting delays at that time. Through an extensive simulation experiment in Section 6.3, we assess the reliability of the model in capturing the incidence-delay effect in the presence of confounding trends, and the implication of ignoring the incidence-delay effects on the model prediction precision and accuracy. Following this, we apply our proposed approach to the Brazilian SARI hospitalisations and Brazilian arbovirus cases in Section 6.4, to provide clearer insights into potential incidence-delay effects and trends than is possible through exploratory plots. We conclude this work with a discussion of our findings and suggest worthwhile avenues for future research in Section 6.5.

6.1 Background

In this Chapter, we will use the same notation as we have previously for disease count data suffering from delayed reporting, first explained in detail in Section 1.1. Briefly recall, from Section 1.1.1, that disease time series data are often available as a total count y_t (cases/ hospitalisations/ deaths) occurring over discrete time steps (e.g. daily/ weekly/ monthly/ yearly) $t = 1, 2, \ldots, T_{now}$. As discussed in Section 1.1.2, due to reporting delays, it is unlikely that y_t will be fully reported at time t. Instead we often find that portions of y_t, $z_{t,d}$, are reported over subsequent time steps, e.g. $t, t+1, t+2, ...$, corresponding to a delay index $d = 0, 1, 2, ..., D_{max}$. Here, D_{max} is the assumed maximum possible delay, such that all the partial counts and thus the total counts, $y_t = \sum_{d=1}^{D_{max}} z_{t,d}$, are fully observed for $t > t + D_{max}$.

Bastos et al. (2019) note that the incidence of a disease may bring about potentially conflicting reactions in the timeliness of those counts being reported. Reporting delays could be expected to decrease during an outbreak due to the increase in awareness and drive to reduce the public health challenge. Alternatively, they may increase due to the burden of high case loads on the reporting system, leading to back-logs in case reporting. The COVID-19 pandemic has further highlighted potential relationships between higher disease incidence levels and longer reporting delays around the world, including Mexico and England (UK) in Gutierrez, Rubli and Tavares (2020), and New York (USA) in Harris (2022). Similarly, Torres, Sippy and Sacoto (2021) analyse data concerning COVID-19 testing delays in Ecuador and Hayashi and Nishiura (2022) investigate case fatality risk for COVID-19 in Japan in relation to case load.

Gutierrez, Rubli and Tavares (2020) investigates the impact of not accounting for reporting delays of COVID-19 deaths on the analyses of SIR (Susceptible - Infectious - Recovered) models, focusing on both Mexico and England. Let $\text{delay}_{s,t}$ be the number of days of delays before an individual COVID-19 death occurring at day t and at location s is reported.

Gutierrez, Rubli and Tavares (2020) assume a Geometric distribution for $\text{delay}_{s,t}$ with a log-linear model for the expected delay $E[\text{delay}_{s,t}]$:

$$\log(E[\text{delay}_{s,t}]+1) = \pi_s + \xi_t + \sum_{q=1}^{Q} \left(\alpha_q \mathbb{1}\left[\text{deaths}_{s,t} = q\right]\right) + \varepsilon_{s,t}. \tag{6.1}$$

Here, the expected reporting delay $E[\text{delay}_{s,t}]$ is determined, at the log scale, by separate spatial and time effects π_s and ξ_t. Then, the term $\varepsilon_{s,t}$ captures the random time and location variability. Meanwhile, coefficients $\alpha_1, \ldots, \alpha_Q$ capture the effect of the binned total number of deaths $\text{deaths}_{s,t}$ on the expected delay. The binned total deaths is implemented as a categorical variable through the indicator function $\mathbb{1}\left[\text{deaths}_{s,t} = q\right]$, which is equal to 1 when $\text{deaths}_{s,t}$ is in category q ($q \in \{1,2,3,4,5+\}$ deaths) and 0 otherwise.

Using this model, Gutierrez, Rubli and Tavares (2020) found that the reporting delays had large spatial heterogeneity and were generally longer at times when there are a greater number of total deaths, for a given location. They suggest that the increase in reporting delay length when average deaths are high could also be related capacity limits. Alternatively, they suggest a higher number of deaths could increase the likelihood of deaths that need further investigation/tests before being reported, increasing the average delay length.

Furthermore, they found that Mexico has longer reporting delays on average than England, and that delays are more strongly influenced by the eventual number of deaths in Mexico. They argue, with relevant evidence, that the difference in reporting delay length between countries is related to relative state capacities. Specifically, Gutierrez, Rubli and Tavares (2020) investigated the effect of state capacity by comparing the average delay in days against metrics for each of the municipalities in Mexico. Metrics for capacity that were found to correspond to longer delays were; fewer health care units per capita, fewer medical staff per capita, and higher patient volume per health care unit. Therefore, they

maintain that there is a need to account for spatial and temporal variability in reporting systems, as well as potential slowdowns in reporting when considering overwhelmed real-time surveillance systems. However, they do not address this within the scope of Gutierrez, Rubli and Tavares (2020).

Similarly, Harris (2022) implements a nowcasting model for COVID-19 deaths in New York city, using an iterative algorithm equivalent to expectation–maximization (EM), where the respective confidence intervals are calculated using a bootstrap method. They model the delay distribution by considering direct model for the partial counts (recall that we call this kind of model a "conditional independence" model, as described in Chapter 2.2.2). They model the partial counts with a Poisson distribution, without formulating a joint model for the total deaths to marginalise over. Their reasoning behind choosing a marginal model is that if separability of total incidence and reporting delay is independent, there is no inherent bias in just modelling the delay distribution. However, they did not compare their marginal model approach to a joint model approach to confirm this claim. Instead, they argue that the independence assumption holds as the delay length of COVID-19 cases are not related to case load, but instead delay length is associated with independent improvements in the testing capacity. However, this is based solely on the observation that delay lengths continued to fall in New York in October 2020 despite COVID-19 cases rising again.

These references are indicative of a growing body of evidence in the literature suggesting a need to account for case load effects in disease surveillance methods. In Section 5.3, we propose a new general approach that addresses this need, based on the Generalized-Dirichlet Multinomial (GDM) that has already proven highly adaptable for new disease surveillance challenges in previous chapters.

6.1.1 Exploratory Analysis

In this Section, we investigate potential evidence for incidence-delay effects, discussed in the previous section, for two operational disease surveillance data sources in Brazil. Statements made to us by expert researchers at Fiocruz, Brazil's leading public health research institute, suggested that these disease data are likely affected by links between the case load and delay in reporting of the disease. Therefore, they can serve as a useful real-world grounding for our investigation into case load effects and as test applications for new models. Furthermore, the expert researchers anticipated that the sign (positive or negative) of the relationship was likely opposite between the two case studies, providing an interesting direction for us to assess the generality of our proposed frameworks.

6.1.1.1 SARI hospitalisations

The first case study in this chapter is SARI hospitalisations in Brazil, published as open data by Ministry of Health Brazil (2022). Recall that we studied this data in Chapter 5, in the context of modelling diseases with nested structures. Details on SARI and its full definition are given in Section 5.1.

The definition of the reporting date for a SARI hospitalisation in Brazil is the date that an individual patient record is digitalised and added to the main database. This task is executed manually and often by hospital staff that also have responsibility for patient care; if the number of patients requiring treatment surges then caring for those patients is likely to be prioritised over digitalisation of records, leading to longer reporting delays. If such an effect is detectable in available data, accounting for it in nowcast models could enable more accurate predictions by accounting for the fact that a surge in cases may not be reflected in initial reports due to longer reporting delays.

Moreover, quantifying this relationship could inform decision makers also allow insights into how the reporting delay process is impacted by potential outbreaks, and possibly lead to targeted intervention to reduce delays when future peaks occur, e.g. employing additional staff/supplies to help increase the reporting capacity or aiding staff efficiency through incentives or specialised training. Alternatively, there could be scope to improve region-specific reporting infrastructure, such as updating technologies or streamlining the reporting process to increase reporting speed.

Here, we investigate signs of any potential relationship in available data between the number of SARI hospitalisations in Brazil and the length of the reporting delays through exploratory plots. In total across the whole of Brazil, Figure 6.1 shows the probit-transformed cumulative proportion of SARI cases reported $\left(C_{t,d} = \frac{\sum_{i=0}^{d} z_{t,i}}{y_t}\right)$ up to and including each week of delay (y-axis), varying with the total number of SARI cases eventually reported (x-axis).

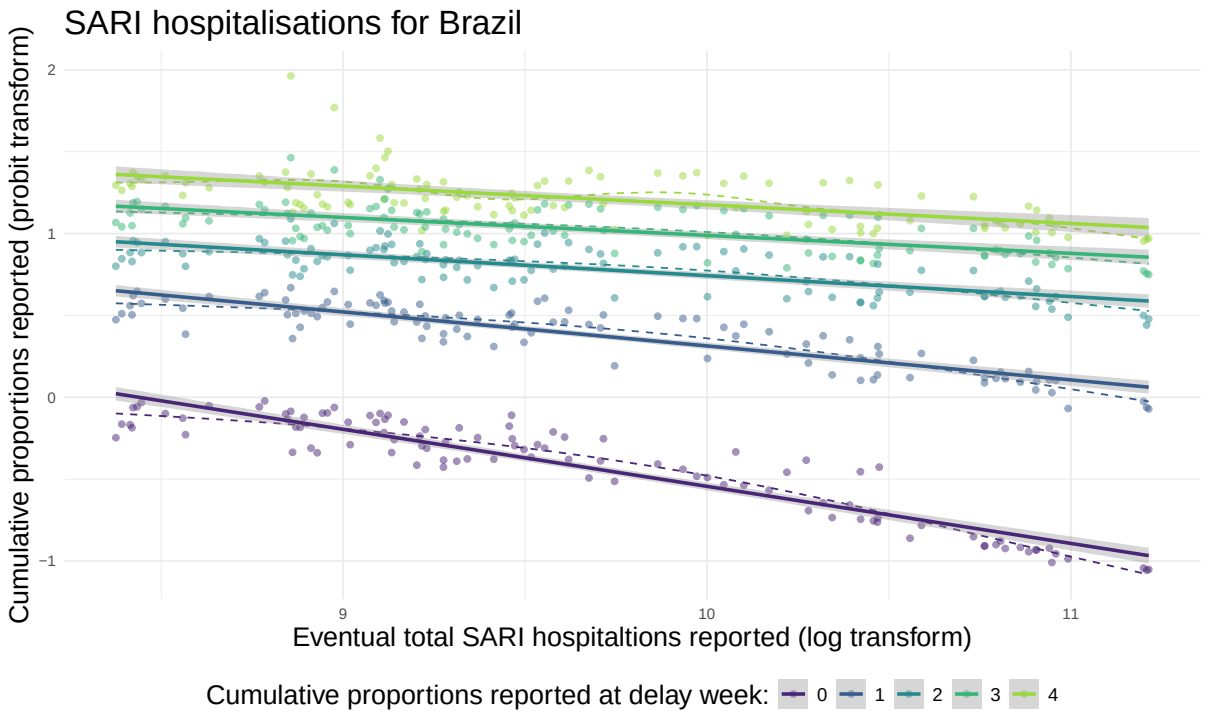

Figure 6.1: Probit-transformed cumulative proportions of SARI hospitalisations reported (probit($C_{t,d}$)) up to and including each week of delay (d), by the log of the number of eventually reported SARI hospitalisations. Points are individual weeks (t). solid lines and associated 95% confidence intervals are from linear regression fits; dashed lines are smooth thin plate splines from Gaussian additive models.

The downward trends of the linear regressions fitted to the data in Figure 6.1 suggest that, on average across Brazil, higher total SARI case counts (x-axis) are generally associated with lower cumulative proportions reported on average (y-axis); lower cumulative proportions suggests that reporting is slower as less cases are being reported after each week of delay.

However, it is worth noting that the trend demonstrated in Figure 6.1 may be a result of a confounding factor over time. As we can see in Figure 6.2, the overall trend in the total SARI hospitalisations across the whole of Brazil is decreasing over time, with each consecutive outbreak being less severe than the previous one, at a national level. This may be due to both pharmaceutical (vaccine uptake) or non-pharmaceutical (social restriction recommendations) interventions related to the COVID-19 positive SARI cases, since COVID-19 largely drove many of peaks in SARI hospitalisations within the time period covered by this data. Meanwhile, the cumulative proportions reported appear to increase steadily over time, potentially indicating long-term improvement in reporting efficiency. The long-term trends in both quantities could, therefore, suggest that an apparent negative relationship between them (in Figure 6.1) is coincidental. However, looking closely at Figure 6.2, we can see that the cumulative proportions do seem to dip below the estimated long-term trend in periods that line up with peaks in the SARI hospitalisations.

Therefore, an ideal modelling framework would be able to separate these potentially confounding or conflicting trends within the data. We investigate whether our proposed framework, introduced in Sections 6.2, is able to do so through our simulation experiments in Section 6.3.

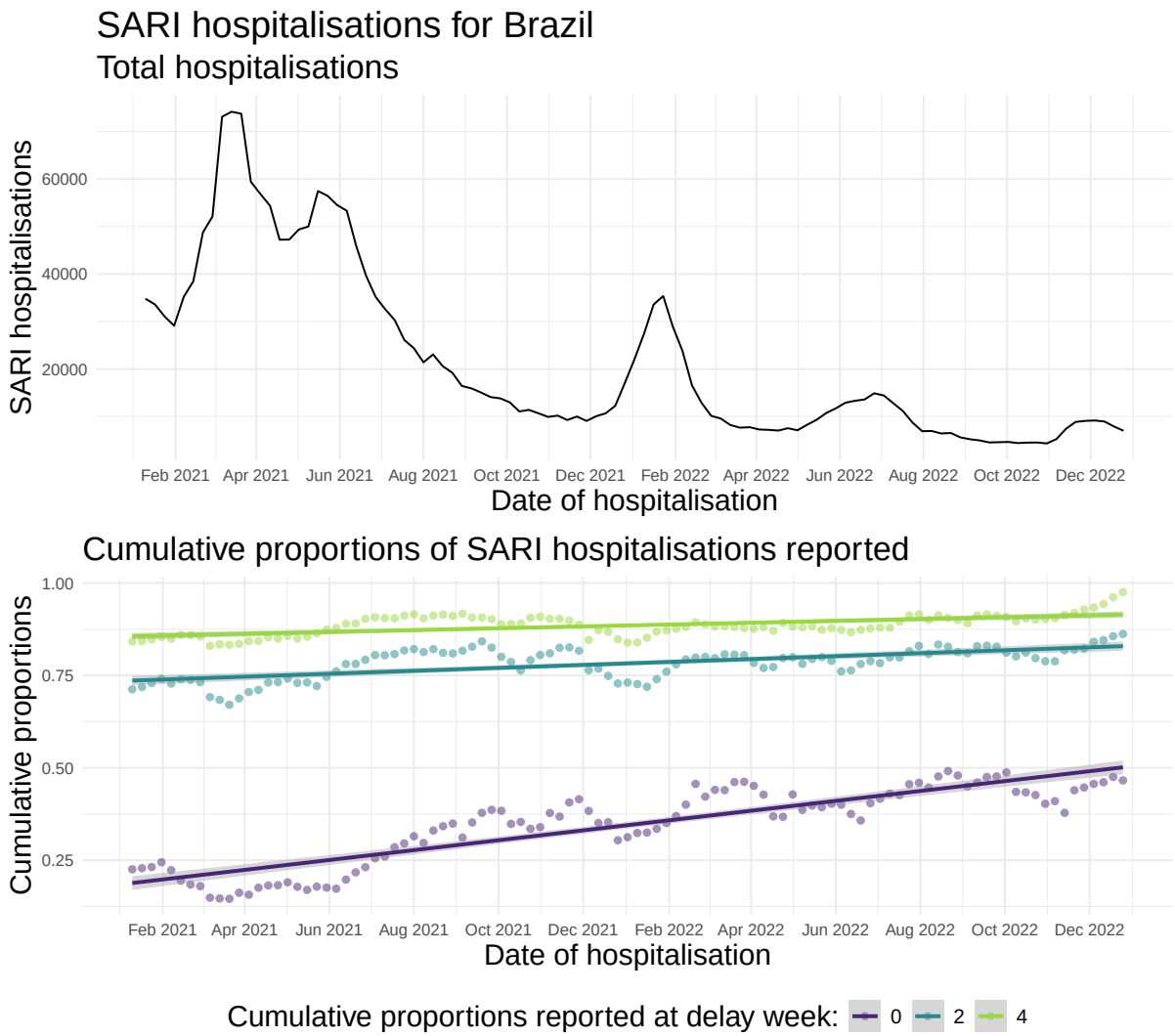

Figure 6.2: The total number of SARI hospitalisations y_t (top panel) and the cumulative proportions reported, $C_{t,d}$, (bottom panel) by date of occurrence t (x-axis). For the bottom panel, lines represent the linear regression of the cumulative proportions fitted to each delay independently, and the shaded regions represent the associated 95% confidence interval.

6.1.1.2 Arbovirus cases

Secondly, we consider the number of arbovirus cases in Brazil using data provided by the Health Problem and Notification Information System Oliveira et al. (2021). In general, the term arbovirus encompasses any virus transmitted by an arthropod vector such as mosquito or ticks (World Health Organization (WHO) (n.d.)). Here we refer to the clinical dataset in Oliveira et al. (2021) that presents confirmed patients with just the mosquito-

borne diseases Dengue or Chikungunya in Brazil, between 2013 to 2020, as well as records where the attributed virus is inconclusive/discarded. Each patient case includes the date of the start of symptoms and the notification data of the record, the difference between these two dates constructs the delay in reporting of each case. We assume that all cases are reported after a maximum delay of $D_{max} = 30$ weeks. The federative unit and city of the health facility related to the identification of each case are provided. In this section, for our exploratory analysis, we consider the cases for the whole of Brazil. In Section 6.4, we apply our proposed framework to the regional data comprising arbovirus cases for each of the 27 federative units of Brazil.

Figure 6.3 shows the probit-tranformed cumulative proportions reported (y-axis) against the log of the total cases (x-axis). For the linear regression of the proportions reported in the first delay ($d = 0$), there appears to be a positive relationship between case load and the proportions. On the other hand, the overall linear trends for the subsequent cumulative proportions reported after 2 and 4 weeks of delay are negative. However, this could be due to the cumulative proportions being more likely to be larger when the total cases are low. Looking at the smooth thin plate regression fitted by the dashed line there is evidence that, after the initial dip, the cumulative proportions start to increase for larger total counts. Alternatively, it could be that the relationship between the log-transformed total cases and probit-transformed cumulative proportions reported is non-linear.

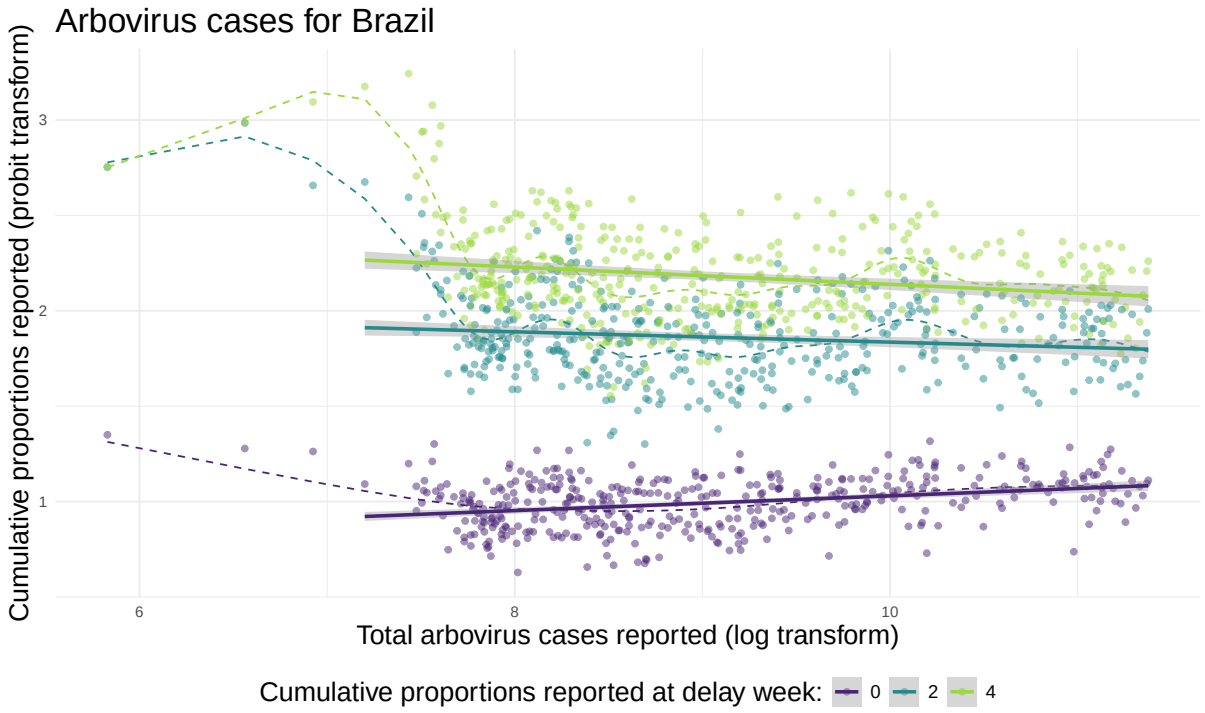

Figure 6.3: The probit-transformed cumulative proportion of arbovirus cases ($\text{probit}(C_{t,d})$) reported by each 0, 2 and 4 weeks delay (d), depicted by colour. Plotted against the log of the number of eventually reported arbovirus cases ($\log(y_t+1)$) for the given time (t). Solid lines represent the fitted values of the normal linear model for total counts, that satisfy $y_t > e^7$, fitted independently to each delay. The shaded regions represent the associated 95% confidence interval of the linear regression. Dashed lines are fitted values from a normal model with a smooth thin plate spline specified by $\text{probit}(C_{t,d}) \sim d + s(\log(y_t), k = 10, by = d)$.

Yet again, Figure 6.4 suggests that the relationship between total cases and the cumulative proportions reported may change over time instead. Here we plot the probit-transform of the proportions reported in the same week of occurrence ($d = 0$), once again, against the log of the total cases in Figure 6.4. We then fit the linear trends separately for five sections of the time series each of length 100 weeks $b_t = \{[1 \leq t < 101 \text{ weeks}], [101 \text{ weeks} \leq t < 201 \text{ weeks}], [201 \text{ weeks} \leq t < 301 \text{ weeks}], [301 \text{ weeks} \leq t < 401 \text{ weeks}], [401 \text{ weeks} \leq t < 501 \text{ weeks}]\}$.

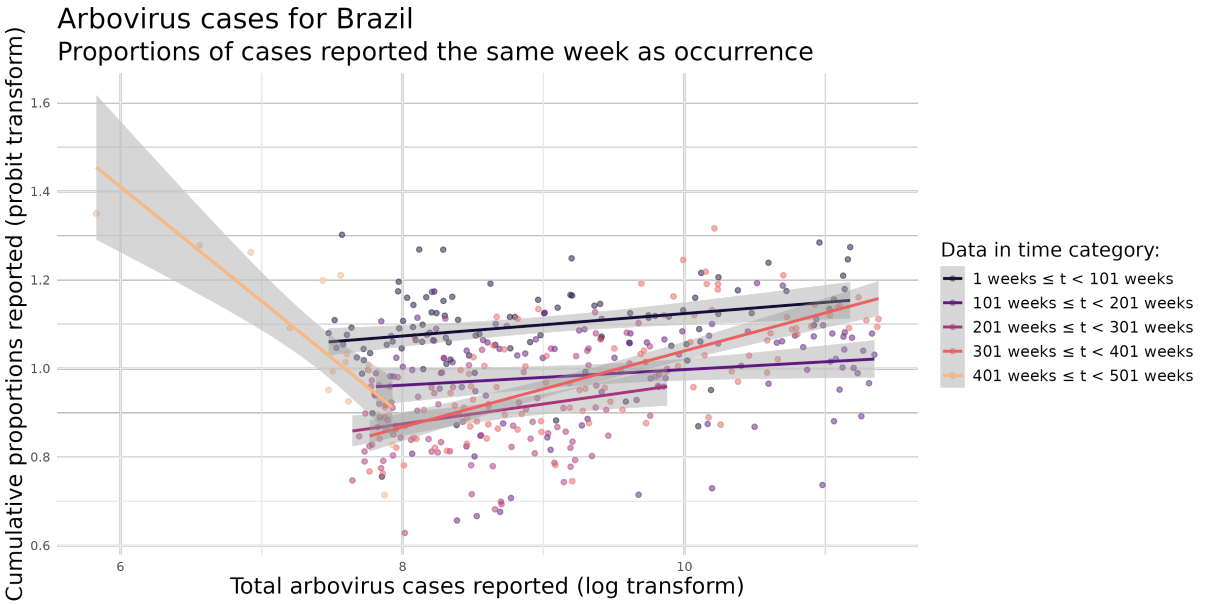

Figure 6.4: The probit-transformed cumulative proportion of arbovirus cases (probit($C_{t,d}$)) reported the same week as the cases occurred (delay $d = 0$) plotted against the log of the number of eventually reported arbovirus cases ($\log(y_t + 1)$). Colour indicates the time category (b_t) that the arbovirus cases fall in measured in weeks from the start of the time series (January 1^{st} 2013). Solid lines represent the fitted values of the normal linear model for each time category and the shaded regions represent the 95% confidence interval.

It is clear from Figure 6.4 that the most recent observed cumulative proportions reported (in the [401 weeks $\leq t <$ 501 weeks] category) have much lower total arbovirus cases reported (x-axis) and appear to have a negative relationship with the magnitude of the proportions reported in the same week as cases occurred (y-axis). On the other hand, all previous time categories appear to have a positive relationship between case load and the proportions reported. A positive relationship suggests for higher log-transformed total arbovirus cases, the probit-transformed cumulative proportions are greater, so more cases are reported after each delay and thus reporting is faster on average.

The scientist at Fiocruz who played a key role in designing the operational surveillance systems InfoDengue and InfoGripe stated to us that public health institutes in Brazil hire additional admin personnel during summer months. In general, arbovirus cases have strong seasonality: in summer, warmer conditions lead to vector breeding and more infections from mosquitoes and ticks (World Health Organization (WHO) (n.d.)). As such, the

additional staff are recruited at this time of year to help manage the reporting/processing of the anticipated higher dengue cases. This may also explain why the relationship between case load and reporting delays of arbovirus cases (which includes dengue cases) does not appear to be as clear as with SARI hospitalisations. There may be two counteracting trends during a surge in arbovirus cases; longer delays could result from the capacity of the reporting system being reached, but otherwise the hiring of additional staff could result in shorter delays. There could also be random variability within this latter trend, depending on when the new hires start and the numbers and efficiency of those that have been hired. This could result in distinct and non-linear trends over time, delay and spatial regions which are hard to capture.

6.2 General framework

The goal of this work is to develop a framework that can correct infectious disease data, where the overall case load directly impacts the length of reporting delays. This is non-trivial to implement into nowcasting models, since the total counts that we may wish to use as an explanatory variable for the reporting delay may not yet be reported.

However, we can frame this problem as an adaption of the modular framework for correcting delayed reporting first introduced in Chapter 1.2 and further explored in the context of joint models in Chapter 2.2.1. Recall that we assume the total counts y_t come from some probabilistic process Y, depending on parameters and random effects $\boldsymbol{\theta}$.

$$y_t \sim Y(\boldsymbol{\theta}). \tag{6.2}$$

Then, a second probabilistic process Z translates the total counts into the partial counts $\boldsymbol{z}_t$, based on some parameters $\boldsymbol{\pi}$:

$$\boldsymbol{z}_t \mid y_t \sim Z(\boldsymbol{\pi}, y_t). \tag{6.3}$$

In the Generalized-Dirichlet Multinomial (GDM) framework, Y is a Negative-Binomial, and Z is a GDM model, given total counts y_t. Conceptually, we could account for a case load effect on the reporting delay, within this framework, by allowing for the potential inclusion of total counts y_t in the parameters of the GDM, rather than just the mean.

This extension is possible for two reasons: first, recall that the Bayesian implementation of the GDM framework in MCMC means that samples for unobserved y_t are generated during the algorithm, meaning they are available for use as explanatory variables in other parts of the model; second, implementation using the NIMBLE R package (de Valpine et al. (2017)) offers great flexibility in considering different adaptations to explicitly add a link between case load and delay.

Two versions of the GDM framework are introduced in Stoner and Economou (2019); the hazard and the survivor. Here we opt for the survivor version, given by:

$$y_t|\lambda_t, \theta \sim \text{Negative-Binomial}(\lambda_t, \theta) \tag{6.4}$$

$$\log(\lambda_t) = f(t) \tag{6.5}$$

$$z_{t,d}|\nu_{t,d}, \phi_{t,d}, y_{t,d}, z_{t,1:(d-1)} \sim \text{Beta-Binomial}\left(\nu_{t,d}, \phi_d, y_t - \sum_{j=1}^{d-1} z_{t,j}\right) \tag{6.6}$$

$$probit(S_{t,d}) = g(t,d) \tag{6.7}$$

$$\nu_{t,d} = \frac{S_{t,d} - S_{t,d-1}}{1 - S_{t,d-1}} \tag{6.8}$$

$$S_{t,d} = E\left[\frac{\sum_{j=1}^{d} z_{t,j}}{y_t}\right], \tag{6.9}$$

which models the expected cumulative proportions reported ($S_{t,d}$), instead of directly modelling the expected relative proportions ($\nu_{t,d}$) as in the hazard version. We focus on modelling the cumulative proportions, motivated by the patterns seen in our exploratory analysis, as summarised in Section 6.1.1. The Beta-Binomial distribution for the partial counts has dispersion parameter ϕ_d and the Negative-Binomial model for the total counts has dispersion parameter θ. In practice these can be assigned any prior or model effects to capture the dispersion in the partial counts and total counts respectively. Here, we assign both a Gamma$(2, 0.02)$ priors for all implementations of our framework, this reflects a prior we would choose in operational settings, as with a mean of 100 and variance of 5000, it is a relatively uninformative prior which captures a wide range of plausible dispersions.

Recall that $f(t)$ and $g(t,d)$ represent general functions of time and delay, which could include linear, non-linear and covariate effects. To achieve a general framework that captures the relationship between case load and reporting delay, i.e. explicitly including total counts y_t in the delay distribution $p(z|y)$, we introduce a new general function $h(y_t)$ in the probit model for the expected cumulative proportion reported, $S_{t,d}$, such that:

$$\text{probit}(S_{t,d}) = g(t,d) + h(y_t) \tag{6.10}$$

The function $h(y_t)$ thus links expected reporting delay lengths to the total counts y_t. Since, this work uses the survivor version of the GDM, $h(y_t)$ is included in linear predictor for the expected cumulative proportions $S_{t,d}$. In theory the function $h(y_t)$ could be any type of modelling effect to capture the relationship between the total number of cases and the delay in reporting, which may be linear or a more complicated non-linear structure.

6.3 Simulation Experiments

In this section we present the design and results of two simulation experiments. The aim of the first simulation study is to determine the effectiveness of our proposed framework in correctly capturing the relationship between the total cases of a disease and the delay in reporting those cases, in the context of temporal confounding. For the second study, we aim to assess predictive accuracy when modelling the effect of case load on reporting delays, compared to existing frameworks that don't account for this.

6.3.1 Data generation

We use the same series of data sets for both simulation experiments, which we design to reflect possible real-world scenarios where the total number of cases that occur on a given day directly affect the reporting efficiency for that count. This allows us to determine the feasibility of our framework for conducting inference on pre-determined parameter values and predictions of simulated total counts.

We must first specify how we generate our simulated data sets. We set the number of time points in the series to be $T = 100$ (e.g. weeks), and set a maximum of $D = 7$ reporting delays, which correspond to cases reported the same week they occur ($d = 0$) and up to 6 weeks after the week they occurred ($d = 1, ..., 6$).

First, we randomly generated dispersion parameter $\theta \sim \text{Gamma}(2, 0.02)$, this reflects the choice of prior for this parameter as discussed in Section 6.2. This allows us to generate the total counts from ag Negative-Binomial model, with a different dispersion parameter θ for each simulation:

$$y_t \mid \lambda_t, \theta \sim \text{Negative-Binomial}(\lambda_t, \theta); \tag{6.11}$$

$$\log(\lambda_t) = \iota + \alpha_t + \zeta t^*. \tag{6.12}$$

To specify the smooth temporal trend in the mean of the total cases (λ_t) we first generate a thin plate spline (α_t),

$$\alpha_t = \boldsymbol{X}_t \boldsymbol{\kappa}_\alpha \tag{6.13}$$

$$\boldsymbol{\Omega}_\alpha = \boldsymbol{S}_1/\sigma^2_{\alpha,1} + \boldsymbol{S}_2/\sigma^2_{\alpha,2} \tag{6.14}$$

$$\boldsymbol{\kappa}_\alpha = \text{Multivariate-Normal}(0, \boldsymbol{\Omega}_\alpha). \tag{6.15}$$

which has a zero mean and no overall linear temporal trend. For all simulations, we set $\sigma_{\alpha,1} = 5$ and $\sigma_{\alpha,2} = 5$. The thin plate spline basis matrix $\boldsymbol{X}_t$, and respective penalty matrices $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$ are generated using the `jagam(.)` function (Wood (2016)). We then combine this with an intercept term ι (simulated from $\iota_{1:D} \sim \text{Normal}(5, 0.25^2)$, such that the mean of the total counts are all centred around a mean of approximately 150) and a linear temporal trend, ζt^*, where $t^* = \{\frac{t - \text{mean}(1:T)}{\text{sd}(1:T)}; t = 1, 2, ..., T\}$ is a scaled and centred time variable. Four examples of the resulting mean temporal trends, λ_t, for $\zeta \in \{-0.1, 0, 0.1\}$ respectively, are given in Figure 6.5.

We then simulate partial counts reported at each delay d $z_{t,d}$ from a GDM model, with dispersion parameters simulated from $\phi_d \sim \text{Gamma}(2, 0.02)$, which is the same distribution as our prior choice for ϕ_d. Here we represent the GDM as a series of Beta-Binomial distributions for the partial counts $z_{t,d}$, where $N_{t,d} = y_t - \sum_{j=1}^{d-1} z_{t,j}$ are the remaining counts

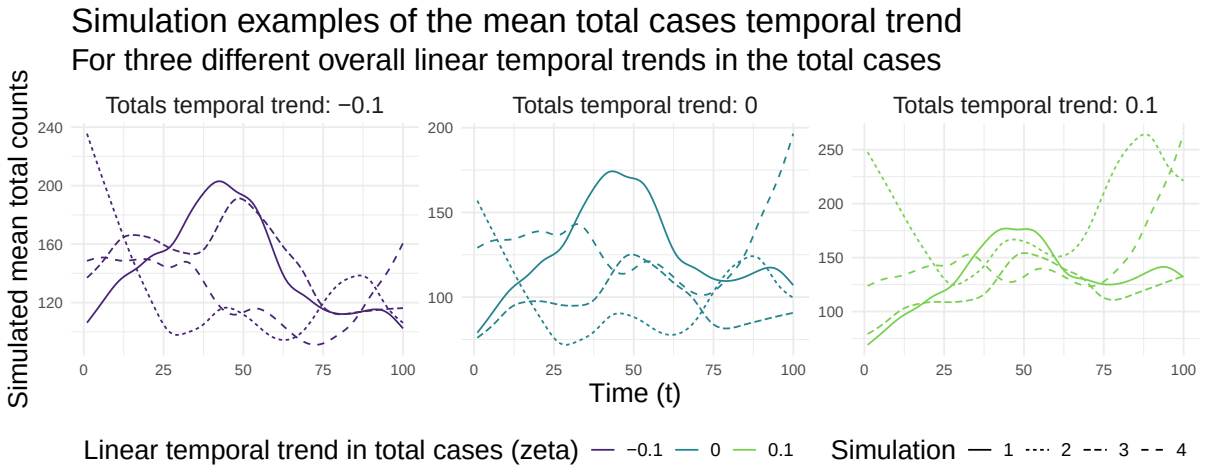

Figure 6.5: The three scenarios, where the overall temporal trend is chosen to be $\zeta \in \{-0.1, 0, 0.1\}$ respectively, for the simulated mean of the total cases ($\log(\lambda_t) = \iota + \alpha_t + \zeta t^*$) and the zero-mean trend (α_t) is a randomly generated thin plate spline.

yet to be reported at delay d:

$$z_{t,d} | \nu_{t,d}, \phi_d, N_{t,d} \sim \text{Beta-Binomial}\left(\nu_{t,d}, \phi_d, N_{t,d} = y_t - \sum_{j=1}^{d-1} z_{t,j}\right) \tag{6.16}$$

$$\nu_{t,d} = \frac{S_{t,d} - S_{t,d-1}}{1 - S_{t,d-1}} \tag{6.17}$$

$$\text{probit}(S_{t,d}) = \psi_d + \eta t^* + \delta y_t^*. \tag{6.18}$$

We opt for the survivor version of the GDM model, introduced in Stoner and Economou (2019), as such the cumulative proportions of the total cases y_t expected to be reported at each delay $S_{t,d}$ are modelled with a probit link function. In Equations (6.16) - (6.18), ψ_d is a intercept term which monotonically increases over delay (as $S_{t,d} > S_{t,d-1}$ for all d). These intercepts are simulated using the distribution $\psi_{1:D} \sim \text{Normal}(0.25, 0.5^2)$, where the D generated values are then sorted in increasing order. This is followed by a linear temporal trend in the reporting rate, characterised by coefficient η multiplied by the scaled and centred time.

Finally, we simulate a linear relationship between the total counts and the delay, this was chosen to initially assess the performance of our framework for a relatively simple correlation before potentially extending it to more complex relationships. Moreover, the exploratory analysis in Section 6.1.1 suggested that a linear relationship may be present

in the two case studies we have considered. To simulate the linear relationship between y_t and the probit-transformed expected cumulative proportion reported $S_{t,1:D}$, we multiply a scaling coefficient δ by $y_t^* = \frac{y_t - \text{mean}(y_{1:T})}{\text{sd}(y_{1:T})}$, a scaled and centred version of the total counts. We used this version of y_t to ensure that the difference in cumulative proportions reported are not systematically higher for positive $\delta = 0.2$ and lower for negative $\delta = -0.2$, skewing our results. Hence, in Figure 6.6, the simulated cumulative proportions reported for both positive and negative incidence-delay effect coefficients are centred around the corresponding cumulative proportions with no incidence-delay effect ($\delta = 0$).

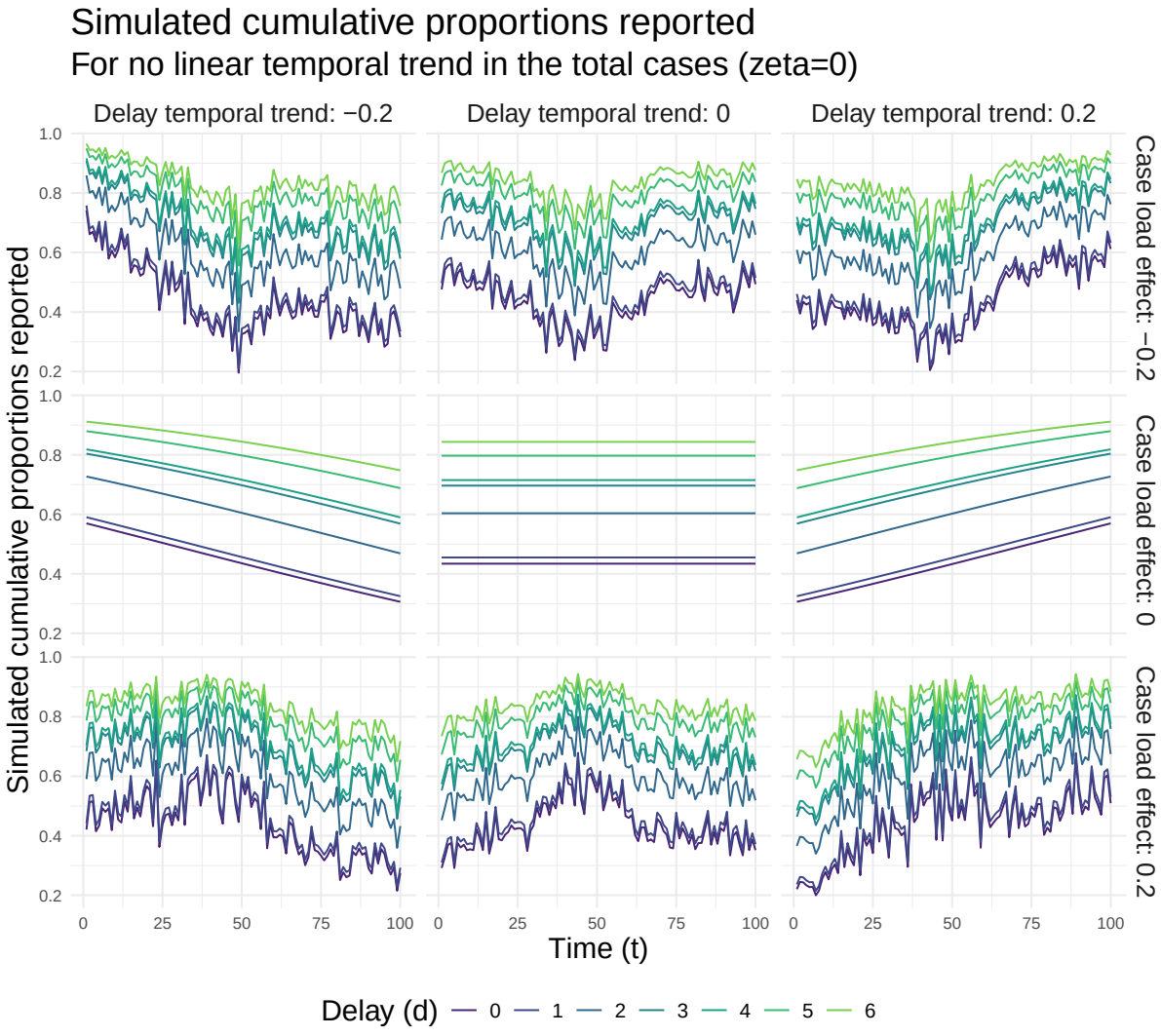

Figure 6.6: The nine scenarios for the temporal trend in the simulated cumulative proportions reported for each delay (given by the colours). Scenarios cover all combinations of the chosen overall temporal trend, $\eta \in \{-0.2, 0, 0.2\}$, and chosen incidence-delay effect on the cumulative proportions, $\delta \in \{-0.2, 0, 0.2\}$ given the linear temporal trend in the totals is $\zeta = 0$.

Our simulation experiment focuses on inference and prediction in the context of potentially confounding trends. As such, we consider different fixed combinations of three of the terms described above: (i) coefficient for trend in the total cases ζ, (ii) the coefficient for the trend in the cumulative proportions reported η, and (iii) the coefficient for the effect of the total counts on the cumulative proportions reported δ. These terms are summarised below for reference:

- ζ: the coefficient of the linear temporal trend in the (log) mean number of hypothetical infectious disease cases. This represents an overall trend of the number of cases increasing or decreasing on average over scaled time.
- η: the coefficient of the linear trend in the cumulative proportions reported over scaled time. Note that the reporting process generally improves over time if $\eta > 0$, the reporting process worsens over time if $\eta < 0$, and the reporting process doesn't change over time on average if $\eta = 0$ (except as a result of the case load effect).
- δ: the incidence-delay coefficient of the linear effect of the scaled total cases on the probit-transformed expected cumulative proportions reported. Note that higher case loads result in: longer reporting delays (smaller cumulative proportions reported) when $\delta < 0$; shorter reporting delays (larger cumulative proportions reported) when $\delta > 0$; and there is no relationship between case load and reporting delay length when $\delta = 0$.

We consider three fixed values for each parameter of interest and all possible combinations between parameters. Hence, evaluating a total of 27 scenarios. For each variable we consider a positive, negative and zero trend. Specifically, we assign $\zeta \in \{-0.1, 0, 0.1\}$, $\eta \in \{-0.2, 0, 0.2\}$ and $\delta \in \{-0.2, 0, 0.2\}$. These values were chosen to give the simulations realistic looking trends that we would expect to see in infectious disease data. They have similar magnitudes, where η and δ are slightly bigger since they are included in a probit link instead of a log link within the model, to ensure they all have a similar impact on the model when considered simultaneously.

We repeat the whole simulation procedure described above 100 times for each of the 27 scenarios to introduce randomness into the simulated data sets. Thus, we have a total of 2700 simulated data sets. For both subsequent simulation experiments, the MCMC method for fitting all models used default NIMBLE (de Valpine et al. (2017)) samplers apart from for a single AF slice sampler that was assigned to the parameters; η, $\psi_{1:D}$, $\sigma_{alpha_{1:2}}$, κ_{alpha}, ι and δ (if included in the model). Two MCMC chains, with a iteration length of $20,000$, a burn-in length of $15,000$ and thinning of 5 were run for each model and simulated data set for both experiments. Convergence was assessed by visually inspecting trace plots and calculating the PSRF (defined by Equation (2.5)) of all model parameters.

6.3.2 Parameter inference experiment

To investigate the ability our proposed framework (Equations (6.4)–(6.10)), to correctly re-capture all the simulated trends, including the case load effect, we assess results from fitting approximately the same model we simulated data from.

In terms of our general framework (Section 6.2), we set $f(t) = \iota + \alpha_t$, where ι is an intercept and α_t is a thin plate spline. Note that α_t can have an overall linear trend, meaning we would expect it to absorb ζt^*. To obtain approximate values of ζ, we fit a linear regression to the posterior samples for α_t, with t^* as the only covariate, and take the slope coefficient. We set the priors for the Negative-Binomial model of the total counts as:

$$\iota \sim \text{Normal}(5, 10^2), \tag{6.19}$$

$$\boldsymbol{\kappa}_\alpha \sim \text{Multivariate-Normal}(0, \boldsymbol{\Omega}_\alpha), \tag{6.20}$$

$$\sigma_{\alpha,1}, \sigma_{\alpha,2} \sim \text{Half-Normal}(0, 10^2), \tag{6.21}$$

$$\theta \sim \text{Gamma}(2, 0.02). \tag{6.22}$$

For the expected cumulative proportions we set $g(t,d) = \psi_d + \eta t^*$, where ψ_d is a random intercept restricted to monotonically increase over delay d and η is a slope coefficient for the scaled time variable t^*. Finally, for the relationship between case load and reporting delay, we fit $h(y_t) = \delta y_t^*$, such that δ is a linear slope coefficient for the scaled total count $y_t^* = \frac{y_t - \text{mean}(y_{1:T})}{\text{sd}(y_{1:T})}$. This is representative of how the data sets have been generated and allows for a direct comparison of the simulated parameters of interest and the model predictions of these parameters. Hence, the priors are:

$$\psi_1 \sim \text{Normal}(-1.5, 5^2), \tag{6.23}$$

$$\psi_d \sim T[\text{Normal}(\psi_{d-1}, 5^2), \psi_{d-1}, \infty], \tag{6.24}$$

$$\eta \sim \text{Normal}(0, 10^2), \tag{6.25}$$

$$\delta \sim \text{Normal}(0, 10^2), \tag{6.26}$$

$$\phi_d \sim \text{Gamma}(2, 0.02). \tag{6.27}$$

Where $T[\text{Normal}(\psi_{d-1}, 5^2), \psi_{d-1}, \infty]$ denotes a Normal distribution that has been truncated between ψ_{d-1} and infinity.

6.3.2.1 Results

Recall that we are considering three parameters of interest, each with three possible chosen values ($\zeta \in \{-0.1, 0, 0.1\}$, $\eta \in \{-0.2, 0, 0.2\}$, $\delta \in \{-0.2, 0, 0.2\}$). Thus, we have 27 scenarios, each with a 100 generated data sets. To summarise parameter inference results, we calculate three main performance metrics for each parameter. The first statistic is the "mean prediction interval width" (PIW), defined by the mean difference between the upper 97.5% and lower 2.5% quantile of the posterior samples. This quantifies the precision/uncertainty of our estimates. The second statistic is the 95% interval "coverage", defined here as the percentage of the true simulated values that fall in the 95% uncertainty interval interval from the posterior samples. This measures how often the model is

able to capture the simulated values in its 95% uncertainty intervals. The final statistic is the "bias", defined as the mean of the true simulated parameter values minus the median of the posterior samples for those parameters. This summarises whether inference is systematically incorrect, i.e. too high or too low compared to the simulated value.

We compute these three summaries across all 100 simulations for a given scenario, determined by the 27 combinations of chosen parameter values. Table 6.1 summarises these three performance metrics, for each of the 27 scenarios, for δ, η and ζ.

Performance metrics for the three parameters of interest (δ,η,ζ).

Table 6.1: Coverage, prediction interval width and bias for the model posterior samples compared to the simulated values of; the linear relationship between the simulated number of cases and the expected cumulative proportion of cases reported at each delay (delta δ), the linear temporal trend in the expected cumulative proportions reported at each delay (eta η), the zero-mean temporal spline of the total number of cases y_t (zeta ζ).

Simulated zeta	Simulated eta	Simulated delta	Coverage (zeta)	PIW (zeta)	Bias (zeta)	Coverage (eta)	PIW (eta)	Bias (eta)	Coverage (delta)	PIW (delta)	Bias (delta)
-0.1	-0.2	-0.2	0.85	0.069	0.000	0.93	0.063	-0.001	0.97	0.061	-0.001
-0.1	-0.2	0	0.93	0.067	0.002	0.96	0.061	0.000	0.91	0.059	0.001
-0.1	-0.2	0.2	0.91	0.065	0.002	0.95	0.061	-0.003	0.94	0.061	0.002
-0.1	0	-0.2	0.90	0.069	0.000	0.96	0.061	0.001	0.95	0.059	-0.001
-0.1	0	0	0.91	0.066	0.001	0.93	0.063	0.000	0.93	0.061	0.001
-0.1	0	0.2	0.93	0.066	0.000	0.91	0.062	-0.001	0.94	0.062	0.001
-0.1	0.2	-0.2	0.94	0.069	-0.002	0.95	0.064	-0.001	0.96	0.061	-0.001
-0.1	0.2	0	0.88	0.065	0.002	0.92	0.062	0.000	0.98	0.060	-0.001
-0.1	0.2	0.2	0.92	0.069	0.001	0.93	0.062	0.001	0.97	0.060	0.000
0	-0.2	-0.2	0.91	0.066	0.000	0.93	0.060	-0.001	0.94	0.058	0.001
0	-0.2	0	0.86	0.070	-0.005	0.97	0.059	0.001	0.98	0.057	0.000
0	-0.2	0.2	0.91	0.069	0.000	0.90	0.060	-0.001	0.94	0.059	-0.003
0	0	-0.2	0.92	0.068	-0.002	0.96	0.061	-0.002	0.95	0.059	0.001
0	0	0	0.92	0.066	0.000	0.97	0.059	0.001	0.93	0.057	-0.001
0	0	0.2	0.87	0.067	-0.002	0.96	0.058	0.002	0.94	0.057	0.000
0	0.2	-0.2	0.89	0.066	0.002	0.92	0.060	0.001	0.98	0.058	0.002
0	0.2	0	0.88	0.066	0.003	0.94	0.061	0.001	0.94	0.059	0.000
0	0.2	0.2	0.91	0.063	-0.001	0.95	0.060	0.001	0.95	0.059	-0.002
0.1	-0.2	-0.2	0.90	0.065	0.000	0.95	0.062	-0.001	0.96	0.061	-0.001
0.1	-0.2	0	0.90	0.069	-0.001	0.95	0.062	-0.001	0.89	0.060	-0.003
0.1	-0.2	0.2	0.91	0.071	-0.001	0.92	0.063	-0.001	0.97	0.062	-0.001
0.1	0	-0.2	0.90	0.071	-0.001	0.95	0.064	0.000	0.93	0.062	-0.003
0.1	0	0	0.97	0.070	-0.001	0.94	0.062	0.000	0.94	0.060	0.001
0.1	0	0.2	0.92	0.070	0.003	0.93	0.060	0.000	0.96	0.060	0.002
0.1	0.2	-0.2	0.87	0.067	-0.003	0.94	0.061	0.000	0.96	0.059	0.000
0.1	0.2	0	0.95	0.069	0.002	0.93	0.062	0.001	0.94	0.060	0.000
0.1	0.2	0.2	0.91	0.068	0.000	0.95	0.060	0.001	0.95	0.059	-0.001

The performance metrics given in Table 6.1 suggest that the model is able to identify and separate the trends in the different parts of the data generating process, allowing for accurate model inference on parameter values. This is evident from very low bias values (about two orders of magnitudes smaller than the simulated coefficient values). The coverages are broadly close to 95%, across all parameters of interest and combinations of their chosen values, suggesting appropriate quantification of uncertainty. Furthermore, Figure 6.7 shows there is no obvious patterns in the bias for each of the combination of chosen parameters values. Hence, the model should be able to separate possible trends in the reporting delay (η), the trend in the total cases (ζ), and the effect of the total cases on the reporting delay (δ). Furthermore, this should be regardless of any combination of compounding or cancelling trends between the three parameters.

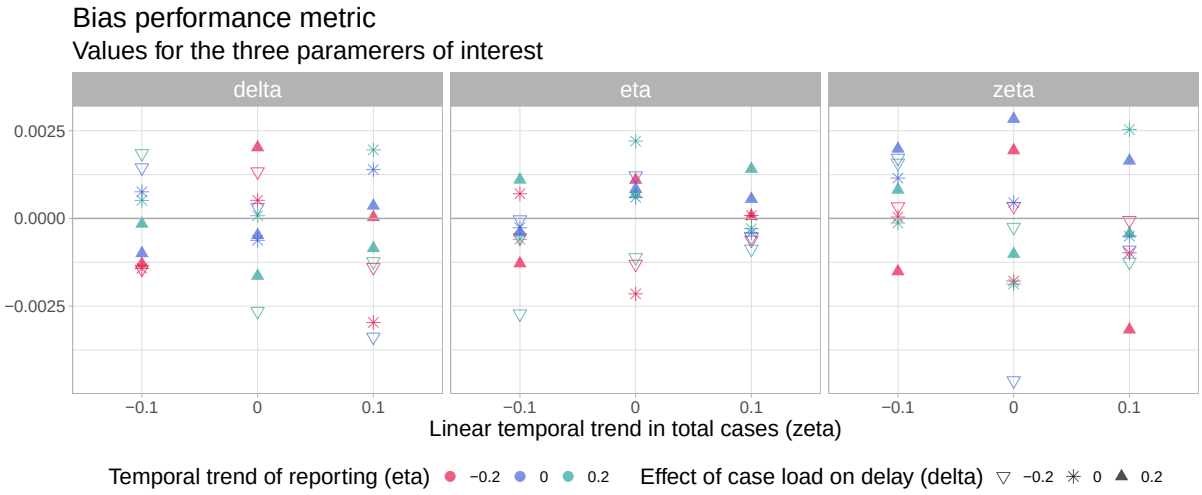

Figure 6.7: The mean bias between simulation posterior medians and the true chosen parameter values for the effect of the total cases on the reporting delay δ (right), the temporal trend in the reporting delay η (centre) and the linear temporal trend in the total cases ζ (left). For each parameter the bias is calculated separately for the 27 combinations of chosen $\delta \in \{-0.2, 0, 0.2\}$, $\eta \in \{-0.2, 0, 0.2\}$ and $\zeta \in \{-0.1, 0, 0.1\}$ values, across the 100 simulated data sets. The x-axis denotes the chosen values of ζ, the colour denotes the chosen values for η and the shape denotes the chosen values for δ.

6.3.3 Prediction performance experiment

To explore the potential gain in predictive performance when a case load effect is appropriately accounted for, for each simulated data set we also carry out a prediction experiment. Hence, we censor each of the simulated data sets such that only the cases that would have been reported by the final time step T_{now} in the series are observed, denoted by $z_{t,d\leq T_{now}-t}$ for $t = (T_{now} - D_{max}), \ldots, T_{now}$. We then fit the following three model versions of our proposed general framework given in Section 6.2.

1. **Survivor model:** Firstly, with $\text{probit}(S_{t,d}) = g(t,d) = \psi_d + \eta t^*$, such that there is no link $h(y_t)$, thus the total number of cases does not inform the length of the reporting delays. This is equivalent to the original GDM model, and allows us to determine if not accounting for the correlation that is present in the simulated data sets will hamper prediction accuracy or precision.
2. **Survivor incidence-delay model:** The model is then fit a second time, with $\text{probit}(S_{t,d}) = g(t,d) + h(y_t) = \psi_d + \eta t^* + \delta \log(y_t + 1)$, such that we are directly modelling the relationship between the total counts and the cumulative proportions reported.This differs from how the data were simulated: recall that the simulated effect was δy_t^*, where y_t^* are scaled total counts. Real-world application of the model for nowcasting involves prediction of unknown y, so consistent scaling is not possible. This version of the model, with $\delta \log(y_t + 1)$, avoids scaling, is less sensitive to extreme case levels due to the log function compared to the alternative δy_t, and is still suitable where $y_t = 0$. Meanwhile, since we are focusing on predictive precision of the total counts in this simulation experiment, it is not imperative that our parameters of interest are comparable to their simulated values. In summary, the model is slightly mis-specified, which reflects the likely result of model design for any real-world data.

3. **Survivor incidence-delay (delta fixed) model:** Finally, we fit the model with $\text{probit}(S_{t,d}) = g(t,d) + h(y_t) = \psi_d + \eta t^* + \delta_{sim} y_t^*$, where δ_{sim} is a constant in the model fixed at the simulated values of delta. Comparing this model to both previous versions allows us to check that including the case load effect does improve predictive performance when exactly captured. Furthermore, this model acts as a "best-case" scenario against which we can compare the mis-specified "Survivor incidence-delay" model to.

As in our previous simulation experiment model in Section 6.3.2, we set $f(t) = \iota + \alpha_t$, where ι is an intercept and α_t is a thin plate spline. All priors for this experiment are the same as for the previous experiment (discussed in Section 6.3.2), as they are once again chosen to reflect the data generating process behind the simulations.

6.3.3.1 Results

To gauge the difference between predictive performance of the three models fitted for our second simulation experiment, outlined in Section 6.3.3, we calculate three performance metrics for the predictions of the total counts y_t. The first metric is the mean absolute error (MAE), defined as the absolute difference of the posterior median predictions minus the true simulated total counts. The second metric is the mean coverage and third is the average prediction interval width (PIW), both as defined in the first simulation experiment but applied to the posterior predictions and simulated true values of the unobserved total counts. These metrics are all calculated for each prediction time difference, such that 0 represent predictions made for the current time period ($T_{now} = 100$) and negative numbers represent time points previous to this. Note that, more partial counts will be available the further back in time you go. Furthermore, in Figure 6.8, we calculate all metrics (rows) across simulations grouped by each possible case load effect value (columns).

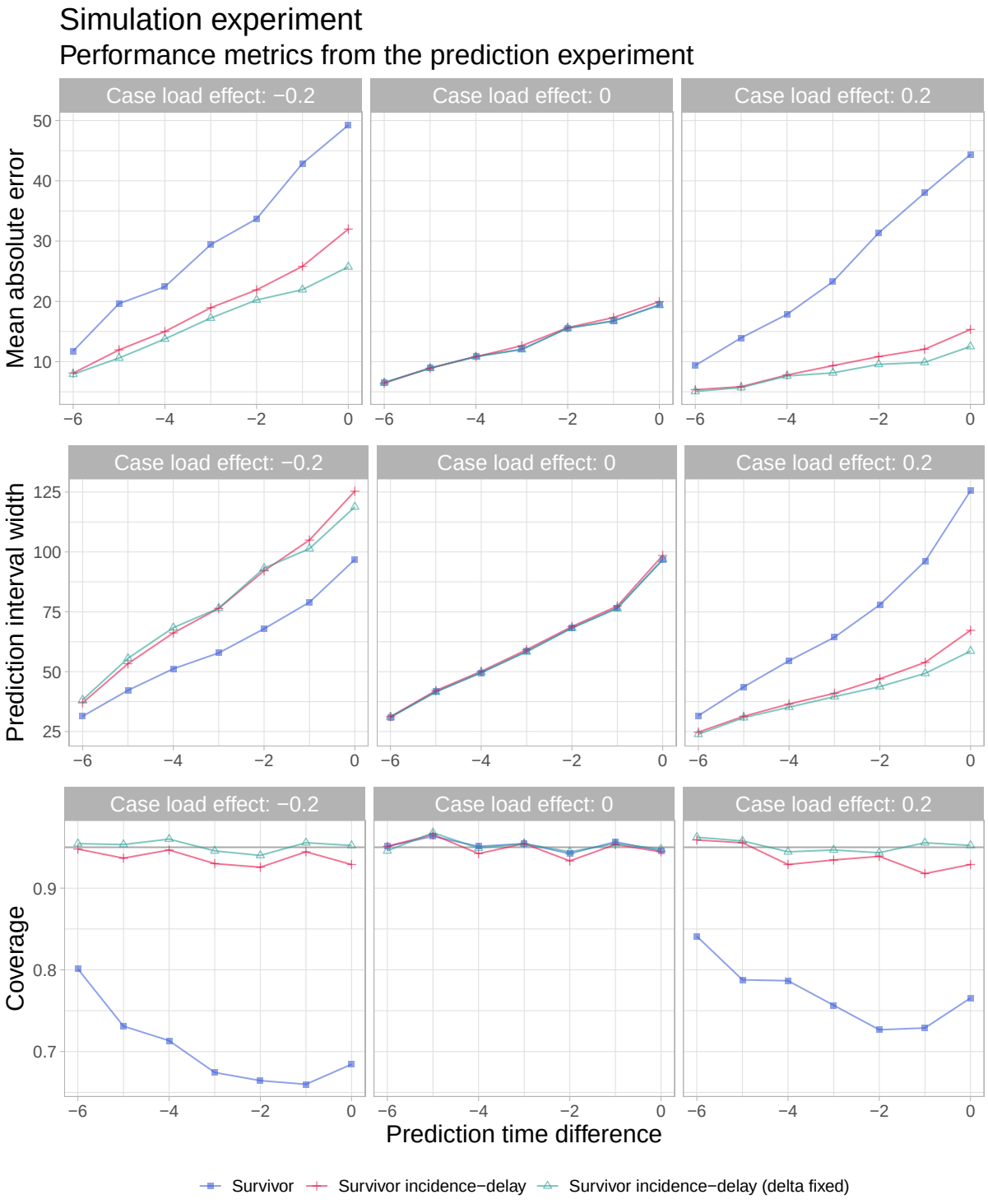

Figure 6.8: The mean absolute error (top panel), prediction interval width (middle panel) and coverage (bottom panel) for model predictions of the simulated total counts. The three models considered in this simulation experiment are the GDM Survivor model (with no incidence-delay effect), the GDM survivor incidence-delay model and the GDM survivor incidence-delay model (delta fixed) where the case load effect coefficient (δ) is set to the chosen values. Models are indicated by both line colour and shapes.

The results in Figure 6.8 clearly show that both the mean absolute error and coverage of the predictions of the unknown total counts is better for the models that include a relationship between the total counts and the delay length, regardless of whether there is a positive or negative case load effect. This indicates that ignoring the case load effect when it is present in the data, such as the "survivor model", could be detrimental to predictive precision of nowcasts and result in higher mean absolute errors and lower coverage. Furthermore, the "survivor incidence-delay" model – where the incidence-delay coefficient is an unknown variable in the model – performs similarly well to the "survivor incidence-delay (delta fixed)" model – where the incidence-delay coefficient is fixed to the known chosen values. This suggests that our proposed framework is able to capture the true relationship between incident and delay well.

However, the prediction interval width of the nowcasts are only smaller than the survivor model (with no incidence-delay effect) when the case load effect is positive ($\delta = 0.2$). The reason for this is potentially due to a negative case load effect ($\delta = -0.2$) introducing identifiability issues into the model: when the case load effect is negative, high disease incidence creates less efficient reporting, and low disease incidence creates more efficient reporting. Therefore, if a smaller than average number of partial counts have been observed so far this could either be a result of a low number of eventual total counts with efficient reporting, or due to a high number of total counts with inefficient reporting. Hence, this introduces more uncertainty into the model.

However, when the model ignores this relationship by not modelling the incidence-delay effect between total counts and reporting delay, this non-identifiability is no longer creating uncertainty in the modelling framework and the prediction interval widths are narrower. Although, removing this link also means the model is no longer able to determine from previous trends which event is more likely, so prediction accuracy is worse as indicated by larger mean absolute errors. Similarly, the 95% coverage of the survivor model with no incidence-delay link is lower than the model with the incidence-delay link for both

a positive or negative case load effect, and far below (65%-85%) what would normally be considered acceptable for a 95% interval. Simply put, even when the case load effect introduces extra uncertainty, its entirely needed to appropriately capture the variability in the data.

On the other hand, when the case load effect is positive high disease incidence creates more efficient reporting and low incidence creates less efficient reporting. Hence, lower observed partial counts occur when total counts are low as reporting will also be less efficient and higher observed partial counts occur when total counts are high and reporting is more efficient. Therefore, a positive case load effect does not introduce identifiability issues and the nowcasts of the total counts have a smaller prediction interval width when the link between case load and reporting delay is included in the modelling framework.

Finally, it is worth noting that the results in Figure 6.8 also suggest that performance of our proposed framework "survivor incidence-delay" is comparable to that of the "Survivor" model when there is no case load effect in the data generating process ($\boldsymbol{\delta} = 0$). This is an important result when considering operational surveillance systems as it suggests that nowcasts will not be less accurate if the case load effect is not present in the data, since its presence can not be certain for real-world data.

6.4 Brazilian case studies

To investigate the effectiveness of the general framework from Section 6.2 in capturing potential trends identified in the exploratory analysis (Section 6.1.1), we fit the model to the SARI and arbovirus case studies from Brazil. Recall from our exploratory analysis and discussion in Section 6.1.1, we expect to capture a negative relationship between the number of SARI hospitalisations and the cumulative proportions reported at each delay and a positive relationship between the number of arbovirus cases and the cumulative proportions reported at each delay.

For arbovirus cases we set $D_{max} = 31$, and model $D = 4$ weeks delay in the GDM since around 98% of cases are reported in the first four delays. For SARI hospitalisations we set $D_{max} = 20$, and model $D = 8$ weeks delay in the GDM since around 95% of the total counts are reported in the first eight delays. For both data sets we fit an independent model to each spatial region in the data set, with no nested hierarchical structure within the splines used. This gives the following version of our general framework:

$$y_{t,s} \mid \lambda_{t,s}, \theta_s \sim \text{Negative-Binomial}(\lambda_{t,s}, \theta_s) \tag{6.28}$$

$$\log(\lambda_{t,s}) = \log(population_s) + \iota_s + \zeta_t^{(s)} + \xi_{weeks[t]}^{(s)} \tag{6.29}$$

$$z_{t,d,s} | \nu_{t,d}, \phi_{d,s}, N_{t,d,s} \sim \text{Beta-Binomial}\left(\nu_{t,d,s}, \phi_{d,s}, N_{t,d,s} = y_{t,s} - \sum_{j=1}^{d-1} z_{t,j,s}\right) \tag{6.30}$$

$$\nu_{t,d,s} = \frac{S_{t,d,s} - S_{t,d-1,s}}{1 - S_{t,d-1,s}} \tag{6.31}$$

$$\text{probit}(S_{t,d,s}) = \psi_{d,s} + \eta_t^{(s)} + \delta_s \log\left(\frac{y_{t,s}+1}{population_s}\right) \tag{6.32}$$

In these models, we include an offset in the log of the expected mean total cases ($\lambda_{t,s}$) of $population_s$, which gives the 2023 population for each respective federative unit s. In addition to this we include an intercept term ι_s and a smooth cubic spline with shrinkage $\zeta_t^{(s)}$ to capture the systematic trend in the total counts, and cyclical cubic regression spline $\xi_{weeks[t]}^{(s)}$ to capture the yearly seasonality in the SARI hospitalisations/arbovirus cases. We

then model the probit-transformed expected cumulative proportions reported as having a monotonically increasing intercept over delay $\psi_{d,s}$, plus a cubic spline with shrinkage to capture the systematic variability $\eta_t^{(s)}$. Finally, δ_s captures the linear relationship of the log of the proportion of total counts in the population, $\log\left(\frac{y_{t,s}+1}{population_s}\right)$, and the probit-transformed expected cumulative proportions. For our models of real-world data we divide the total counts for each region $y_{t,s}$ by the respective population of the regions, so that we consider the effect of the rate of incidence, since the case load is likely to be proportional to population. Hence, any comparison of the incidence-delay effect coefficient (δ_s) between regions will account for the difference in population.

$$\Omega_\eta^{(s)} = \frac{S_t}{\sigma_{\eta,s}^2}, \tag{6.33}$$

$$\kappa_\eta^{(s)} \sim \text{Multivariate-Normal}\left(\mathbf{0}, \Omega_\eta^{(s)}\right), \tag{6.34}$$

$$\eta_{t,s} = X_t \kappa_\eta^{(s)}, \tag{6.35}$$

$$\Omega_\xi^{(s)} = \frac{S_w}{\sigma_{\xi,s}^2}, \tag{6.36}$$

$$\kappa_\xi^{(s)} \sim \text{Multivariate-Normal}\left(\mathbf{0}, \Omega_\xi^{(s)}\right), \tag{6.37}$$

$$\xi_{\text{week[t]},s} = X_w \kappa_\xi^{(s)}, \tag{6.38}$$

$$\Omega_\zeta^{(s)} = \frac{S_t}{\sigma_{\zeta,s}^2}, \tag{6.39}$$

$$\kappa_\zeta^{(s)} \sim \text{Multivariate-Normal}\left(\mathbf{0}, \Omega_\zeta^{(s)}\right), \tag{6.40}$$

$$\zeta_{t,s} = X_t \kappa_\zeta^{(s)}, \tag{6.41}$$

$$\sigma_{\zeta,s}, \sigma_{\eta,s}, \sigma_{\xi,s} \sim \text{Half-Normal}(0, 10^2), \tag{6.42}$$

$$\psi_{1,s} \sim \text{Normal}(0, 5^2), \quad \psi_{d,s} \sim \text{T}[\text{Normal}(\psi_{d-1,s}, 5^2), \psi_{d-1,s}, \infty] \tag{6.43}$$

$$\iota_s \sim \text{Half-Normal}(-10, 10^2), \tag{6.44}$$

$$\delta_s \sim \text{Normal}(0, 5^2), \tag{6.45}$$

$$\theta_s, \phi_{s,d} \sim \text{Gamma}(2, 0.02). \tag{6.46}$$

Here X_t and S_t are the temporal cubic spline basis and penalty matrices. The number of knots for the cubic temporal splines are determined by the number of weeks of data the model if fit to T_{now}, such that the number of knots is $\frac{T_{now}}{20}$ rounded down to the nearest whole number. The knots are evenly placed between weeks 1 and T_{now}. Similarly, X_w and S_w are the cyclic cubic spline for week of the year, where 7 knots are evenly placed to capture the closed seasonal cycle over weeks 1 to 52.

Since this model is a directed acyclic graph (DAG), as demonstrated in Appendix C.3, it can be fit in NIMBLE (de Valpine et al. (2017)). Two AF slice samplers were set in each independent model; one for the parameters ι_s, $\kappa_\zeta^{(s)}$, $\kappa_x^{(s)}i$, $\sigma_{\zeta,s}$ and $\sigma_{\xi,s}$ for each region s, and one for the parameters $\psi_{1:D,s}$, $\kappa_\eta^{(s)}$, $\sigma_{\eta,s}$ and δ_s (when included in the model) for each region. Otherwise, the default NIMBLE MCMC samplers were used as described in de Valpine et al. (2021). For each model, two MCMC chains with an iteration length of $10,000$, burn-in of $5,000$ and a thinning of 5 were run. Convergence in both cases was assessed by inspecting trace plots of key parameters and calculating the PSRF for all parameters in the model.

We split this section into two separate investigations and in each we consider data for both SARI hospitalisations and arbovirus cases in Brazil. First, in Section 6.4.1 we investigate the case load effect through parameter inference. Second, in Section 6.4.2 we perform a rolling prediction experiment to explore potential benefits in predictive precision of our proposed models.

6.4.1 Investigating case load effects

In this section, we fit our proposed model (Equations (6.28)–(6.32)) to all of the observed data available for each case study data set. This allows us to investigate the features of the case load effects, using as much information as possible.

6.4.1.1 SARI hospitalisations

Figure 6.9 shows that the linear estimated relationship between case load and reporting delay in the Brazilian SARI hospitalisations data mostly agrees with the negative relationship we discovered in our exploratory analysis in Figure 6.1. Despite estimating these 27 coefficients independently, the only region in Figure 6.9 that has a positive posterior median for the incidence-delay coefficient (δ_s) is Rio Grande do Norte (RN).

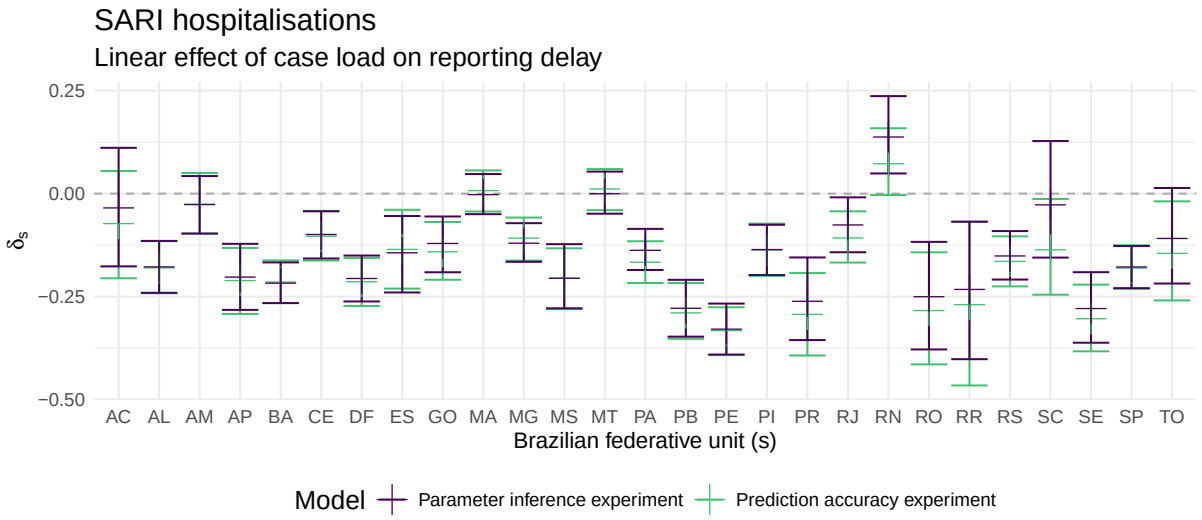

Figure 6.9: Posterior medians (middle line) and 95% prediction intervals (between upper and lower lines) of the coefficient for the linear relationship between case load and reporting delay (δ_s). For each federal unit in Brazil (x-axis) captured by the GDM model fitted to SARI hospitalisations data. Colour indicates whether results are for the parameter inference experiment model fitted to fully observed data or the prediction accuracy experiment model where data is censored for delayed reporting. In both cases results are from data up to the 27th March 2021.

Ideally, we would like to identify possible regional factors that might be contributing to the magnitude and sign of the effect of case load on reporting delays to gain greater insight into the driving force behind this effect. We plot the median posterior predictions for the case load effect on the cumulative proportions in Figure 6.10 however there is no clear spatial pattern. We also investigated if there was a relationship between $\boldsymbol{\delta}$ and the average number of SARI cases, the population density, the gross domestic product (GDP) and the government health expenditure of the regions but there were no obvious trends. However,

of the four metrics, there could be an argument for a slight downward trend between population density and the incidence-delay coefficient δ_s. This could indicate hospitals in more built up areas are more likely to have a bigger increase in reporting delays when there is an increase in the number of cases. However, a more in depth analysis is needed to confirm whether this may be the case.

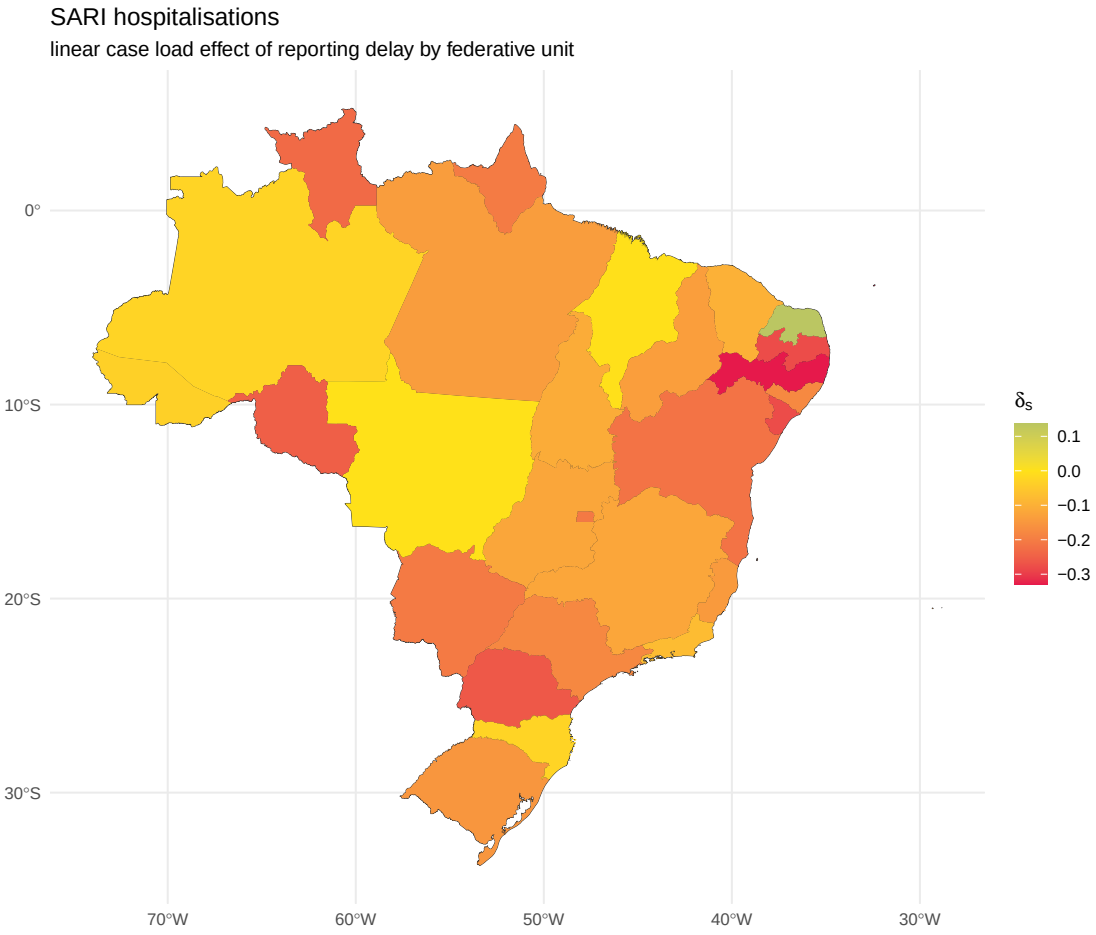

Figure 6.10: A map of the posterior median coefficients for the incidence-delay effect on the cumulative proportions (δ_s) for SARI hospitalisations for each federative unit in Brazil.

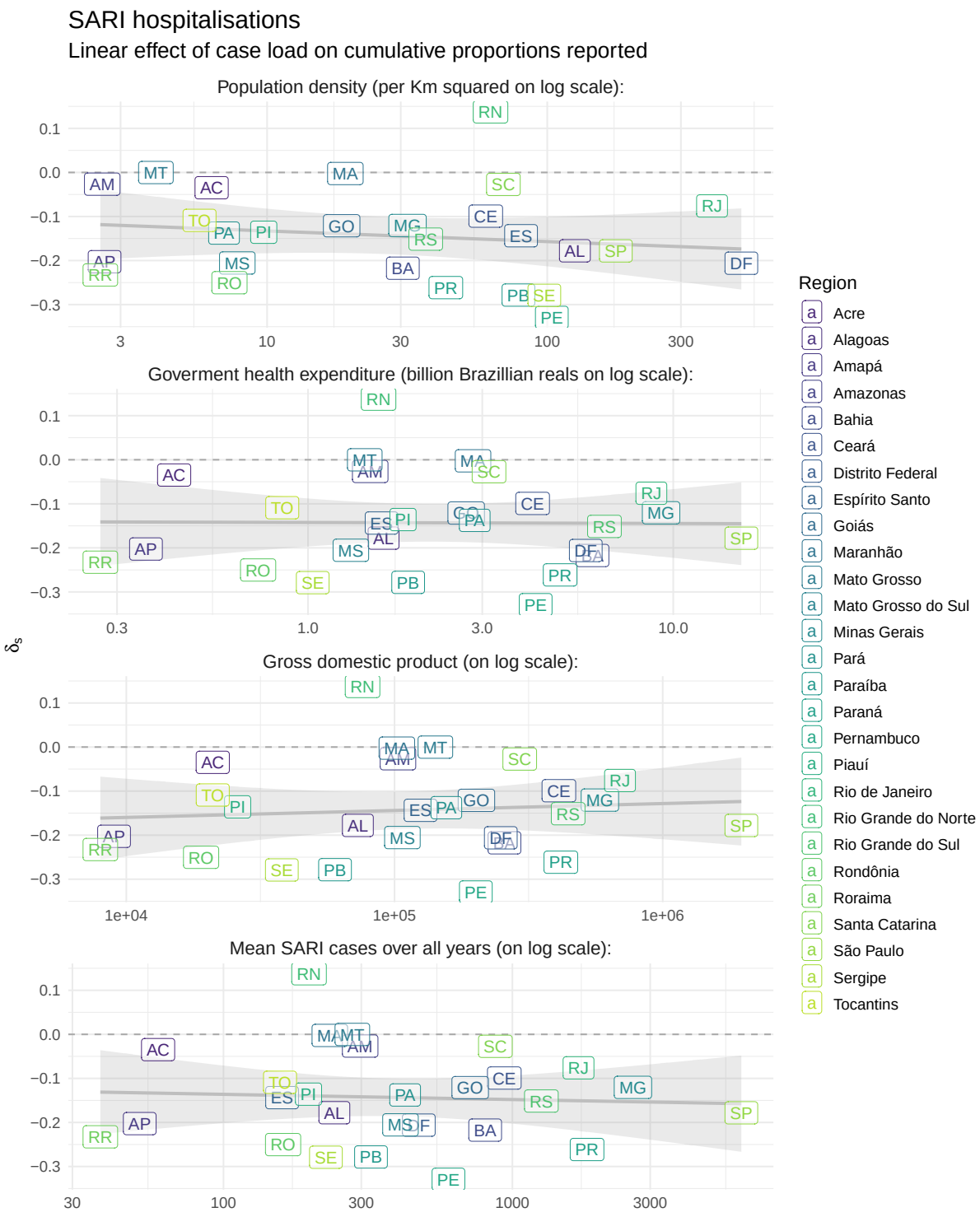

Figure 6.11: The posterior median coefficients for the incidence-delay effect of SARI hospitalisations for each federative unit (δ_s) plotted against regional metrics to identify potential influential factors on the magnitude of the effect. The regional metrics from the top are; the population density (1st panel), the government health expenditure (2nd panel), the Gross domestic product (3rd panel) and the mean number of SARI cases for the entire observed data set (4th panel). The solid lines represent a normal linear regression, and the shaded areas are the respective 95% confidence intervals.

6.4.1.2 Arbovirus cases

Unlike the negative relationships seen for SARI, the linear effect of case load on the cumulative proportions reported is estimated to be positive for arbovirus cases for the majority of Brazilian federative units as shown by Figure 6.12. Therefore, whilst an increase in the number of SARI hospitalisations increases the length of reporting delays, conversely an increase in the number of arbovirus cases may decrease the length of the reporting delay. Despite the majority of case load effects for arboviruses being opposite in sign to that of the SARI case load effect, they appear to have a similar range of magnitude over regions.

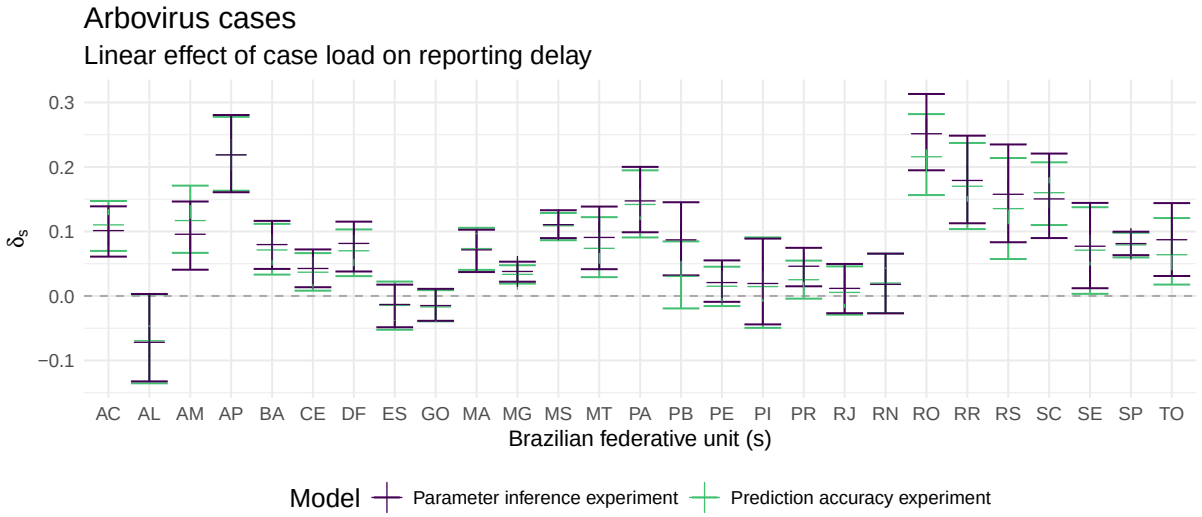

Figure 6.12: Posterior medians (middle line) and 95% prediction intervals (between upper and lower lines) of the coefficient for the linear relationship between case load and reporting delay (δ_s). For each federal unit in Brazil (x-axis) captured by the GDM model fitted to arbovirus cases data. Colour indicates whether results are for the parameter inference experiment model fitted to fully observed data or the prediction accuracy experiment model where data is censored for delayed reporting. In both cases results are from data up to the 11th January 2014.

Similarly to the SARI case study, we wish to gain insight from the difference in the dengue case load effect across the Brazil to help answer questions such as why the federative unit Alagoas (AL) doesn't exhibit a positive case load effect. Figure 6.13 suggests that the case load effect is more strongly positive in the north-western regions, which are typically less densely populated, rural regions within Brazil where arbovirus cases are likely to be

greater. Hence, in Figure 6.14 regions with higher population densities appear to have a smaller case load effect which we see in the more populated south-eastern regions of Figure 6.13. Furthermore, in Figure 6.14 we see a slight negative correlation between the incidence-delay coefficient δ_s and both government health expenditure and GDP which are both likely to be higher in larger regions with larger populations and cities. Finally, the last panel in Figure 6.14 shows there is no strong relationship between the mean number of arbovirus cases in a region and the case load effect.

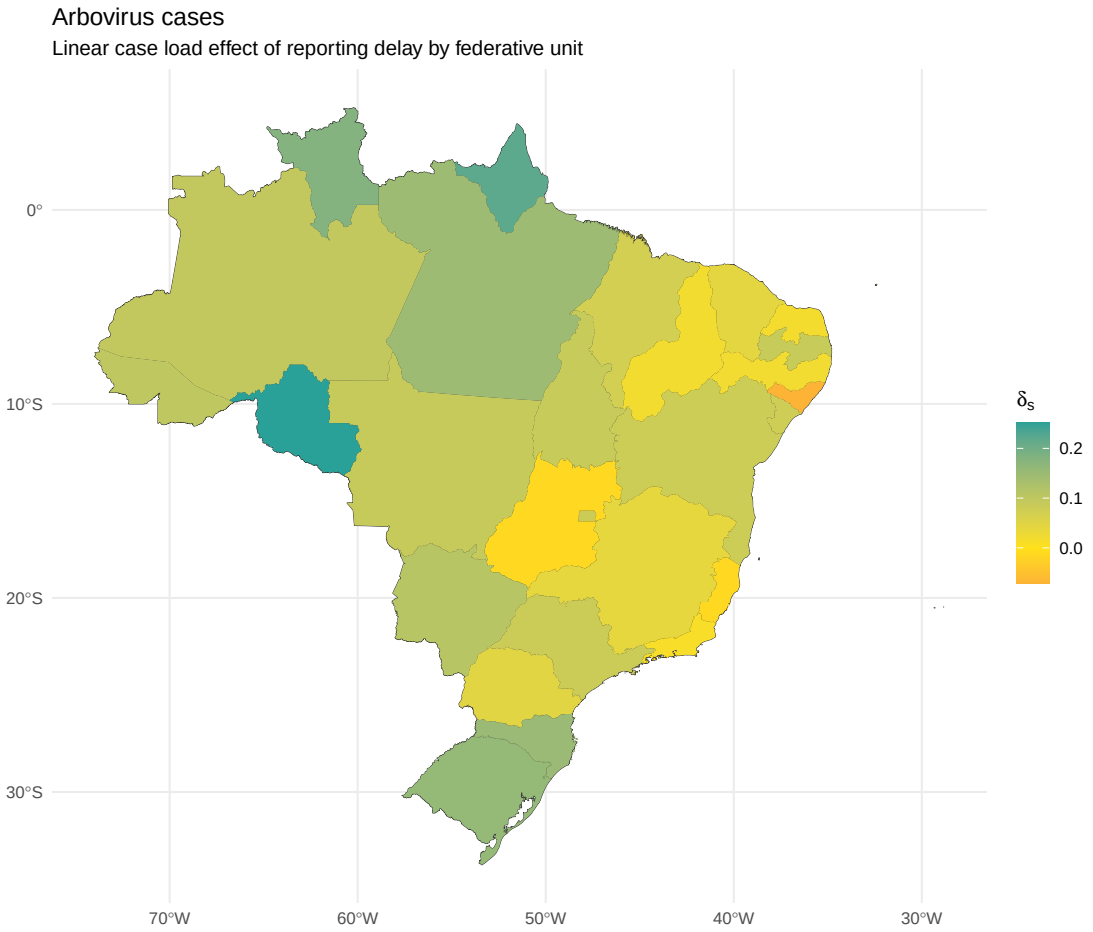

Figure 6.13: A map of posterior median coefficients for incidence-delay effect on the cumulative proportions (δ_s) for arbovirus cases for each federative unit in Brazil.

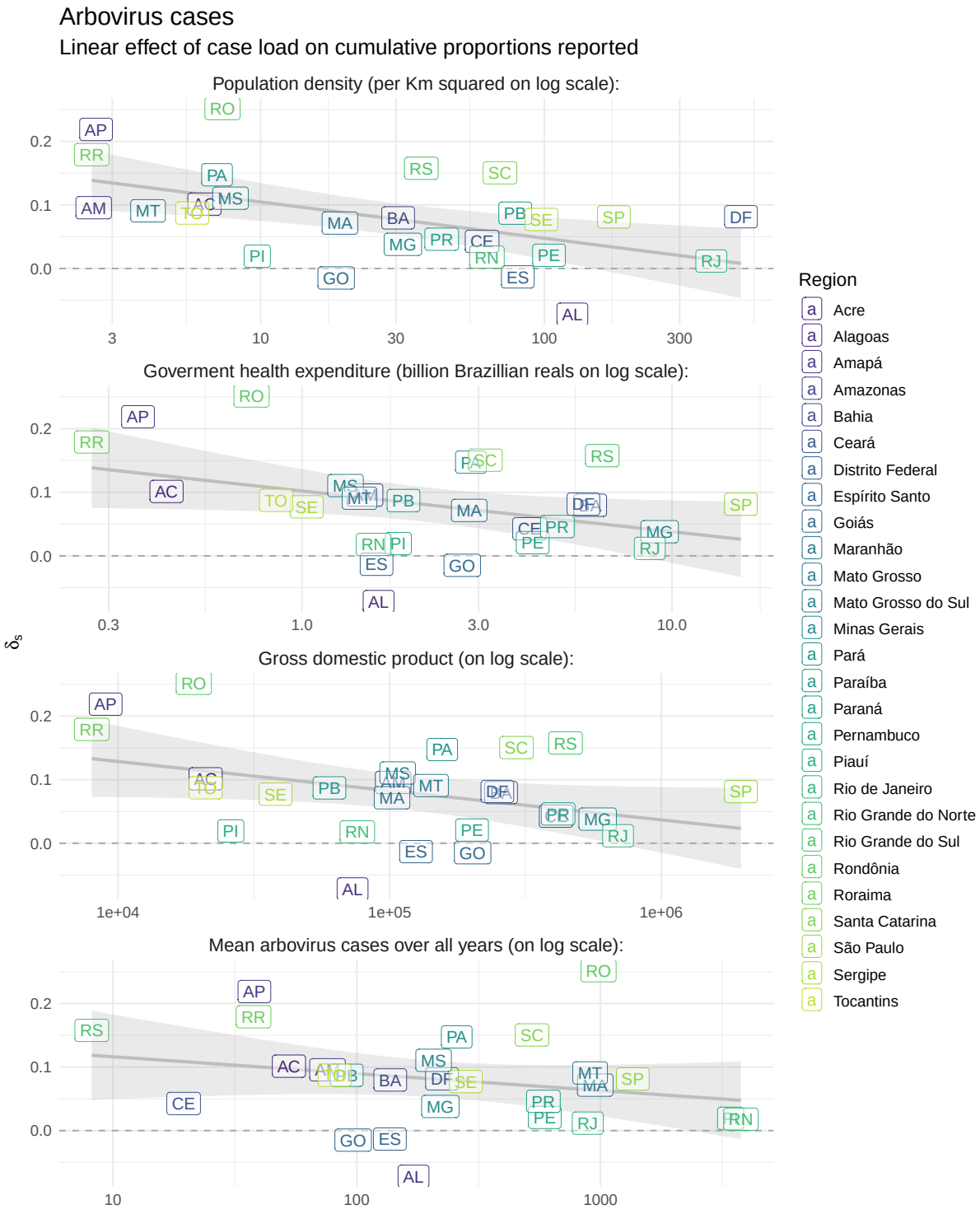

Figure 6.14: Posterior median coefficients for the incidence-delay effect of arbovirus cases for each federative unit (δ_s) plotted against regional metrics to identify potential influential factors on the magnitude of the effect. The regional metrics from the top are; the population density (1st panel), the government health expenditure (2nd panel), the Gross domestic product (3rd panel) and the mean number of arbovirus cases for the entire observed data set (4th panel). The solid lines represent a normal linear regression, and the shaded areas are the respective 95% confidence intervals.

6.4.2 Rolling prediction experiment

In this section, we carry out a rolling prediction experiment on both the SARI and arbovirus case studies. This is to determine any potential gains in prediction precision compared to a GDM model that does not account for the incidence-delay effect. We carry out nowcasts for a series of four rolling nowcast dates for both the SARI hospitalisations (31-07-2022, 04-09-2022, 16-10-2022 & 27-11-2022) and the arbovirus cases (23-12-2014, 13-09-2016, 05-06-2018 & 03-03-2020). Hence, for each nowcast date, we construct the respective censored data sets from the observed historic data we have for both SARI and arbovirus cases in Brazil. Theses are modified such that the cases that would be unknown at the given nowcast date due to reporting delays are removed from the data set.

We fit the our proposed framework, Equations (6.28)–(6.32), to these censored data sets, which allows us to obtain predictions for the "unknown" total counts. These can then be compared to the true total counts that would eventually be reported which we know since we are working with historic data. Next, we perform the same rolling prediction experiment again, but with a model where there is no link between total cases and the cumulative proportions reported, such that there is no $h(y_t, s)$, so we have $\text{probit}(S_{t,d,s}) = g(t,d,s)$.

6.4.2.1 SARI hospitalisations

For our rolling prediction experiment we averaged three performance metrics over time and region; the mean absolute error, the prediction interval width and the coverage. Figure 6.15 plots these three performance metrics for SARI hospitalisations predictions against the prediction time difference (PTD). Where a PTD of zero represents predictions made for the week for which we have the most recent data for and a negative PTD is the number of weeks prior the the most recent week for which we made predictions for. Although the

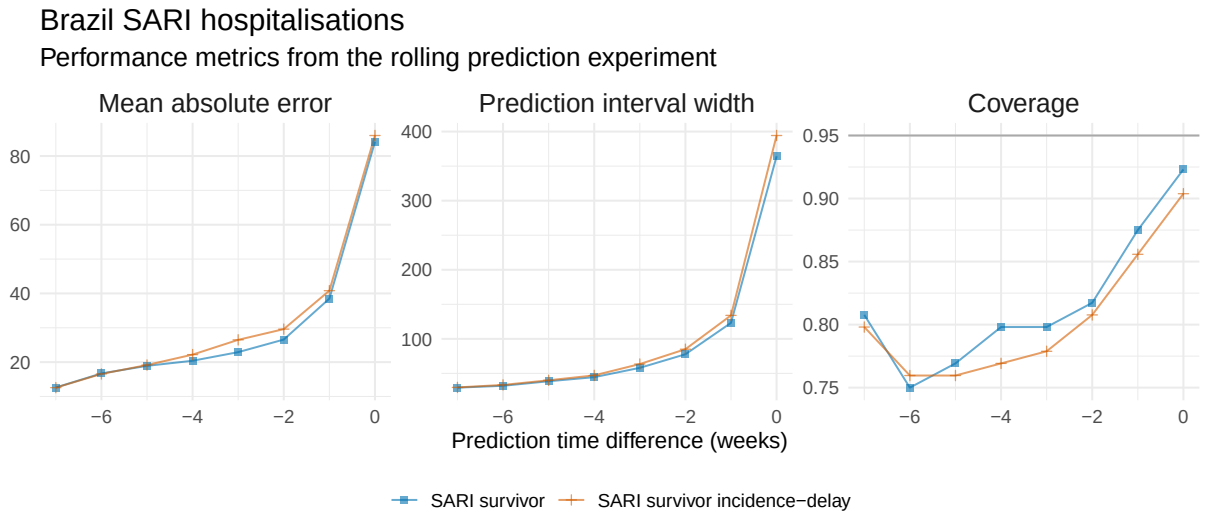

Figure 6.15: The mean absolute error, mean prediction interval width and 95% prediction interval coverage of predicted total SARI hospitalisations in Brazil, from the rolling prediction experiment. The GDM survivor incidence-delay model that models the case load effect in the SARI data is compared to the GDM survivor model which doesn't explicitly model this effect.

model still picks up the significant relationships between case load and reporting delay, as evident in Figure 6.9, this does not seem to translate into an improvement of prediction precision. In fact, Figure 6.15 shows that all prediction performance metrics are marginally better when the model does not include the explicit link between case load and reporting delay.

6.4.2.2 Arbovirus cases

For the rolling prediction experiment on arbovirus cases, we compare three models in Figure 6.16. First, the "Survivor" model without link between case load and delay. Second, the "survivor incidence-delay" model which explicitly models the case load effect. Third, the "survivor incidence-delay (time category)" model. This was motivated by the exploratory plot for arbovirus cases, Figure 6.4, where we see that the relationship between the cumulative proportions and the total arbovirus cases systematically varies given time category $b_t = \{[1 \leq t < 101 \text{ weeks}], [101 \text{ weeks} \leq t < 201 \text{ weeks}], [201 \text{ weeks} \leq t <$

301 weeks], [301 weeks $\leq t <$ 401 weeks], [401 weeks $\leq t <$ 501 weeks]}. Therefore, for this final model the relationship between total cases and the proportion of cases reported at each delay is a separate incidence-delay coefficient for each time category b_t such that $h(y_t, s) = \delta_{b_t} \log(\frac{y_t+1}{population_s})$.

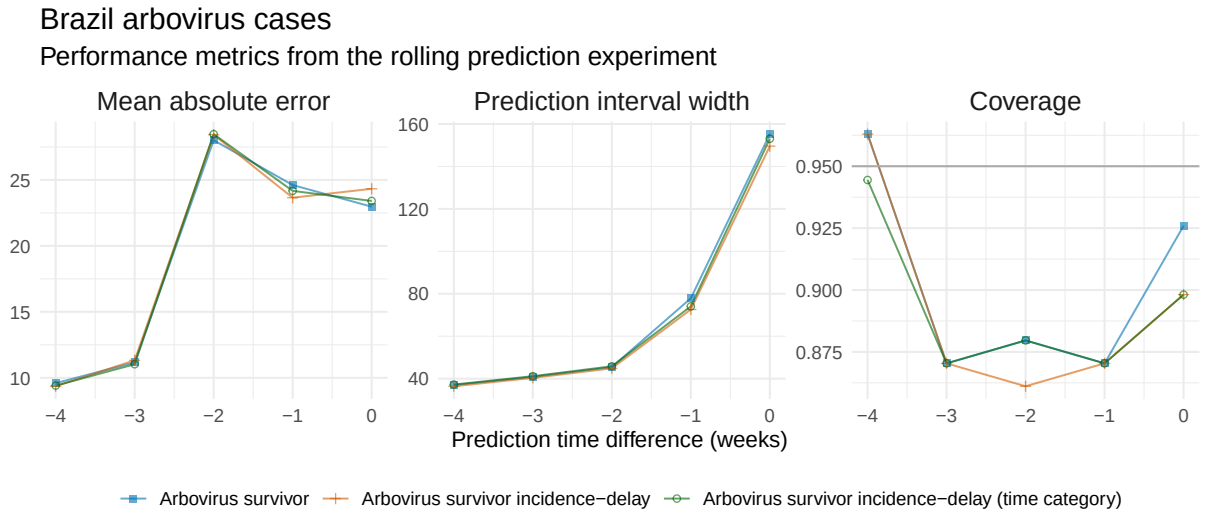

Figure 6.16: The mean absolute error, prediction interval width and coverage of predicted total arbovirus cases in Brazil for the rolling prediction experiment. The x-axis gives the prediction time difference in weeks which is the difference between the "current" time of the rolling prediction experiment and the week that the cases are being nowcasted for. This plot compares the GDM survivor incidence-delay model that models the case load effect in the SARI data, the GDM survivor incidence-delay (time category) model, where the incidence-delay effect δ_{b_t} is fixed given time category b_t, and the GDM survivor model which doesn't explicitly model the effect.

The results in Figure 6.16 suggest that there is not much difference between these three models. For the most current nowcasts (PTD=0 weeks), the model with no incidence-delay effect performs best, followed closely by the time category dependent incidence-delay effect, and then closely again by the incidence-delay effect model (where the link is constant for all times). The mean absolute error also shows a spike at $PTD = -2$, usually we see a trend of prediction accuracy improving further back from the nowcast date T_{now} as more data is seen by the model. Closer inspection showed this is due to the predictions in the Paraná (PR) region of Brazil underestimating the arbovirus cases two days prior ($PTD = -2$) to the nowcast date 03-03-2020 (375 weeks) by approximately 2031 cases (on average compared to the posterior medians) across all models considered.

This is potentially due to an outlier or anomaly in the data for this point which made it difficult for all models to accurately predict. It's also worth noting that these means are only calculated across four nowcast dates, so we would expect these sorts of discrepancies in the data to have less of an effect on the overall trend of the mean absolute errors if more nowcast dates were considered in this experiment. For prediction interval width the two models that explicitly model the case load effect have similar and slightly narrower interval widths than the model which doesn't consider the case load effect. However, for coverage the model with no link has the highest coverage of the three models for all prediction time differences (PTD).

6.5 Discussion

In this chapter we investigated and sought to address potential relationships between the number of cases of a disease occurring in a given time period and the length of reporting delays for cases in that period. In Section 6.1, we considered literature that discuss such effects but have not yet proposed any approach to correcting for them. We also explored two data sets for Brazil that suggest a potential positive relationship for SARI hospitalisations, where a higher disease incidence is associated with longer delays, and a potential negative relationship for arbovirus cases, where a higher disease incidence is associated with shorter delays.

To address the lack of a method for correcting case load effects, we presented a framework in Section 6.2, based on the GDM method, that is able to capture such relationships in a general way.

Through an extensive simulation study, we demonstrated that this approach is able to offer reliable parameter inference in the presence of possible compounding or cancelling trends (Section 6.3.2.1). Our simulation study also suggested that predictive accuracy of the unknown total counts and uncertainty quantification are both improved when explicitly modelling the link between incidence and reporting delay if it is present in the data generating process.

Applying our framework to both SARI hospitalisations and arbovirus cases in Brazil (Section 6.4.1), we were able to find significant and consistent effects of case load on regional reporting processes. Therefore, a by-product of our proposed GDM model is the identification of areas with substandard reporting delays related to disease incidence, which can help to direct possible reform within the reporting process if used in an operational surveillance system.

However, through rolling prediction experiments for this real-world data, in Section 6.4.2, we found that modelling this relationship with a linear coefficient for the log of the total counts in the probit-transformed cumulative proportions reported has little impact on the prediction precision when performing nowcasts on both SARI hospitalisation and arbovirus case data from Brazil. There are two possible explanations for this; the relationship between case load and delay does not affect the predictions of the total counts in these two case studies, or the modelling assumptions we have made are not appropriate for capturing the relationship that does exist in these data sets which could then improve prediction precision or accuracy.

In our proposed framework, we have assumed that the relationship between the number of total counts and the cumulative proportions reported is constant over delay, case load and time (for our models of SARI hospitalisations). However, more flexibility in capturing the case load relationship in the data may be beneficial to the model. It is possible to design more flexible versions of our presented framework, as demonstrated by having the incidence-delay effect vary over different time categories for our arbovirus case study in Section 6.4.2.2.

This model is fully identifiable when $\boldsymbol{y}$ is known, when $\boldsymbol{y}$ is unknown (as in nowcasting) it is relying on what has been learned from previously observed values of $\boldsymbol{y}$. A flexible function for the case load effect may not be well informed or constrained when proposed y_t are beyond the range of the previously observed y_t, leading to non-identifiability. Therefore, care must be taken to ensure the model remains identifiable. For example, the linear coefficient $\boldsymbol{\delta}_s$ we use avoids this issue since its behaviour in the extremes of the range of y_t is well determined by past y_t in the middle of the range. On the other hand, if $\boldsymbol{\delta}$ were time-varying in a very flexible way it could also become non-identifiable when nowcasting. One way to increase model flexibility while maintaining identifiability is to use a penalised regression spline with a strong informative prior on the smoothness penalty parameter. This prior controls the variance of the non-linear component of the spline. A strong prior on the smoothness parameter forces the variance of the non-linear trend towards zero unless there is enough data to support non-linearity. This ensures that the model does not introduce unwarranted flexibility in poorly informed regions of y_t, improving stability in nowcasting scenarios.

Moreover, we assumed that the case load effect can be characterised as only depending on contemporaneous total counts. Instead, reporting delays could be influenced by total counts from previous weeks, which would require lagged effect(s) in the model. Similarly, reporting delay length may be related to the the rate of change of the total counts over time, e.g. a steep increase in cases could cause slower reporting, or to the difference between current case levels and the long-term expected trend (e.g. an expected seasonal

pattern). However, implementing these new structures should also be supplemented with additional simulation experiments to asses our models ability to capture potential delay dependent and non-linear relationships. It is also worth noting that in doing so there may be a trade-off in computational speed as one introduces more parameters into the model. Currently, the models with a link between case load and reporting delay take 8 minutes to run for the SARI data and 1.9 hours for the arbovirus data.

Alternatively, there may be other methods of incorporating the relationship between case load and delay that we have not explored here, that may result in improvements in predictive performance. Therefore, more work is needed to develop a framework that could utilise the potential incident-delay relationship in real data to improve upon existing nowcasting models for operational warning systems.

Chapter 7

Conclusion

Disease epidemics can result in economic and public health costs, such as low quality welfare and loss of life, by requiring expensive resources and putting a strain on health care systems. In order to effectively manage outbreaks, and minimise these costs, surveillance systems are often implemented. These typically monitor count data that represent disease outcomes, such as cases or deaths, which act as a quantitative indicator for the severity of a prevailing disease. Enabling the clear communication and deeper understanding of the threats to a community. These are needed to help public health authorities allocate resources and inform policies such as pharmaceutical and non-pharmaceutical interventions (World Health Organization (2020b)). A barrier to effective disease monitoring is the presence of reporting delays in these disease outcomes, which prevent an up-to-date overview of the true levels of disease outcomes across a region.

The overall goal of this thesis was to develop nowcasting tools to correct for these reporting delays. This can help identify trends in disease data and facilitate up to date surveillance systems. This was achieved by first reviewing the existing literature in Chapter 2, which highlighted the relative benefits and limitations of existing approaches. To summarise, for theoretical peak predictive capabilities we require approaches that exhibit the flexibility to separate and capture all sources of variability in the data which we outline in Section 1.1. In Section 2.2, we broadly summarise the existing approaches as two main groups. The first group of approaches implement conditionally independent models of the observed partial counts, which don't take into account that the partial counts have a compositional structure that sums to the total counts. The second group of approaches, which jointly model the partial and total counts, address the compositional structure by implementing a Multinomial model for the partial counts. However, they often still lacks the flexibility to capture the random variability of the partial counts due to the lack of variance parameters.

Hence, in Section 2.3, the Generalised-Dirichlet Multinomial (GDM) model is presented as an approach that can model the complexities in the covariance structure of the partial counts by extending the Multinomial framework with additional parameters. However, as a consequence of the underlying methods used for the GDM, we noted that the GDM is

relatively slow compared to competing approaches for complex cases, such as hierarchical models for the total and partial counts in delayed disease reporting. We discuss the underlying methods used to fit both the GDM and competing approaches at the start of Chapter 2.

Also, through critiquing potential extensions of nowcasting frameworks in Section 2.5, we emphasize that flexibility and usability can improve model practicality for real-world applications. Furthermore, no frameworks thus far consider the implications of the incidence of the disease on the length of reporting delays, or nested structures which are often common in disease data.

In Chapter 3 we sought to address the issue of the GDM model's inefficient computational speed, compared to alternative approaches, which may limit its wider potential for being an operational nowcasting tool. For two case studies, we showed that implementing a series of alterations to how MCMC sampling is executed approximately halved the run time of the GDM model. We also investigated two other more computationally complex routes for achieving quicker predictions. Firstly, we considered a modelling framework that is suitable for fitting using INLA (Lindgren and Rue (2015)) which attempted to approximate the GDM's ability to capture the covariance structure of the partial counts using a Multivariate Normal random effect. It was shown, for a given simulation study, the GDM approximation INLA model had less uncertainty than the Bastos et al. (2019) INLA model for the predictions of the total counts. However, the approximation in INLA was unable to separate the uncertainty in the total counts and the partial counts as consistently as the GDM approach. Our second route created a technique for optimising the join posterior of the GDM framework. This could then be used to not only inform the initial values of the MCMC chains, to allow convergence in fewer iterations, but also to potentially give decision makers a timely approximate estimate of the eventual MCMC posterior point predictions of the total counts. However, we recognise that this technique may have unknown repercussions on the final predictions which needs to be more thoroughly investigated prior to advocating its general use.

Next, in Chapter 4 we implemented the GDM model with our additional computational advancements (developed in Chapter 3) to COVID-19 deaths in England. This application helped asses the GDM's capabilities for real-time disease surveillance compared to competing frameworks. Section 4.1 outlines the results of the published paper this work resulted in, Stoner, Halliday and Economou (2022), and highlighted my personal contributions to it. We showed that the GDM model produced more accurate prediction compared to alternative approaches when nowcasting COVID-19 deaths in England. Furthermore, I carried out a simulation study to determine that the GDM model allowed for parameter inference as it was able to recapture covariate effects in the artificially generated data.

Chapter 5 went on to extend the GDM framework to make it suitable for novel applications, by allowing the model to capture nested structures in disease data. For example, if a particular variant of a virus is responsible for a nested proportion or subset of the total number of infections. This framework allowed for the prediction of COVID-19 positive SARI hospitalisations in Brazil which would otherwise be unattainable using existing nowcasting methods, due to the delay length of the COVID-19 test results being unknown. Additionally, we presented a new choice of link function for the relative proportions reported, that are modelled in the GDM framework, using the centralised log ratio (CLR) transform of the absolute proportions reported. Although, this was not found to be perform better than the existing survivor or hazard link functions introduced in Stoner and Economou (2019) for this application, it could have benefits when the data generating process results in complex structures in the absolute proportions reported.

Finally, in Chapter 6 we again extended the original GDM framework to include the potential relationship between the level of incidence of a disease and the length of the delays in reporting the disease. This was possible in the GDM framework thanks to the separation of the total counts and partial counts as well as the flexibility of fitting Bayesian models in the package NIMBLE (de Valpine et al. (2017)) affords. To the best of our knowledge, this was the first nowcasting framework that incorporates this into the predictions of the unknown total counts. Through simulation experiments, in Section 6.3, we showed that

when a case load relationship is present in the data and not explicitly modelled within GDM framework, predictions of the total counts were less accurate (larger mean absolute errors) compared to our proposed incidence-delay GDM model. In Section 6.4, we applied both the original and our proposed framework again, to two real-world case studies in Brazil; SARI hospitalisations and arbovirus cases. However, for both data sets we found our proposed incidence-delay GDM model, that did explicitly model the incidence-delay effect, was comparable to the original GDM model in terms of prediction precision and uncertainty. Hence, the modelling assumptions of our proposed framework may not be suitable for the underlying data generating processes present in the real-world data, i.e. the process responsible for the observed correlation between number of counts and length of delay.

All data and code that supplements this thesis is available at `https://github.com/AlbaMH/Thesis`.

7.1 Discussion of the results

In general this thesis has developed frameworks that we have shown to have considerable benefits over existing approaches. Hence, if applied in an operational disease surveillance setting this could result in improvements on the timeliness and effectiveness of public health response to epidemic outbreaks. For instance, we have updated the GDM model to be more computationally efficient in Chapter 3. This means that, for any application, the computational costs will be lower, giving users more time to analyse and utilise the results. In particular, if the GDM is implemented in a real-time surveillance system, these improvements could be crucial in providing timely results that could inform public health policies. We showcased the GDM as a highly competitive option for such disease surveillance, in terms of both prediction precision and parameter inference, in Chapter 4.

Chapter 5 could also facilitate public health impact: by jointly modelling the total SARI hospitalisations and the subgroup of these that test positive for COVID-19, we allow for nowcasts of these COVID-19 positive counts to be possible. This is due to the unknown delays in awaiting for COVID-19 test results to confirm cases rendering nowcasting models which would just consider the COVID-19 counts to be unsuitable. Unknown reporting delay length is a common issue within reporting processes subject to delayed reporting and is often challenging to mitigate. Hence, the impact of this work has the potential to spread beyond the specific application that motivated its development.

This is also true for the novel GDM version we formulated in Section 5.3.3 which models systematic variability in the reporting delay by linking to expected proportion reported at each delay to covariate and/or random effects (including penalised regression splines and tensor product smooths). This has ramifications when using the GDM framework as the choice of link function can have a non-trivial impact on both posterior predictions and computational efficiency. Since previously existing GDM formulations don't allow the expected absolute proportions to be modelled directly, they may be limiting in applications where users wish to capture trends identified in these proportions. This may especially be the case if there is a strong intuition for the trends in the data generating process or if unconstrained effects are required to capture the trends. Hence, in some scenarios, the CLR version of the GDM could lead to more accurate and quicker results. However, it is worth noting that performance will also depend on the choice of model effects the link function is paired with.

Finally, the results in Chapter 6 have the potential to lead to a new class of nowcasting model that considers the effect of case load on the delay process. However, there will be a stronger argument for this if it is shown to provide greater improvements in a real-world application over the existing class of model that do not take this relationship into account. Despite this, we have shown through simulation studies that this is likely to be

the case if the model assumptions match the data generating process behind this case load effect. Furthermore, inference gained through explicitly modelling this link using our framework may still be informative to public health scientist investigating reporting delays by providing information about the magnitude of this effect across different regions.

7.2 Potential future work

Here we outline potential avenues for future work in improving upon the nowcasting models for operational infectious disease surveillance that we have developed in this thesis. We critique nowcasting approaches for in two main areas; predictive performance and operational ability.

The GDM model, which this thesis has focused on, enables higher predictive accuracy and precision by separating the four sources of variability in the data, summarised in Section 1.1.3, compared to competing approaches. Furthermore, it captures the potentially complex covariance structure between the proportion of the observed total counts reported at each delay, as we discussed in Section 1.1, owing to the flexibility provided by additional model parameters. But, improvements could be gained by considering alternative sources of information to help better inform predictions. For example, through informative covariates as we discuss in Section 2.5.2.

We incorporated the polynomial effect of age distribution when modelling the nested disease structure of SARI hospitalisations that are COVID-19 positive in Section 5.4. This helped inform the model predictions and allowed for inference on the likelihood of different age groups within SARI hospitalisation testing positive for COVID-19. Hence, other informative covariates and their impact on nowcasting frameworks could be considered, e.g. weather forecasts or social media trends. Moreover, prediction performance and inference

of spatial effects could be improved by considering more complex spatial structures both for the GDM and more generally for all surveillance systems. This could be achieved by modelling disease data on a finer spatial resolution and capturing the more intricate spatial dependencies, allowing for a more detailed overview of the progression of an outbreak and potentially more localised interventions which could be more cost effective.

As for operationally ability, the major drawbacks of the GDM are due to its slow computational speed and requirement for specialised knowledge to essentially hand-write the model code, and to implement the MCMC sampling methods, including setting prior distributions, iteration lengths and executing different random, fixed, seasonal and spatial effects as needed. We start to address this in this thesis by improving the speed at which the GDM model produces posterior predictions in Section 3.2. Despite this, it is still one of the slowest nowcasting models among the most cited approaches, as we compare in Chapter 4. Hence, there is still scope to close this gap in computational speed. This may include developing more efficient MCMC sampling methods that are suitable for the GDM. Alternatively, further work could attempt to justify the technique of optimising the joint posterior to set the initial values of MCMC sampling, which we investigated in Section 3.3, for operational use. Similarly, there may be a way to design an INLA model which can capture the covariance structure of the partial counts to allow for precision and confidence in predictions comparable to the GDM, as the approach we present in Section 3.1 is currently unable to do so.

Furthermore, we have not addressed the issue of usability, which evaluates the ease of implementing the model to novel applications, especially by individuals who may not be familiar with the method. In Section 7.2.0.1, we discuss potential future work of developing an R package for the GDM framework. This could greatly improve the accessibility of the GDM model, allowing it to be more generally applied to various real-world data sets that require modelling of compositional count data subject to reporting delays.

Additionally, correcting for reporting delays is only one barrier to disease surveillance. As we discuss in Section 2.5.3 there may be a need to correct for under-reporting in conjunction with correcting for delayed reporting. There is likely to be an element of under-reporting in any disease monitoring system due to capacity limits and imperfections in the reporting system. For example, this could include resource limitations that restrict the number of tests that can be carried out or the number of patients that can be hospitalised. On the other hand, admin errors may mean that cases are missed. Similarly, due to the design of the reporting system individuals may not be counted if they are asymptomatic, do not seek medical care, or seek medical care from alternative sources such as unregistered or private health care providers.

We attempt to develop a framework in Section **??** that exploits the fact that deaths are less likely to be under-reported than cases, to help quantify the under-reporting in positive test results for COVID-19 in England where both are monitored. However, this needs additional information about COVID-19 infection fatality rates in order to overcome identifiability issues and be valid for reliable inference. In Section 7.2.0.2, we outline an additional extension we considered for the GDM model, which considers the potential effect of disease incidence on testing capacity and hence under-reporting, but were unable to investigate further within the time-frame of this thesis.

Finally, a deeper investigation into the case load effect may be required for our GDM framework which explicitly models the link between disease counts and reporting delay length. As of present, in real-world application our proposed model (Section 6.2) appears to have little to no effect on the predictive precision of the resulting nowcasts compared to models which do not account for this relationship. We believe this is likely due to modelling assumptions about the nature of this relationship being unsuitable.

7.2.0.1 R package

The Generalized-Dirichlet Multinomial (GDM) method for disease surveillance or other general compositional data applications (where data is expressed as parts summing to a total) is a Bayesian hierarchical model that can currently only be implemented using the Markov Chain Monte Carlo (MCMC) method. We implement the method using flexible software NIMBLE de Valpine et al. (2017), which is suitable for modelling complex hierarchical structures. However, adapting the method for new applications relies on experience in Bayesian hierarchical models and MCMC methods. This is a barrier to adoption of the method by the wider community of public health practitioners and epidemiological researchers, therefore limiting the potential impact of work. Other nowcasting approaches have developed R packages so that end users can implement their methods with simple functions, increasing their ease of use. For example, McGough et al. (2020) created the R package **NobBS** and the package **Nowcaster** is based on the work by Bastos et al. (2019). However, we review both these approaches in Chapter 2 and compare them to the GDM for nowcasting COIVD-19 deaths in Chapter 4, and found them to be lacking reliable predictive precision compared to the GDM.

Developing the GDM model into an R package would be challenging due to its generality, meaning that there are a multitude of ways in which it can be customised for specific problems. Furthermore, it also can take time to calibrate the GDM so that it runs optimally, as MCMC methods require user specified parameters, such as MCMC chain; iteration length, burn-in length, thinning and intial values. Moreover, users have to ensure convergence has been achieved, by inspecting trace plots and the potential scale reduction factor (PSRF), preferably without the cost of excessively long run times. Hence, an R-package could allow for users to carry out model fitting using more familiar interfaces, avoiding

knowledge of NIMBLE, such as the popular formula syntax used by many R packages including INLA (Lindgren and Rue (2015)) and JAGAM (Wood (2016)). Additionally, some sort of internal selection process for both modelling terms and MCMC parameters could be beneficial to reduce the amount of user expertise required.

Primarily, the goal of the package would be to allow users to fit the GDM model to the general case of disease surveillance application suffering from delayed reporting. This would involve a total count $y_{t,s}$ for a given time step t across spatial regions s, where the partial counts reported at each delay d, $z_{t,d,s}$, sum to the aforementioned total count. There could also be scope to customise the package further, to include more bespoke versions of the GDM. For example, in Section 5.3, we introduced our framework that extends the GDM model so that it can capture nested structures which are often observed in disease surveillance data. Similarly, Section 6.2 introduced our framework for accounting for potential case load effect on reporting delays.

7.2.0.2 Testing capacity

Torres, Sippy and Sacoto (2021) investigate potential under-reporting of COVID-19 due to limited testing capacity in Ecuador. This was initiated as Ecuador had relatively few confirmed cases despite having one of the highest global death rates. A driving factor of this was a backlog in test processing due to limited imported laboratory supplies and gaps in the workforce. They also note that testing burden at laboratory level also influenced delays in processing. Hence, capturing the relationship between number of cases and the reporting delay, for example by using the framework we introduce in Section 6.2, could be especially informative if the reporting delay includes the time taken to receive a laboratory test result which is likely to be the case for many infectious disease surveillance systems.

Another avenue of investigation, which could help improve the decision makers knowledge of the true level of a disease present in a population, would be to consider the extent to which cases may be under-reported. Under-reporting due to limited testing capacity is likely the reason for the low numbers of reported confirmed COVID-19 cases in Torres, Sippy and Sacoto (2021) compared to fatality rates. This relationship may also be significant in Brazil case study we consider in Chapter 5, where the number of severe COVID-19 cases in Brazil, defined as SARI hospitalisations where the respiratory virus has been identified to be COVID-19 by laboratory tests. If not all SARI hospitalisations are given COVID-19 tests, then individuals with COVID-19 may not be identified out of the total number of SARI cases, meaning that the observed proportion of SARI hospitalisations that test COVID-19 positive may be smaller than the 'true' proportion.

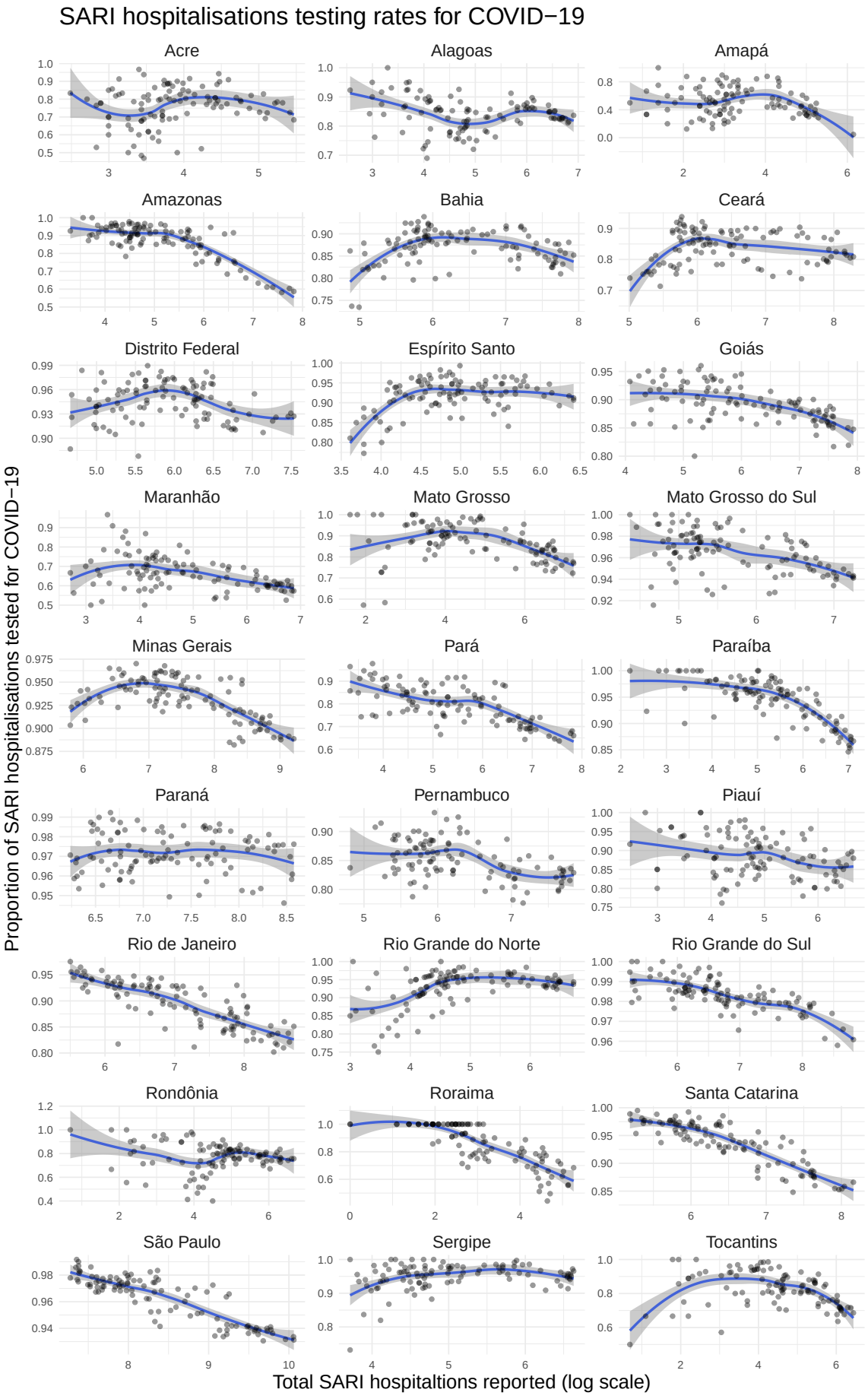

Figure 7.1: The proportion of SARI cases reported that are recorded as having been tested (either antigen or PCR test) plotted against the total number of SARI cases.

We explore this by considering how testing rates for COVID-19 seem to change over the total number of SARI hospitalisations, shown in Figure 7.1. When there is a greater number of SARI cases, which indicates a peak in an epidemic outbreak, there appears to be a lower testing rate in some regions. This may be due to limitations in tests available or staff that are able to administer and run the tests. Similarly, some regions exhibit a dip in testing for low numbers of SARI cases as well, this could be due to lack of awareness in periods between COVID-19 outbreaks or due to lack of testing availability. Alternatively, these trends could be due to a lack of testing both at the start and the end of the time series we have available due to both the availability of tests and priority of identifying COVID-19 cases being higher during in the larger outbreaks in 2021/22. Therefore, the rationale and protocol behind how healthcare providers decide whether to run a test will have to be carefully considered when investigating this, so expert insights will be crucial.

7.3 Final remarks

Disease surveillance is a key aspect of public health monitoring and management. In this thesis we have presented Bayesian hierarchical models that can improve upon existing surveillance systems for correcting disease data subject to delayed reporting. These models could give decision makers more information to work with through more accurate and precise estimates of disease outcome indicators. This has been achieved, in the first instance, by addressing concerns about timeliness of existing nowcasting approaches and, in the second instance, by developing novel frameworks for real-world situations where existing approaches are unsuitable.

The next steps for the work covered in this thesis is to bridge the gap between methodological advancements and real-world impact. This will be achieved by collaborations between the University of Glasgow and the Oswaldo Cruz Foundation (Fiocruz) which I have been appointed to help facilitate. The goal of this collaboration is to implement the

frameworks developed here for novel applications in Brazil into a real-time surveillance system. This includes and extensive knowledge exchange, and an R package as discussed in Section 7.2.0.1. The R package will implement the more computationally efficient version of the original GDM model (Stoner and Economou (2019)) which we updated in Chapter 3, for general applications with limited user input required. Additionally, an operational prototype of the GDM model for nested structures, developed in Chapter 5, could highlight the more reliable and accurate real-time nowcast predictions, for both SARI and severe COVID-19 hospitalisations, our framework can achieve compared to those currently operational in Brazil.

Moreover, this collaboration could result in a GDM modelling framework which reforms the one we proposed in Chapter 6, by enabling the case load effect of real-world disease to be captured in a way which improves prediction performance. This is something we were unsuccessful in achieving here (Section 6.4). Although our proposed framework has the theoretical capability to capture any type of relationship between case load and delay, we currently lack the understanding of how this is generated in real-world data and hence opted for potentially restrictive effects. However, researchers at Fiocruz may hold key insights in the reporting process which could be crucial in enhancing the design of our method.

Appendices

A Definition of distributions

In this section, we define the probability distributions employed in the Generalized Dirichlet Multinomial (GDM) model.

A.1 Poisson Distribution

The Poisson distribution is a discrete probability distribution used to model the number of events occurring within a fixed interval of time, space, or another continuous domain. It is parameterised by the rate parameter $\lambda > 0$, which represents the average number of events in the interval.

The probability mass function of the Poisson distribution $X \sim \text{Poisson}(\lambda)$ is given by:

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}, \quad k = 0, 1, 2, \ldots \tag{1}$$

where k is the number of events in the interval. The mean and variance are both determined by the rate parameter, $\mathbb{E}[X] = \mathrm{Var}[X] = \lambda$.

A.2 Beta Distribution

The Beta distribution is a continuous probability distribution defined on the interval $[0,1]$. It is widely used to model probabilities, proportions, and rates. The Beta distribution is parameterised in two common ways: using the standard shape parameters α, β and an alternative form in terms of the expected proportion ν and dispersion parameter ϕ.

The Beta distribution with shape parameters $\alpha > 0$ and $\beta > 0$ has the following probability density function (PDF):

$$f(x;\alpha,\beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}, \quad 0 < x < 1, \tag{2}$$

where $B(\alpha,\beta)$ is the Beta function:

$$B(\alpha,\beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}dt = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}. \tag{3}$$

For this parameterisation the mean is $\mathbb{E}[X] = \frac{\alpha}{\alpha+\beta}$ and the variance is $\mathrm{Var}[X] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$.

An alternative and often more interpretable parameterisation of the Beta distribution expresses it in terms of the expected proportion (mean) ν and a dispersion parameter ϕ that controls the variability of X. The relationship between the two parameterisations is given by $\alpha = \nu\phi$ and $\beta = (1-\nu)\phi$. Using this formulation, the probability density

function can be rewritten as:

$$f(x;\nu,\phi) = \frac{x^{\nu\phi-1}(1-x)^{(1-\nu)\phi-1}}{B(\nu\phi,(1-\nu)\phi)}. \tag{4}$$

For this parameterisation the mean is $\mathbb{E}[X] = \nu$ and the variant is $\text{Var}[X] = \frac{\nu(1-\nu)}{1+\phi}$.

A.3 Gamma Distribution

The Gamma distribution is a continuous probability distribution often used to model waiting times or non-negative skewed data. In this body of work we use two parameterisations. Firstly, when using the Gamma distribution as a prior distribution, we use $X \sim \text{Gamma}(\theta, r)$, where $\theta > 0$ is the **shape** parameter and $r > 0$ is the **rate** parameter, the probability density function (PDF) is given by:

$$f(x;\theta,r) = \frac{r^{\theta}}{\Gamma(\theta)} x^{\theta-1} e^{-rx}, \quad x > 0, \tag{5}$$

where $\theta > 0$ is controlling the distribution's skewness, $r > 0$ is controlling the scale (inverse of the mean) and $\Gamma(\theta)$ is the gamma function, defined as $\Gamma(\theta) = \int_0^\infty t^{\theta-1} e^{-t}\, dt$. The mean can then be calculated by $\mathbb{E}[X] = \frac{\theta}{r}$ and variance by $\text{Var}[X] = \frac{\theta}{r^2}$.

For the alternative parameterisation, which we use when defining a Negative-Binomial as a Poisson-Gamma mixture model (which we derive in Section B.1), we include a mean parameter λ which we substitute the rate parameter for by $r = \frac{\theta}{\lambda}$. The probability density function (PDF) is then given by:

$$f(x;\theta,\lambda) = \frac{\frac{\theta}{\lambda}^{\theta}}{\Gamma(\theta)} x^{\theta-1} e^{-\frac{\theta}{\lambda}x}, \quad x > 0. \tag{6}$$

A.4 Inverse-Gamma Distribution

The Inverse-Gamma distribution is a continuous probability distribution used primarily as a conjugate prior for variance parameters in Bayesian analysis. In NIMBLE, the `dinvgamma` function parameterises the distribution in terms of $\alpha > 0$, the shape parameter and $\beta > 0$, the scale parameter. A random variable with $X \sim \text{Inverse-Gamma}(\alpha, \beta)$ distribution has a probability density function (PDF) given by:

$$f(x;\alpha,\beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{-\alpha-1} e^{-\frac{\beta}{x}}, \quad x > 0, \tag{7}$$

where $\Gamma(\alpha)$ is the Gamma function:

$$\Gamma(\alpha) = \int_0^{\infty} t^{\alpha-1} e^{-t} dt. \tag{8}$$

The mean is $\mathbb{E}[X] = \frac{\beta}{\alpha-1}$ and the variance is $\text{Var}[X] = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)}$, for $\alpha > 2$.

A.5 Negative-Binomial Distribution

The Negative-Binomial distribution is commonly used to model overdispersed count data. In this thesis we use the parameterisation in terms of mean ($\lambda > 0$) and dispersion ($\theta > 0$). For this paramterisation the probability mass function of $X \sim \text{Negative-Binomial}(\lambda, \theta)$ is given by:

$$P(X = k) = \frac{\Gamma(k+\theta)}{\Gamma(\theta)k!} \left(\frac{\theta}{\theta+\lambda}\right)^{\theta} \left(\frac{\lambda}{\theta+\lambda}\right)^{k}. \tag{9}$$

In the above equation $k \geq 0$ is the count of failures before achieving θ successes and $\Gamma(\cdot)$ is the gamma function.

The mean of this distribution can then be calculated by $\mathbb{E}[X] = \lambda$, and the variance can be calculated by $\text{Var}[X] = \lambda + \frac{\lambda^2}{\theta}$. As $\theta \to \infty$, the distribution approaches a Poisson distribution, hence demonstrating that θ models the over-dispersion relative to the Poisson distribution.

A.6 Dirichlet Distribution

The Dirichlet distribution is a multivariate generalization of the beta distribution. For a random vector $\mathbf{X} = (X_1, \ldots, X_k)$ that follows a Dirichlet distribution, $\mathbf{X} \sim \text{Dir}(\boldsymbol{\alpha})$, the density function is:

$$f(\mathbf{X}; \boldsymbol{\alpha}) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} X_i^{\alpha_i - 1}, \tag{10}$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_k)$ are the concentration parameters, $\alpha_i > 0$, and $\sum_{i=1}^{k} X_i = 1$. Each α_i influences the expected proportion and variability of X_i.

A.7 Generalized-Dirichlet Distribution

The Generalized-Dirichlet (GD) distribution is an extension of the Dirichlet distribution that allows for more flexible modelling of dependencies between components (Connor and Mosimann (1969)). For a random vector $\mathbf{X} = (X_1, \ldots, X_k)$ that follows a Generalized Dirichlet distribution, denoted as $\mathbf{X} \sim \text{GD}(\boldsymbol{\alpha}, \boldsymbol{\beta})$, the density function is:

$$f(\mathbf{X};\boldsymbol{\alpha},\boldsymbol{\beta}) = \prod_{i=1}^{k-1} \frac{1}{B(\alpha_i,\beta_i)} x_i^{\alpha_i-1} \left(1-\sum_{j=1}^{i} x_j\right)^{\beta_i-\alpha_{i+1}-1} \left(1-\sum_{j=1}^{k-1} x_j\right)^{\alpha_k-1}. \tag{11}$$

In the above equation, $\boldsymbol{\alpha} = (\alpha_1,\ldots,\alpha_k)$ are the shape parameters with $\alpha_i > 0$, and $\boldsymbol{\beta} = (\beta_1,\ldots,\beta_k)$ are the secondary shape parameters. Then, $x_i \in (0,1)$ for $i = 1,\ldots,k$, and $x_{k+1} = 1 - \sum_{i=1}^{k} x_i$.

The GD distribution generalizes the Dirichlet distribution by introducing additional parameters $\boldsymbol{\beta}$ that allow for increased flexibility in modelling correlations and marginal distributions of the components.

A.8 Beta-Binomial Distribution

The Beta-Binomial distribution is a compound distribution where the success probability of a Binomial distribution is itself a random variable following a Beta distribution.

Often the parameterisation of the Beta-Binomial distribution is given by two shape parameters, ($\alpha > 0$) and ($\beta > 0$), which control the mean and variance of the success probability, and n, the total number of trials.

In this thesis we re-parameterise the distribution as $X \sim$ Beta-Binomial model(ν,ϕ,n), where $\nu \in (0,1)$ represents the expected proportion of successes, $\phi > 0$ controls over-dispersion and $n \geq 0$ represents the number of trials. The parameters α and β of the underlying Beta distribution can be expressed by $\alpha = \nu\phi$ and $\beta = (1-\nu)\phi$.

The probability mass function of the Beta-Binomial distribution is:

$$P(X=k)=\binom{n}{k}\frac{B(k+\nu\phi,n-k+(1-\nu)\phi)}{B(\nu\phi,(1-\nu)\phi)}, \tag{12}$$

In the above equation $k \geq 0$ is the number of successes in n trials and $B(\cdot,\cdot)$ is the beta function.

Using this parameterisation the mean of the Beta-Binomial distribution can be calculated by $\mathbb{E}[X]=n\nu$ and the variance can be calculated by $\text{Var}[X]=n\nu(1-\nu)\left(1+\frac{n-1}{\phi+1}\right)$.

A.9 Multinomial Distribution

The multinomial distribution is a generalisation of the binomial distribution that describes the probability of counts across multiple categories. When parameterised as $\mathbf{X} \sim$ Multinomial($\mathbf{p},n$) by the probability vector $\mathbf{p}=(p_1,p_2,\ldots,p_k)$ and the number of trials $n>0$, the probability mass function (PMF) of a multinomial distribution is given by:

$$P(X_1=x_1,X_2=x_2,\ldots,X_k=x_k)=\frac{n!}{x_1!x_2!\cdots x_k!}\prod_{i=1}^{k}p_i^{x_i}, \tag{13}$$

In this equation $n>0$ is the total number of trials, x_i is the number of observations in category $i=1,\ldots,k$ and p_i is the probability of an observation falling into category $i=1,...,k$. These parameters are subject to the following conditions:

- $x_i \geq 0$ for all $i=1,\ldots,k$,
- $\sum_{i=1}^{k} x_i = n$ (total counts equal the number of trials),
- $\sum_{i=1}^{k} p_i = 1$ (probabilities sum to 1),
- $1 \geq p_i \geq 0$ for all $i=1,\ldots,k$.

The mean of the distribution can be calculated by $\mathbb{E}[X_i] = np_i$ and the variance can be calculated by $\text{Var}[X_i] = np_i(1-p_i)$, both for $i = 1,\ldots,k$. Also, the covariance can be calculated by $\text{Cov}[X_i, X_j] = -np_ip_j$, for $i \neq j$.

To obtain the marginalised likelihood $P(X_1 = x_1)$ we can marginalise the full likelihood given in Equation (13) over $X_2,\ldots,X_k$, and sum over all configurations such that $x_2 + \cdots + x_k = n - x_1$:

$$P(X_1 = x_1) = \sum_{x_2+\cdots+x_k=n-x_1} \frac{n!}{x_1!x_2!\cdots x_k!} p_1^{x_1} \prod_{i=2}^{k} p_i^{x_i}. \tag{14}$$

After marginalising, the result simplifies to:

$$P(X_1 = x_1) = \binom{n}{x_1} p_1^{x_1}(1-p_1)^{n-x_1}, \tag{15}$$

where $1 - p_1 = \sum_{i=2}^{k} p_i$.

B Derivations

B.1 Marginal distribution of a Poisson-Gamma mixture model is a Negative-Binomial distribution

In this section we show that for the hierarchical Poisson-Gamma mixture model model:

$$y_t \mid \kappa_t \sim \text{Poisson}(\kappa_t), \quad \kappa_t \mid \lambda_t \sim \text{Gamma}(\theta, \lambda_t), \tag{16}$$

the marginal distribution of y_t follows a Negative-Binomial distribution:

$$y_t \sim \text{Negative-Binomial}(\lambda_t, \theta), \tag{17}$$

where λ_t is the mean and θ is the dispersion parameter. The PMF of the Poisson distribution with rate parameter $\kappa_t > 0$ is:

$$p(y_t \mid \kappa_t) = \frac{\kappa_t^{y_t} e^{-\kappa_t}}{y_t!}, \quad y_t = 0, 1, 2, \ldots \tag{18}$$

The PDF of the Gamma distribution with shape parameter θ and rate parameter t is, where the mean is given by $\lambda_t = \frac{\theta}{r_t}$:

$$p(\kappa_t \mid r_t, \theta) = \frac{r_t^{\theta}}{\Gamma(\theta)} \kappa_t^{\theta-1} e^{-r_t \kappa_t}, \quad \kappa_t > 0. \tag{19}$$

The marginal distribution of y_t is obtained by integrating out κ_t:

$$p(y_t \mid r_t, \theta) = \int_0^{\infty} p(y_t \mid \kappa_t) p(\kappa_t \mid r_t, \theta)\, d\kappa_t. \tag{20}$$

Substituting the Poisson and Gamma densities:

$$p(y_t \mid r_t, \theta) = \int_0^{\infty} \frac{\kappa_t^{y_t} e^{-\kappa_t}}{y_t!} \cdot \frac{r_t^{\theta}}{\Gamma(\theta)} \kappa_t^{\theta-1} e^{-r_t \kappa_t}\, d\kappa_t. \tag{21}$$

Combine terms:

$$p(y_t \mid r_t, \theta) = \frac{r_t^{\theta}}{y_t! \Gamma(\theta)} \int_0^{\infty} \kappa_t^{y_t+\theta-1} e^{-(r_t+1)\kappa_t}\, d\kappa_t. \tag{22}$$

The integral can be given by the normalising constant of a Gamma distribution with shape parameter $\theta + y_t$ and rate parameter $r_t + 1$:

$$\int_0^{\infty} \kappa_t^{\theta+y_t-1} e^{-(r_t+1)\kappa_t}\, d\kappa_t = \frac{\Gamma(\theta + y_t)}{(r_t+1)^{\theta+y_t}}. \tag{23}$$

Substitute this result back into the expression in Equation (22) for $p(y_t \mid r_t, \theta)$:

$$p(y_t \mid r_t, \theta) = \frac{r_t^{\theta}\, \Gamma(\theta + y_t)}{y_t! \Gamma(\theta)(r_t+1)^{\theta+y_t}}. \tag{24}$$

This can be simplified to:

$$p(y_t \mid r_t, \theta) = \frac{\Gamma(\theta + y_t)}{y_t! \Gamma(\theta)} \cdot \left(\frac{r_t}{r_t + 1}\right)^{\theta} \left(\frac{1}{r_t + 1}\right)^{y_t}. \tag{25}$$

In order to reparameterise in term of the mean parameter we can substitute in $r_t = \frac{\theta}{\lambda_t}$:

$$p(y_t \mid \lambda_t, \theta) = \frac{\Gamma(\theta + y_t)}{y_t! \Gamma(\theta)} \cdot \left(\frac{\frac{\theta}{\lambda_t}}{\frac{\theta}{\lambda_t} + 1}\right)^{\theta} \left(\frac{1}{\frac{\theta}{\lambda_t} + 1}\right)^{y_t}. \tag{26}$$

This expression can then be simplified to the Negative-Binomial form:

$$p(y_t \mid \lambda_t, \theta) = \frac{\Gamma(\theta + y_t)}{y_t! \Gamma(\theta)} \cdot \left(\frac{\theta}{\theta + \lambda_t}\right)^{\theta} \left(\frac{\lambda_t}{\theta + \lambda_t}\right)^{y_t}. \tag{27}$$

This is the PMF of a Negative-Binomial distribution $y_t \sim \text{Negative-Binomial}(\lambda_t, \theta)$, defined in Section A.5, with dispersion parameter θ and mean parameter λ_t.

B.2 Marginal distribution of a Poisson-Multinomial mixture model is a Poisson distribution

Here we show that the Poisson-Multinomial mixture model:

$$y_t \mid \kappa_t \sim \text{Poisson}(\kappa_t), \quad \mathbf{z}_t \mid \mathbf{p}_t, y_t \sim \text{Multinomial}(\mathbf{p}_t, y_t), \tag{28}$$

has a marginal distribution of

$$z_{t,d} \mid \kappa_t, p_{t,d} \sim \text{Poisson}(p_{t,d}\kappa_t), \tag{29}$$

where the rate parameter is $p_{t,d}\kappa_t > 0$.

Since $z_{t,d}$ is a component of the latent variable y_t, the marginal distribution is calculated by:

$$p(z_{t,d} \mid \kappa_t, p_{t,d}) = \sum_{y_t=z_{t,d}}^{\infty} p(z_{t,d} \mid y_t, p_{t,d}) p(y_t \mid \kappa_t). \tag{30}$$

The marginal multinomial likelihood for a single category d, after we marginalise out all other counts in $\mathbf{z_t}$, where $\sum_{d=1}^{D} z_{t,d} = y_t$ is:

$$p(z_{t,d} \mid y_t, p_{t,d}) = \binom{y_t}{z_{t,d}} p_{t,d}^{z_{t,d}} (1 - p_{t,d})^{y_t - z_{t,d}}, \tag{31}$$

where $\binom{y_t}{z_{t,d}} = \frac{y_t!}{z_{t,d}!(y_t - z_{t,d})!}$. The parameter $0 \leq p_{t,d} \leq 1$ is the probability such that $\sum_{d=1}^{D} p_{t,d} = 1$, and y_t is the number of trials.

The Poisson PMF of y_t with rate parameter $\kappa_t > 0$ is:

$$p(y_t \mid \kappa_t) = \frac{\kappa_t^{y_t} e^{-\kappa_t}}{y_t!}. \tag{32}$$

Substituting these into the marginal distribution gives:

$$p(z_{t,d} \mid \kappa_t, p_{t,d}) = \sum_{y_t=z_{t,d}}^{\infty} \binom{y_t}{z_{t,d}} p_{t,d}^{z_{t,d}} (1 - p_{t,d})^{y_t - z_{t,d}} \frac{\kappa_t^{y_t} e^{-\kappa_t}}{y_t!}. \tag{33}$$

The Binomial term can then be simplified by substituting $\binom{y_t}{z_{t,d}} = \frac{y_t!}{z_{t,d}!(y_t - z_{t,d})!}$, and noting that $y_t!/y_t! = 1$:

$$p(z_{t,d} \mid \kappa_t, p_{t,d}) = \frac{p_{t,d}^{z_{t,d}}}{z_{t,d}!} \sum_{y_t=z_{t,d}}^{\infty} \frac{(1 - p_{t,d})^{y_t - z_{t,d}} \kappa_t^{y_t} e^{-\kappa_t}}{(y_t - z_{t,d})!}. \tag{34}$$

If we substitute $y_t = z_{t,d} + k$, where $k = y_t - z_{t,d}$, so that $y_t - z_{t,d} = k$ and $y_t = z_{t,d} + k$, the summation becomes:

$$p(z_{t,d} \mid \kappa_t, p_{t,d}) = \frac{p_{t,d}^{z_{t,d}} \kappa_t^{z_{t,d}} e^{-\kappa_t}}{z_{t,d}!} \sum_{k=0}^{\infty} \frac{\big((1 - p_{t,d}) \kappa_t\big)^k}{k!}. \tag{35}$$

This summation is the Taylor expansion of e^x with $x = \kappa_t(1 - p_{t,d})$ where:

$$\sum_{k=0}^{\infty} \frac{\left(\kappa_t(1-p_{t,d})\right)^k}{k!} = e^{\kappa_t(1-p_{t,d})}. \tag{36}$$

Substituting this into the expression gives:

$$p(z_{t,d} \mid \kappa_t, p_{t,d}) = \frac{p_{t,d}^{z_{t,d}} \kappa_t^{z_{t,d}} e^{-\kappa_t} e^{\kappa_t(1-p_{t,d})}}{z_{t,d}!}. \tag{37}$$

Combine the exponential terms:

$$p(z_{t,d} \mid \kappa_t, p_{t,d}) = \frac{\left(p_{t,d}\kappa_t\right)^{z_{t,d}} e^{-p_{t,d}\kappa_t}}{z_{t,d}!}. \tag{38}$$

This is the PMF of a Poisson distribution with rate parameter $p_{t,d}\kappa_t$. Therefore, we have $z_{t,d} \mid \kappa_t, p_{t,d} \sim \text{Poisson}(p_{t,d}\kappa_t)$.

Since each $z_{t,d}$ follows a Poisson distribution with its own rate parameter $p_{t,d}\kappa_t$, and the derivation does not introduce any dependency between different d, it follows that:

$$z_{t,d} \perp z_{t,d'} \mid \kappa_t, p_{t,d}, p_{t,d'} \quad \forall d \neq d'. \tag{39}$$

Poisson-distributed variables with different rates are independent. This conditional independence result follows from the additive property of Poisson-distributed variables. Thus, the Poisson-Multinomial mixture model yields conditionally independent $_{t,d}$, given κ_t and $p_{t,d}$. Moreover, by the additive property, summing these independent Poisson-distributed counts reconstructs the total.

B.3 Marginal distribution of a Negative-Binomial-Multinomial mixture model is a Negative-Binomial distribution

We consider the hierarchical model:

$$y_t \mid \lambda_t \sim \text{Negative-Binomial}(\lambda_t, \theta), \tag{40}$$

$$z_{t,d} \mid \mathbf{p}_t, y_t \sim \text{Multinomial}(\mathbf{p}_t, y_t), \tag{41}$$

where for the multinomial distribution $\mathbf{p}_t = (p_{t,1}, \ldots, p_{t,D})$ are the probabilities of each category and y_t is the total number of trials. For the Negative-Binomial distribution λ_t is the mean and θ is the dispersion parameter. Here we derive the marginal distribution of $z_{t,d}$ by integrating out y_t.

The joint distribution of y_t and $z_{t,d}$ is given by:

$$P(z_{t,d} \mid \lambda_t, \theta, p_{t,d}) = \sum_{y_t=z_{t,d}}^{\infty} P(z_{t,d} \mid y_t, p_{t,d}) P(y_t \mid \lambda_t, \theta). \tag{42}$$

Given y_t, the multinomial distribution specifies (see Appendix A.9):

$$P(z_{t,d} \mid y_t, p_{t,d}) = \binom{y_t}{z_{t,d}} p_{t,d}^{z_{t,d}} (1 - p_{t,d})^{y_t - z_{t,d}}. \tag{43}$$

The Negative Binomial distribution for $y_t \mid \lambda_t, \theta$ is:

$$P(y_t \mid \lambda_t, \theta) = \binom{y_t + \theta - 1}{y_t} \left(\frac{\lambda_t}{\lambda_t + \theta}\right)^{y_t} \left(\frac{\theta}{\lambda_t + \theta}\right)^{\theta}. \tag{44}$$

Substituting the above expressions into the joint distribution:

$$P(z_{t,d} \mid \lambda_t, \theta, p_{t,d}) = \sum_{y_t=z_{t,d}}^{\infty} \binom{y_t}{z_{t,d}} p_{t,d}^{z_{t,d}} (1-p_{t,d})^{y_t - z_{t,d}} \times \binom{y_t + \theta - 1}{y_t} \left(\frac{\lambda_t}{\lambda_t + \theta}\right)^{y_t} \left(\frac{\theta}{\lambda_t + \theta}\right)^{\theta}. \tag{45}$$

Substituting the expressions for $P(z_{t,d} \mid y_t, p_{t,d})$ and $P(y_t \mid \lambda_t, \theta)$ into the summation:

$$P(z_{t,d} \mid \lambda_t, \theta, p_{t,d}) = \sum_{y_t=z_{t,d}}^{\infty} \binom{y_t}{z_{t,d}} p_{t,d}^{z_{t,d}} (1-p_{t,d})^{y_t - z_{t,d}} \times \binom{y_t + \theta - 1}{y_t} \left(\frac{\lambda_t}{\lambda_t + \theta}\right)^{y_t} \left(\frac{\theta}{\lambda_t + \theta}\right)^{\theta}. \tag{46}$$

Expanding the binomial coefficient $\binom{y_t}{z_{t,d}}$, we have:

$$\binom{y_t}{z_{t,d}} = \frac{y_t!}{z_{t,d}!(y_t - z_{t,d})!}.$$

Substituting this into the expression:

$$P(z_{t,d} \mid \lambda_t, \theta, p_{t,d}) = \frac{p_{t,d}^{z_{t,d}}}{z_{t,d}!} \sum_{y_t=z_{t,d}}^{\infty} \frac{y_t!}{(y_t - z_{t,d})!} (1-p_{t,d})^{y_t - z_{t,d}} \times \binom{y_t + \theta - 1}{y_t} \left(\frac{\lambda_t}{\lambda_t + \theta}\right)^{y_t} \left(\frac{\theta}{\lambda_t + \theta}\right)^{\theta}. \tag{47}$$

Next, let $m = y_t - z_{t,d}$, which implies $y_t = m + z_{t,d}$. Making the substitution, we get:

$$P(z_{t,d} \mid \lambda_t, \theta, p_{t,d}) = \frac{p_{t,d}^{z_{t,d}}}{z_{t,d}!} \sum_{m=0}^{\infty} \frac{(m + z_{t,d})!}{m!} (1-p_{t,d})^{m} \times \binom{m + z_{t,d} + \theta - 1}{m + z_{t,d}} \left(\frac{\lambda_t}{\lambda_t + \theta}\right)^{m + z_{t,d}} \left(\frac{\theta}{\lambda_t + \theta}\right)^{\theta}. \tag{48}$$

Splitting the factorial and rearranging terms:

$$P(z_{t,d} \mid \lambda_t, \theta, p_{t,d}) = \frac{p_{t,d}^{z_{t,d}}}{z_{t,d}!} \left(\frac{\lambda_t}{\lambda_t + \theta} \right)^{z_{t,d}} \times \sum_{m=0}^{\infty} \binom{m + z_{t,d} + \theta - 1}{m} \left((1 - p_{t,d}) \frac{\lambda_t}{\lambda_t + \theta} \right)^{m} \left(\frac{\theta}{\lambda_t + \theta} \right)^{\theta}. \quad (49)$$

The summation now matches the Negative Binomial generating function. Using this property, the marginal distribution is:

$$P(z_{t,d} \mid \lambda_t, \theta, p_{t,d}) = \binom{z_{t,d} + \theta - 1}{z_{t,d}} \left(\frac{p_{t,d}\lambda_t}{p_{t,d}\lambda_t + \theta} \right)^{z_{t,d}} \left(\frac{\theta}{p_{t,d}\lambda_t + \theta} \right)^{\theta}. \quad (50)$$

Thus, the marginal distribution of $z_{t,d}$ is:

$$z_{t,d} \sim \text{Negative-Binomial}(p_{t,d}\lambda_t, \theta), \quad (51)$$

where $p_{t,d}\lambda_t$ is the mean and θ is the dispersion parameter.

B.4 Derivation of the Generalized-Dirichlet Multinomial (GDM)

Given the total count y_t, the compositional counts $\mathbf{z}_t = (z_{t,1}, \ldots, z_{t,D})$ follow a multinomial distribution:

$$\mathbf{z}_t \mid \mathbf{p}_t, y_t \sim \text{Multinomial}(\mathbf{p}_t, y_t), \quad (52)$$

where $\mathbf{p}_t = (p_{t,1}, p_{t,2}, \ldots, p_{t,D})$ represents the probability vector for the $d = 1, \ldots, D$ categories. The probabilities $\mathbf{p}_t$ are modelled using the Generalized-Dirichlet (GD) distribution:

$$\mathbf{p}_t \sim \text{Generalized-Dirichlet}(\boldsymbol{\alpha}_t, \boldsymbol{\beta}_t). \quad (53)$$

The multinomial likelihood is given by:

$$P(\mathbf{z}_t \mid \mathbf{p}_t, y_t) = \frac{y_t!}{z_{t,1}! \cdots z_{t,D}!} \prod_{d=1}^{D} p_{t,d}^{z_{t,d}}. \tag{54}$$

The Generalized-Dirichlet prior on $\mathbf{p}_t$ has the density function:

$$P(\mathbf{p}_t \mid \boldsymbol{\alpha}_t, \boldsymbol{\beta}_t) = \prod_{d=1}^{D-1} \frac{1}{B(\alpha_{t,d}, \beta_{t,d} - \alpha_{t,d})} p_{t,d}^{\alpha_{t,d}-1} (1 - \sum_{j=1}^{d} p_{t,j})^{\beta_{t,d}-\alpha_{t,d}-1}, \tag{55}$$

where the Beta function $B(a,b)$ is defined as:

$$B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}. \tag{56}$$

To obtain the Generalized-Dirichlet Multinomial (GDM) distribution, we integrate out $\mathbf{p}_t$:

$$P(\mathbf{z}_t \mid y_t, \boldsymbol{\alpha}_t, \boldsymbol{\beta}_t) = \int P(\mathbf{z}_t \mid \mathbf{p}_t, y_t) P(\mathbf{p}_t \mid \boldsymbol{\alpha}_t, \boldsymbol{\beta}_t) d\mathbf{p}_t. \tag{57}$$

Substituting the expressions for $P(\mathbf{z}_t \mid \mathbf{p}_t, y_t)$ and $P(\mathbf{p}_t \mid \boldsymbol{\alpha}_t, \boldsymbol{\beta}_t)$, the integral follows a standard result that yields the Generalized-Dirichlet-Multinomial (GDM) distribution:

$$P(\mathbf{z}_t \mid y_t, \boldsymbol{\alpha}_t, \boldsymbol{\beta}_t) = \frac{y_t!}{z_{t,1}! \cdots z_{t,D}!} \frac{\prod_{d=1}^{D-1} B(z_{t,d} + \alpha_{t,d}, y_t - \sum_{j=1}^{d} z_{t,j} + \beta_{t,d} - \alpha_{t,d})}{\prod_{d=1}^{D-1} B(\alpha_{t,d}, \beta_{t,d} - \alpha_{t,d})}. \tag{58}$$

Instead of using the shape parameters $\boldsymbol{\alpha}_t$ and $\boldsymbol{\beta}_t$, we can reparameterise the model in terms of the expected proportions $\nu_{t,d}$ and a dispersion parameter $\phi_{t,d}$, which controls the variability of category proportions. The expected proportion are defined as:

$$\nu_{t,d} = \frac{\alpha_{t,d}}{\alpha_{t,d} + \beta_{t,d}}. \tag{59}$$

The dispersion parameter is introduced to control the spread of the proportions:

$$\phi_{t,d} = \alpha_{t,d} + \beta_{t,d}. \tag{60}$$

Using this reparameterisation, we express $\alpha_{t,d}$ and $\beta_{t,d}$ as:

$$\alpha_{t,d} = \nu_{t,d}\phi_{t,d}, \tag{61}$$

$$\beta_{t,d} = (1-\nu_{t,d})\phi_{t,d}. \tag{62}$$

Substituting these into the GDM density function gives:

$$P(\mathbf{z}_t \mid y_t, \boldsymbol{\nu}_t, \phi) = \frac{y_t!}{z_{t,1}! \cdots z_{t,D}!} \frac{\prod_{d=1}^{D-1} B\left(z_{t,d} + \nu_{t,d}\phi, y_t - \sum_{j=1}^{d} z_{t,j} + (1-\nu_{t,d})\phi\right)}{\prod_{d=1}^{D-1} B\left(\nu_{t,d}\phi, (1-\nu_{t,d})\phi\right)}. \tag{63}$$

Thus, integrating out $\mathbf{p}_t$ results in the Generalized-Dirichlet-Multinomial (GDM) distribution:

$$\mathbf{z}_t \mid y_t \sim \text{GDM}(\boldsymbol{\nu}_t, \boldsymbol{\phi}_t, y_t). \tag{64}$$

This generalizes the Dirichlet-Multinomial (DM) by allowing additional flexibility in the covariance structure of $\mathbf{z}_t$.

B.5 Derivation of the Generalized-Dirichlet Multinomial (GDM) as a Beta-Binomial Series

For the Generalized-Dirichlet Multinomial (GDM) equations, the $\mathbf{p}_t \sim \text{Generalized-Dirichlet}(\boldsymbol{\alpha}_t, \boldsymbol{\beta}_t)$ distribution can be constructed as a series of independent Beta distributions:

$$p_{t,1} \sim \text{Beta}(\nu_{t,1}\phi_{t,1}, (1-\nu_{t,1})\phi_{t,1}); \tag{65}$$

$$\frac{p_{t,2}}{1-p_{t,1}} \sim \text{Beta}(\nu_{t,2}\phi_{t,2}, (1-\nu_{t,2})\phi_{t,2}); \tag{66}$$

$$\dots$$

$$\frac{p_{t,d}}{1-\sum_{i=1}^{d-1} p_{t,i}} \sim \text{Beta}(\nu_{t,d}\phi_{t,d}, (1-\nu_{t,d})\phi_{t,d}); \tag{67}$$

$$\dots$$

$$p_{t,D_{\max}} = 1 - \sum_{i=1}^{D_{\max}-1} p_{t,i}. \tag{68}$$

Here, the Beta distributions ensure that each fraction of the remaining probability mass follows a hierarchical allocation process. Similarly, the Multinomial $\mathbf{z}_t \mid \mathbf{p}_t, y_t \sim \text{Multinomial}(\mathbf{p}_t, y_t)$, can be rewritten as a series of conditional Binomial distributions:

$$z_{t,1} \mid p_{t,1}, y_t \sim \text{Binomial}(y_t, p_{t,1}); \tag{69}$$

$$z_{t,2} \mid p_{t,2}, z_{t,1}, y_t \sim \text{Binomial}(y_t - z_{t,1}, p_{t,2}); \tag{70}$$

$$\dots$$

$$z_{t,d} \mid p_{t,d}, z_{t,<d}, y_t \sim \text{Binomial}(N_{t,d}, p_{t,d}); \tag{71}$$

$$\dots$$

$$z_{t,D_{\max}} = N_{t,D_{\max}}, \tag{72}$$

where we define the remaining sample size at step d as:

$$N_{t,d} = y_t - \sum_{j=1}^{d-1} z_{t,j}. \tag{73}$$

This formulation expresses the counts as a series of nested Binomial processes, where each count $z_{t,d}$ is a Binomial draw from the remaining counts $N_{t,d}$. The density function for the Binomial distribution is:

$$P(z_{t,d} \mid p_{t,d}, N_{t,d}) = \binom{N_{t,d}}{z_{t,d}} p_{t,d}^{z_{t,d}} (1-p_{t,d})^{N_{t,d}-z_{t,d}}. \tag{74}$$

The density function for the Beta prior is given by:

$$P(p_{t,d}) = \frac{p_{t,d}^{\nu_{t,d}\phi_{t,d}-1}(1-p_{t,d})^{(1-\nu_{t,d})\phi_{t,d}-1}}{B(\nu_{t,d}\phi_{t,d}, (1-\nu_{t,d})\phi_{t,d})}. \tag{75}$$

To obtain the marginalised Beta-Binomial formulation, we integrate out $p_{t,d}$.

$$P(z_{t,d} \mid z_{t,<d}, N_{t,d}) = \int_0^1 P(z_{t,d} \mid p_{t,d}, N_{t,d}) P(p_{t,d}) dp_{t,d}. \tag{76}$$

Substituting in the probability density functions as well as the following Beta function:

$$B(a,b) = \int_0^1 t^{a-1}(1-t)^{b-1} dt = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}, \tag{77}$$

where a and b are set as

$$a = z_{t,d} + \nu_{t,d}\phi_{t,d}, \tag{78}$$

$$b = (N_{t,d} - z_{t,d}) + (1-\nu_{t,d})\phi_{t,d}. \tag{79}$$

Hence, the integral simplifies to:

$$\frac{B(z_{t,d} + \nu_{t,d}\phi_{t,d}, N_{t,d} - z_{t,d} + (1-\nu_{t,d})\phi_{t,d})}{B(\nu_{t,d}\phi_{t,d}, (1-\nu_{t,d})\phi_{t,d})}. \tag{80}$$

The probability distribution is then:

$$P(z_{t,d} \mid z_{t,<d}, N_{t,d}) = \binom{N_{t,d}}{z_{t,d}} \frac{B(z_{t,d} + \nu_{t,d}\phi_{t,d}, N_{t,d} - z_{t,d} + (1-\nu_{t,d})\phi_{t,d})}{B(\nu_{t,d}\phi_{t,d}, (1-\nu_{t,d})\phi_{t,d})}. \tag{81}$$

Thus, the marginal distribution of $z_{t,d}$ is:

$$z_{t,d} \mid z_{t,<d}, N_{t,d} \sim \text{Beta-Binomial}(N_{t,d}, \nu_{t,d}, \phi_{t,d}). \tag{82}$$

C Directed acyclic graphs

C.1 The Generalised-Dirichlet Multinomial (GDM) model

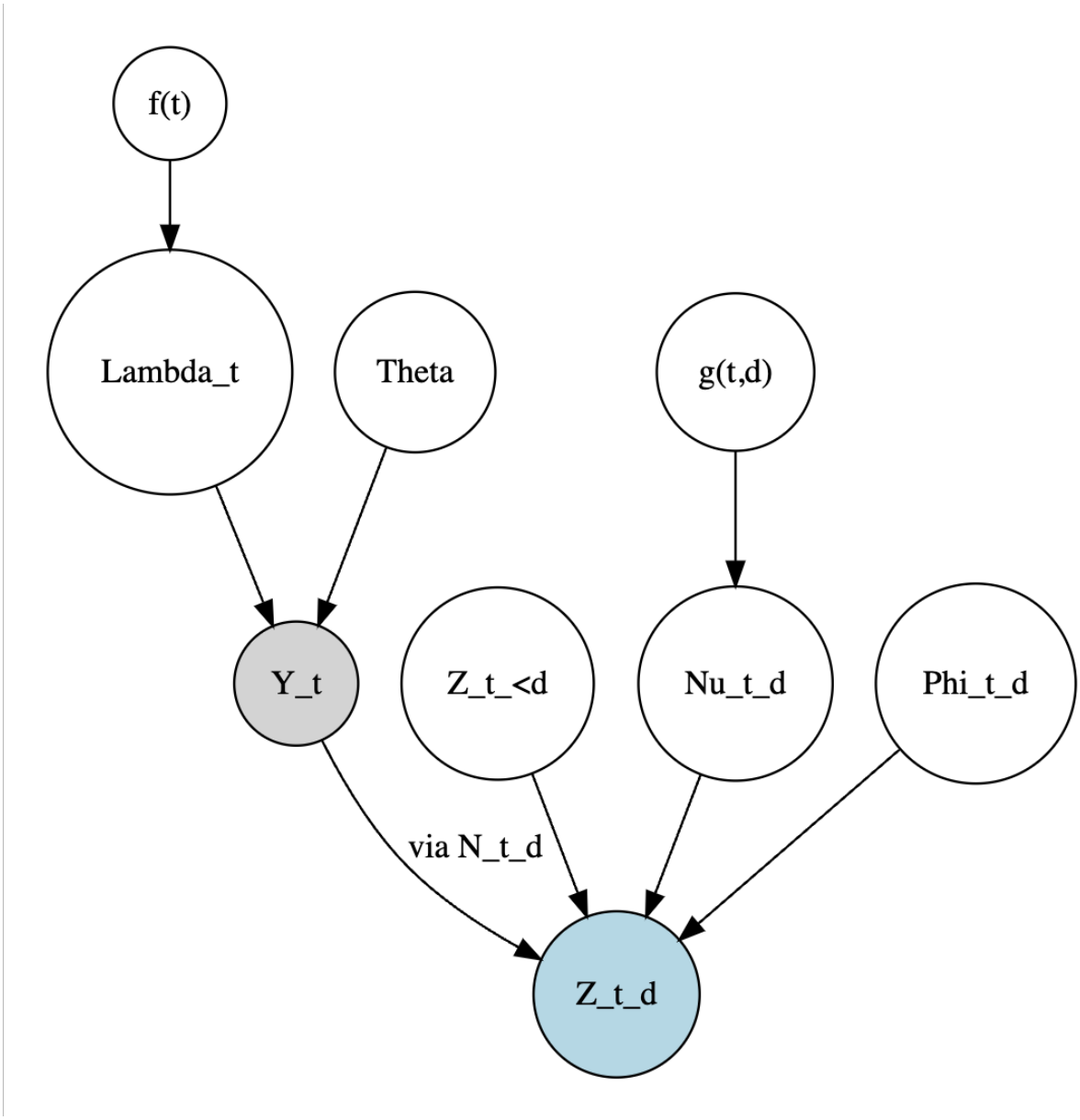

Figure 6: Directed acyclic graph of the GDM model, given by Equations (2.69)–(2.76).

C.2 The nested GDM model

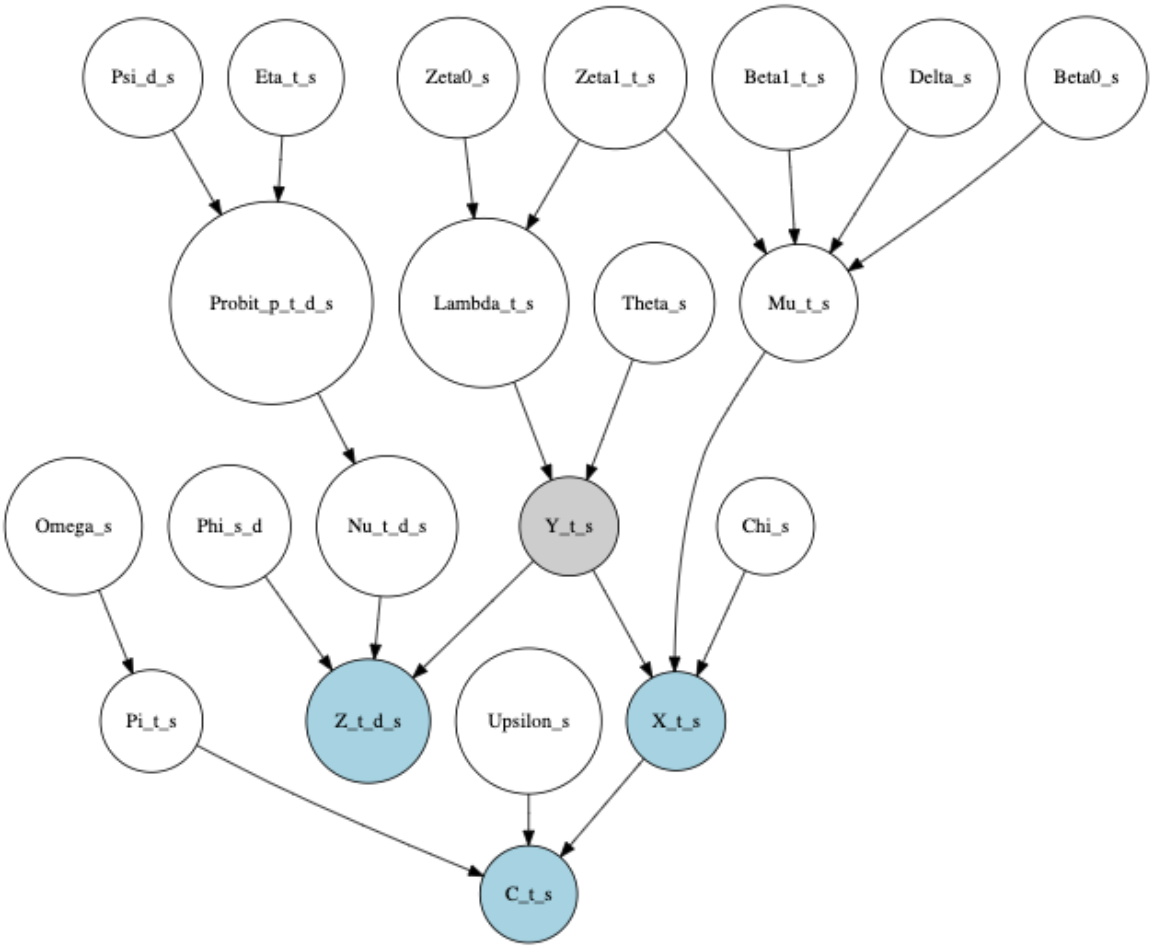

Figure 7: Directed acyclic graph of the nested GDM model, given by Equations (5.6)–(5.12).

C.3 The caseload effect GDM model

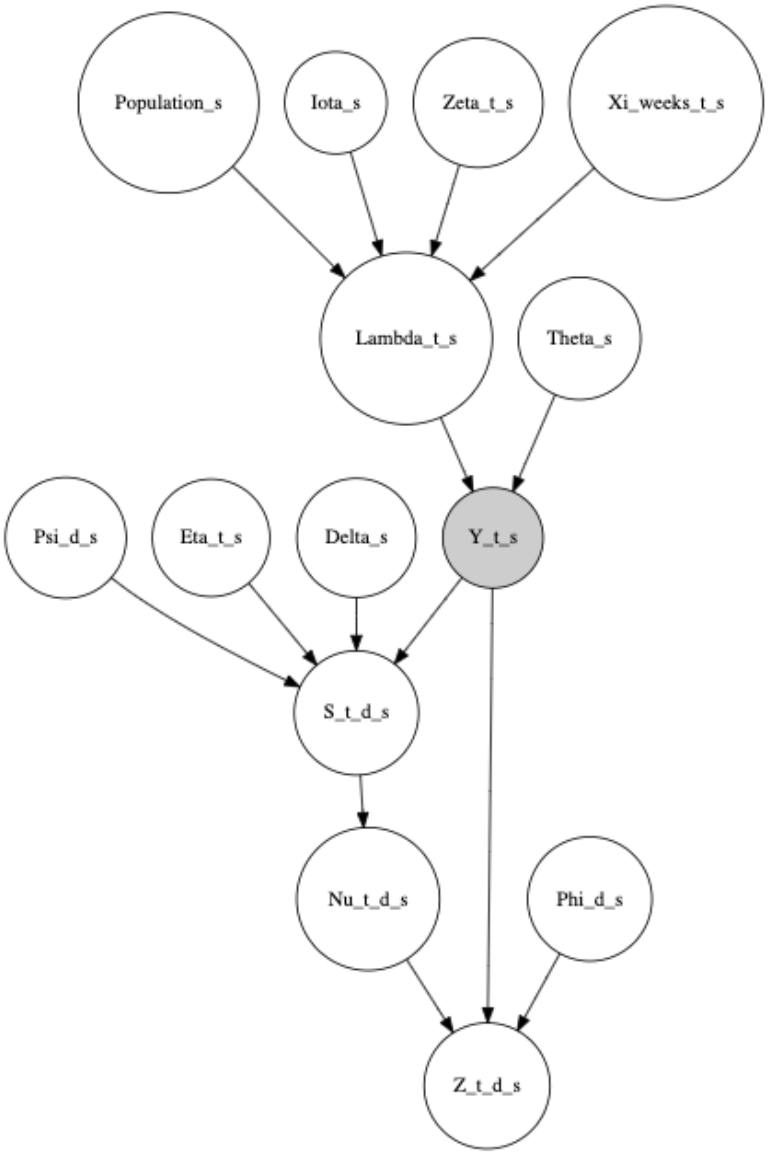

Figure 8: Directed acyclic graph of the case load effect GDM model, given by Equations (6.28)–(6.32).

Bibliography

[1] Emily L. Aiken, Sarah F. McGough, Maimuna S. Majumder, Gal Wachtel, Andre T. Nguyen, Cecile Viboud and Mauricio Santillana. 'Real-time estimation of disease activity in emerging outbreaks using internet search information'. In: *PLoS Computational Biology* 16.8 (2020), pp. 1–19. ISSN: 15537358. DOI: 10.1371/journal.pcbi.1008117.

[2] Serena Arima, Silvia Polettini, Giuseppe Pasculli, Loreto Gesualdo, Francesco Pesce and Deni-Aldo Procaccini. 'A Bayesian nonparametric approach to correct for underreporting in count data'. In: *Biostatistics* (Sept. 2023), kxad027. ISSN: 1465-4644. DOI: 10.1093/biostatistics/kxad027.

[3] Trevor C. Bailey, Marilia S. Carvalho, Thiago M. Lapa, Wander V. Souza and Mark J. Brewer. 'Modeling of Under-detection of Cases in Disease Surveillance'. In: *Annals of Epidemiology* 15.5 (2005), pp. 335–343. ISSN: 1047-2797. DOI: https://doi.org/10.1016/j.annepidem.2004.09.013.

[4] Leonardo S. Bastos, Theodoros Economou, Marcelo F.C. Gomes, Daniel A.M. Villela, Flavio C. Coelho, Oswaldo G. Cruz, Oliver Stoner, Trevor Bailey and Claudia T. Codeço. 'A modelling approach for correcting reporting delays in disease surveillance data'. In: *Statistics in Medicine* 38.22 (2019), pp. 4363–4377. ISSN: 10970258. DOI: 10.1002/sim.8303.

[5] Michael Braun. *trustOptim: An R Package for Trust Region Optimization with Sparse Hessians.* R package version 0.8.4. 2014.

[6] Michael Braun. 'trustOptim: An R package for trust region optimization with sparse Hessians'. In: *Journal of Statistical Software* 60.4 (2014), pp. 1–16. ISSN: 15487660. DOI: 10.18637/jss.v060.i04.

[7] Ron. Brookmeyer and Mitchell. H. Gail. *AIDS epidemiology : a quantitative approach*. eng. Monographs in epidemiology and biostatistics ; v. 22. New York: Oxford University Press, 1994. Chap. 7. ISBN: 0195076419.

[8] Stephen P. Brooks and Andrew Gelman. 'General methods for monitoring convergence of iterative simulations'. In: *Journal of Computational and Graphical Statistics* 7.4 (1998), pp. 434–455. ISSN: 15372715. DOI: 10.1080/10618600.1998.10474787.

[9] Gareth P. Campbell, James M. Curran, Gordon M. Miskelly, Sally Coulson, Gregory M. Yaxley, Eric C. Grunsky and Simon C. Cox. 'Compositional data analysis for elemental data in forensic science'. In: *Forensic Science International* 188.1 (2009), pp. 81–90. ISSN: 0379-0738. DOI: https://doi.org/10.1016/j.forsciint.2009.03.018.

[10] Bob Carpenter, Andrew Gelman, Matthew D. Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li and Allen Riddell. 'Stan: A Probabilistic Programming Language'. In: *Journal of Statistical Software* 76.1 (2017), pp. 1–32. DOI: 10.18637/jss.v076.i01.

[11] Jun Chen and Hongzhe Li. 'Variable selection for sparse Dirichlet-multinomial regression with an application to microbiome data analysis'. In: *The Annals of Applied Statistics* 7.1 (2013), pp. 418 –442. DOI: 10.1214/12-AOAS592.

[12] Robert J. Connor and James E. Mosimann. 'Concepts of Independence for Proportions with a Generalization of the Dirichlet Distribution'. In: *Journal of the American Statistical Association* 64.325 (1969), pp. 194–206. DOI: 10.1080/01621459.1969.10500963.

[13] Microsoft Corporation and Steve Weston. *doParallel: Foreach Parallel Adaptor for the 'parallel' Package*. R package version 1.0.17. 2022.

[14] Adele Cutler, D. Richard Cutler and John R Stevens. 'Random Forrests'. In: *Ensemble Machine Learning* (2012), pp. 157–175. DOI: 10.1007/978-1-4419-9326-7.

[15] Perry de Valpine, Christopher Paciorek, Daniel Turek, Nicholas Michaud, Christian Anderson-Bergman, Franz Obermeyer, César Wehrhahn Cortes, Andrés Rodríguez, Duncan Temple Lang, Weixuan Zhang, Stefano Paganin, Jonathan Hug and Peter van Dam-Bates. *NIMBLE User Manual*. Version 1.2.1. 2021. DOI: 10.5281/zenodo.1211190.

[16] Perry de Valpine, Daniel Turek, Christopher J. Paciorek, Clifford Anderson-Bergman, Duncan Temple Lang and Rastislav Bodik. 'Programming With Models: Writing Statistical Algorithms for General Model Structures With NIMBLE'. In: *Journal of Computational and Graphical Statistics* 26.2 (2017), pp. 403–413. DOI: 10.1080/10618600.2016.1172487.

[17] André Ricardo Ribas Freitas and Maria Rita Donalisio. 'Respiratory syncytial virus seasonality in Brazil: implications for the immunisation policy for at-risk populations'. In: *Memórias do Instituto Oswaldo Cruz* 111.5 (2016), pp. 294–301. DOI: 10.1590/0074-02760150341.

[18] Andrew Gelman, John B. Carlin, Hal S. Stern and Donald B. Rubin. *Bayesian Data Analysis*. 3rd. CRC Press LLC, 2013.

[19] Andrew Gelman and Jennifer Hill. 'Data Analysis Using Regression and Multilevel/Hierarchical Models'. In: Analytical Methods for Social Research. Cambridge University Press, 2006, pp. 235–236.

[20] Andrew Gelman and Donald B. Rubin. 'Inference from Iterative Simulation Using Multiple Sequences'. In: *Statistical Science* 7.4 (1992), pp. 457 –472.

[21] Andrew Gelman and Donald B. Rubin. 'Markov chain Monte Carlo methods in biostatistics'. In: *Statistical Methods in Medical Research* 5.4 (1996), pp. 339–355. ISSN: 09622802. DOI: 10.1177/096228029600500402.

[22] Paul Gilbert and Ravi Varadhan. *numDeriv: Accurate Numerical Derivatives*. R package version 2016.8-1.1. 2019.

[23] Felix Günther, Andreas Bender, Katharina Katz, Helmut Küchenhoff and Michael Höhle. 'Nowcasting the COVID-19 pandemic in Bavaria'. In: *Biometrical Journal* October (2020), pp. 1–13. ISSN: 15214036. DOI: 10.1002/bimj.202000112.

[24] Emilio Gutierrez, Adrian Rubli and Tiago Tavares. 'Delays in Death Reports and their Implications for Tracking the Evolution of COVID-19'. In: *SSRN* (2020), pp. 1 –31.

[25] Emilio Gutierrez, Adrian Rubli and Tiago Tavares. 'Information and behavioral responses during a pandemic: Evidence from delays in Covid-19 death reports.' In: *Journal of development economics* 154 (May 2022), p. 102774. DOI: doi:10.1016/j.jdeveco.2021.102774.

[26] Alba Halliday, Oliver Stoner and Bastos Leonardo. *Modelling diseases with nested structures and delayed reporting using a hierarchical framework.* Version pre-submission. June 2025. DOI: 10.5281/zenodo.8099078.

[27] Robbert L. Harms and Alard Roebroeck. 'Robust and fast Markov chain Monte Carlo sampling of diffusion MRI microstructure models'. In: *Frontiers in Neuroinformatics* 12.December (2018), pp. 1–18. ISSN: 16625196. DOI: 10.3389/fninf.2018.00097.

[28] Sean M. Harrington, Van Wishingrad and Robert C Thomson. 'Properties of Markov Chain Monte Carlo Performance across Many Empirical Alignments'. In: *Molecular Biology and Evolution* 38.4 (2021), pp. 1627–1640. ISSN: 15371719. DOI: 10.1093/molbev/msaa295.

[29] Jeffrey E. Harris. 'Timely epidemic monitoring in the presence of reporting delays: anticipating the COVID-19 surge in New York City, September 2020'. In: *BMC Public Health* 22.1 (May 2022), article number 871.

[30] Katsuma Hayashi and Hiroshi Nishiura. 'Time-dependent risk of COVID-19 death with overwhelmed health-care capacity in Japan, 2020–2022.' In: *BMC Infectious Diseases* 22 (Dec. 2022), p. 933. DOI: https://doi.org/10.1186/s12879-022-07929-8.

[31] Michael Höhle and Matthias An Der Heiden. 'Bayesian nowcasting during the STEC O104: H4 outbreak in Germany, 2011'. In: *Biometrics* 70.4 (2014), pp. 993–1002. ISSN: 15410420. DOI: 10.1111/biom.12194.

[32] David Kline, Ayaz Hyder, Enhao Liu, Michael Rayo, Samuel Malloy and Elisabeth Root. 'A Bayesian Spatiotemporal Nowcasting Model for Public Health Decision-Making and Surveillance'. In: *American Journal of Epidemiology* 191.6 (Feb. 2022), pp. 1107–1115. ISSN: 0002-9262. DOI: 10.1093/aje/kwac034.

[33] Matthew D. Koslovsky. 'A Bayesian zero-inflated Dirichlet-multinomial regression model for multivariate compositional count data'. In: *Biometrics* n/a.n/a (2023). DOI: https://doi.org/10.1111/biom.13853.

[34] John K. Kruschke. *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan*. Academic Press. Academic Press, 2015. ISBN: 9780124058880.

[35] Jerald F. Lawless. 'Adjustments for reporting delays and the prediction of occurred but not reported events'. In: *Canadian Journal of Statistics* 22.1 (1994), pp. 15–31. ISSN: 1708945X. DOI: 10.2307/3315826.n1.

[36] Finn Lindgren and Håvard Rue. 'Bayesian Spatial Modelling with R-INLA'. In: *Journal of Statistical Software, Articles* 63.19 (2015), pp. 1–25. ISSN: 1548-7660. DOI: 10.18637/jss.v063.i19.

[37] Sarah F. McGough, Michael A. Johansson, Marc Lipsitch and Nicolas A. Menzies. 'Nowcasting by Bayesian smoothing: A flexible, generalizable model for real-time epidemic tracking'. In: *PLoS Computational Biology* 16.4 (2020), pp. 1–20. ISSN: 15537358. DOI: 10.1371/journal.pcbi.1007735.

[38] Ministry of Health Brazil. *OpenDataSUS*. Downloaded: 2022-12-22. 2022. URL: https://opendatasus.saude.gov.br/en/dataset/srag-2021-e-2022/resource/62803c57-0b2d-4bcf-b114-380c392fe825.

[39] Iain Murray, Ryan Prescott Adams and David J. C. MacKay. 'Elliptical slice sampling'. English. In: *PMLR*. Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, (2010), pp. 541–548.

[40] Radford M. Neal. 'Slice sampling'. In: *The Annals of Statistics* 31.3 (2003), pp. 705 –767. DOI: 10.1214/aos/1056562461.

[41] NHS England. Downloaded: 2021-11-25.

[42] Observatório COVID-19 BR. *Informações Técnicas*. https://covid19br.github.io. [Online; accessed 16-Jan-2024]. 2024.

[43] Thomás Oliveira, Igor Teixeira, Leonides Medeiros Neto, Theo Lynn, Sebastião Silva Neto, Vanderson Sampaio and Patricia Endo. *Arbovirus clinical data, Brazil, 2013–2020*. Version V2. 2021. DOI: 10.17632/2d3kr8zynf.2.

[44] Jasmina Panovska-Griffiths. 'Can mathematical modelling solve the current Covid-19 crisis?' In: *BMC Public Health* 20.1 (2020), pp. 1–3. ISSN: 14712458. DOI: 10.1186/s12889-020-08671-z.

[45] Roger D. Peng. *R Programming for Data Science*. Bookdown Publishing, 2020.

[46] Martyn Plummer. 'JAGS: A Program for Analysis of Bayesian Graphical Models Using Gibbs Sampling'. In: *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)*. Vienna, Austria, 2003.

[47] R Core Team. *parallel: Support for Parallel Computation in R*. R package version 4.3.1. 2023.

[48] R Development Core Team. 'R: A Language and Environment for Statistical Computing.' In: *R Foundation for Statistical Computing*. (2011).

[49] Caitlin Rivers, Jean Paul Chretien, Steven Riley, Julie A. Pavlin, Alexandra Woodward, David Brett-Major, Irina Maljkovic Berry, Lindsay Morton, Richard G. Jarman, Matthew Biggerstaff, Michael A. Johansson, Nicholas G. Reich, Diane Meyer, Michael R. Snyder and Simon Pollett. 'Using "outbreak science" to strengthen the use of models during epidemics'. In: *Nature Communications* 10.1 (2019), pp. 9–11. ISSN: 20411723. DOI: 10.1038/s41467-019-11067-2.

[50] Håvard Rue, Sara Martino and Nicolas Chopin. 'Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations'. In: *Journal of the Royal Statistical Society. Series B: Statistical Methodology* 71.2 (2009), pp. 319–392. ISSN: 13697412. DOI: 10.1111/j.1467-9868.2008.00700.x.

[51] Saumya Yashmohini Sahai, Saket Gurukar, Wasiur R. KhudaBukhsh, Srinivasan Parthasarathy and Grzegorz A. Rempała. 'A machine learning model for nowcasting epidemic incidence'. In: *Mathematical Biosciences* 343 (2022), p. 108677.

[52] Maëlle Salmon, Dirk Schumacher, Klaus Stark and Michael Höhle. 'Bayesian outbreak detection in the presence of reporting delays'. In: *Biometrical Journal* 57.6 (2015), pp. 1051–1067. ISSN: 15214036. DOI: 10.1002/bimj.201400159.

[53] Shaun R. Seaman, Pantelis Samartsidis, Meaghan Kall and Daniela De Angelis. 'Nowcasting COVID-19 Deaths in England by Age and Region'. In: *Journal of the Royal Statistical Society Series C: Applied Statistics* 71.5 (June 2022), pp. 1266–1281. ISSN: 0035-9254. DOI: 10.1111/rssc.12576.

[54] Lone Simonsen, Julia R Gog, Don Olson and Cécile Viboud. 'Infectious Disease Surveillance in the Big Data Era: Towards Faster and Locally Relevant Systems'. In: *The Journal of Infectious Diseases* 214.Suppl 4 (2016), S380–5. DOI: 10.1093/infdis/jiw376.

[55] Ann Singleton. 'Migration and asylum data for policy-making in the European Union. The Problem with Numbers.' In: *Bruss. CEPS Pap. Lib. Secur. Eur.* (2016).

[56] Oliver Stoner. 'Bayesian Hierarchical Modelling Frameworks for Flawed Data in Environment and Health'. PhD thesis. Mathematics and Statistics, University of Exeter, 2019.

[57] Oliver Stoner and Theo Economou. 'Multivariate hierarchical frameworks for modeling delayed reporting in count data'. In: *Biometrics* 76.3 (2019), pp. 789–798. ISSN: 15410420. DOI: 10.1111/biom.13188.

[58] Oliver Stoner, Theo Economou and Gabriela Drummond Marques da Silva. 'A Hierarchical Framework for Correcting Under-Reporting in Count Data'. In: *Journal of the American Statistical Association* 114.528 (2019), pp. 1481–1492. ISSN: 1537274X. DOI: 10.1080/01621459.2019.1573732.

[59] Oliver Stoner, Alba Halliday and Theo Economou. 'Correcting delayed reporting of COVID-19 using the generalized-Dirichlet-multinomial method'. In: *Biometrics.* (2022), pp. 1–45. DOI: 10.1111/biom.13810.

[60] Oliver Stoner, Gavin Shaddick, Theo Economou, Sophie Gumy, Jessica Lewis, Itzel Lucio, Giulia Ruggeri and Heather Adair-Rohani. 'Global Household Energy Model: A Multivariate Hierarchical Approach to Estimating Trends in The Use of Polluting and Clean Fuels for Cooking'. In: *Journal of the Royal Statistical Society Series C: Applied Statistics* 69.4 (July 2020), pp. 815–839. ISSN: 0035-9254. DOI: 10.1111/rssc.12428.

[61] Zheng-Zheng Tang and Guanhua Chen. 'Zero-inflated generalized Dirichlet multinomial regression model for microbiome compositional data analysis'. In: *Biostatistics* 20.4 (June 2018), pp. 698–713. ISSN: 1465-4644. DOI: 10.1093/biostatistics/kxy025.

[62] Kasturi Thulasiraman and Mohan N.S. Swamy. *Graphs: Theory and Algorithms*. John Wiley Sons, Ltd, 1992. Chap. 1, pp. 1–30. ISBN: 9781118033104. DOI: https://doi-org.ezproxy2.lib.gla.ac.uk/10.1002/9781118033104.ch1.

[63] Matthew M. Tibbits, Chris Groendyke, Murali Haran and John C. Liechty. 'Automated Factor Slice Sampling'. In: *Journal of Computational and Graphical Statistics* 23.2 (2014), pp. 543–563. ISSN: 10618600.

[64] Luke Tierney and Joseph B. Kadane. 'Accurate Approximations for Posterior Moments and Marginal Densities'. In: *Journal of the American Statistical Association* 81.393 (1986), pp. 82–86. DOI: 10.1080/01621459.1986.10478240.

[65] Irene Torres, Rachel Sippy and Fernando Sacoto. 'Estimating the cumulative incidence of COVID-19 in the United States using influenza surveillance, virologic testing, and mortality data: Four complementary approaches'. In: *BMC Public Health* 21.1 (Apr. 2021), article number 637.

[66] UK Government. *COVID-19 Data*. https://coronavirus.data.gov.uk. Downloaded: 2021-07-23. 2023.

[67] University of Exeter. *Statistical model enables improved disease management in Brazil*. https://results2021.ref.ac.uk/impact/ab5200bd-f30e-4b76-9ee0-7a6bf78ff119?page=1. REF 2021, Impact Case Study. 2021.

[68] Janet Van Niekerk, Elias Krainski, Denis Rustand and Håvard Rue. 'A new avenue for Bayesian inference with INLA'. In: *Computational Statistics Data Analysis* 181 (2023), p. 107692. ISSN: 0167-9473. DOI: https://doi.org/10.1016/j.csda.2023.107692.

[69] Dootika Vats, James M. Flegal and Galin L. Jones. 'Multivariate output analysis for Markov chain Monte Carlo'. In: *Biometrika* 106.2 (2019), pp. 321–337. ISSN: 14643510. DOI: 10.1093/biomet/asz002.

[70] Jay M. Ver Hoef, Erin E. Peterson, Mevin B. Hooten, Ephraim M. Hanks and Marie Josèe Fortin. 'Spatial autoregressive models for statistical inference from ecological data'. In: *Ecological Monographs* 88.1 (2018), pp. 36–59. ISSN: 15577015. DOI: 10.1002/ecm.1283.

[71] Sjoerd de Vos. 'The analysis of compositional effects as exemplified by the study of elections'. In: *GeoJournal* 44.1 (1998), pp. 43–49. DOI: 10.1023/A:1006820202876.

[72] Dennis Wesselbaum. 'Environmental drivers of delays in reporting crimes'. In: *Policing: An International Journal* 46.2 (2023), pp. 328–346. DOI: https://doi.org/10.1108/PIJPSM-09-2022-0124.

[73] Simon N. Wood. *Generalized Additive Models: An Introduction with R, Second Edition*. Chapman & Hall/CRC Texts in Statistical Science. CRC Press, 2017. ISBN: 9781498728348.

[74] Simon N. Wood. 'Just Another Gibbs Additive Modeller: Interfacing JAGS and mgcv'. English. In: *Journal of Statistical Software* 75.7 (Dec. 2016). ISSN: 1548-7660. DOI: 10.18637/jss.v075.i07.

[75] World Health Organization. 'International Statistical Classification of Diseases and Related Health Problems (ICD)'. In: *https://www.who.int/classifications/classification-of-diseases* (2020).

[76] World Health Organization. *Report of the WHO-China Joint Mission on Coronavirus Disease 2019 (COVID-19)*. https://www.who.int/docs/default-source/coronaviruse/who-china-joint-mission-on-covid-19-final-report.pdf. 2020. URL: https://www.who.int/docs/default-source/coronaviruse/who-china-joint-mission-on-covid-19-final-report.pdf (visited on 14/11/2023).

[77] World Health Organization (WHO). *Arboviral diseases*. https://tdr.who.int/our-work/research-for-implementation/vector-borne-diseases-and-climate-changes/arboviral-diseases. Accessed: 2024-09-13. n.d.

[78] Fan Xia, Jun Chen, Wing Kam Fung and Hongzhe Li. 'A Logistic Normal Multinomial Regression Model for Microbiome Compositional Data Analysis'. In: *Biometrics* 69.4 (2013), pp. 1053–1063. DOI: https://doi.org/10.1111/biom.12079.

[79] Jelmer Ypma. *nloptr: R Interface to NLopt*. R package version 1.0.4. 2014.

[80] Scott L. Zeger, Lai-Chu See and Peter J. Diggle. 'Statistical methods for monitoring the AIDS epidemic'. In: *Statistics in Medicine* 8.1 (1989), pp. 3–21. DOI: https://doi.org/10.1002/sim.4780080104.

[81] Jingru Zhang and Wei Lin. 'Scalable estimation and regularization for the logistic normal multinomial model'. In: *Biometrics* 75.4 (2019), pp. 1098–1108. DOI: https://doi.org/10.1111/biom.13071.